NEW YORK CABS DATA: CASE STUDY

Katerina Dimitrova, Jose Romero, Sergi Munoz

March 18, 2018

Previous work

Load requiered packages

Select 5000 samples

We generate a sample from a random seed (Sergi's birthday) and store it in df var. Since we have already done it before and we don't want to risk that some random effect could have any impact on our data generation, we will just load the initial data raw (without any previous pre-processing execution)

```
#Load samples
#df<-read.table("green_tripdata_2016-01.csv",header=T, sep=",")
#set.seed(03101994)
#sam<-as.vector(sort(sample(1:nrow(df),5000)))
#df<-df[sam,]
#load("C:/Users/Sergi/Desktop/Sergi/ADEI/Taxi5000_raw_initial.RData")
load("Taxi5000_raw_initial.RData")
```

Load usefull functions

```
# Useful function
calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q2=s.x[3],        q3=s.x[5], ma
}


countX <- function(x,X) {
  n_x <- NULL
  for (j in 1:ncol(x)) {n_x[j] <- sum(x[,j]==X) }
  n_x <- as.data.frame(n_x)
  rownames(n_x) <- names(x)
  nx_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {nx_i <- nx_i + as.numeric(x[,j]==X) }
  list(nx_col=n_x,nx_ind=nx_i) }

countNA <- function(x) {
  mis_x <- NULL
  for (j in 1:ncol(x)) {mis_x[j] <- sum(is.na(x[,j])) }
  mis_x <- as.data.frame(mis_x)
  rownames(mis_x) <- names(x)
  mis_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {mis_i <- mis_i + as.numeric(is.na(x[,j])) }
  list(mis_col=mis_x,mis_ind=mis_i) }
```

```
man.dist.manual <- function(p1Lat, p1Lon, p2Lat, p2Lon) {
  #return(abs(pointDistance(c(p1$lon, p1$lat), c(p1$lon, p2$lat), longlat=TRUE)) + abs(pointDistance(c(
  R = 6371
  lat1 = degrees.to.radians(p1Lat)
  lon1 = degrees.to.radians(p1Lon)
  lat2 = degrees.to.radians(p2Lat)
  lon2 = degrees.to.radians(p2Lon)
  A_lat = lat2 - lat1
  A_lon = lon2 - lon1
  a = sin(A_lat/2)^2
  c = 2 * atan2(sqrt(a), sqrt(1-a))
  dist_lat = R * c
  a = sin(A_lon/2)^2
  c = 2 * atan2(sqrt(a), sqrt(1-a))
  dist_lon = R * c
  abs(dist_lat) + abs(dist_lon)
  return(abs(dist_lat) + abs(dist_lon))
}

degrees.to.radians<-function(value) {
  return(value*0.0174532925)
}
```

Delete unnecessary attributes

```
table(df$Ehail_fee) ##Delete unnecessary row
```

```
## < table of extent 0 >
```

```
df$Ehail_fee<-NULL
```

Initialize counters for Missing Values and Outliers

```
names(df)
```

```
##  [1] "VendorID"              "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime" "Store_and_fwd_flag"
##  [5] "RateCodeID"            "Pickup_longitude"
##  [7] "Pickup_latitude"       "Dropoff_longitude"
##  [9] "Dropoff_latitude"      "Passenger_count"
## [11] "Trip_distance"         "Fare_amount"
## [13] "Extra"                 "MTA_tax"
## [15] "Tip_amount"            "Tolls_amount"
## [17] "improvement_surcharge" "Total_amount"
## [19] "Payment_type"          "Trip_type"
```

```
vars_data_qual <-names(df)[c(6:18)]
length(vars_data_qual)
```

```
## [1] 13
```

```
# Missing data

imis<-rep(0,nrow(df))  # rows - trips
jmis<-rep(0,length(vars_data_qual))  # columns - variables


# Outliers for numerical variables

iouts<-rep(0,nrow(df))  # rows - trips
jouts<-rep(0,length(vars_data_qual))  # columns - variables

# Errors

ierr<-rep(0,nrow(df))  # rows - trips
jerr<-rep(0,length(vars_data_qual))  # columns - variables
```

Conversion of qualitative variables

Numeric variables corresponding to qualitative concepts are converted to factors. ## VendorID

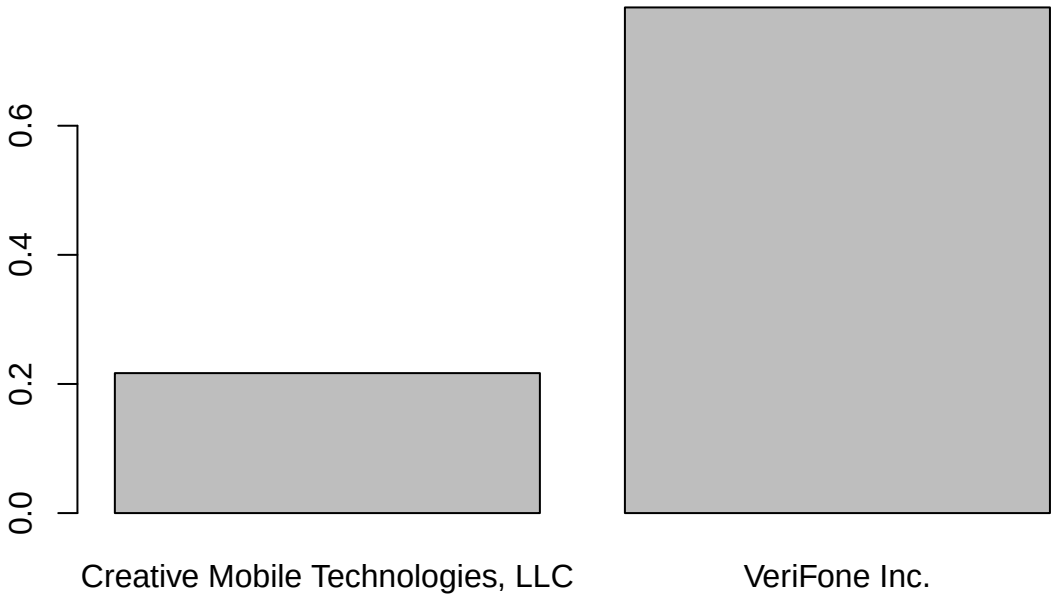
```
#No missing Data
missingData<-which(is.na(df$VendorID));length(missingData)
```

```
## [1] 0
```

```
df$VendorID<-factor(df$VendorID,labels=c("Creative Mobile Technologies, LLC","VeriFone Inc."))
table(df$VendorID)
```

```
##
## Creative Mobile Technologies, LLC                   VeriFone Inc.
##                              1084                            3916
```

```
barplot(prop.table(table(df$VendorID)))
```



RateCodeID

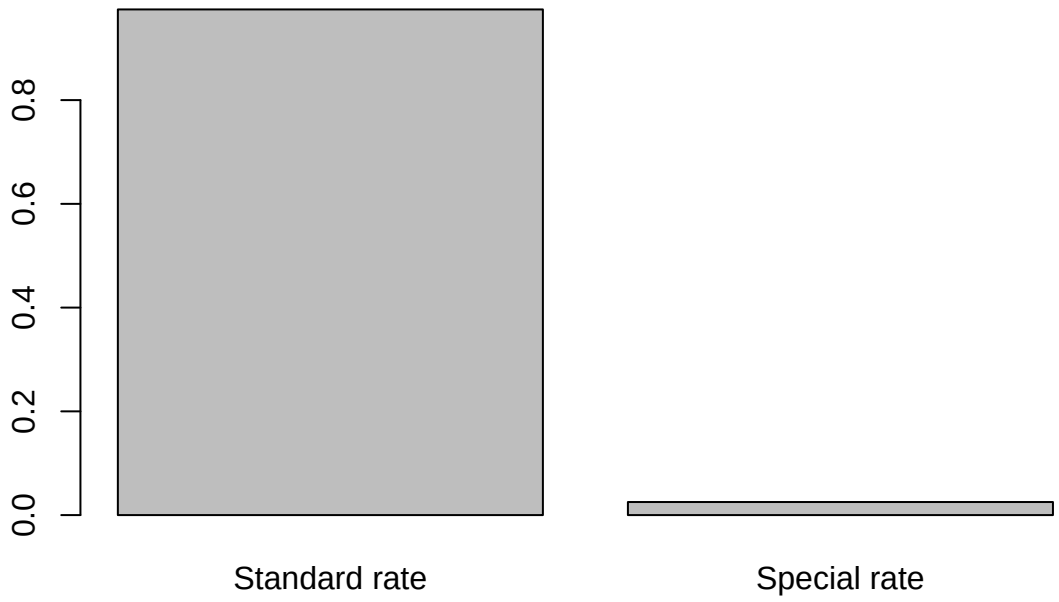
```
#No missing Data
missingData<-which(is.na(df$RateCodeID));length(missingData)
```

```
## [1] 0
```

```
df$RateCodeID<-factor(df$RateCodeID,labels=c("Standard rate","JFK","Newark","Nassau or Westchester","Neg
levels(df$RateCodeID)[levels(df$RateCodeID)=="Newark"] <- "Special rate"
levels(df$RateCodeID)[levels(df$RateCodeID)=="Nassau or Westchester"] <- "Special rate"
levels(df$RateCodeID)[levels(df$RateCodeID)=="Negotiated fare"] <- "Special rate"
levels(df$RateCodeID)[levels(df$RateCodeID)=="JFK"] <- "Special rate"
table(df$RateCodeID)
```

```
##
## Standard rate  Special rate
##          4874           126
```

```
barplot(prop.table(table(df$RateCodeID)))
```



Store_and_fwd_flag

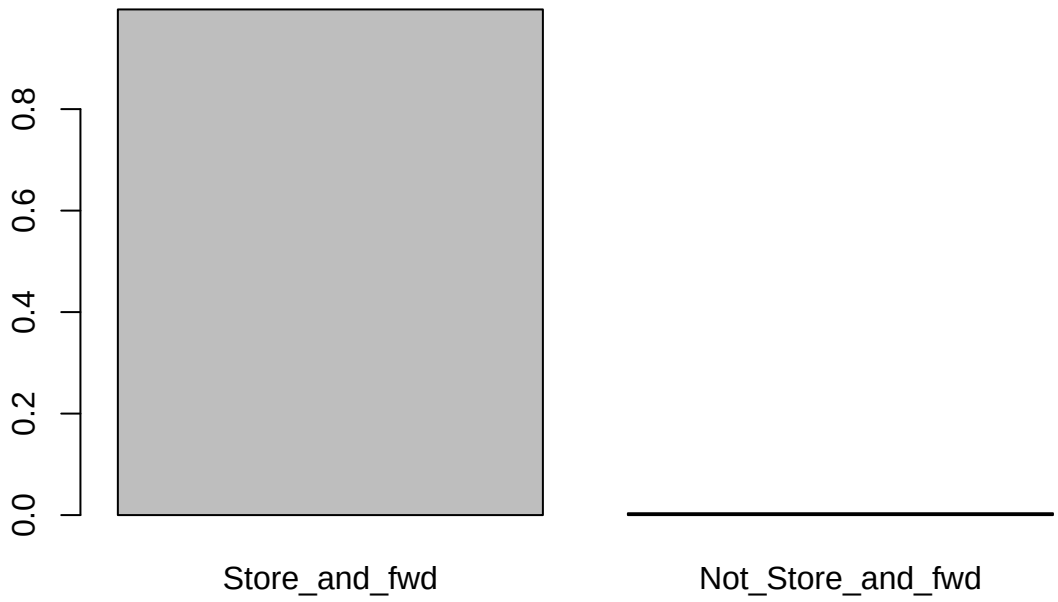
```
#No missing Data
missingData<-which(is.na(df$Store_and_fwd_flag));length(missingData)
```

```
## [1] 0
```

```
df$Store_and_fwd_flag<-factor(df$Store_and_fwd_flag,labels=c("not a store and forward trip","store and f
levels(df$Store_and_fwd_flag) <- c("Store_and_fwd","Not_Store_and_fwd")
table(df$Store_and_fwd_flag)
```

```
##
##     Store_and_fwd Not_Store_and_fwd
##              4982                18
```

```
barplot(prop.table(table(df$Store_and_fwd_flag)))
```



Payment_type

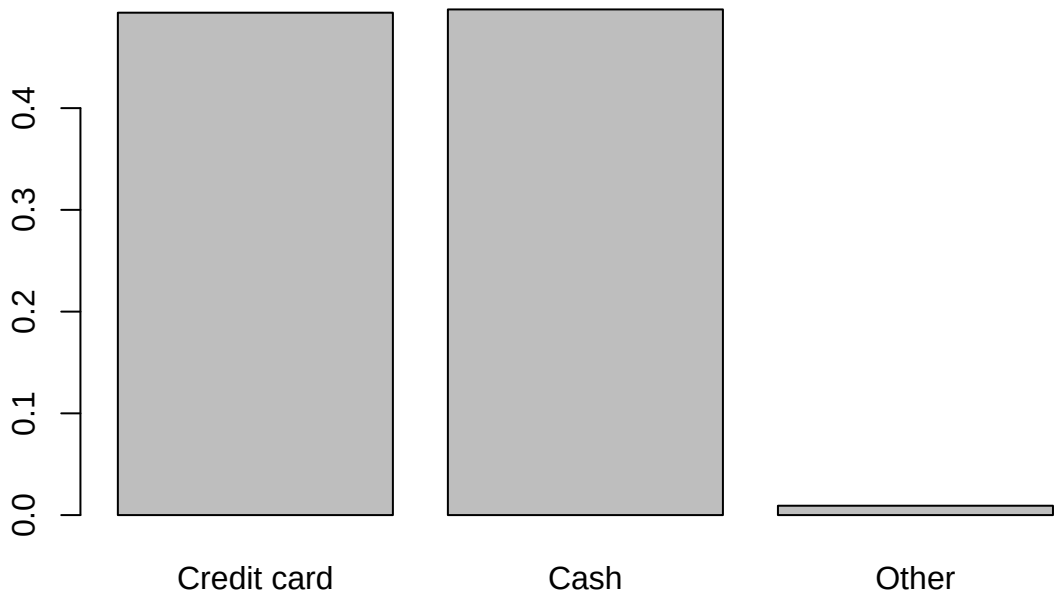
```
#No missing Data
missingData<-which(is.na(df$Trip_type));length(missingData)
```

```
## [1] 0
```

```
df$Payment_type<-factor(df$Payment_type,labels=c("Credit card","Cash", "No charge", "Dispute"))
levels(df$Payment_type)[levels(df$Payment_type)=="No charge"] <- "Other"
levels(df$Payment_type)[levels(df$Payment_type)=="Dispute"] <- "Other"
table(df$Payment_type)
```

```
##
## Credit card        Cash       Other
##        2469        2485          46
```

```
barplot(prop.table(table(df$Payment_type)))
```



Trip_type

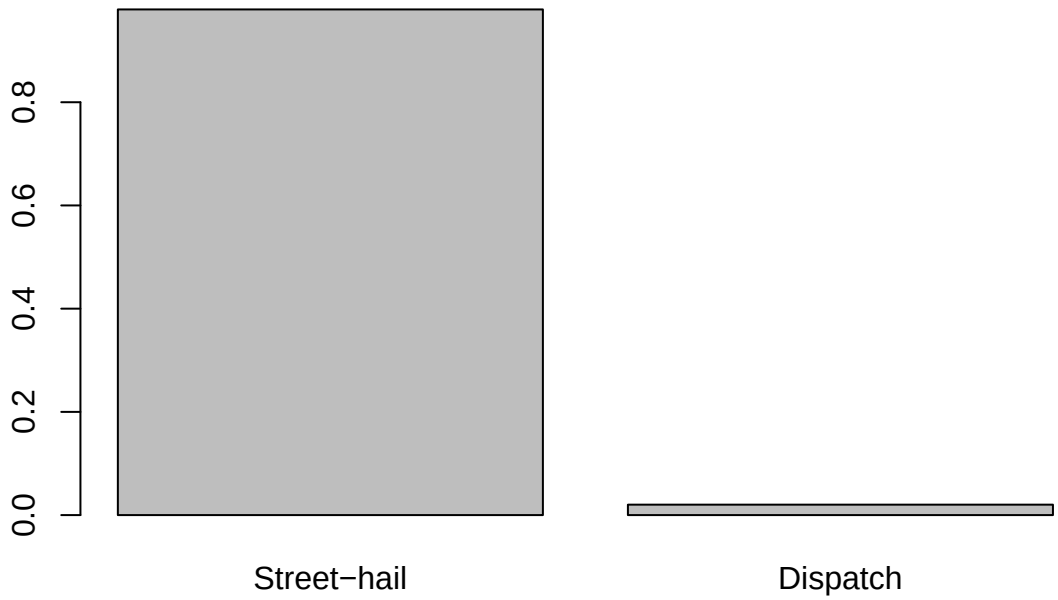
```
#No missing Data
missingData<-which(is.na(df$Trip_type));length(missingData)
```

```
## [1] 0
```

```
df$Trip_type<-factor(df$Trip_type,labels=c("Street-hail","Dispatch"))
table(df$Trip_type)
```

```
##
## Street-hail    Dispatch
##        4899         101
```

```
barplot(prop.table(table(df$Trip_type)))
```



Univariant Descriptive Analysis

Pickup_longitude

```
missingData<-which(is.na(df$Trip_distance));length(missingData) #No missing Data
```

```
## [1] 0
```

```
#min and max longitudes for New York city boundaries
min_long <- -74.15
max_long <- -73.7004

errors<-which(df$Pickup_longitude< min_long);length(errors)
```

```
## [1] 1
```

```
errors<-c(errors,which(df$Pickup_longitude> max_long));length(errors)
```

```
## [1] 7
```

```
errors<-c(errors,which(df$Pickup_longitude==0.0));length(errors)
```

```
## [1] 12
```

```
jerr[1] = jerr[1]+length(errors)
df[errors,"Pickup_longitude"]<-(-9999)
```

```r
ll<-which(df$Pickup_longitude == -9999);ll
```

```
## [1] 1580 1652 2639 3197 3221 4305 4639
```

```r
if(length(ll)>0){
  dfaux<-df[-ll,]
}

iqrvar<-IQR(dfaux$Pickup_longitude)
quantil3<-quantile(dfaux$Pickup_longitude, .75);quantil3 #get 3rd quartile
```

```
##       75%
## -73.91782
```

```r
quantil1<-quantile(dfaux$Pickup_longitude, .25);quantil1 #get 1st quartile
```

```
##       25%
## -73.96023
```

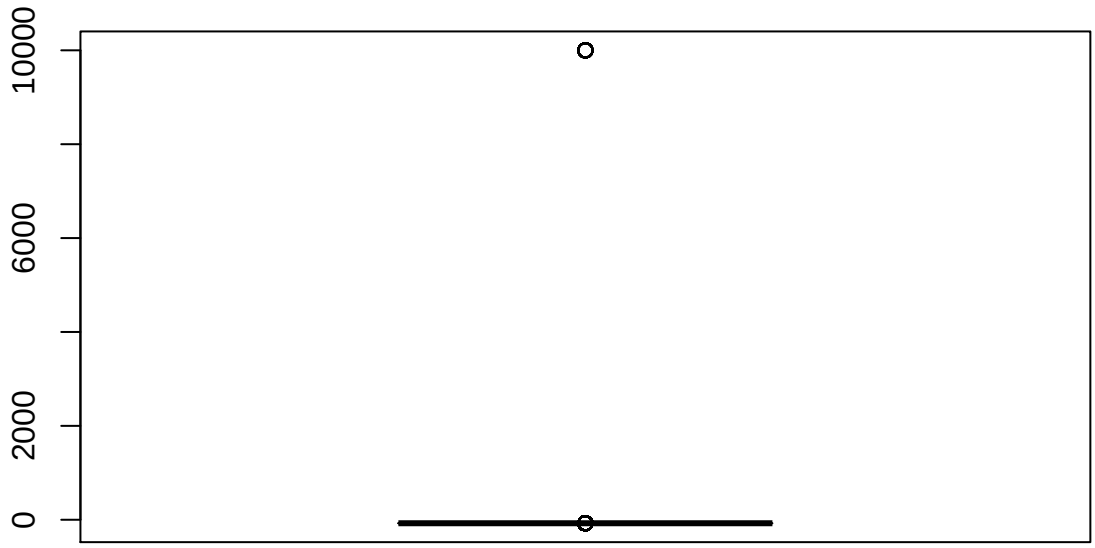
```r
UpperOutlier<-which(df$Pickup_longitude>quantil3+(iqrvar*3));length(UpperOutlier) #14 extreme UpperOutl
```

```
## [1] 14
```

```r
LowerOutlier<-which(df$Pickup_longitude<quantil1-(iqrvar*3));length(LowerOutlier) #1 extreme LowerOutli
```

```
## [1] 8
```

```r
df[UpperOutlier,"Pickup_longitude"]<- 9999
df[LowerOutlier,"Pickup_longitude"]<- 9999
boxplot(df$Pickup_longitude)
```




```
summary(df$Pickup_longitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -74.04  -73.96  -73.95  -29.61  -73.92 9999.00
```

```
jerr[1] = jerr[1]+length(errors)
jmis[1] = jmis[1]+length(missingData)
jouts[1] = jouts[1]+length(UpperOutlier)+length(LowerOutlier)
```

Pickup_latitude

```
missingData<-which(is.na(df$Pickup_latitude));length(missingData) #No missing Data
```

```
## [1] 0
```

```
#we need to add here error control (what if longitude is out of scope?) and outlier management

summary(df$Pickup_latitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   40.69   40.75   40.70   40.80   40.92
```

```
#min and max latitudes for New York city boundaries
min_lat <- 40.5774
max_lat <- 40.9176

errors<-which(df$Pickup_latitude< min_lat);length(errors)
```

```
## [1] 11
```

```
errors<-c(errors,which(df$Pickup_latitude> max_lat));length(errors)
```

```
## [1] 12
```

```
errors<-c(errors,which(df$Pickup_latitude==0.0));length(errors)
```

```
## [1] 17
```

```
df[errors,"Pickup_latitude"]<-(-9999) #17 errors

ll<-which(df$Pickup_latitude == -9999);ll
```

```
##  [1]  179 1580 2110 2241 2354 2639 2971 3197 3221 4305 4635 4639
```

```
if(length(ll)>0){
  dfaux<-df[-ll,]
}

iqrvar<-IQR(dfaux$Pickup_latitude)
quantil3<-quantile(dfaux$Pickup_latitude, .75);quantil3 #get 3rd quartile
```

```
##      75%
## 40.79892
```

```
quantil1<-quantile(dfaux$Pickup_latitude, .25);quantil1 #get 1st quartile
```

```
##      25%
## 40.69458
```

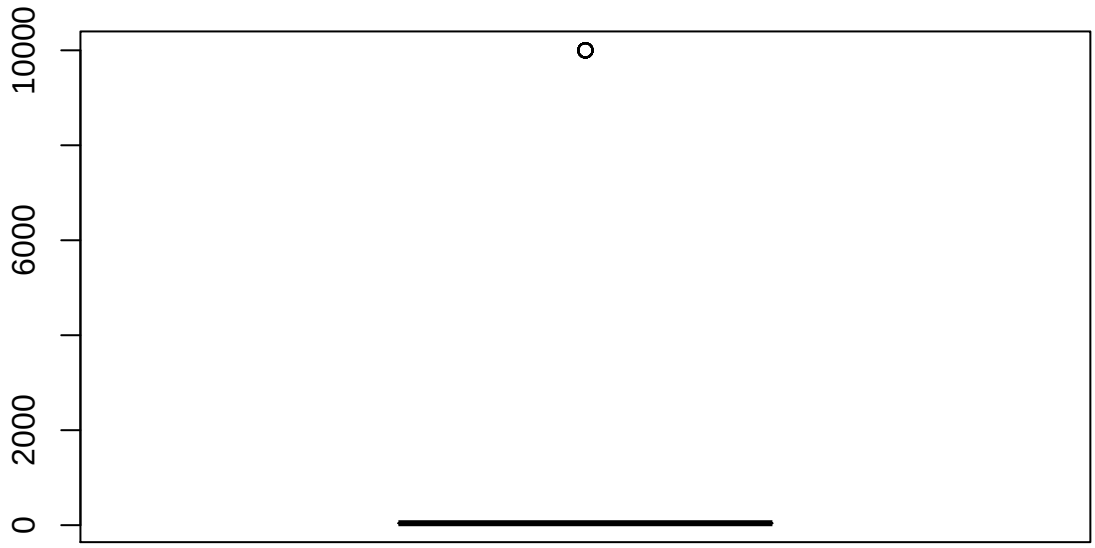
```
UpperOutlier<-which(df$Pickup_latitude>quantil3+(iqrvar*3));length(UpperOutlier) #0 extreme UpperOutlie
```

```
## [1] 0
```

```
LowerOutlier<-which(df$Pickup_latitude<quantil1-(iqrvar*3));length(LowerOutlier) #0 extreme LowerOutlie
```

```
## [1] 12
```

```
df[UpperOutlier,"Pickup_latitude"]<-9999
df[LowerOutlier,"Pickup_latitude"]<-9999
boxplot(df$Pickup_latitude)
```




```
summary(df$Pickup_latitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.58   40.70   40.75   64.65   40.80 9999.00
```

```
jerr[2] = jerr[2]+length(errors)
jmis[2] = jmis[2]+length(missingData)
jouts[2] = jouts[2]+length(UpperOutlier)+length(LowerOutlier)
```

Dropoff_longitude

```
missingData<-which(is.na(df$Dropoff_longitude));length(missingData) #No missing Data
```

```
## [1] 0
```

```
errors<-c(errors,which(df$Dropoff_longitude==0.0));length(errors) #26 errors
```

```
## [1] 26
```

```
df[errors,"Dropoff_longitude"]<-(-9999)

ll<-which(df$Dropoff_longitude == -9999);ll
```

```
##  [1]  179  638 1580 1713 1986 2026 2110 2241 2354 2639 2698 2971 3109 3197
## [15] 3221 4097 4285 4305 4635 4639
```

```r
if(length(ll)>0){
  dfaux<-df[-ll,]
}

iqrvar<-IQR(dfaux$Dropoff_longitude)
quantil3<-quantile(dfaux$Dropoff_longitude, .75);quantil3 #get 3rd quartile
```

```
##       75%
## -73.91151
```

```r
quantil1<-quantile(dfaux$Dropoff_longitude, .25);quantil1 #get 1st quartile
```

```
##      25%
## -73.9675
```

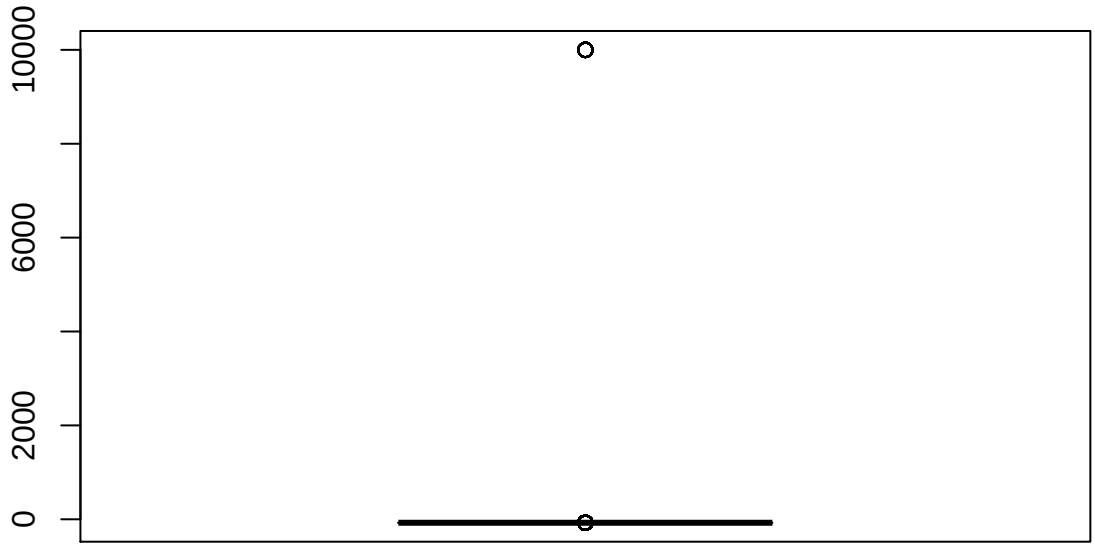
```r
UpperOutlier<-which(df$Dropoff_longitude>quantil3+(iqrvar*3));length(UpperOutlier) #0 extreme UpperOutl
```

```
## [1] 7
```

```r
LowerOutlier<-which(df$Dropoff_longitude<quantil1-(iqrvar*3));length(LowerOutlier) #0 extreme LowerOutl
```

```
## [1] 25
```

```r
df[UpperOutlier,"Dropoff_longitude"]<-9999
df[LowerOutlier,"Dropoff_longitude"]<-9999

boxplot(df$Dropoff_longitude)
```



```
jerr[3] = jerr[3]+length(errors)
jmis[3] = jmis[3]+length(missingData)
jouts[3] = jouts[3]+length(UpperOutlier)+length(LowerOutlier)
```

Dropoff_latitude

```
missingData<-which(is.na(df$Dropoff_latitude));length(missingData) #No missing Data
```

```
## [1] 0
```

```
errors<-c(errors,which(df$Dropoff_latitude==0.0));length(errors) #35 errors
```

```
## [1] 35
```

```
df[errors,"Dropoff_latitude"]<-(-9999)

ll<-which(df$Dropoff_latitude == -9999);ll
```

```
##  [1]  179  638 1580 1713 1986 2026 2110 2241 2354 2639 2698 2971 3109 3197
## [15] 3221 4097 4285 4305 4635 4639
```

```
if(length(ll)>0){
  dfaux<-df[-ll,]
}

iqrvar<-IQR(dfaux$Dropoff_latitude)
quantil3<-quantile(dfaux$Dropoff_latitude, .75);quantil3 #get 3rd quartile
```

```
##      75%
## 40.78581
```

```
quantil1<-quantile(dfaux$Dropoff_latitude, .25);quantil1 #get 1st quartile
```

```
##      25%
## 40.69629
```

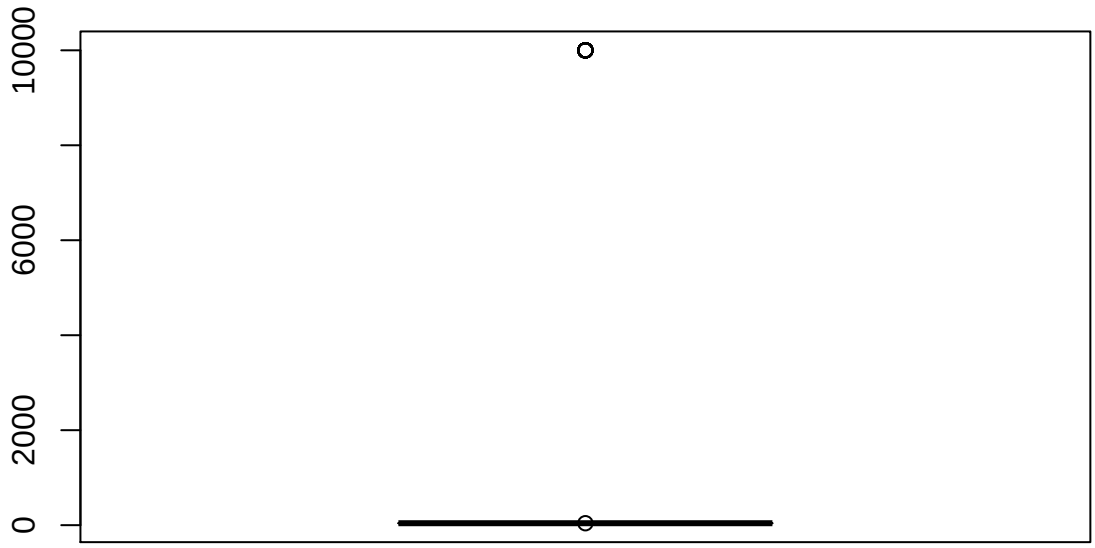
```
UpperOutlier<-which(df$Dropoff_latitude>quantil3+(iqrvar*3));length(UpperOutlier) #0 extreme UpperOutli
```

```
## [1] 0
```

```
LowerOutlier<-which(df$Dropoff_latitude<quantil1-(iqrvar*3));length(LowerOutlier) #0 extreme LowerOutli
```

```
## [1] 20
```

```
df[UpperOutlier,"Dropoff_latitude"]<-9999
df[LowerOutlier,"Dropoff_latitude"]<-9999

boxplot(df$Dropoff_latitude)
```




```
jerr[4] = jerr[4]+length(errors)
jmis[4] = jmis[4]+length(missingData)
jouts[4] = jouts[4]+length(UpperOutlier)+length(LowerOutlier)
```

Passenger_count

```
missingData<-which(is.na(df$Passenger_count));length(missingData) #No missing Data
```

```
## [1] 0
```

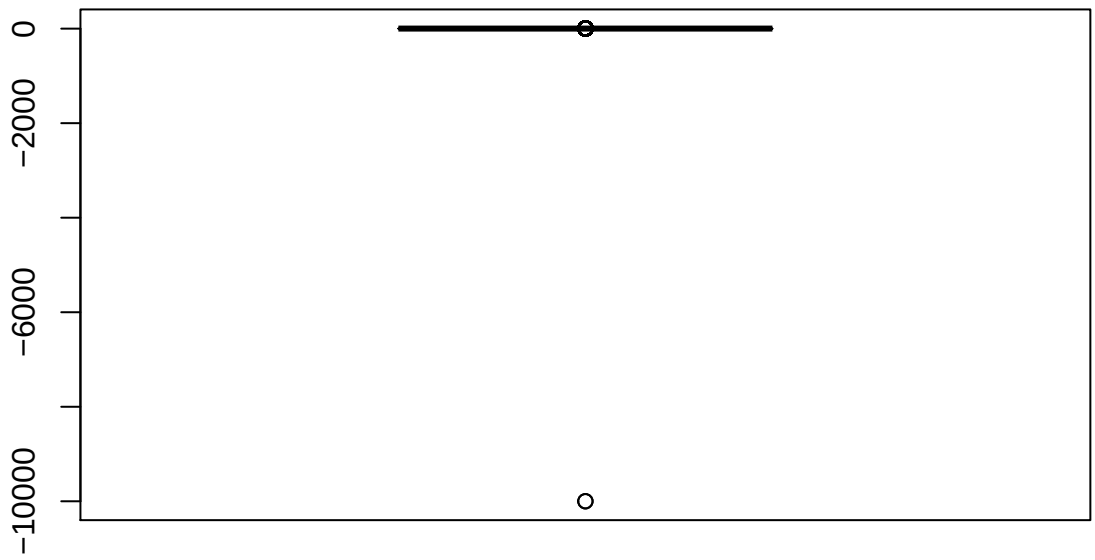
```
errors<-which(df$Passenger_count<=0.0);length(errors) #2 errors
```

```
## [1] 2
```

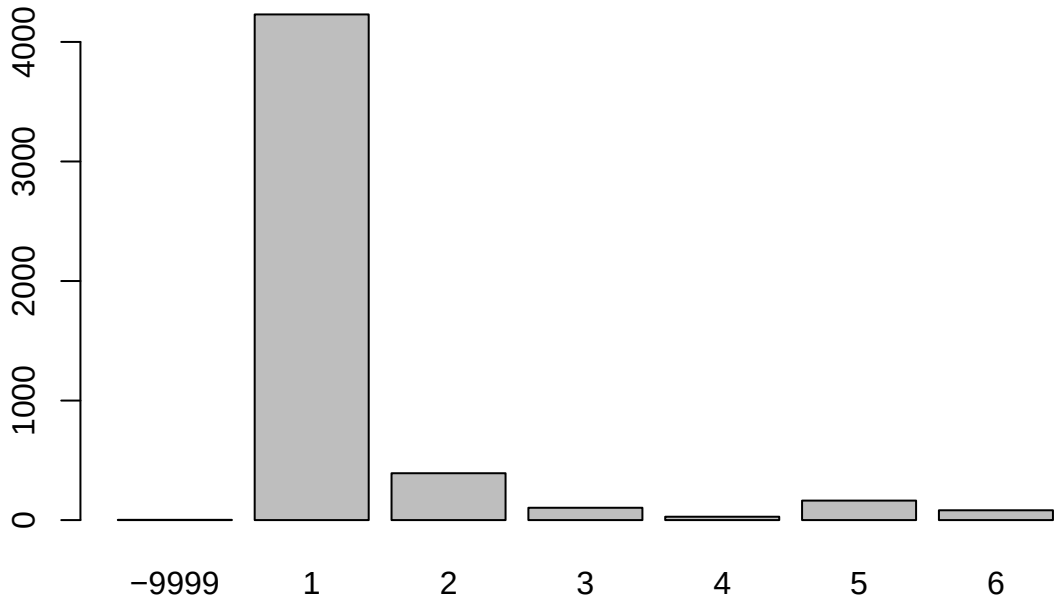
```
outliers<-which(df$Passenger_count>6.0);length(outliers) #0 outlier
```

```
## [1] 0
```

```
df[errors,"Passenger_count"]<-(-9999)
df[outliers,"Passenger_count"]<-9999
boxplot(df$Passenger_count)
```




```
barplot(table(df$Passenger_count))
```




```
jerr[5] = jerr[5]+length(errors)
jmis[5] = jmis[5]+length(missingData)
jouts[5] = jouts[5]+length(outliers)
```

Trip_distance

We distinguish as an outlier for Trip_distance those elements which value over 20, since we believe it is big enough to our study case and we can not be too severes applying Interquartile Range (we keep a long right queue distribution).

```
missingData<-which(is.na(df$Trip_distance));length(missingData) #No missing Data
```

```
## [1] 0
```

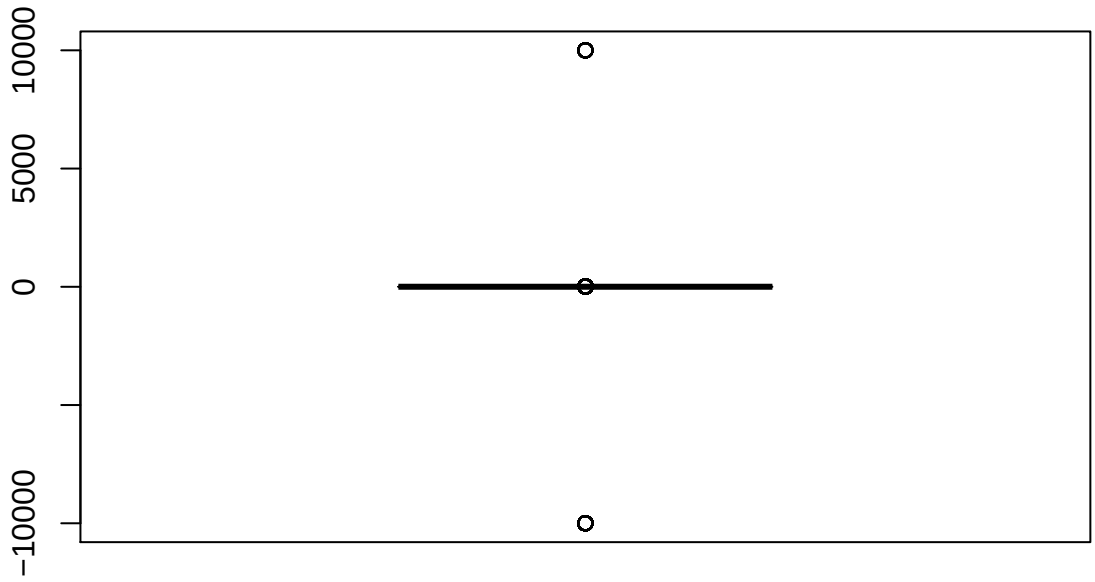
```
errors<-which(df$Trip_distance<=0.0);length(errors) #59 errors
```

```
## [1] 59
```

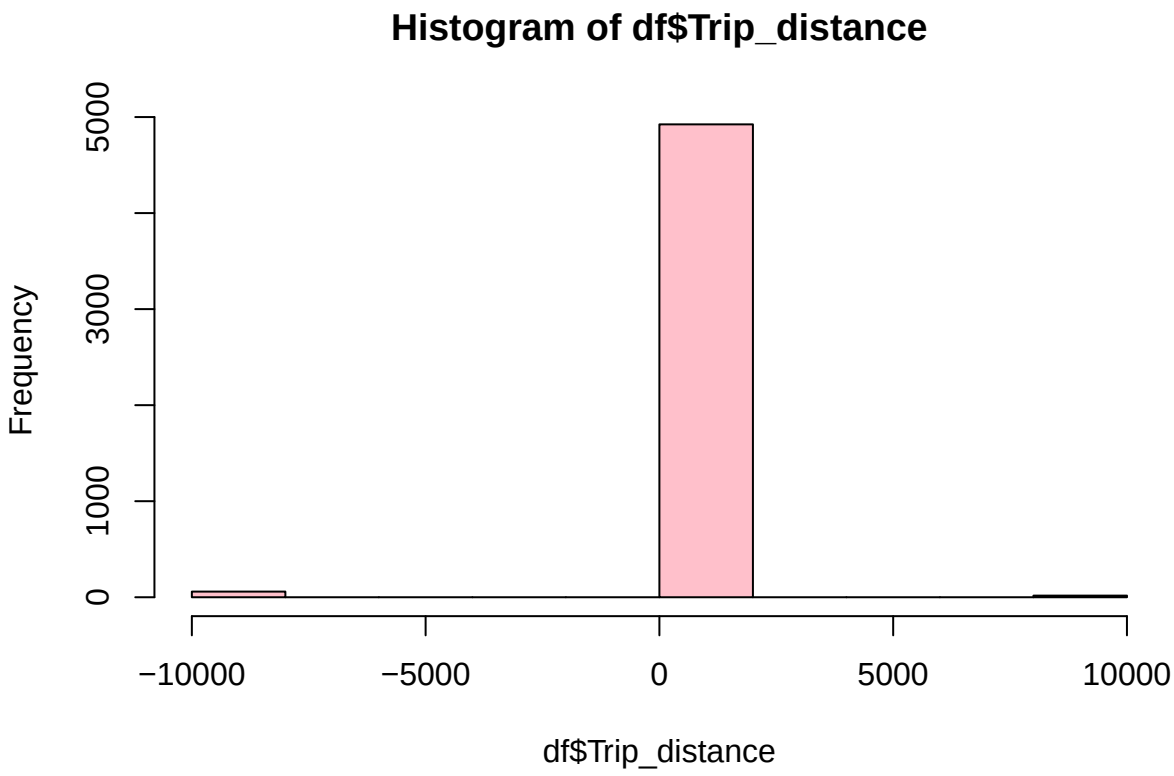
```
outliers <- which(df$Trip_distance>20.0);length(outliers) #17 outlier
```

```
## [1] 17
```

```
df[outliers,"Trip_distance"]<-9999
df[errors,"Trip_distance"]<-(-9999)
boxplot(df$Trip_distance)
```




```
hist(df$Trip_distance, col="pink")
```




```
summary(df$Trip_distance)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -9999.000     1.000     1.810   -81.272     3.413  9999.000
```

```
jerr[6] = jerr[6]+length(errors)
jmis[6] = jmis[6]+length(missingData)
jouts[6] = jouts[6]+length(outliers)
```

Fare_amount

We distinguish as an outlier for Fare_amount those elements which value over 60, since we believe it is big enough to our study case and we can not be too severes applying Interquartile Range (we keep a long right queue distribution).

```
missingData<-which(is.na(df$Fare_amount));length(missingData) #No missing Data
```

```
## [1] 0
```

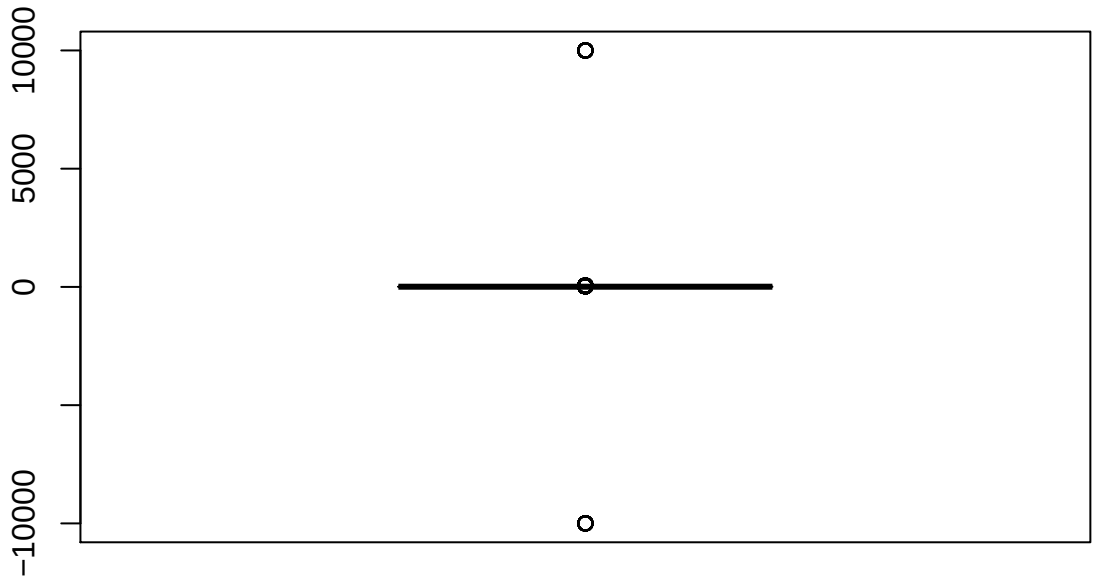
```
#23 errors
sel<-which(df$Fare_amount<=0.0);length(sel)
```

```
## [1] 23
```

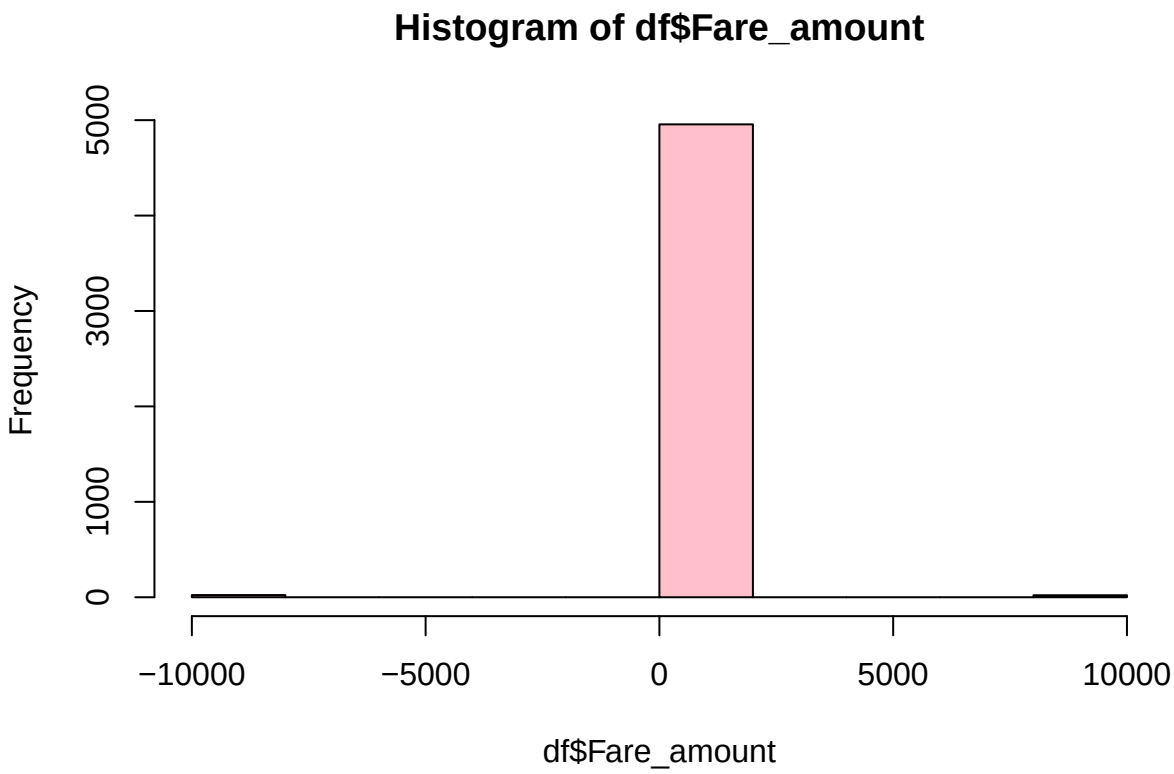
```
#21 outlier
outlier<-which(df$Fare_amount>60);length(outlier)
```

```
## [1] 21
```

```
df[sel,"Fare_amount"]<-(-9999)
df[outlier,"Fare_amount"]<-9999
boxplot(df$Fare_amount)
```




```
hist(df$Fare_amount, col="pink")
```




```
summary(df$Fare_amount)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -9999.000     6.000     9.000     7.667    14.500  9999.000
```

```
jerr[7] = jerr[7]+length(sel)
jmis[7] = jmis[7]+length(missingData)
jouts[7] = jouts[7]+length(outlier)
```

Extra

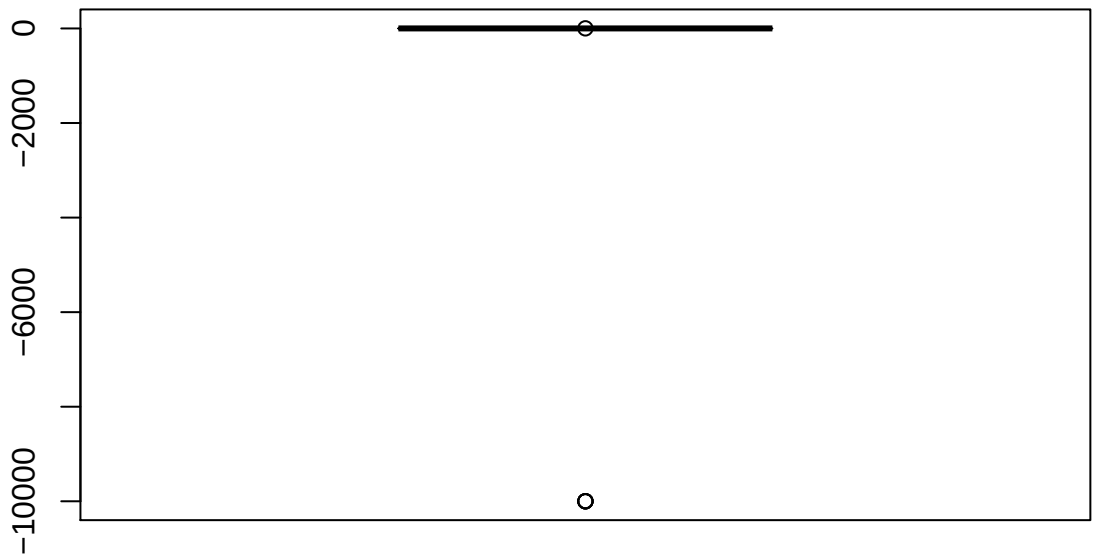
```
missingData<-which(is.na(df$Extra));length(missingData) #No missing Data
```

```
## [1] 0
```

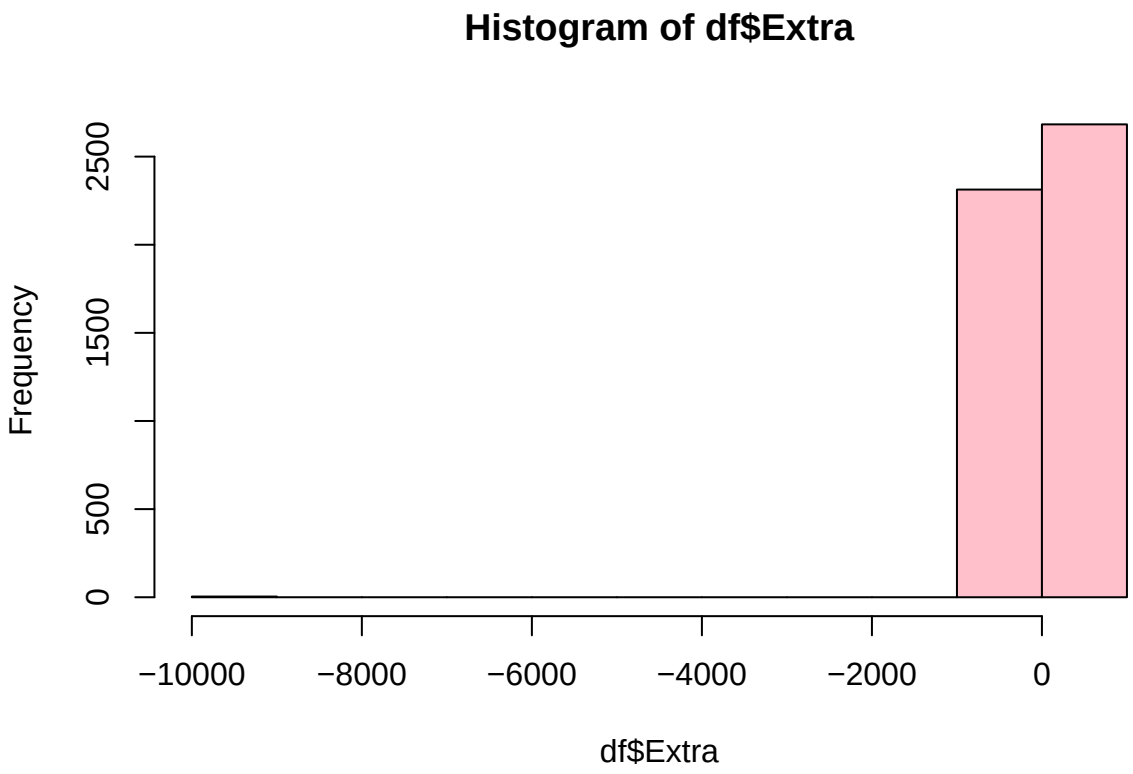
```
sel<-which(df$Extra<0.0);length(sel) #4 error
```

```
## [1] 4
```

```
df[sel,"Extra"]<-(-9999)
boxplot(df$Extra)
```




```
hist(df$Extra, col="pink")
```




```
jerr[8] = jerr[8]+length(sel)
jmis[8] = jmis[8]+length(missingData)
```

MTA_tax

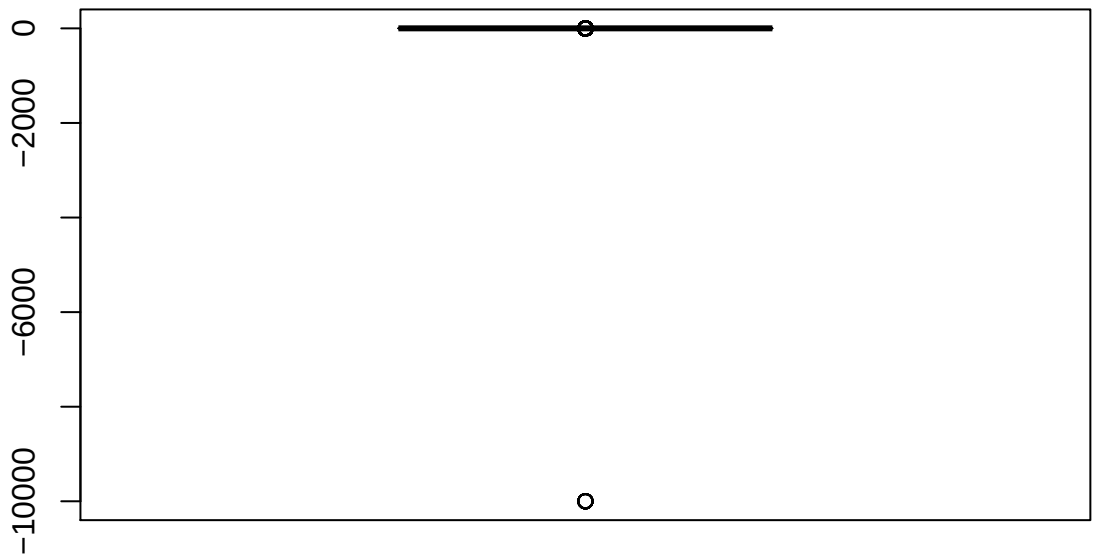
```
missingData<-which(is.na(df$MTA_tax));length(missingData) #No missing Data
```

```
## [1] 0
```

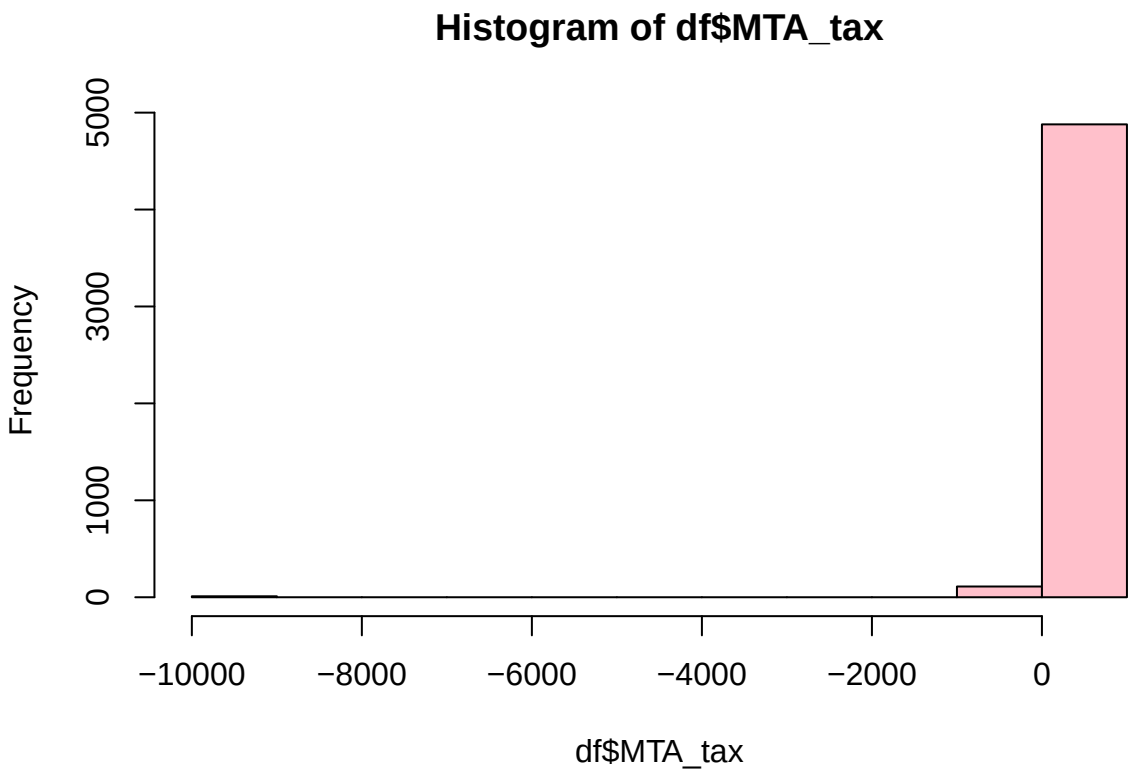
```
sel<-which(df$MTA_tax<0.0);length(sel) #10 error
```

```
## [1] 10
```

```
df[sel,"MTA_tax"]<-(-9999)
boxplot(df$MTA_tax)
```




```
hist(df$MTA_tax, col="pink")
```




```
jerr[9] = jerr[9]+length(sel)
jmis[9] = jmis[9]+length(missingData)
```

Tip_amount

We distinguish as an outlier for Tip_amount those elements which value over 25, since we believe it is big enough to our study case and we can not be too severes applying Interquartile Range (we keep a long right queue distribution).

```
missingData<-which(is.na(df$Tip_amount));length(missingData) #No missing Data
```

```
## [1] 0
```

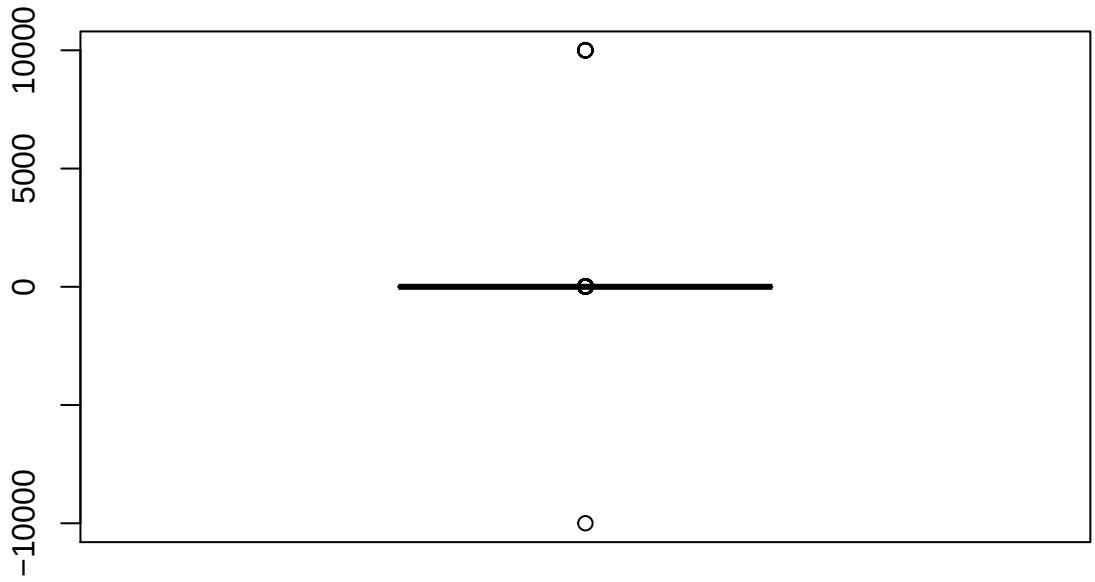
```
sel<-which(df$Tip_amount<0.0);length(sel) #1 error
```

```
## [1] 1
```

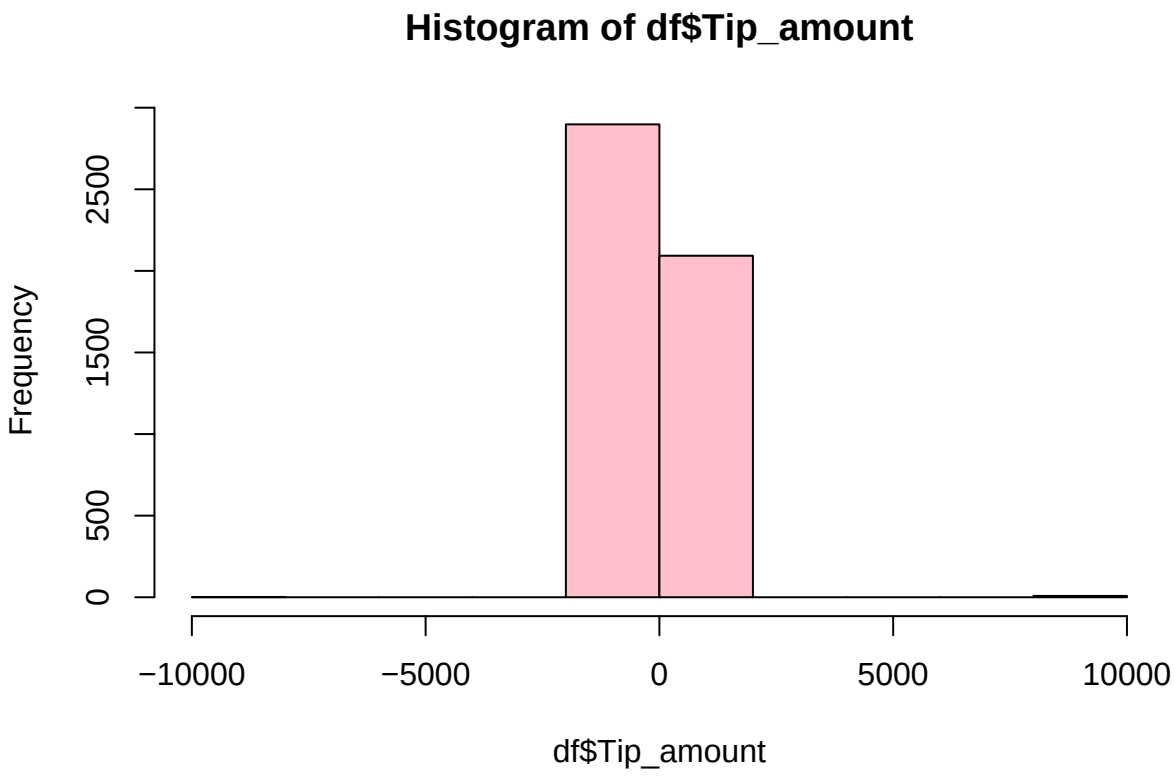
```
outlier<-which(df$Tip_amount>25.0);length(outlier) #8 outliers
```

```
## [1] 8
```

```
df[outlier,"Tip_amount"]<-9999
df[sel,"Tip_amount"]<-(-9999)
boxplot(df$Tip_amount)
```




```
hist(df$Tip_amount, col="pink")
```




```
jerr[10] = jerr[10]+length(sel)
jmis[10] = jmis[10]+length(missingData)
jouts[10] = jouts[10]+length(outlier)
```

Tolls_amount

We don't get rid of any row for this variable because of its own nature. We are lack of criteria to detect what would be an outlier and it does not seem to achieve very high values in its distribution (histogram plot).

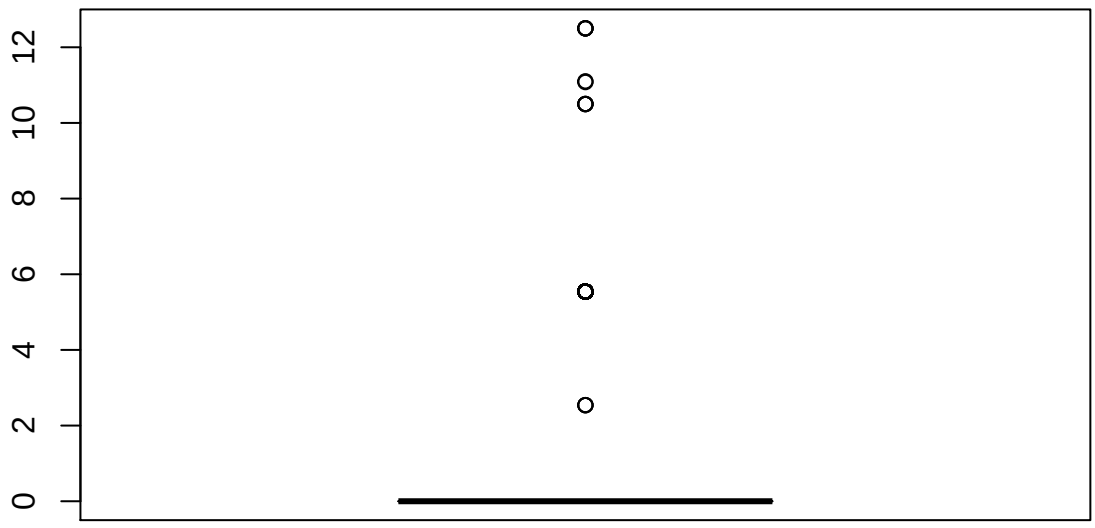
```
missingData<-which(is.na(df$Tolls_amount));length(missingData) #No missing Data
```

```
## [1] 0
```

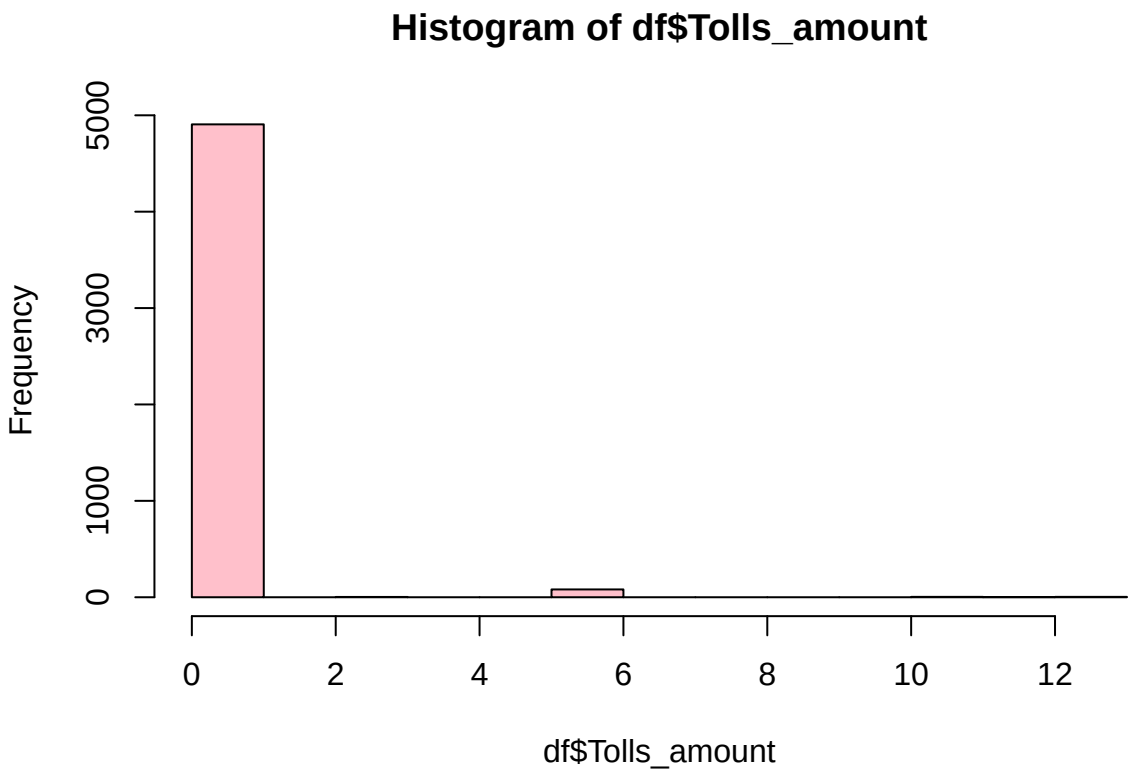
```
sel<-which(df$Tolls_amount<0.0);length(sel) #0 errors
```

```
## [1] 0
```

```
df[sel,"Tolls_amount"]<-(-9999)
boxplot(df$Tolls_amount)
```




```r
hist(df$Tolls_amount, col="pink")
```




```
summary(df$Tolls_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1141  0.0000 12.5000
```

```
jerr[11] = jerr[11]+length(sel)
jmis[11] = jmis[11]+length(missingData)
```

Improvement_surcharge

```
missingData<-which(is.na(df$improvement_surcharge));length(missingData) #No missing Data
```

```
## [1] 0
```

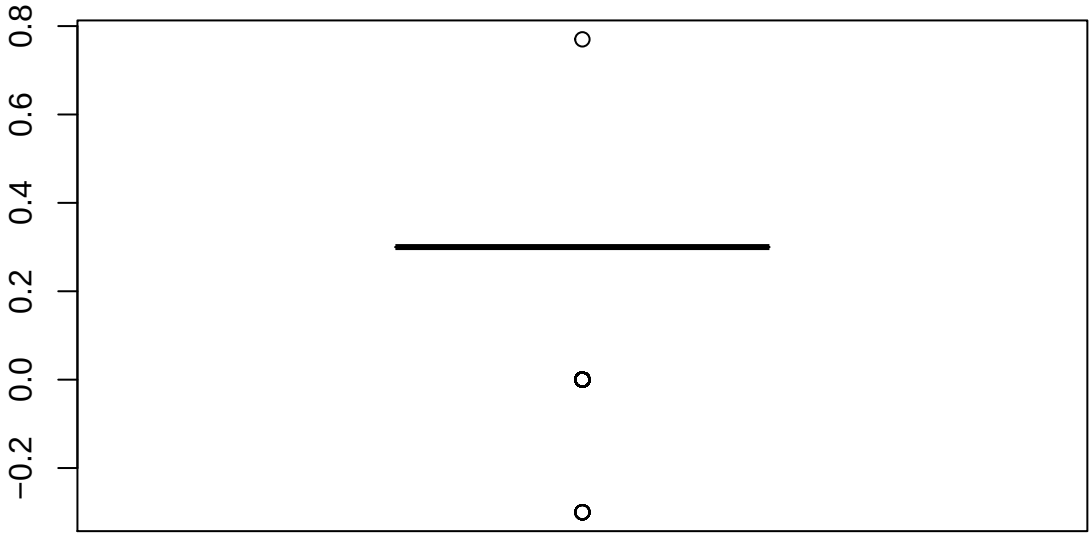
```
sel<-which(df$improvement_surcharge<0.0);length(sel) #10 errors
```

```
## [1] 10
```

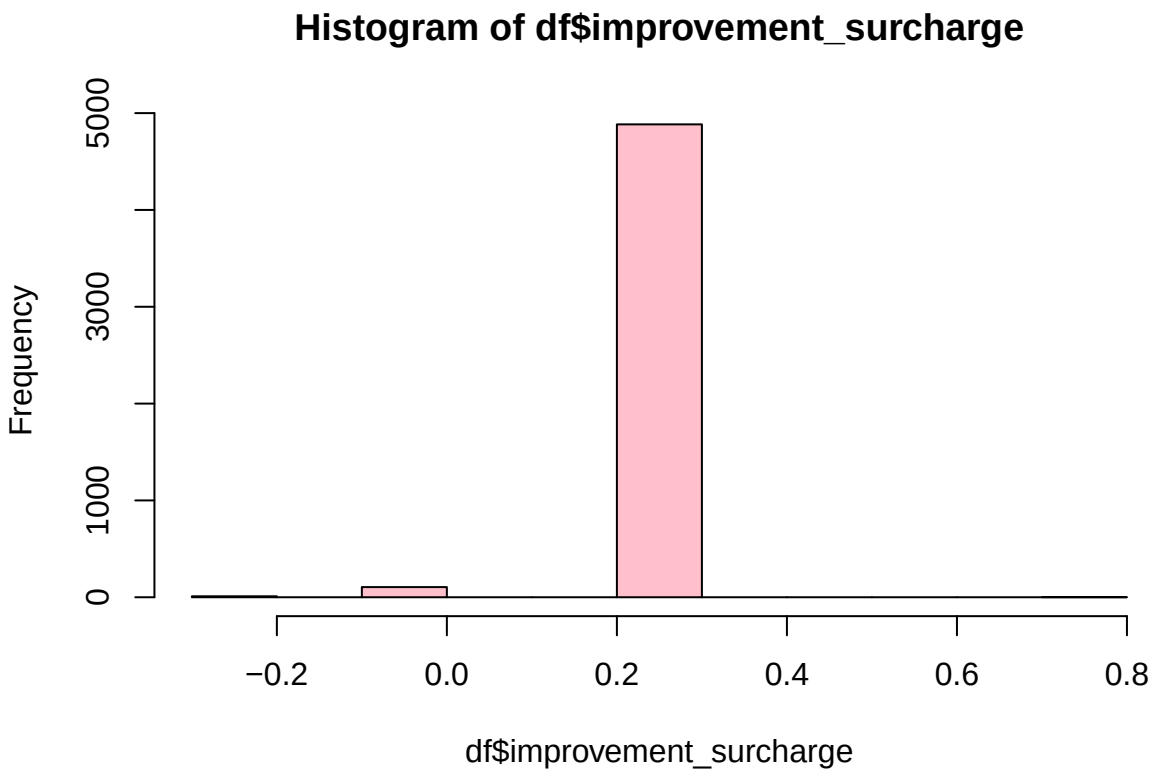
```
df[sel,"improvement_surcharge"]<(-9999)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
boxplot(df$improvement_surcharge)
```




```
hist(df$improvement_surcharge, col="pink")
```




```
jerr[12] = jerr[12]+length(sel)
jmis[12] = jmis[12]+length(missingData)
```

Total_amount (Target)

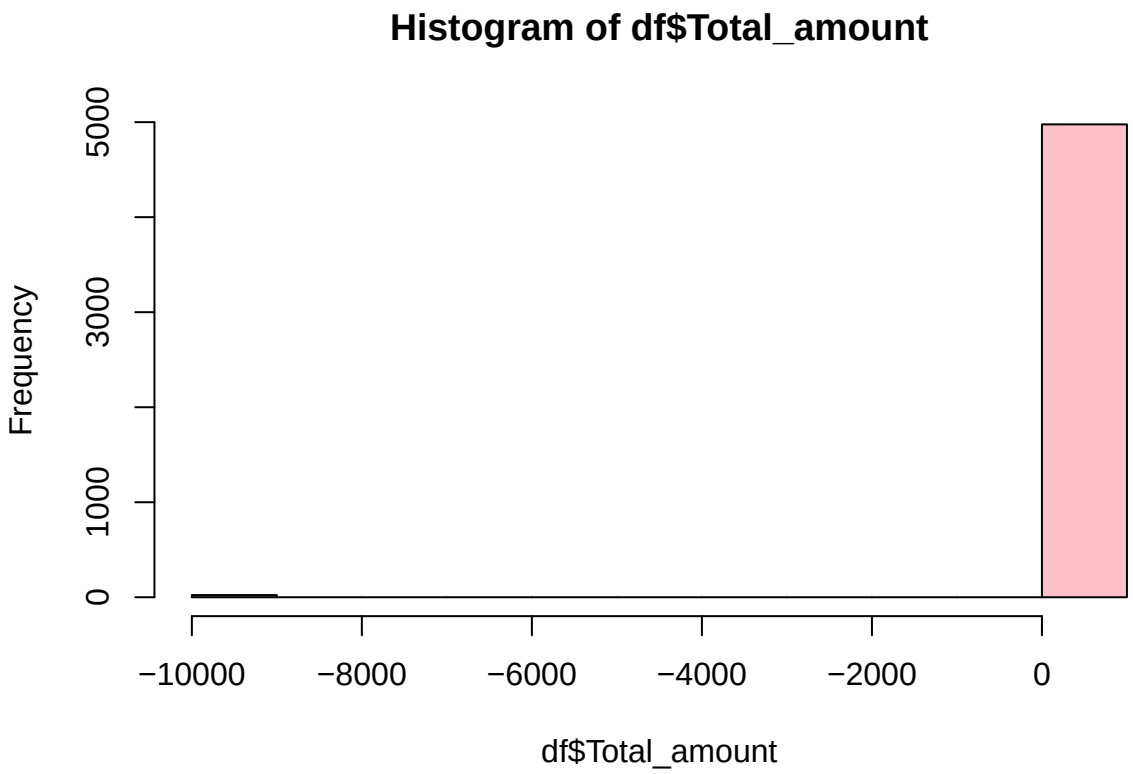
```
missingData<-which(is.na(df$Total_amount));length(missingData) #No missing Data
```

```
## [1] 0
```

```
sel<-which(df$Total_amount<=0.0);length(sel) #23 errors
```

```
## [1] 23
```

```
df[sel,"Total_amount"]<-(-9999)

hist(df$Total_amount, col="pink")
```




```
iqrvar<-IQR(dfaux$Total_amount)
quantil3<-quantile(dfaux$Total_amount, .75) #get 3rd quartile
severePoint<-(iqrvar*3)+quantil3; severePoint; #it's too severe, we won't use this method
```

```
##  75%
## 45.6
```

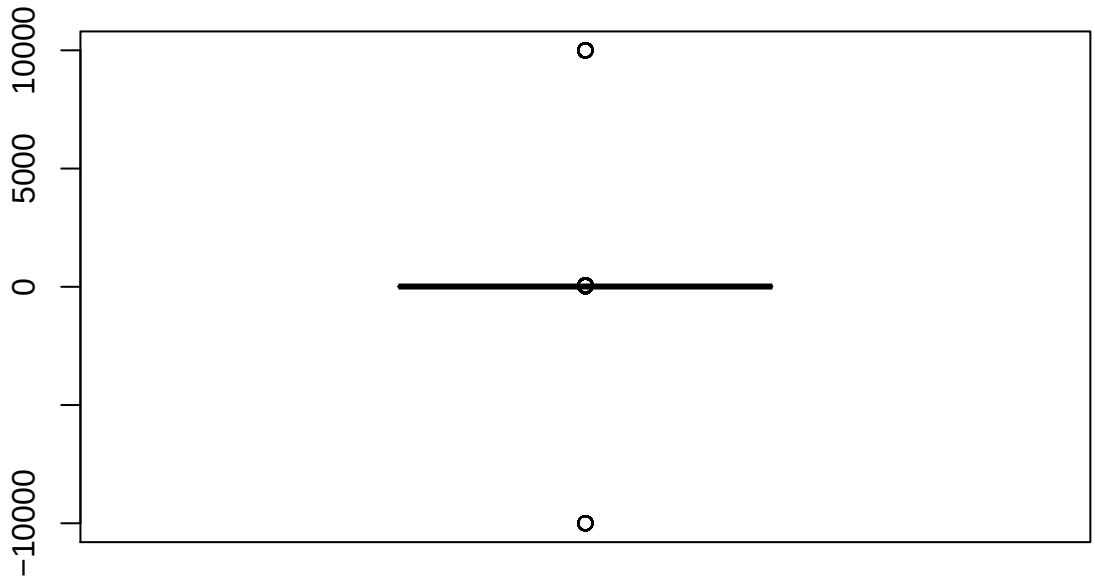
```
outlier<- which(df$Total_amount>severePoint);length(outlier) #111 extreme outliers
```

```
## [1] 111
```

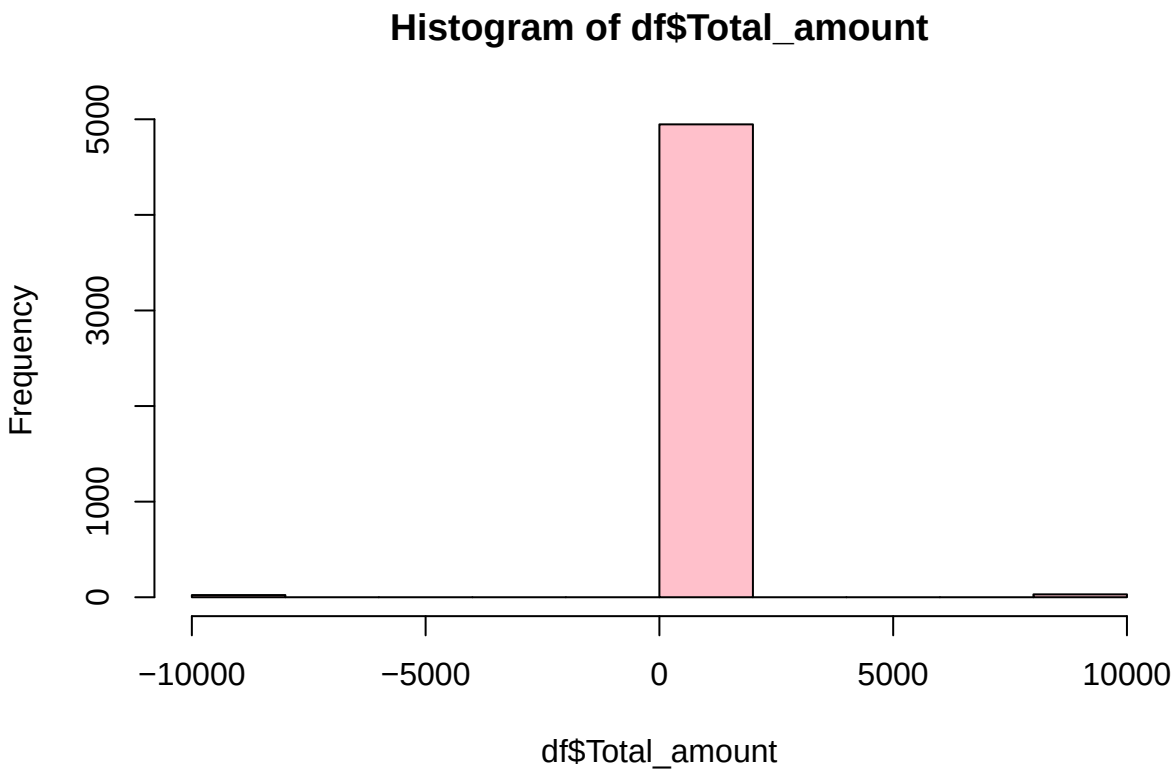
```
outlier<- which(df$Total_amount>70);length(outlier) #30 extreme outliers
```

```
## [1] 30
```

```
df[outlier,"Total_amount"]<-9999
boxplot(df$Total_amount)
```




```
hist(df$Total_amount, col="pink")
```




```
jerr[13] = jerr[13]+length(sel)
jmis[13] = jmis[13]+length(missingData)
jouts[13] = jouts[13]+length(outlier)
```

Data Quality report

```
#Variables
cols <- vars_data_qual

dqT <- data.frame(cols, jerr, jmis, jouts)
#print
print(dqT[order(dqT$jerr, decreasing = TRUE),], row.names = FALSE)
```

```
##                   cols jerr jmis jouts
##          Trip_distance   59    0    17
##       Dropoff_latitude   35    0    20
##      Dropoff_longitude   26    0    32
##       Pickup_longitude   24    0    22
##            Fare_amount   23    0    21
##           Total_amount   23    0    30
##        Pickup_latitude   17    0    12
##                MTA_tax   10    0     0
##  improvement_surcharge   10    0     0
##                  Extra    4    0     0
```

```
##        Passenger_count    2    0     0
##             Tip_amount    1    0     8
##           Tolls_amount    0    0     0
```

```
#Individuals Outlier
outQuality<-countX(df, 9999)
attributes(outQuality)
```

```
## $names
## [1] "nx_col" "nx_ind"
```

```
df$out_ind <- outQuality$nx_ind
sort(outQuality$nx_col$n_x,decreasing = TRUE)
```

```
##  [1] 32 30 22 21 20 17 12  8  0  0  0  0  0  0  0  0  0  0  0  0
```

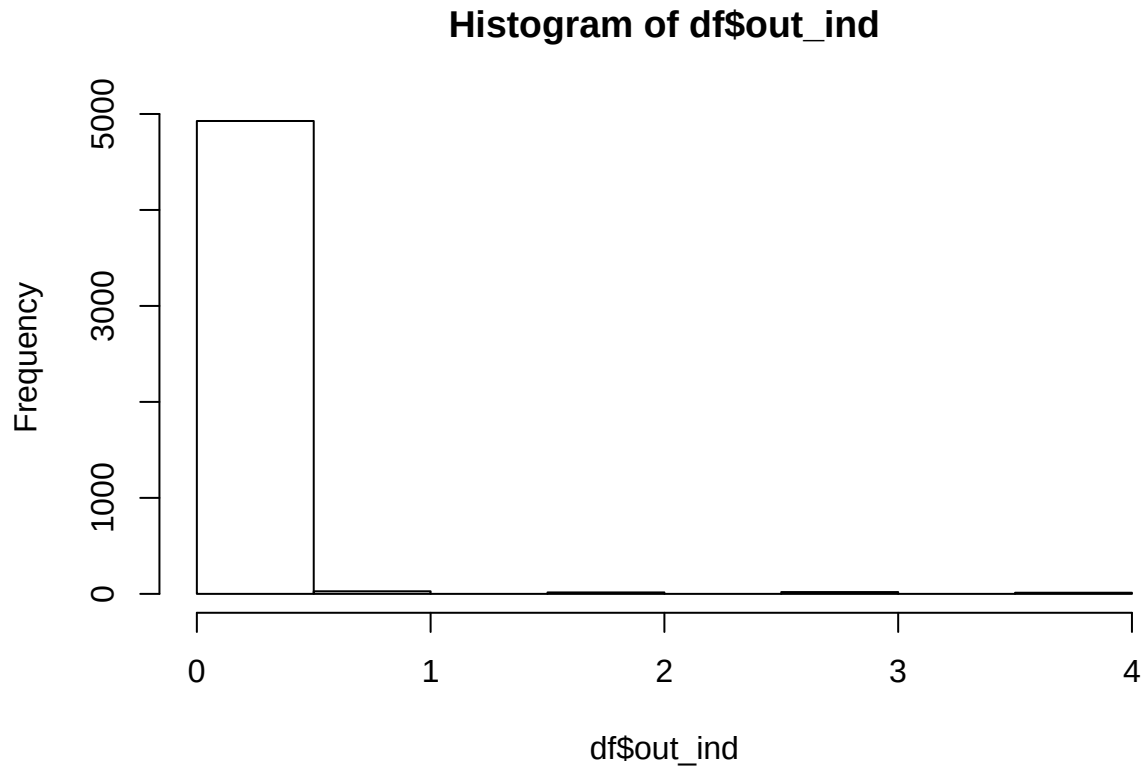
```
outQuality$mis_col
```

```
## NULL
```

```
summary(df$out_ind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.0324  0.0000  4.0000
```

```
hist(df$out_ind)
```




```
f.out<-factor(cut(df$out_ind,breaks=c(-1,0,1,2,3,4)))
summary(f.out)
```

```
## (-1,0]  (0,1]  (1,2]  (2,3]  (3,4]
##   4927     27     15     19     12
```

```
#Individuals Error
errQuality<-countX(df, -9999)
attributes(errQuality)
```

```
## $names
## [1] "nx_col" "nx_ind"
```

```
df$err_ind <- errQuality$nx_ind
sort(outQuality$nx_col$n_x,decreasing = TRUE)
```

```
##  [1] 32 30 22 21 20 17 12  8  0  0  0  0  0  0  0  0  0  0  0  0
```

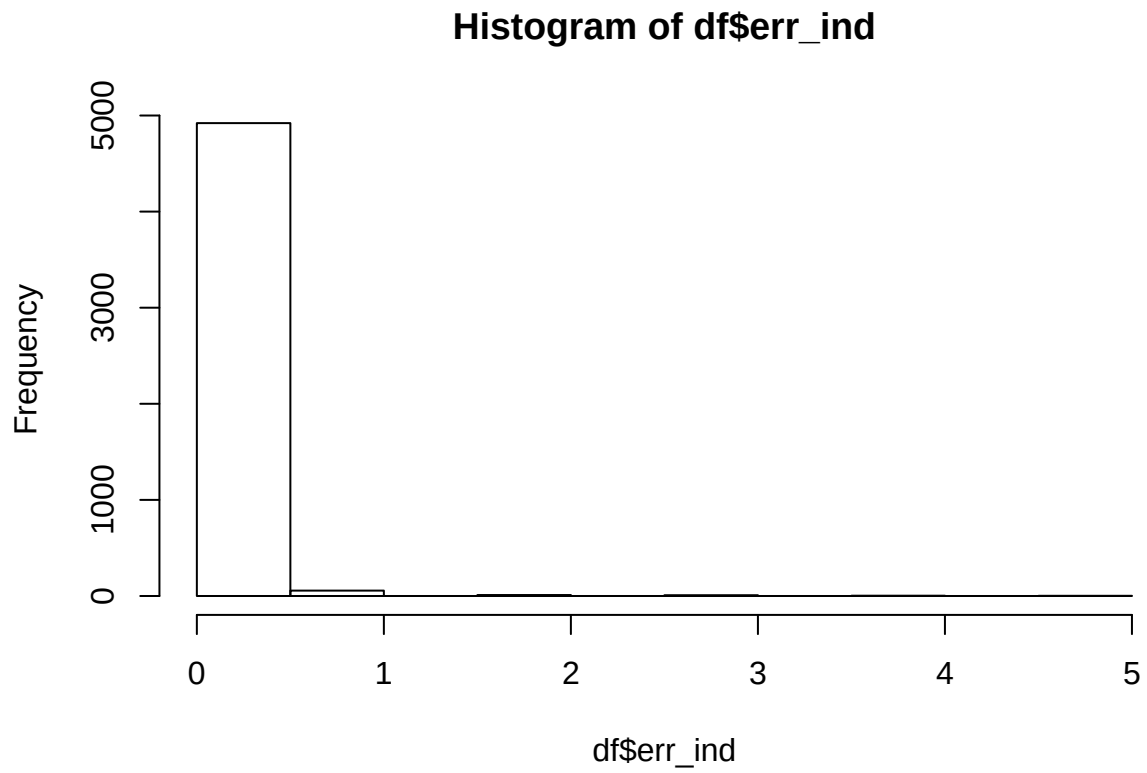
```
outQuality$mis_col
```

```
## NULL
```

```
summary(df$err_ind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.0244  0.0000  5.0000
```

```
hist(df$err_ind)
```




```
f.err<-factor(cut(df$err_ind,breaks=c(-1,0,1,2,3,4,5,6,7,8,9)))
summary(f.err)
```

```
## (-1,0]  (0,1]  (1,2]  (2,3]  (3,4]  (4,5]
```

```
##   4921     56     10      8      3      2
```

```
#No missing Data
mis1<-countNA(df)
summary(mis1$mis_ind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

Travel time in minutes

```
b1<-as.POSIXlt(df$lpep_pickup_datetime)
b2<-as.POSIXlt(df$Lpep_dropoff_datetime)
df$travel_time<-as.double(difftime(b2,b1,units='min'))
error<-which(df$travel_time< 1.0);length(error) #No errors
```

```
## [1] 90
```

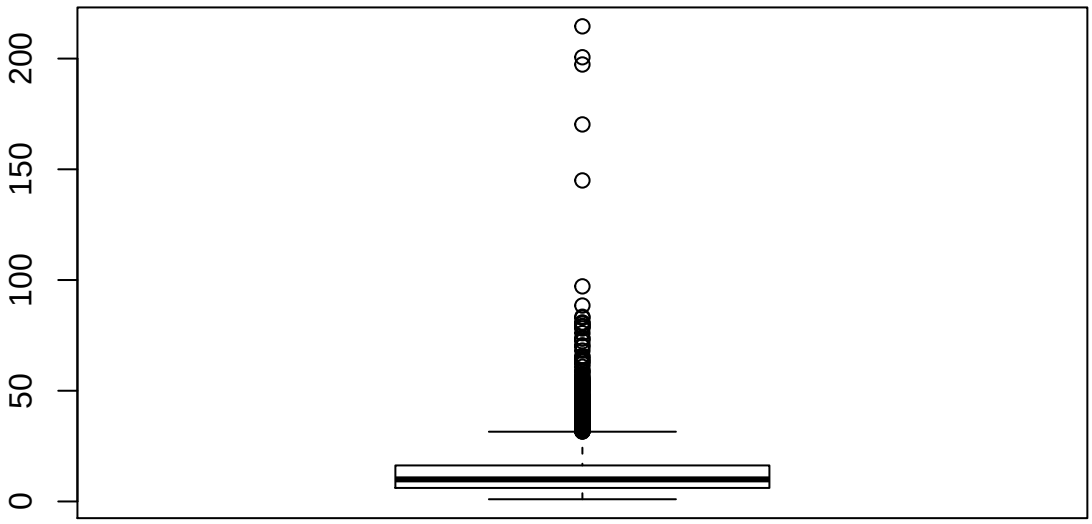
```
df[error,"travel_time"]<-NA
outlier<-which(df$travel_time>400.0);length(outlier) #
```

```
## [1] 29
```

```
df[outlier,"travel_time"]<-NA
summary(df$travel_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.100   9.967  12.944  16.267 214.550     119
```

```
boxplot(df$travel_time)
```

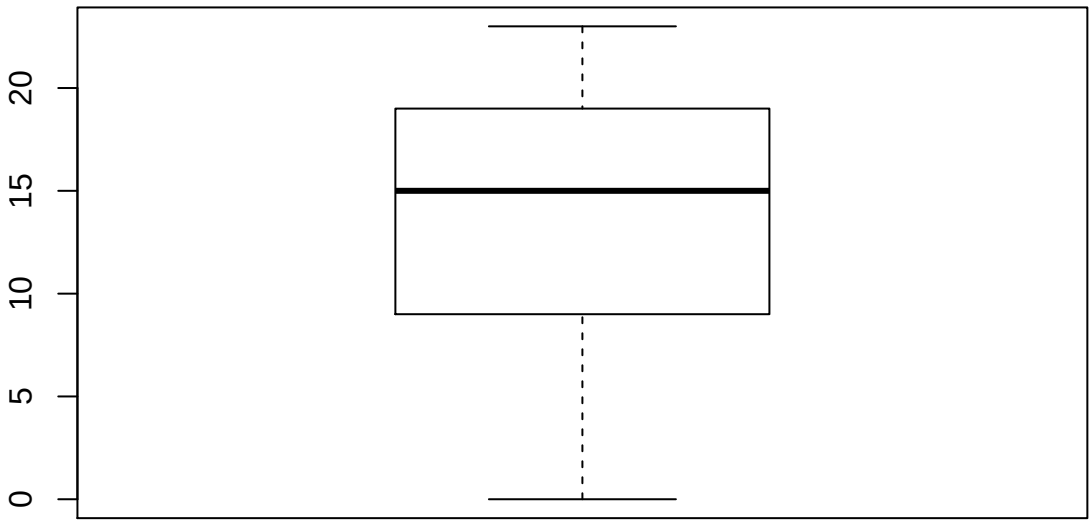


Pick__up__hour

```
mydate <- as.POSIXlt(df$lpep_pickup_datetime)
df$pick_up_hour <- mydate$hour

summary(df$pick_up_hour)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   15.00   13.47   19.00   23.00
```

```
boxplot(df$pick_up_hour)
```



Pick__up__period

```
# night, morning, valley and afternoon

df$pick_up_period= cut(df$pick_up_hour, breaks = c(-1, 5, 11, 17, 23), labels= c("night", "morning", "va

summary(df$pick_up_period)
```

```
##     night   morning    valley afternoon
##       807      1011      1411      1771
```

Subtitute outlier codes for NA

```
library(naniar)
df <- replace_with_na_all(df, condition = ~.x == -9999)
df <- replace_with_na_all(df, condition = ~.x == 9999)
df <- as.data.frame(df)
```

Declaring vectors of data

```
names(df)
```

```
##  [1] "VendorID"              "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime" "Store_and_fwd_flag"
##  [5] "RateCodeID"            "Pickup_longitude"
##  [7] "Pickup_latitude"       "Dropoff_longitude"
##  [9] "Dropoff_latitude"      "Passenger_count"
## [11] "Trip_distance"         "Fare_amount"
## [13] "Extra"                 "MTA_tax"
## [15] "Tip_amount"            "Tolls_amount"
## [17] "improvement_surcharge" "Total_amount"
## [19] "Payment_type"          "Trip_type"
## [21] "out_ind"               "err_ind"
## [23] "travel_time"           "pick_up_hour"
## [25] "pick_up_period"
```

```
vars_con<-names(df)[c(6:9, 11,12,15,16,23)] #24:26,29
vars_con
```

```
## [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
## [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
## [7] "Tip_amount"        "Tolls_amount"      "travel_time"
```

```
vars_dis<-names(df)[c(1,4,5,10,13,14,17,19)]#,20,27,28,30:44
vars_dis
```

```
## [1] "VendorID"              "Store_and_fwd_flag"    "RateCodeID"
## [4] "Passenger_count"       "Extra"                 "MTA_tax"
## [7] "improvement_surcharge" "Payment_type"
```

```
vars_res<-names(df)[c(18)]#,23
vars_res
```

```
## [1] "Total_amount"
```

Imputation

Remove observations with NA at targets

Getting rows coded

```
ll<-which(is.na(df$Passenger_count));length(ll)
```

```
## [1] 2
```

```
if(length(ll)>0){
  df<-df[-ll,]
}

ll<-which(is.na(df$Total_amount));length(ll)
```

```
## [1] 52
```

```
if(length(ll)>0){
  df<-df[-ll,]
}

# ll<-which(is.na(df$AnyTip));length(ll)
# if(length(ll)>0){
#   df<-df[-ll,]
# }
```

Imputation of numeric variables

```
library(missMDA)
names(df)
```

```
##  [1] "VendorID"              "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime" "Store_and_fwd_flag"
##  [5] "RateCodeID"            "Pickup_longitude"
##  [7] "Pickup_latitude"       "Dropoff_longitude"
##  [9] "Dropoff_latitude"      "Passenger_count"
## [11] "Trip_distance"         "Fare_amount"
## [13] "Extra"                 "MTA_tax"
## [15] "Tip_amount"            "Tolls_amount"
## [17] "improvement_surcharge" "Total_amount"
## [19] "Payment_type"          "Trip_type"
## [21] "out_ind"               "err_ind"
## [23] "travel_time"           "pick_up_hour"
## [25] "pick_up_period"
```

```
res.comp <- imputePCA(df[,vars_con], ncp=4)
df[,"Pickup_longitude"]<-res.comp$completeObs[,"Pickup_longitude"]
df[,"Pickup_latitude"]<-res.comp$completeObs[,"Pickup_latitude"]
df[,"Dropoff_longitude"]<-res.comp$completeObs[,"Dropoff_longitude"]
df[,"Dropoff_latitude"]<-res.comp$completeObs[,"Dropoff_latitude"]
df[,"Trip_distance"]<-res.comp$completeObs[,"Trip_distance"]
# df[,"trip_distance_km"]<-res.comp$completeObs[,"trip_distance_km"]
# df[,"trip_length"]<-res.comp$completeObs[,"trip_length"]
df[,"Fare_amount"]<-res.comp$completeObs[,"Fare_amount"]
df[,"travel_time"]<- res.comp$completeObs[,"travel_time"]
df[,"Tip_amount"]<-res.comp$completeObs[,"Tip_amount"]
df[,"Tolls_amount"]<-res.comp$completeObs[,"Tolls_amount"]
summary(df)
```

```
##     VendorID     lpep_pickup_datetime Lpep_dropoff_datetime
##  Min.   :1.000   Min.   :      1      Min.   :      1
##  1st Qu.:2.000   1st Qu.: 254881      1st Qu.: 254539
##  Median :2.000   Median : 516405      Median : 515358
##  Mean   :1.785   Mean   : 520620      Mean   : 519927
##  3rd Qu.:2.000   3rd Qu.: 795588      3rd Qu.: 794506
##  Max.   :2.000   Max.   :1044592      Max.   :1042914
##  Store_and_fwd_flag   RateCodeID   Pickup_longitude Pickup_latitude
##  Min.   :1.000      Min.   :1.00   Min.   :-74.04   Min.   :40.58
##  1st Qu.:1.000      1st Qu.:1.00   1st Qu.:-73.96   1st Qu.:40.69
##  Median :1.000      Median :1.00   Median :-73.95   Median :40.75
```

```
##  Mean   :1.004      Mean   :1.02   Mean   :-73.94   Mean   :40.75
##  3rd Qu.:1.000      3rd Qu.:1.00   3rd Qu.:-73.92   3rd Qu.:40.80
##  Max.   :2.000      Max.   :2.00   Max.   :-73.79   Max.   :40.91
##  Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
##  Min.   :-74.05    Min.   :40.57    Min.   :1.000   Min.   : 0.010
##  1st Qu.:-73.97    1st Qu.:40.70    1st Qu.:1.000   1st Qu.: 1.020
##  Median :-73.95    Median :40.75    Median :1.000   Median : 1.830
##  Mean   :-73.94    Mean   :40.74    Mean   :1.348   Mean   : 2.737
##  3rd Qu.:-73.91    3rd Qu.:40.79    3rd Qu.:1.000   3rd Qu.: 3.410
##  Max.   :-73.75    Max.   :41.02    Max.   :6.000   Max.   :19.780
##   Fare_amount       Extra           MTA_tax         Tip_amount
##  Min.   : 0.1   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 6.0   1st Qu.:0.0000   1st Qu.:0.5000   1st Qu.: 0.000
##  Median : 9.0   Median :0.5000   Median :0.5000   Median : 0.000
##  Mean   :11.7   Mean   :0.3483   Mean   :0.4909   Mean   : 1.199
##  3rd Qu.:14.0   3rd Qu.:0.5000   3rd Qu.:0.5000   3rd Qu.: 2.000
##  Max.   :60.0   Max.   :2.0000   Max.   :0.5000   Max.   :25.000
##   Tolls_amount      improvement_surcharge  Total_amount    Payment_type
##  Min.   : 0.00000   Min.   :0.0000        Min.   : 0.10   Min.   :1.000
##  1st Qu.: 0.00000   1st Qu.:0.3000        1st Qu.: 7.80   1st Qu.:1.000
##  Median : 0.00000   Median :0.3000        Median :11.16   Median :2.000
##  Mean   : 0.09345   Mean   :0.2948        Mean   :14.16   Mean   :1.513
##  3rd Qu.: 0.00000   3rd Qu.:0.3000        3rd Qu.:17.16   3rd Qu.:2.000
##  Max.   :12.50000   Max.   :0.7700        Max.   :70.00   Max.   :3.000
##    Trip_type        out_ind           err_ind         travel_time
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :  1.00
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:  6.05
##  Median :1.000   Median :0.00000   Median :0.00000   Median :  9.90
##  Mean   :1.017   Mean   :0.01597   Mean   :0.01092   Mean   : 12.67
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.: 16.18
##  Max.   :2.000   Max.   :4.00000   Max.   :1.00000   Max.   :214.55
##   pick_up_hour   pick_up_period
##  Min.   : 0.00   Min.   :1.00
##  1st Qu.: 9.00   1st Qu.:2.00
##  Median :15.00   Median :3.00
##  Mean   :13.47   Mean   :2.83
##  3rd Qu.:19.00   3rd Qu.:4.00
##  Max.   :23.00   Max.   :4.00
```

Creating synthethic variables and doing their analysis

Creating AnyTip

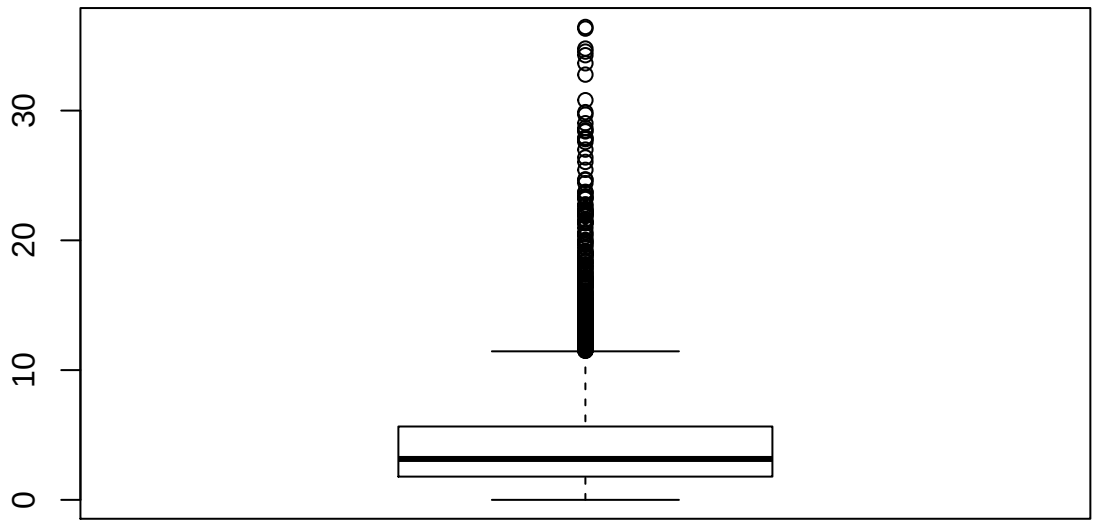
```
df$AnyTip<-ifelse(df$Tip_amount<0.0001,0,1)
df$AnyTip<-factor(df$AnyTip,labels=paste("AnyTip",c("No","Yes")))
```

Trip length

```
for (i in 1:nrow(df)){
  df$trip_length[i] <- man.dist.manual(df$Pickup_latitude[i],df$Pickup_longitude[i],df$Dropoff_latitude
}

# outlier <- which(df$trip_length>100);length(outlier) #No missing Data
# df[outlier,"trip_length"]<-NA

summary(df$trip_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.792   3.152   4.486   5.653  36.444
```

```
boxplot(df$trip_length)
```

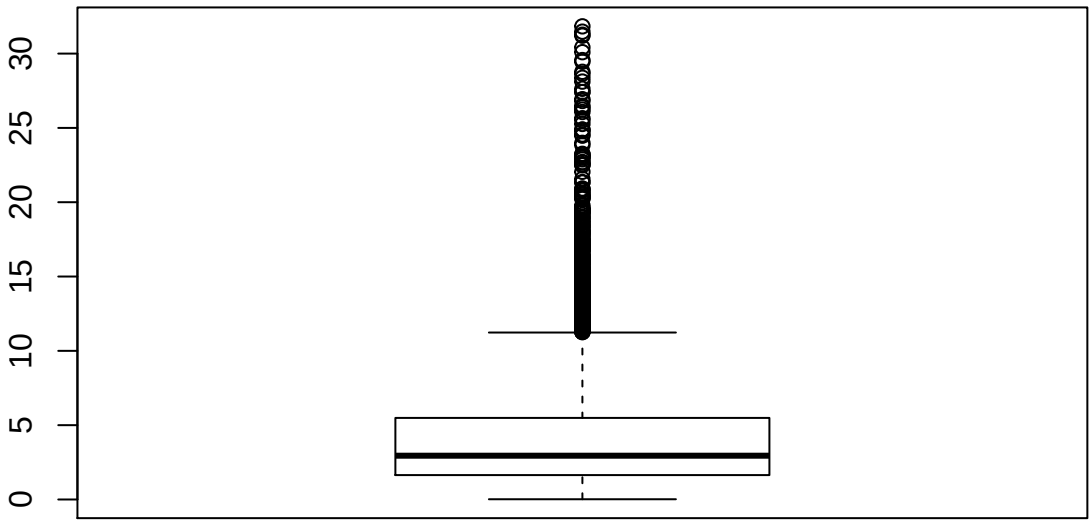


Trip distance in km

```
df$trip_distance_km<-df$Trip_distance*1.609344 # Miles to km
# outlier <- which(df$trip_length>100);length(outlier) #No missing Data
# df[outlier,"trip_length"]<-NA
summary(df$trip_distance_km)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.01609  1.64153  2.94510  4.40431  5.48786 31.83282
```

```
boxplot(df$trip_distance_km)
```



Espeed (km/h)

```
#efective speed : trigonometric distance between pickup point and dropoff point divided by travel time

for (i in 1:nrow(df)){
  df$espeed[i] <- df$trip_length[i]/(df$travel_time[i]/60)
}

summary(df$espeed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   15.11   20.26   22.01   27.07  145.27
```

```
names(df)
```

```
##  [1] "VendorID"              "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime" "Store_and_fwd_flag"
##  [5] "RateCodeID"            "Pickup_longitude"
##  [7] "Pickup_latitude"       "Dropoff_longitude"
##  [9] "Dropoff_latitude"      "Passenger_count"
## [11] "Trip_distance"         "Fare_amount"
## [13] "Extra"                 "MTA_tax"
## [15] "Tip_amount"            "Tolls_amount"
## [17] "improvement_surcharge" "Total_amount"
## [19] "Payment_type"          "Trip_type"
```

```
## [21] "out_ind"               "err_ind"
## [23] "travel_time"           "pick_up_hour"
## [25] "pick_up_period"        "AnyTip"
## [27] "trip_length"           "trip_distance_km"
## [29] "espeed"
```

```
vars_con <- names(df)[c(6:9, 11,12,15,16,23,27:29)]
outliers<-which(df$espeed>100.0);length(outliers) #0 outlier
```

```
## [1] 3
```

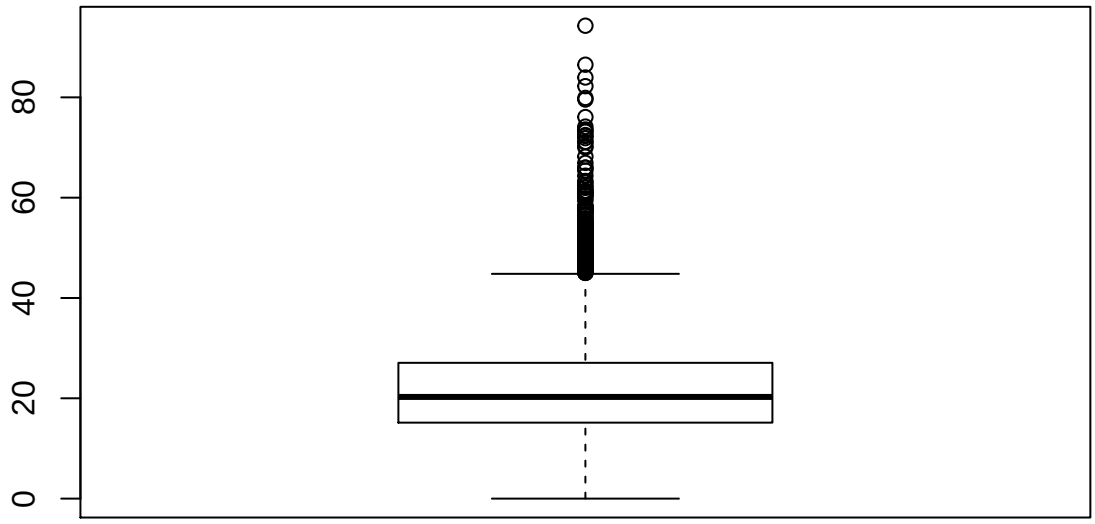
```
df[outliers,"espeed"]<-NA
errors<-which(df$espeed<=0.0);length(errors) #0 outlier
```

```
## [1] 7
```

```
df[errors,"espeed"]<-NA
res.comp <- imputePCA(df[,vars_con], ncp=4)
df[,"espeed"]<- res.comp$completeObs[,"espeed"]
summary(df$espeed)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00125 15.15233 20.27299 21.98925 27.07058 94.28724
```

```
boxplot(df$espeed)
```



Creating factors

f.passenger

```
df$f.passenger<-factor(cut(df$Passenger_count,breaks=c(0,1,6)))
summary(df$f.passenger)
```

```
## (0,1] (1,6]
##  4192   754
```

f.distance

```
df$f.distance<-factor(cut(df$Trip_distance,breaks=c(0,1.01,1.8,3.31,19.8)))
summary(df$f.distance)
```

```
##    (0,1.01]  (1.01,1.8]  (1.8,3.31] (3.31,19.8]
##        1228        1222        1202        1294
```

f.pickup_longitude

```
df$f.pickup_longitude<-factor(cut(df$Pickup_longitude,breaks=c(-74.1,-73.96,-73.947,-73.918,-73.79)))
summary(df$f.pickup_longitude)
```

```
##  (-74.1,-73.96] (-73.96,-73.95] (-73.95,-73.92] (-73.92,-73.79]
##            1245            1168            1289            1244
```

f.pickup_latitude

```
df$f.pickup_latitude<-factor(cut(df$Pickup_latitude,breaks=c(40.5,40.695,40.744,40.8,40.92)))
summary(df$f.pickup_latitude)
```

```
##  (40.5,40.7] (40.7,40.74] (40.74,40.8] (40.8,40.92]
##         1253         1157         1332         1204
```

f.dropoff_longitude

```
df$f.dropoff_longitude<-factor(cut(df$Dropoff_longitude,breaks=c(-74.1,-73.97,-73.945,-73.91,-73.75)))
summary(df$f.dropoff_longitude)
```

```
##  (-74.1,-73.97] (-73.97,-73.94] (-73.94,-73.91] (-73.91,-73.75]
##            1143            1388            1221            1194
```

f.dropoff_latitude

```
df$f.dropoff_latitude<-factor(cut(df$Dropoff_latitude,breaks=c(40.53,40.7,40.75,40.79,41.5)))
summary(df$f.dropoff_latitude)
```

```
##  (40.53,40.7]  (40.7,40.75] (40.75,40.79]  (40.79,41.5]
##          1308          1358          1101          1179
```

f.fare_amount

```
df$f.fare_amount<-factor(cut(df$Fare_amount,breaks=c(0,6,9,14,60.5)))
summary(df$f.fare_amount)
```

```
##     (0,6]     (6,9]    (9,14] (14,60.5]
##      1252      1255      1203      1236
```

f.extra

```
df$f.extra<-factor(cut(df$Extra,breaks=c(-0.1,0.5,2)))
summary(df$f.extra)
```

```
## (-0.1,0.5]    (0.5,2]
##       4175        771
```

f.MTA_tax

```
df$f.MTA_tax<-factor(cut(df$MTA_tax,breaks=c(-0.1,0.4,0.5)))
summary(df$f.MTA_tax) #11 NA's -> values of -0.5 => Outliers?
```

```
## (-0.1,0.4]  (0.4,0.5]
##         90       4856
```

f.Improvement_surcharge

```
df$f.Improvement_surcharge<-factor(cut(df$improvement_surcharge, breaks=c(-0.1,0.1,0.8)))
summary(df$f.Improvement_surcharge) #11 NA's -> values of -0.3 => Outliers?
```

```
## (-0.1,0.1]  (0.1,0.8]
##         87       4859
```

f.tip_amount

```
df$f.tip_amount<-factor(cut(df$Tip_amount, breaks=c(-0.1,1,25.1)))
summary(df$f.tip_amount)
```

```
## (-0.1,1] (1,25.1]
##     3076     1870
```

f.tolls_amount

```
df$f.toll<-factor(cut(df$Tolls_amount,breaks=c(-1,1,50)))
summary(df$f.toll)
```

```
## (-1,1] (1,50]
##   4866     80
```

f.total_amount

```
df$f.total<-factor(cut(df$Total_amount,breaks=c(-1,7.8,11,16.6,70.1)))
summary(df$f.total)
```

```
##    (-1,7.8]    (7.8,11]   (11,16.6] (16.6,70.1]
##        1253        1187        1203        1303
```

f.ttime

```
df$f.ttime<-factor(cut(df$travel_time,breaks=c(-1,6,9.78,15.7,415)))
summary(df$f.ttime)
```

```
##      (-1,6]    (6,9.78] (9.78,15.7]  (15.7,415]
##        1223        1211        1216        1296
```

f.espeed

```
df$f.espeed<-factor(cut(df$espeed,breaks=c(0.0,15.3,20.1,26.2,95)))
summary(df$f.espeed)
```

```
##    (0,15.3] (15.3,20.1] (20.1,26.2]   (26.2,95]
##        1280        1156        1137        1373
```

Re-declaring vectors of data

```
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "out_ind"                 "err_ind"
## [23] "travel_time"             "pick_up_hour"
## [25] "pick_up_period"          "AnyTip"
## [27] "trip_length"             "trip_distance_km"
## [29] "espeed"                  "f.passenger"
## [31] "f.distance"              "f.pickup_longitude"
## [33] "f.pickup_latitude"       "f.dropoff_longitude"
## [35] "f.dropoff_latitude"      "f.fare_amount"
```

```
## [37] "f.extra"                 "f.MTA_tax"
## [39] "f.Improvement_surcharge" "f.tip_amount"
## [41] "f.toll"                  "f.total"
## [43] "f.ttime"                 "f.espeed"
```

```
vars_con <-names(df)[c(6:9, 11,12,15,16,23,27:29)]
vars_con
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Tip_amount"        "Tolls_amount"      "travel_time"
## [10] "trip_length"       "trip_distance_km"  "espeed"
```

```
vars_dis<-names(df)[c(1,4,5,10,13,14,17,19,20,24,25,27,28,30:44)]
vars_dis
```

```
##  [1] "VendorID"                "Store_and_fwd_flag"
##  [3] "RateCodeID"              "Passenger_count"
##  [5] "Extra"                   "MTA_tax"
##  [7] "improvement_surcharge"   "Payment_type"
##  [9] "Trip_type"               "pick_up_hour"
## [11] "pick_up_period"          "trip_length"
## [13] "trip_distance_km"        "f.passenger"
## [15] "f.distance"              "f.pickup_longitude"
## [17] "f.pickup_latitude"       "f.dropoff_longitude"
## [19] "f.dropoff_latitude"      "f.fare_amount"
## [21] "f.extra"                 "f.MTA_tax"
## [23] "f.Improvement_surcharge" "f.tip_amount"
## [25] "f.toll"                  "f.total"
## [27] "f.ttime"                 "f.espeed"
```

```
vars_res<-names(df)[c(18,26)]
vars_res
```

```
## [1] "Total_amount" "AnyTip"
```

Profiling

```
library(FactoMineR)
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "out_ind"                 "err_ind"
## [23] "travel_time"             "pick_up_hour"
## [25] "pick_up_period"          "AnyTip"
## [27] "trip_length"             "trip_distance_km"
```

```
## [29] "espeed"                   "f.passenger"
## [31] "f.distance"               "f.pickup_longitude"
## [33] "f.pickup_latitude"        "f.dropoff_longitude"
## [35] "f.dropoff_latitude"       "f.fare_amount"
## [37] "f.extra"                  "f.MTA_tax"
## [39] "f.Improvement_surcharge" "f.tip_amount"
## [41] "f.toll"                   "f.total"
## [43] "f.ttime"                  "f.espeed"
```

```
# Target Total Amount
vars_con_profiling <- c(vars_con,(names(df)[c(18)]))
condes(df[, vars_con_profiling], num.var=12)
```

```
## $quanti
##                   correlation       p.value
## trip_length        0.54315273  0.000000e+00
## trip_distance_km   0.35486908 1.009055e-146
## Trip_distance      0.35486908 1.009055e-146
## Fare_amount        0.19370723  5.093862e-43
## Total_amount       0.18719417  3.057565e-40
## Dropoff_longitude  0.15154907  8.436416e-27
## Tolls_amount       0.11307710  1.510950e-15
## Pickup_longitude   0.09101690  1.429513e-10
## Pickup_latitude    0.07163739  4.575539e-07
## Tip_amount         0.05581404  8.588750e-05
## travel_time       -0.07997169  1.783043e-08
```

```
# Binary Target AnyTip
vars_con;vars_dis
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Tip_amount"        "Tolls_amount"      "travel_time"
## [10] "trip_length"       "trip_distance_km"  "espeed"
```

```
##  [1] "VendorID"                 "Store_and_fwd_flag"
##  [3] "RateCodeID"               "Passenger_count"
##  [5] "Extra"                    "MTA_tax"
##  [7] "improvement_surcharge"    "Payment_type"
##  [9] "Trip_type"                "pick_up_hour"
## [11] "pick_up_period"           "trip_length"
## [13] "trip_distance_km"         "f.passenger"
## [15] "f.distance"               "f.pickup_longitude"
## [17] "f.pickup_latitude"        "f.dropoff_longitude"
## [19] "f.dropoff_latitude"       "f.fare_amount"
## [21] "f.extra"                  "f.MTA_tax"
## [23] "f.Improvement_surcharge" "f.tip_amount"
## [25] "f.toll"                   "f.total"
## [27] "f.ttime"                  "f.espeed"
```

```
names(df)
```

```
##  [1] "VendorID"                 "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"    "Store_and_fwd_flag"
##  [5] "RateCodeID"               "Pickup_longitude"
##  [7] "Pickup_latitude"          "Dropoff_longitude"
##  [9] "Dropoff_latitude"         "Passenger_count"
```

```
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "out_ind"                 "err_ind"
## [23] "travel_time"             "pick_up_hour"
## [25] "pick_up_period"          "AnyTip"
## [27] "trip_length"             "trip_distance_km"
## [29] "espeed"                  "f.passenger"
## [31] "f.distance"              "f.pickup_longitude"
## [33] "f.pickup_latitude"       "f.dropoff_longitude"
## [35] "f.dropoff_latitude"      "f.fare_amount"
## [37] "f.extra"                 "f.MTA_tax"
## [39] "f.Improvement_surcharge" "f.tip_amount"
## [41] "f.toll"                  "f.total"
## [43] "f.ttime"                 "f.espeed"
```

```
vars_con_and_dis <- c(vars_con,vars_dis)
vars_con_catdes <- c(vars_con_and_dis,(names(df)[c(26)]))
#if p.value is under 0.5 then we have to reject the hipothesis of no relation between variable and targ
catdes(df[,vars_con_catdes], num.var = 41)
```

```
## $test.chi2
##                               p.value df
## f.tip_amount             0.000000e+00  1
## f.total                 2.575626e-108  3
## f.dropoff_longitude      6.355313e-67  3
## f.pickup_longitude       1.389244e-57  3
## f.distance               2.793033e-24  3
## f.dropoff_latitude       7.275499e-24  3
## f.fare_amount            7.310100e-21  3
## f.pickup_latitude        1.716898e-20  3
## f.ttime                  8.952886e-20  3
## f.Improvement_surcharge  2.105245e-08  1
## f.MTA_tax                2.593100e-08  1
## f.toll                   1.057770e-06  1
##
## $category
## $category$`AnyTip No`
##                                      Cla/Mod     Mod/Cla    Global
## f.tip_amount=(-0.1,1]               93.20546 100.0000000 62.191670
## f.total=(-1,7.8]                    82.20271  35.9260551 25.333603
## f.pickup_longitude=(-73.92,-73.79]  74.03537  32.1241716 25.151638
## f.dropoff_longitude=(-73.91,-73.75] 73.11558  30.4499477 24.140720
## f.dropoff_latitude=(40.79,41.5]     70.48346  28.9850017 23.837444
## f.ttime=(-1,6]                      67.70237  28.8803627 24.727052
## f.distance=(0,1.01]                 67.42671  28.8803627 24.828144
## f.dropoff_longitude=(-73.94,-73.91] 67.23997  28.6362051 24.686615
## f.fare_amount=(0,6]                 66.77316  29.1594001 25.313385
## f.pickup_latitude=(40.74,40.8]      65.01502  30.2057900 26.930853
## f.Improvement_surcharge=(-0.1,0.1]  87.35632   2.6508546  1.758997
## f.MTA_tax=(-0.1,0.4]                86.66667   2.7206139  1.819652
## f.pickup_latitude=(40.8,40.92]      64.20266  26.9619812 24.342903
## f.toll=(-1,1]                       58.40526  99.1280084 98.382531
```

```
## f.pickup_longitude=(-73.95,-73.92]  63.07215  28.3571678 26.061464
## f.distance=(1.01,1.8]               62.35679  26.5783048 24.706834
## f.fare_amount=(6,9]                 61.03586  26.7178235 25.374040
## f.espeed=(20.1,26.2]                55.05717  21.8346704 22.988273
## f.distance=(1.8,3.31]               54.49251  22.8461807 24.302467
## f.dropoff_latitude=(40.53,40.7]     54.35780  24.7994419 26.445613
## f.toll=(1,50]                       31.25000   0.8719916  1.617469
## f.total=(11,16.6]                   51.62095  21.6602721 24.322685
## f.pickup_latitude=(40.7,40.74]      51.33967  20.7185211 23.392641
## f.MTA_tax=(0.4,0.5]                 57.43410  97.2793861 98.180348
## f.Improvement_surcharge=(0.1,0.8]   57.43980  97.3491454 98.241003
## f.pickup_latitude=(40.5,40.7]       50.59856  22.1137077 25.333603
## f.dropoff_latitude=(40.7,40.75]     50.88365  24.1018486 27.456531
## f.dropoff_longitude=(-73.97,-73.94] 49.71182  24.0669690 28.063081
## f.ttime=(15.7,415]                  49.22840  22.2532264 26.202992
## f.pickup_longitude=(-73.96,-73.95]  48.11644  19.6023718 23.615042
## f.fare_amount=(14,60.5]             48.05825  20.7185211 24.989891
## f.distance=(3.31,19.8]              48.06801  21.6951517 26.162556
## f.pickup_longitude=(-74.1,-73.96]   45.86345  19.9162888 25.171856
## f.dropoff_longitude=(-74.1,-73.97]  42.25722  16.8468783 23.109584
## f.total=(16.6,70.1]                 39.67767  18.0327869 26.344521
## f.tip_amount=(1,25.1]                0.00000   0.0000000 37.808330
##                                          p.value     v.test
## f.tip_amount=(-0.1,1]               0.000000e+00        Inf
## f.total=(-1,7.8]                    5.250180e-97  20.900939
## f.pickup_longitude=(-73.92,-73.79]  1.007321e-41  13.532363
## f.dropoff_longitude=(-73.91,-73.75] 2.670773e-35  12.398255
## f.dropoff_latitude=(40.79,41.5]     4.764958e-24  10.114489
## f.ttime=(-1,6]                      1.013652e-15   8.025194
## f.distance=(0,1.01]                 5.449104e-15   7.816086
## f.dropoff_longitude=(-73.94,-73.91] 2.328935e-14   7.631026
## f.fare_amount=(0,6]                 1.787752e-13   7.363779
## f.pickup_latitude=(40.74,40.8]      8.703531e-10   6.131535
## f.Improvement_surcharge=(-0.1,0.1]  2.274384e-09   5.976889
## f.MTA_tax=(-0.1,0.4]                3.202421e-09   5.920870
## f.pickup_latitude=(40.8,40.92]      4.074847e-07   5.065428
## f.toll=(-1,1]                       1.369623e-06   4.829375
## f.pickup_longitude=(-73.95,-73.92]  1.459756e-05   4.334675
## f.distance=(1.01,1.8]               3.255890e-04   3.594041
## f.fare_amount=(6,9]                 1.063920e-02   2.554331
## f.espeed=(20.1,26.2]                2.385808e-02  -2.259407
## f.distance=(1.8,3.31]               5.156341e-03  -2.797101
## f.dropoff_latitude=(40.53,40.7]     2.105198e-03  -3.074976
## f.toll=(1,50]                       1.369623e-06  -4.829375
## f.total=(11,16.6]                   3.312774e-07  -5.104724
## f.pickup_latitude=(40.7,40.74]      2.046168e-07  -5.195094
## f.MTA_tax=(0.4,0.5]                 3.202421e-09  -5.920870
## f.Improvement_surcharge=(0.1,0.8]   2.274384e-09  -5.976889
## f.pickup_latitude=(40.5,40.7]       1.140156e-09  -6.088441
## f.dropoff_latitude=(40.7,40.75]     6.246703e-10  -6.184077
## f.dropoff_longitude=(-73.97,-73.94] 2.564523e-13  -7.315486
## f.ttime=(15.7,415]                  1.526778e-13  -7.384804
## f.pickup_longitude=(-73.96,-73.95]  8.337542e-15  -7.762342
## f.fare_amount=(14,60.5]             5.170105e-16  -8.107431
```

```
## f.distance=(3.31,19.8]              6.560467e-17  -8.354690
## f.pickup_longitude=(-74.1,-73.96]   2.551434e-23  -9.948839
## f.dropoff_longitude=(-74.1,-73.97]  3.100341e-34 -12.200207
## f.total=(16.6,70.1]                 2.695101e-54 -15.516209
## f.tip_amount=(1,25.1]               0.000000e+00       -Inf
##
## $category$`AnyTip Yes`
##                                        Cla/Mod    Mod/Cla    Global
## f.tip_amount=(1,25.1]                100.000000 89.9470899 37.808330
## f.total=(16.6,70.1]                   60.322333 37.8066378 26.344521
## f.dropoff_longitude=(-74.1,-73.97]    57.742782 31.7460317 23.109584
## f.pickup_longitude=(-74.1,-73.96]     54.136546 32.4194324 25.171856
## f.distance=(3.31,19.8]                51.931994 32.3232323 26.162556
## f.fare_amount=(14,60.5]               51.941748 30.8802309 24.989891
## f.pickup_longitude=(-73.96,-73.95]    51.883562 29.1486291 23.615042
## f.ttime=(15.7,415]                    50.771605 31.6498316 26.202992
## f.dropoff_longitude=(-73.97,-73.94]   50.288184 33.5738336 28.063081
## f.dropoff_latitude=(40.7,40.75]       49.116348 32.0827321 27.456531
## f.pickup_latitude=(40.5,40.7]         49.401437 29.7739298 25.333603
## f.Improvement_surcharge=(0.1,0.8]     42.560198 99.4708995 98.241003
## f.MTA_tax=(0.4,0.5]                   42.565898 99.4227994 98.180348
## f.pickup_latitude=(40.7,40.74]        48.660328 27.0803271 23.392641
## f.total=(11,16.6]                     48.379052 27.9942280 24.322685
## f.toll=(1,50]                         68.750000  2.6455026  1.617469
## f.dropoff_latitude=(40.53,40.7]       45.642202 28.7157287 26.445613
## f.distance=(1.8,3.31]                 45.507488 26.3107263 24.302467
## f.espeed=(20.1,26.2]                  44.942832 24.5791246 22.988273
## f.fare_amount=(6,9]                   38.964143 23.5209235 25.374040
## f.distance=(1.01,1.8]                 37.643208 22.1260221 24.706834
## f.pickup_longitude=(-73.95,-73.92]    36.927851 22.8956229 26.061464
## f.toll=(-1,1]                         41.594739 97.3544974 98.382531
## f.pickup_latitude=(40.8,40.92]        35.797342 20.7311207 24.342903
## f.MTA_tax=(-0.1,0.4]                  13.333333  0.5772006  1.819652
## f.Improvement_surcharge=(-0.1,0.1]    12.643678  0.5291005  1.758997
## f.pickup_latitude=(40.74,40.8]        34.984985 22.4146224 26.930853
## f.fare_amount=(0,6]                   33.226837 20.0096200 25.313385
## f.dropoff_longitude=(-73.94,-73.91]   32.760033 19.2400192 24.686615
## f.distance=(0,1.01]                   32.573290 19.2400192 24.828144
## f.ttime=(-1,6]                        32.297629 18.9995190 24.727052
## f.dropoff_latitude=(40.79,41.5]       29.516539 16.7388167 23.837444
## f.dropoff_longitude=(-73.91,-73.75]   26.884422 15.4401154 24.140720
## f.pickup_longitude=(-73.92,-73.79]    25.964630 15.5363155 25.151638
## f.total=(-1,7.8]                      17.797287 10.7263107 25.333603
## f.tip_amount=(-0.1,1]                  6.794538 10.0529101 62.191670
##                                           p.value     v.test
## f.tip_amount=(1,25.1]                 0.000000e+00        Inf
## f.total=(16.6,70.1]                   2.695101e-54  15.516209
## f.dropoff_longitude=(-74.1,-73.97]    3.100341e-34  12.200207
## f.pickup_longitude=(-74.1,-73.96]     2.551434e-23   9.948839
## f.distance=(3.31,19.8]                6.560467e-17   8.354690
## f.fare_amount=(14,60.5]               5.170105e-16   8.107431
## f.pickup_longitude=(-73.96,-73.95]    8.337542e-15   7.762342
## f.ttime=(15.7,415]                    1.526778e-13   7.384804
## f.dropoff_longitude=(-73.97,-73.94]   2.564523e-13   7.315486
```

```
## f.dropoff_latitude=(40.7,40.75]     6.246703e-10   6.184077
## f.pickup_latitude=(40.5,40.7]       1.140156e-09   6.088441
## f.Improvement_surcharge=(0.1,0.8]   2.274384e-09   5.976889
## f.MTA_tax=(0.4,0.5]                 3.202421e-09   5.920870
## f.pickup_latitude=(40.7,40.74]      2.046168e-07   5.195094
## f.total=(11,16.6]                   3.312774e-07   5.104724
## f.toll=(1,50]                       1.369623e-06   4.829375
## f.dropoff_latitude=(40.53,40.7]     2.105198e-03   3.074976
## f.distance=(1.8,3.31]               5.156341e-03   2.797101
## f.espeed=(20.1,26.2]                2.385808e-02   2.259407
## f.fare_amount=(6,9]                 1.063920e-02  -2.554331
## f.distance=(1.01,1.8]               3.255890e-04  -3.594041
## f.pickup_longitude=(-73.95,-73.92]  1.459756e-05  -4.334675
## f.toll=(-1,1]                       1.369623e-06  -4.829375
## f.pickup_latitude=(40.8,40.92]      4.074847e-07  -5.065428
## f.MTA_tax=(-0.1,0.4]                3.202421e-09  -5.920870
## f.Improvement_surcharge=(-0.1,0.1]  2.274384e-09  -5.976889
## f.pickup_latitude=(40.74,40.8]      8.703531e-10  -6.131535
## f.fare_amount=(0,6]                 1.787752e-13  -7.363779
## f.dropoff_longitude=(-73.94,-73.91] 2.328935e-14  -7.631026
## f.distance=(0,1.01]                 5.449104e-15  -7.816086
## f.ttime=(-1,6]                      1.013652e-15  -8.025194
## f.dropoff_latitude=(40.79,41.5]     4.764958e-24 -10.114489
## f.dropoff_longitude=(-73.91,-73.75] 2.670773e-35 -12.398255
## f.pickup_longitude=(-73.92,-73.79]  1.007321e-41 -13.532363
## f.total=(-1,7.8]                    5.250180e-97 -20.900939
## f.tip_amount=(-0.1,1]               0.000000e+00       -Inf
##
##
## $quanti.var
##                              Eta2      P-value
## Tip_amount            0.500939145 0.000000e+00
## Payment_type          0.722151968 0.000000e+00
## Dropoff_longitude     0.046864643 1.541030e-53
## Pickup_longitude      0.042753976 6.725337e-49
## Fare_amount           0.016353184 1.743519e-19
## Pickup_latitude       0.015984669 4.450313e-19
## travel_time           0.015704831 9.065624e-19
## trip_distance_km      0.015136353 3.846528e-18
## trip_distance_km.1    0.015136353 3.846528e-18
## Trip_distance         0.015136353 3.846528e-18
## Dropoff_latitude      0.014820733 8.580711e-18
## trip_length           0.013899949 8.912869e-17
## trip_length.1         0.013899949 8.912869e-17
## Trip_type             0.006575412 1.124663e-08
## RateCodeID            0.006521759 1.290286e-08
## MTA_tax               0.006265747 2.486107e-08
## improvement_surcharge 0.005852982 7.166351e-08
## Tolls_amount          0.004532967 2.146027e-06
##
## $quanti
## $quanti$`AnyTip No`
##                            v.test Mean in category Overall mean
## Payment_type            59.758192       1.88454831   1.51273757
```

```
## Dropoff_longitude      15.223195     -73.92660752 -73.93548178
## Pickup_longitude       14.540234     -73.92894640 -73.93623052
## Pickup_latitude         8.890680      40.75244348  40.74641788
## Dropoff_latitude        8.560872      40.75024649  40.74438270
## Trip_type               5.702229       1.02615975   1.01718560
## RateCodeID              5.678917       1.02964772   1.02001617
## Tolls_amount           -4.734503       0.05024067   0.09344521
## improvement_surcharge  -5.379870       0.29221137   0.29481803
## MTA_tax                -5.566338       0.48639693   0.49090174
## trip_length.1          -8.290672       4.06425461   4.48565422
## trip_length            -8.290672       4.06425461   4.48565422
## trip_distance_km.1     -8.651547       3.95905964   4.40431069
## trip_distance_km       -8.651547       3.95905964   4.40431069
## Trip_distance          -8.651547       2.46004561   2.73671178
## travel_time            -8.812513      11.52797395  12.66848470
## Fare_amount            -8.992580      10.81056189  11.70334027
## Tip_amount            -49.770916       0.00000000   1.19907138
##                      sd in category  Overall sd      p.value
## Payment_type             0.35572231  0.51380016 0.000000e+00
## Dropoff_longitude        0.04860204  0.04813886 2.481132e-52
## Pickup_longitude         0.04296490  0.04136903 6.736174e-48
## Pickup_latitude          0.05635126  0.05596742 6.073620e-19
## Dropoff_latitude         0.05905423  0.05656269 1.120174e-17
## Trip_type                0.15961020  0.12996253 1.182510e-08
## RateCodeID               0.16961347  0.14005544 1.355503e-08
## Tolls_amount             0.54517836  0.75357188 2.195921e-06
## improvement_surcharge    0.04901179  0.04001134 7.453982e-08
## MTA_tax                  0.08134182  0.06683077 2.601485e-08
## trip_length.1            3.89132416  4.19733974 1.126101e-16
## trip_length              3.89132416  4.19733974 1.126101e-16
## trip_distance_km.1       3.99445531  4.24992174 5.080599e-18
## trip_distance_km         3.99445531  4.24992174 5.080599e-18
## Trip_distance            2.48203946  2.64077894 5.080599e-18
## travel_time              9.53903345 10.68733338 1.223711e-18
## Fare_amount              7.73654261  8.19839994 2.414937e-19
## Tip_amount               0.00000000  1.98947805 0.000000e+00
##
## $quanti$`AnyTip Yes`
##                         v.test Mean in category Overall mean
## Tip_amount             49.770916        2.8526248   1.19907138
## Fare_amount             8.992580       12.9345070  11.70334027
## travel_time             8.812513       14.2412814  12.66848470
## Trip_distance           8.651547        3.1182423   2.73671178
## trip_distance_km.1      8.651547        5.0183245   4.40431069
## trip_distance_km        8.651547        5.0183245   4.40431069
## trip_length.1           8.290672        5.0667762   4.48565422
## trip_length             8.290672        5.0667762   4.48565422
## MTA_tax                 5.566338        0.4971140   0.49090174
## improvement_surcharge   5.379870        0.2984127   0.29481803
## Tolls_amount            4.734503        0.1530255   0.09344521
## RateCodeID             -5.678917        1.0067340   1.02001617
## Trip_type              -5.702229        1.0048100   1.01718560
## Dropoff_latitude       -8.560872       40.7362964  40.74438270
## Pickup_latitude        -8.890680       40.7381084  40.74641788
```

```
## Pickup_longitude      -14.540234      -73.9462755 -73.93623052
## Dropoff_longitude     -15.223195      -73.9477196 -73.93548178
## Payment_type          -59.758192        1.0000000   1.51273757
##                       sd in category  Overall sd      p.value
## Tip_amount                2.16778255  1.98947805 0.000000e+00
## Fare_amount               8.64569402  8.19839994 2.414937e-19
## travel_time              11.91557124 10.68733338 1.223711e-18
## Trip_distance             2.80071566  2.64077894 5.080599e-18
## trip_distance_km.1        4.50731495  4.24992174 5.080599e-18
## trip_distance_km          4.50731495  4.24992174 5.080599e-18
## trip_length.1             4.52200435  4.19733974 1.126101e-16
## trip_length               4.52200435  4.19733974 1.126101e-16
## MTA_tax                   0.03787707  0.06683077 2.601485e-08
## improvement_surcharge     0.02176398  0.04001134 7.453982e-08
## Tolls_amount              0.96694479  0.75357188 2.195921e-06
## RateCodeID                0.08178423  0.14005544 1.355503e-08
## Trip_type                 0.06918720  0.12996253 1.182510e-08
## Dropoff_latitude          0.05185831  0.05656269 1.120174e-17
## Pickup_latitude           0.05434870  0.05596742 6.073620e-19
## Pickup_longitude          0.03676585  0.04136903 6.736174e-48
## Dropoff_longitude         0.04468995  0.04813886 2.481132e-52
## Payment_type              0.00000000  0.51380016 0.000000e+00
##
##
## attr(,"class")
## [1] "catdes" "list "
```

Previous work

Load requiered packages

PCA analysis

1. The Kaiser rule is to drop all components with eigenvalues under 1.0 for a normalized PcA According to the Elbow rule when the drop ceases and the curve makes an elbow toward less steep declinewe should drop all further components after the one starting the elbow.

I. I. Eigenvalues and axes

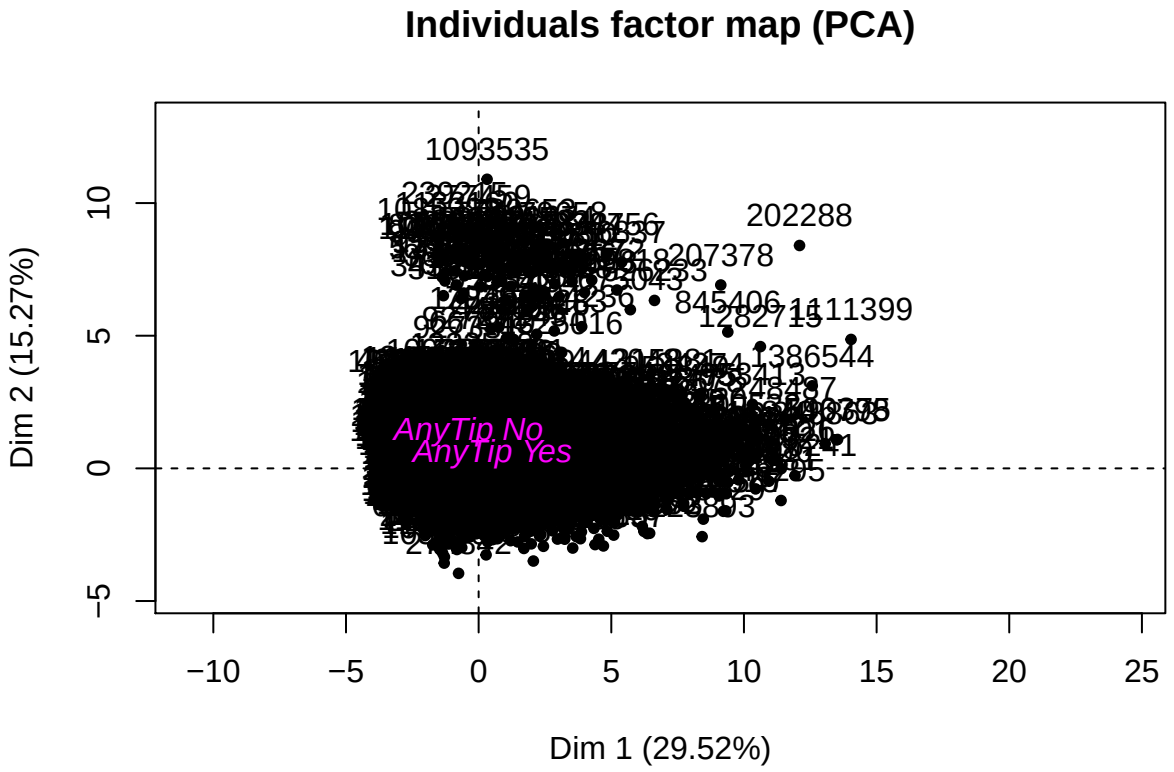
For the PCA analysis we take all numerical variables as active, where TotalAmount and Anytip are suplementary.

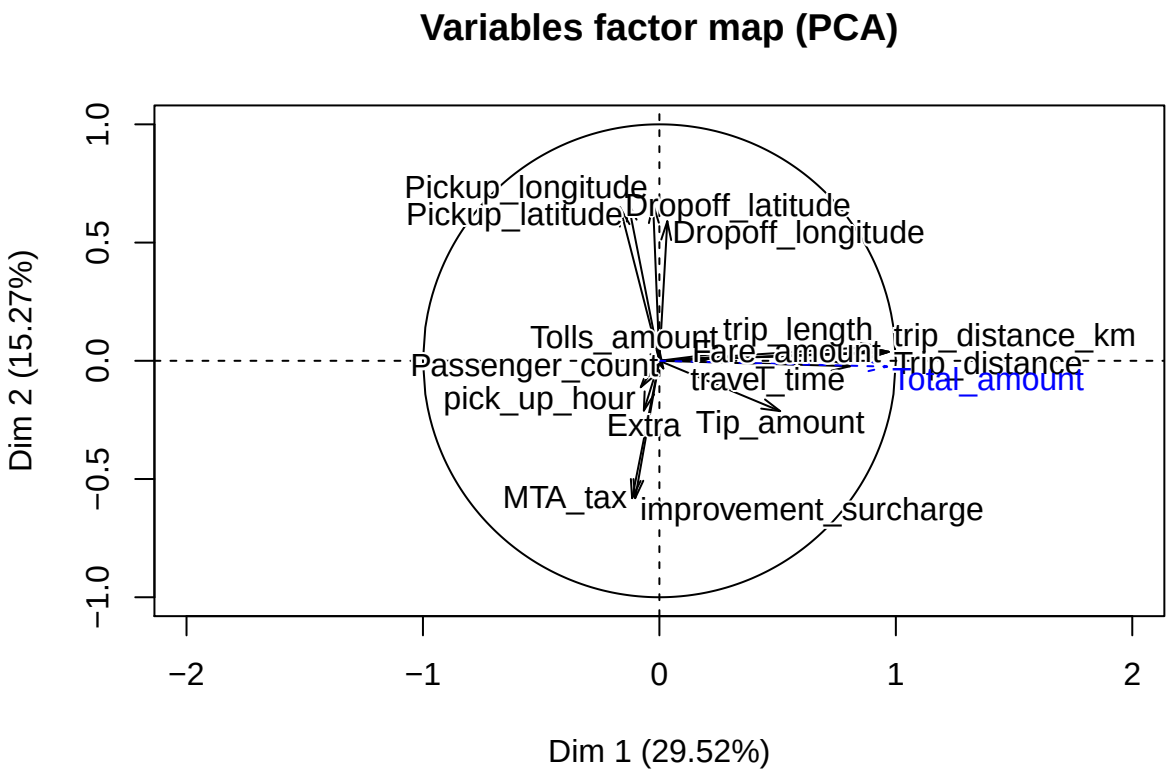
```
load("Taxi5000_raw_DataDefinitivev2.RData")
library(FactoMineR)
names (df)
```

```
##  [1] "VendorID"               "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"  "Store_and_fwd_flag"
##  [5] "RateCodeID"             "Pickup_longitude"
##  [7] "Pickup_latitude"        "Dropoff_longitude"
##  [9] "Dropoff_latitude"       "Passenger_count"
```

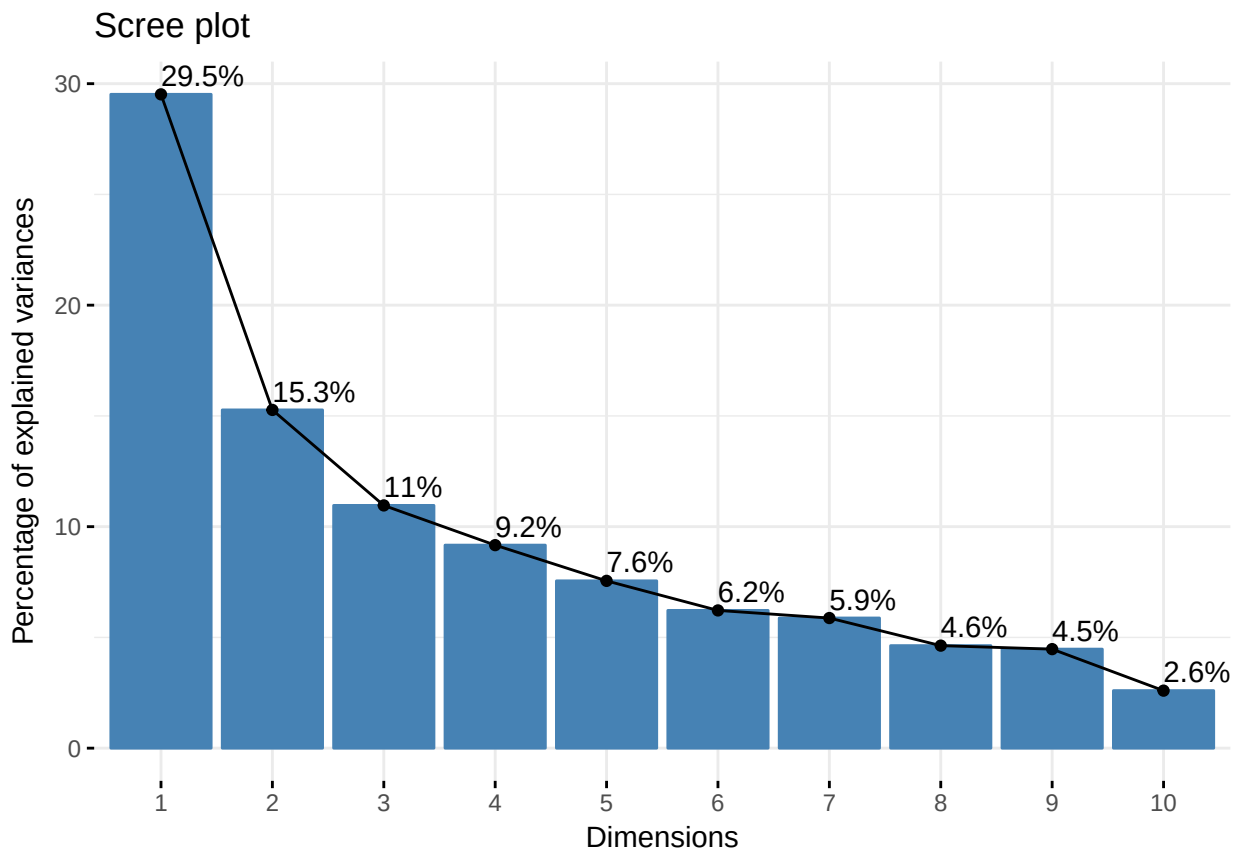
```
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"
```

```
vars_con_pca<-c(6,7,8,9,10,11,12,13,14,15,16,17,18,22,23,24,25,26)

#From te plot we see that the variables "Trip_distance", "Trip_length", "Travel_time" and "Fare_amount"
res.pca<-PCA(df[,vars_con_pca], quanti.sup = 13, quali.sup = 14, ncp = 6 ) # TotalAmount and AnyTip
```

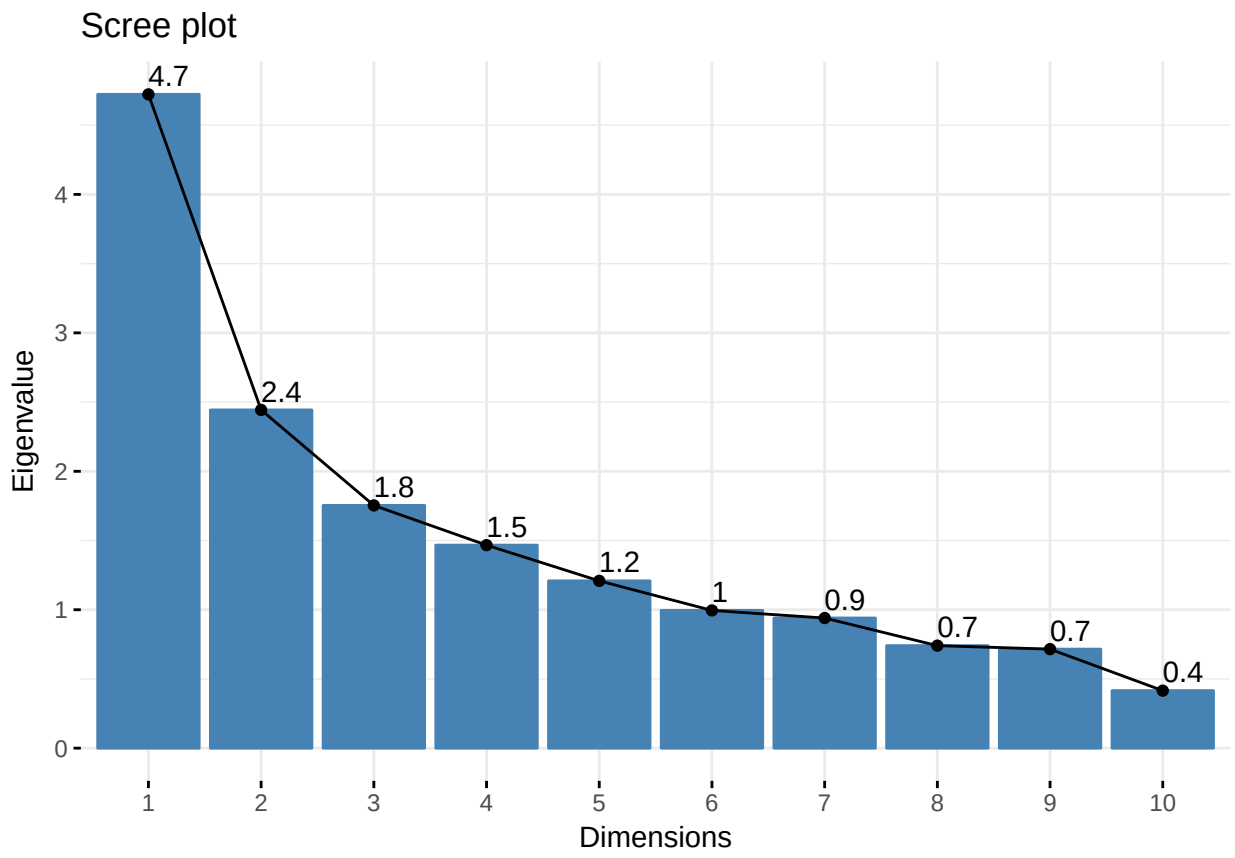





```
fviz_eig(res.pca, addlabels = TRUE)
```




```
fviz_eig(res.pca, choice = "eigenvalue",addlabels = TRUE)
```



II. Individuals point of view

Look at variables that are too contributive

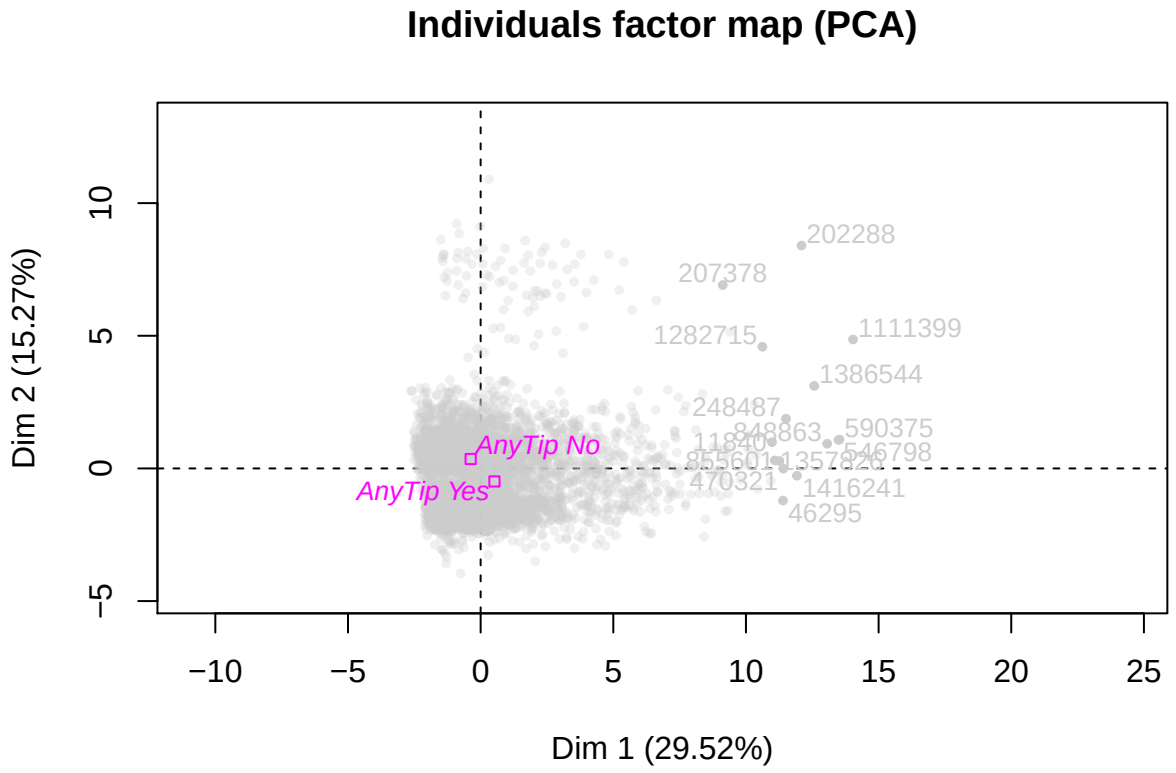
```
summary(res.pca, dig = 2, nbelements = 17, nbind=3, ncp=4)
```

```
##
## Call:
## PCA(X = df[, vars_con_pca], ncp = 6, quanti.sup = 13, quali.sup = 14)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               4.723   2.443   1.753   1.467   1.208   0.995
## % of var.             29.519  15.270  10.959   9.166   7.553   6.219
## Cumulative % of var.  29.519  44.789  55.747  64.913  72.466  78.685
##                        Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance               0.940   0.741   0.715   0.416   0.229   0.138
## % of var.              5.875   4.629   4.471   2.597   1.433   0.860
## Cumulative % of var.  84.559  89.188  93.659  96.256  97.689  98.549
##                       Dim.13  Dim.14  Dim.15  Dim.16
## Variance               0.089   0.078   0.065   0.000
## % of var.              0.558   0.487   0.407   0.000
## Cumulative % of var.  99.106  99.593 100.000 100.000
```

```
##
## Individuals (the 3 first)
##                          Dist    Dim.1    ctr   cos2    Dim.2    ctr
## 285                   |  3.103 |  0.914  0.004  0.087 | -0.117  0.000
## 307                   |  2.916 |  1.081  0.005  0.137 |  0.729  0.004
## 401                   |  2.410 |  0.516  0.001  0.046 | -0.325  0.001
##                        cos2    Dim.3    ctr   cos2    Dim.4    ctr   cos2
## 285                   0.001 |  0.070  0.000  0.001 |  1.789  0.044  0.333
## 307                   0.063 |  0.746  0.006  0.065 | -0.448  0.003  0.024
## 401                   0.018 | -0.016  0.000  0.000 |  0.355  0.002  0.022
##
## 285                   |
## 307                   |
## 401                   |
##
## Variables
##                          Dim.1    ctr   cos2    Dim.2    ctr   cos2
## Pickup_longitude      | -0.027  0.015  0.001 |  0.662 17.915  0.438 |
## Pickup_latitude       | -0.132  0.371  0.018 |  0.670 18.370  0.449 |
## Dropoff_longitude     |  0.033  0.024  0.001 |  0.590 14.254  0.348 |
## Dropoff_latitude      | -0.166  0.586  0.028 |  0.649 17.220  0.421 |
## Passenger_count       |  0.017  0.006  0.000 | -0.033  0.043  0.001 |
## Trip_distance         |  0.970 19.917  0.941 |  0.038  0.059  0.001 |
## Fare_amount           |  0.958 19.447  0.918 |  0.035  0.051  0.001 |
## Extra                 | -0.066  0.091  0.004 | -0.211  1.814  0.044 |
## MTA_tax               | -0.115  0.282  0.013 | -0.581 13.813  0.337 |
## Tip_amount            |  0.511  5.526  0.261 | -0.213  1.856  0.045 |
## Tolls_amount          |  0.270  1.548  0.073 |  0.034  0.048  0.001 |
## improvement_surcharge | -0.102  0.222  0.011 | -0.581 13.813  0.337 |
## trip_length           |  0.925 18.128  0.856 |  0.061  0.152  0.004 |
## trip_distance_km      |  0.970 19.917  0.941 |  0.038  0.059  0.001 |
## travel_time           |  0.807 13.789  0.651 | -0.023  0.022  0.001 |
## pick_up_hour          | -0.079  0.131  0.006 | -0.112  0.511  0.012 |
##                         Dim.3    ctr   cos2    Dim.4    ctr   cos2
## Pickup_longitude       0.312  5.535  0.097 |  0.558 21.264  0.312 |
## Pickup_latitude        0.394  8.872  0.156 | -0.518 18.312  0.269 |
## Dropoff_longitude      0.304  5.261  0.092 |  0.654 29.126  0.427 |
## Dropoff_latitude       0.392  8.756  0.154 | -0.538 19.754  0.290 |
## Passenger_count        0.037  0.076  0.001 |  0.111  0.834  0.012 |
## Trip_distance          0.080  0.361  0.006 |  0.004  0.001  0.000 |
## Fare_amount            0.030  0.052  0.001 | -0.020  0.029  0.000 |
## Extra                  0.139  1.098  0.019 |  0.309  6.526  0.096 |
## MTA_tax                0.769 33.696  0.591 |  0.005  0.001  0.000 |
## Tip_amount            -0.015  0.012  0.000 | -0.151  1.552  0.023 |
## Tolls_amount           0.127  0.917  0.016 | -0.159  1.722  0.025 |
## improvement_surcharge  0.775 34.241  0.600 |  0.009  0.005  0.000 |
## trip_length            0.112  0.711  0.012 |  0.030  0.062  0.001 |
## trip_distance_km       0.080  0.361  0.006 |  0.004  0.001  0.000 |
## travel_time           -0.027  0.043  0.001 | -0.041  0.117  0.002 |
## pick_up_hour          -0.012  0.009  0.000 |  0.101  0.692  0.010 |
##
## Supplementary continuous variable
##                          Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
## Total_amount          |  0.963  0.928 | -0.025  0.001 |  0.046  0.002 |
```

```
##                          Dim.4   cos2
## Total_amount           -0.050  0.002 |
##
## Supplementary categories
##                               Dist     Dim.1    cos2  v.test     Dim.2
## AnyTip No                |   0.719 |  -0.376   0.274 -14.306 |   0.355
## AnyTip Yes               |   0.993 |   0.520   0.274  14.306 |  -0.490
##                             cos2  v.test     Dim.3    cos2  v.test     Dim.4
## AnyTip No                  0.244  18.769 |   0.048   0.005   3.022 |   0.171
## AnyTip Yes                 0.244 -18.769 |  -0.067   0.005  -3.022 |  -0.236
##                             cos2  v.test
## AnyTip No                  0.057  11.672 |
## AnyTip Yes                 0.057 -11.672 |
```

```
#The summary confirms the correlations between the variables that we already interpreted from the plots
#The plot show us that individuals that had to pay more tend to leave a tip.
plot.PCA(res.pca, choix=c("ind"),cex=0.8,col.ind="grey80",select="contrib15",axes=c(1,2))
```




```
#DIMENSION1
#Since the multivariant detection didnt manage to find outliers well enogh we are going to obtain them

#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.6199 -1.4527 -0.6964  0.0000  0.7312 14.0375
```

```
iqrvar<-IQR(res.pca$ind$coord[,1])
quantil3<-quantile(res.pca$ind$coord[,1], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.7312285
```

```
outliers<-which(res.pca$ind$coord[,1]>(iqrvar*3)+quantil3);length(outliers)
```

```
## [1] 74
```

```
df$f.outlierPCAd1<-0
df[outliers,"f.outlierPCAd1"]<-1
df$f.outlierPCAd1<-factor(df$f.outlierPCAd1,labels=c("NoOutDim1", "YesOutDim1"))
summary(df$f.outlierPCAd1)
```

```
##  NoOutDim1 YesOutDim1
##       4869         74
```

```
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
```

```
#catdes(,names(df)[c(22)])

#DIMENSION2
#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9574 -1.2216  0.2102  0.0000  0.8422 10.8960
```

```
iqrvar<-IQR(res.pca$ind$coord[,2])
quantil3<-quantile(res.pca$ind$coord[,2], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.8421855
```

```
outliers2<-which(res.pca$ind$coord[,2]>(iqrvar*3)+quantil3);length(outliers2)
```

```
## [1] 46
```

```
df$f.outlierPCAd2<-0
df[outliers2,"f.outlierPCAd2"]<-1
df$f.outlierPCAd2<-factor(df$f.outlierPCAd2,labels=c("NoOutDim2", "YesOutDim2"))
summary(df$f.outlierPCAd2)
```

```
##  NoOutDim2 YesOutDim2
##       4897         46
```

```
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"
```

```
#DIMENSION3
#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,3])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.3871  -0.5798   0.3402   0.0000   0.6577   5.1735
```

```
iqrvar<-IQR(res.pca$ind$coord[,3])
quantil3<-quantile(res.pca$ind$coord[,3], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.6577473
```

```
outliers3<-which(res.pca$ind$coord[,3]>(iqrvar*3)+quantil3);length(outliers3)
```

```
## [1] 1
```

```
df$f.outlierPCAd3<-0
df$f.outlierPCAd3<-factor(df$f.outlierPCAd3,labels=c("NoOutDim3"))
summary(df$f.outlierPCAd3)
```

```
## NoOutDim3
##      4943
```

```
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"          "f.outlierPCAd3"
```

```
#DIMENSION4
#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,4])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.85510 -0.96172 -0.07117  0.00000  0.69198  4.48903
```

```
iqrvar<-IQR(res.pca$ind$coord[,4])
quantil3<-quantile(res.pca$ind$coord[,4], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.6919838
```

```
outliers4<-which(res.pca$ind$coord[,4]>(iqrvar*3)+quantil3);length(outliers4)
```

```
## [1] 0
```

```
df$f.outlierPCAd4<-0
df$f.outlierPCAd4<-factor(df$f.outlierPCAd4,labels=c("NoOutDim4"))
summary(df$f.outlierPCAd4)
```

```
## NoOutDim4
##      4943
```

```
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
```

```
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"          "f.outlierPCAd3"
## [47] "f.outlierPCAd4"
```

```
#Finally we obtained 121 extreme outliers.
llvout<- c(outliers, outliers2, outliers3, outliers4);length(llvout)
```

```
## [1] 121
```

```
df$f.outlierPCA<-0
df[llvout,"f.outlierPCA"]<-1
df$f.outlierPCA<-factor(df$f.outlierPCA,labels=c("NoOut", "YesOut"))
summary(df$f.outlierPCA)
```

```
##  NoOut YesOut
##   4824    119
```

```
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
```

```
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"          "f.outlierPCAd3"
## [47] "f.outlierPCAd4"          "f.outlierPCA"
```

```
catdes_var <-c(1,4:48)
# We see that the most linked variables to the outliers are pickup and droppoff latitude and longtitude
catdes(df[,catdes_var], num.var = 46)
```

```
## $test.chi2
##                               p.value df
## f.outlierPCAd1           0.000000e+00  1
## f.outlierPCAd2           0.000000e+00  1
## f.outlierPCAd3           0.000000e+00  1
## f.outlierPCAd4           0.000000e+00  1
## RateCodeID              3.085549e-276  1
## Trip_type               2.907539e-246  1
## f.MTA_tax               1.148177e-241  1
## f.Improvement_surcharge 2.928586e-240  1
## f.ttime                  4.384970e-36  3
## f.distance               1.281262e-34  3
## f.fare_amount            7.583543e-33  3
## f.total                  9.124908e-26  3
## f.dropoff_longitude      6.772431e-18  3
## f.toll                   3.836078e-16  1
## f.espeed                 3.644908e-10  3
## f.pickup_longitude       3.708334e-08  3
## f.pickup_latitude        5.033577e-08  3
## f.dropoff_latitude       5.258935e-06  3
## f.extra                  3.202594e-03  1
##
## $category
## $category$NoOut
##                                      Cla/Mod     Mod/Cla     Global
## f.outlierPCAd1=NoOutDim1            99.07579 100.0000000 98.5029334
## f.outlierPCAd2=NoOutDim2            98.50929 100.0000000 99.0693911
## RateCodeID=Standard rate            98.69942  99.1086235 97.9971677
## f.MTA_tax=(0.4,0.5]                 98.57820  99.1708126 98.1792434
## Trip_type=Street-hail               98.55908  99.2537313 98.2803965
## f.Improvement_surcharge=(0.1,0.8]   98.55848  99.2122720 98.2399353
## f.fare_amount=(0,6]                 99.92006  25.9121061 25.3085171
## f.dropoff_longitude=(-73.97,-73.94] 99.71140  28.6484245 28.0396520
## f.ttime=(6,9.78]                    99.75166  24.9792703 24.4386000
## f.total=(-1,7.8]                    99.68077  25.8913765 25.3489784
## f.distance=(1.01,1.8]               99.59050  25.2072968 24.7015982
## f.toll=(-1,1]                       97.82028  98.6111111 98.3815497
## f.espeed=(15.3,20.1]                99.22945  24.0257048 23.6293749
## f.ttime=(9.78,15.7]                 99.17763  25.0000000 24.6004451
## f.distance=(1.8,3.31]               99.16667  24.6683250 24.2767550
## f.fare_amount=(6,9]                 99.12281  25.7669983 25.3692090
## f.pickup_longitude=(-73.96,-73.95]  99.05741  23.9635158 23.6091442
## f.total=(11,16.6]                   99.00249  24.6890547 24.3374469
## f.distance=(0,1.01]                 98.94137  25.1865672 24.8432126
## f.pickup_latitude=(40.74,40.8]      98.87387  27.3009950 26.9471981
## f.ttime=(-1,6]                      98.85901  25.1451078 24.8229820
```

```
## f.extra=(0.5,2]                      99.08973  15.7960199 15.5573538
## f.dropoff_latitude=(40.7,40.75]      98.52616  27.7155887 27.4529638
## f.dropoff_latitude=(40.75,40.79]     98.63760  22.5124378 22.2739227
## pick_up_period=afternoon             98.29060  35.7587065 35.5047542
## f.pickup_latitude=(40.7,40.74]       98.44291  23.5903814 23.3866073
## f.extra=(-0.1,0.5]                   97.31672  84.2039801 84.4426462
## f.dropoff_latitude=(40.79,41.5]      95.75912  23.4038143 23.8519118
## f.pickup_latitude=(40.8,40.92]       95.43189  23.8184080 24.3576775
## f.pickup_longitude=(-73.92,-73.79]   95.41801  24.6061360 25.1669027
## f.toll=(1,50]                        83.75000   1.3888889  1.6184503
## f.espeed=(26.2,95]                   95.32037  27.4461028 28.1003439
## f.dropoff_longitude=(-73.91,-73.75] 94.30962  23.3623549 24.1756019
## f.total=(16.6,70.1]                  93.69231  25.2487562 26.2998179
## f.fare_amount=(14,60.5]              93.04207  23.8391376 25.0050577
## f.distance=(3.31,19.8]               92.96754  24.9378109 26.1784341
## f.ttime=(15.7,415]                   92.87926  24.8756219 26.1379729
## f.Improvement_surcharge=(-0.1,0.1]   43.67816   0.7877280  1.7600647
## Trip_type=Dispatch                   42.35294   0.7462687  1.7196035
## f.MTA_tax=(-0.1,0.4]                 44.44444   0.8291874  1.8207566
## RateCodeID=Special rate              43.43434   0.8913765  2.0028323
## f.outlierPCAd2=YesOutDim2             0.00000   0.0000000  0.9306089
## f.outlierPCAd1=YesOutDim1             0.00000   0.0000000  1.4970666
##                                            p.value     v.test
## f.outlierPCAd1=NoOutDim1            3.561163e-133  24.557901
## f.outlierPCAd2=NoOutDim2             1.834601e-79  18.874903
## RateCodeID=Standard rate             6.661489e-70  17.673911
## f.MTA_tax=(0.4,0.5]                  2.888593e-61  16.514434
## Trip_type=Street-hail                4.407319e-61  16.488924
## f.Improvement_surcharge=(0.1,0.8]    2.280921e-60  16.389290
## f.fare_amount=(0,6]                  2.219549e-14   7.637224
## f.dropoff_longitude=(-73.97,-73.94]  1.382264e-12   7.085817
## f.ttime=(6,9.78]                     2.436425e-11   6.677145
## f.total=(-1,7.8]                     6.896731e-11   6.522886
## f.distance=(1.01,1.8]                1.403104e-09   6.055123
## f.toll=(-1,1]                        4.223017e-08   5.481262
## f.espeed=(15.3,20.1]                 3.973737e-06   4.612751
## f.ttime=(9.78,15.7]                  5.201243e-06   4.556502
## f.distance=(1.8,3.31]                7.384629e-06   4.482290
## f.fare_amount=(6,9]                  8.147041e-06   4.461283
## f.pickup_longitude=(-73.96,-73.95]   5.021972e-05   4.054602
## f.total=(11,16.6]                    7.556073e-05   3.958061
## f.distance=(0,1.01]                  1.358503e-04   3.815604
## f.pickup_latitude=(40.74,40.8]       1.451682e-04   3.799193
## f.ttime=(-1,6]                       3.772301e-04   3.555522
## f.extra=(0.5,2]                      1.264493e-03   3.223919
## f.dropoff_latitude=(40.7,40.75]      6.322005e-03   2.730595
## f.dropoff_latitude=(40.75,40.79]     7.188075e-03   2.688003
## pick_up_period=afternoon             1.569337e-02   2.415970
## f.pickup_latitude=(40.7,40.74]       2.620900e-02   2.223101
## f.extra=(-0.1,0.5]                   1.264493e-03  -3.223919
## f.dropoff_latitude=(40.79,41.5]      1.010526e-05  -4.414909
## f.pickup_latitude=(40.8,40.92]       1.606255e-07  -5.239944
## f.pickup_longitude=(-73.92,-73.79]   6.680935e-08  -5.399547
## f.toll=(1,50]                        4.223017e-08  -5.481262
```

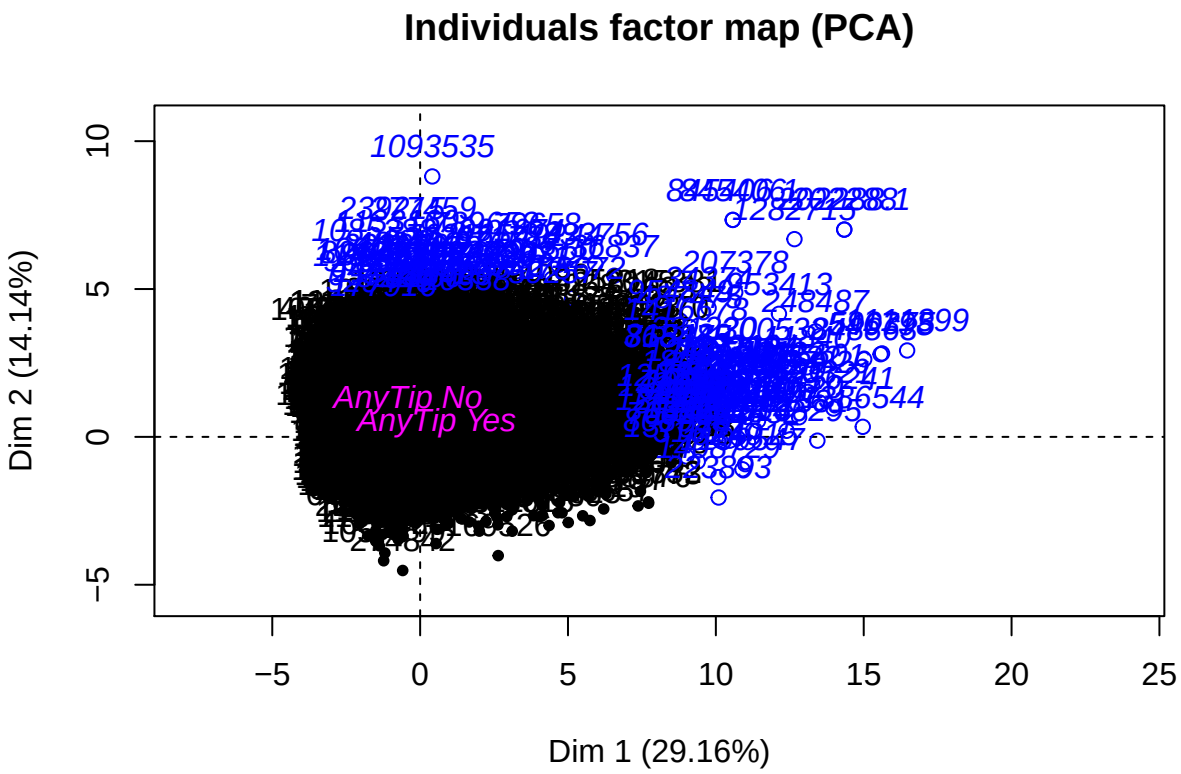
```
## f.espeed=(26.2,95]                    9.017320e-10  -6.125899
## f.dropoff_longitude=(-73.91,-73.75]  8.464512e-15  -7.760425
## f.total=(16.6,70.1]                   1.005865e-22  -9.811377
## f.fare_amount=(14,60.5]               7.543120e-28 -10.938508
## f.distance=(3.31,19.8]                5.369243e-31 -11.577330
## f.ttime=(15.7,415]                    4.705796e-32 -11.784252
## f.Improvement_surcharge=(-0.1,0.1]   2.280921e-60 -16.389290
## Trip_type=Dispatch                    4.407319e-61 -16.488924
## f.MTA_tax=(-0.1,0.4]                  2.888593e-61 -16.514434
## RateCodeID=Special rate               6.661489e-70 -17.673911
## f.outlierPCAd2=YesOutDim2             1.834601e-79 -18.874903
## f.outlierPCAd1=YesOutDim1            3.561163e-133 -24.557901
##
## $category$YesOut
##                                          Cla/Mod    Mod/Cla     Global
## f.outlierPCAd1=YesOutDim1           100.00000000 62.1848739  1.4970666
## f.outlierPCAd2=YesOutDim2           100.00000000 38.6554622  0.9306089
## RateCodeID=Special rate              56.56565657 47.0588235  2.0028323
## f.MTA_tax=(-0.1,0.4]                 55.55555556 42.0168067  1.8207566
## Trip_type=Dispatch                   57.64705882 41.1764706  1.7196035
## f.Improvement_surcharge=(-0.1,0.1]   56.32183908 41.1764706  1.7600647
## f.ttime=(15.7,415]                    7.12074303 77.3109244 26.1379729
## f.distance=(3.31,19.8]                7.03245750 76.4705882 26.1784341
## f.fare_amount=(14,60.5]               6.95792880 72.2689076 25.0050577
## f.total=(16.6,70.1]                   6.30769231 68.9075630 26.2998179
## f.dropoff_longitude=(-73.91,-73.75]   5.69037657 57.1428571 24.1756019
## f.espeed=(26.2,95]                    4.67962563 54.6218487 28.1003439
## f.toll=(1,50]                        16.25000000 10.9243697  1.6184503
## f.pickup_longitude=(-73.92,-73.79]    4.58199357 47.8991597 25.1669027
## f.pickup_latitude=(40.8,40.92]        4.56810631 46.2184874 24.3576775
## f.dropoff_latitude=(40.79,41.5]       4.24088210 42.0168067 23.8519118
## f.extra=(-0.1,0.5]                    2.68327743 94.1176471 84.4426462
## f.pickup_latitude=(40.7,40.74]        1.55709343 15.1260504 23.3866073
## pick_up_period=afternoon              1.70940171 25.2100840 35.5047542
## f.dropoff_latitude=(40.75,40.79]      1.36239782 12.6050420 22.2739227
## f.dropoff_latitude=(40.7,40.75]       1.47383935 16.8067227 27.4529638
## f.extra=(0.5,2]                       0.91027308  5.8823529 15.5573538
## f.ttime=(-1,6]                        1.14099430 11.7647059 24.8229820
## f.pickup_latitude=(40.74,40.8]        1.12612613 12.6050420 26.9471981
## f.distance=(0,1.01]                   1.05863192 10.9243697 24.8432126
## f.total=(11,16.6]                     0.99750623 10.0840336 24.3374469
## f.pickup_longitude=(-73.96,-73.95]    0.94258783  9.2436975 23.6091442
## f.fare_amount=(6,9]                   0.87719298  9.2436975 25.3692090
## f.distance=(1.8,3.31]                 0.83333333  8.4033613 24.2767550
## f.ttime=(9.78,15.7]                   0.82236842  8.4033613 24.6004451
## f.espeed=(15.3,20.1]                  0.77054795  7.5630252 23.6293749
## f.toll=(-1,1]                         2.17972445 89.0756303 98.3815497
## f.distance=(1.01,1.8]                 0.40950041  4.2016807 24.7015982
## f.total=(-1,7.8]                      0.31923384  3.3613445 25.3489784
## f.ttime=(6,9.78]                      0.24834437  2.5210084 24.4386000
## f.dropoff_longitude=(-73.97,-73.94]   0.28860029  3.3613445 28.0396520
## f.fare_amount=(0,6]                   0.07993605  0.8403361 25.3085171
## f.Improvement_surcharge=(0.1,0.8]     1.44151565 58.8235294 98.2399353
## Trip_type=Street-hail                 1.44092219 58.8235294 98.2803965
```

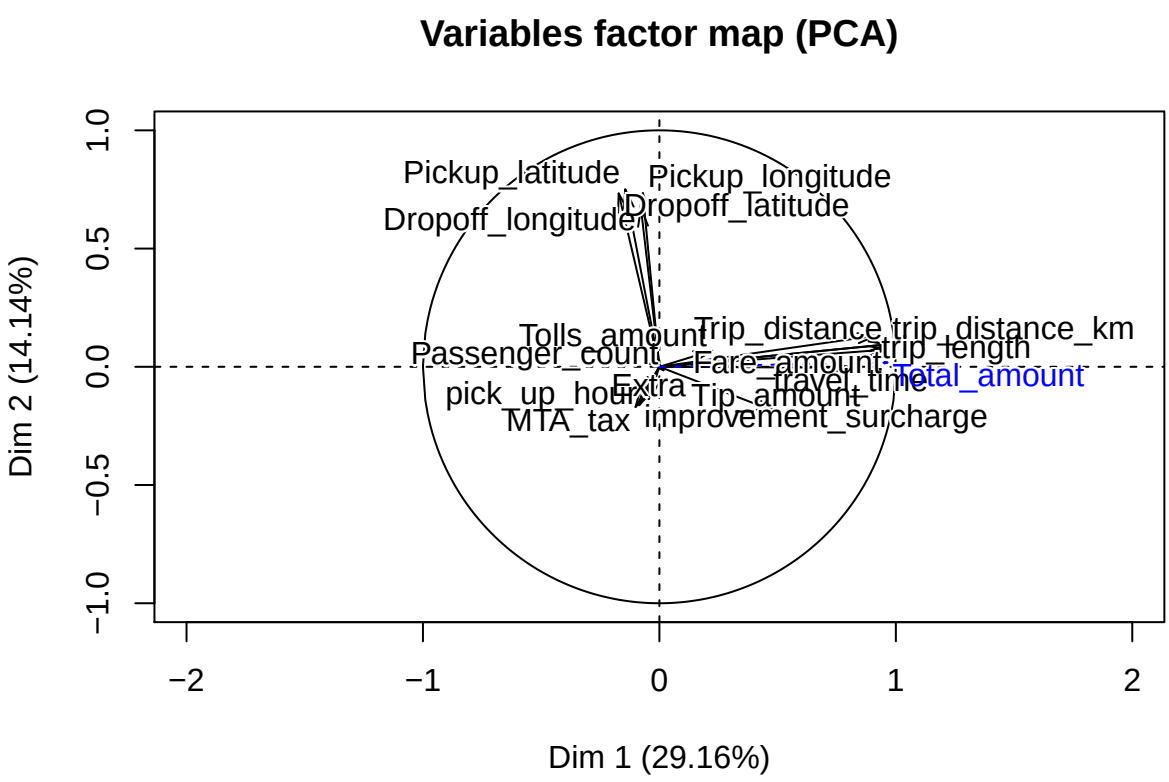
```
## f.MTA_tax=(0.4,0.5]                    1.42180095 57.9831933 98.1792434
## RateCodeID=Standard rate               1.30057803 52.9411765 97.9971677
## f.outlierPCAd2=NoOutDim2               1.49070860 61.3445378 99.0693911
## f.outlierPCAd1=NoOutDim1               0.92421442 37.8151261 98.5029334
##                                            p.value     v.test
## f.outlierPCAd1=YesOutDim1           3.561163e-133  24.557901
## f.outlierPCAd2=YesOutDim2            1.834601e-79  18.874903
## RateCodeID=Special rate              6.661489e-70  17.673911
## f.MTA_tax=(-0.1,0.4]                 2.888593e-61  16.514434
## Trip_type=Dispatch                   4.407319e-61  16.488924
## f.Improvement_surcharge=(-0.1,0.1]   2.280921e-60  16.389290
## f.ttime=(15.7,415]                   4.705796e-32  11.784252
## f.distance=(3.31,19.8]               5.369243e-31  11.577330
## f.fare_amount=(14,60.5]              7.543120e-28  10.938508
## f.total=(16.6,70.1]                  1.005865e-22   9.811377
## f.dropoff_longitude=(-73.91,-73.75]  8.464512e-15   7.760425
## f.espeed=(26.2,95]                   9.017320e-10   6.125899
## f.toll=(1,50]                        4.223017e-08   5.481262
## f.pickup_longitude=(-73.92,-73.79]   6.680935e-08   5.399547
## f.pickup_latitude=(40.8,40.92]       1.606255e-07   5.239944
## f.dropoff_latitude=(40.79,41.5]      1.010526e-05   4.414909
## f.extra=(-0.1,0.5]                   1.264493e-03   3.223919
## f.pickup_latitude=(40.7,40.74]       2.620900e-02  -2.223101
## pick_up_period=afternoon             1.569337e-02  -2.415970
## f.dropoff_latitude=(40.75,40.79]     7.188075e-03  -2.688003
## f.dropoff_latitude=(40.7,40.75]      6.322005e-03  -2.730595
## f.extra=(0.5,2]                      1.264493e-03  -3.223919
## f.ttime=(-1,6]                       3.772301e-04  -3.555522
## f.pickup_latitude=(40.74,40.8]       1.451682e-04  -3.799193
## f.distance=(0,1.01]                  1.358503e-04  -3.815604
## f.total=(11,16.6]                    7.556073e-05  -3.958061
## f.pickup_longitude=(-73.96,-73.95]   5.021972e-05  -4.054602
## f.fare_amount=(6,9]                  8.147041e-06  -4.461283
## f.distance=(1.8,3.31]                7.384629e-06  -4.482290
## f.ttime=(9.78,15.7]                  5.201243e-06  -4.556502
## f.espeed=(15.3,20.1]                 3.973737e-06  -4.612751
## f.toll=(-1,1]                        4.223017e-08  -5.481262
## f.distance=(1.01,1.8]                1.403104e-09  -6.055123
## f.total=(-1,7.8]                     6.896731e-11  -6.522886
## f.ttime=(6,9.78]                     2.436425e-11  -6.677145
## f.dropoff_longitude=(-73.97,-73.94]  1.382264e-12  -7.085817
## f.fare_amount=(0,6]                  2.219549e-14  -7.637224
## f.Improvement_surcharge=(0.1,0.8]    2.280921e-60 -16.389290
## Trip_type=Street-hail                4.407319e-61 -16.488924
## f.MTA_tax=(0.4,0.5]                  2.888593e-61 -16.514434
## RateCodeID=Standard rate             6.661489e-70 -17.673911
## f.outlierPCAd2=NoOutDim2             1.834601e-79 -18.874903
## f.outlierPCAd1=NoOutDim1            3.561163e-133 -24.557901
##
##
## $quanti.var
##                               Eta2       P-value
## MTA_tax                0.222964048 5.192282e-273
## improvement_surcharge  0.215679649 5.419249e-263
```

```
## Trip_distance         0.188881014 6.455986e-227
## trip_distance_km      0.188881014 6.455986e-227
## Fare_amount           0.172571834 1.544587e-205
## trip_length           0.162297813 2.769369e-192
## Total_amount          0.157624268 2.617730e-186
## travel_time           0.096019080 1.796679e-110
## Dropoff_longitude     0.026064082  3.221637e-30
## Tip_amount            0.022721292  1.635468e-26
## espeed                0.014665713  1.301418e-17
## Tolls_amount          0.014340480  2.973345e-17
## Pickup_longitude      0.008814615  3.747067e-11
## mis_ind               0.007463466  1.174318e-09
## Extra                 0.006675267  8.800261e-09
## Pickup_latitude       0.003228785  6.409570e-05
## Dropoff_latitude      0.001891703  2.224009e-03
##
## $quanti
## $quanti$NoOut
##                           v.test Mean in category Overall mean
## MTA_tax                33.194703       0.49585406   0.49089622
## improvement_surcharge  32.647953       0.29773425   0.29481489
## Extra                   5.743620       0.35271559   0.34806797
## Dropoff_latitude       -3.057580      40.74404515  40.74443156
## Pickup_latitude        -3.994578      40.74596616  40.74646566
## mis_ind                -6.073257       0.02300995   0.02629982
## Pickup_longitude       -6.600138     -73.93683412 -73.93622358
## Tolls_amount           -8.418471       0.07932421   0.09350192
## espeed                 -8.513398      22.08472780  22.28520498
## Tip_amount            -10.596633       1.15217454   1.19928788
## Dropoff_longitude     -11.349392     -73.93669909 -73.93547744
## travel_time           -21.783624      12.13813362  12.65794487
## Total_amount          -27.910198      13.55080846  14.13787376
## trip_length           -28.320943       4.30691367   4.57452370
## Fare_amount           -29.203596      11.17300580  11.70873776
## Trip_distance         -30.552414       2.55628323   2.73668393
## trip_distance_km      -30.552414       4.11393908   4.40426587
##                       sd in category  Overall sd       p.value
## MTA_tax                   0.04534071  0.06685067 1.283715e-241
## improvement_surcharge     0.02737907  0.04002327 8.567615e-234
## Extra                     0.36255737  0.36218118  9.267368e-09
## Dropoff_latitude          0.05557847  0.05656589  2.231319e-03
## Pickup_latitude           0.05525463  0.05596793  6.480951e-05
## mis_ind                   0.23235665  0.24245901  1.253417e-09
## Pickup_longitude          0.04078401  0.04140391  4.107739e-11
## Tolls_amount              0.68675249  0.75379701  3.814303e-17
## espeed                   10.19814341 10.54007035  1.689085e-17
## Tip_amount                1.86240658  1.99002040  3.089027e-26
## Dropoff_longitude         0.04666352  0.04817875  7.468044e-30
## travel_time               9.44475500 10.68063515 3.317601e-105
## Total_amount              8.10688270  9.41466867 2.006558e-171
## trip_length               3.56906208  4.22937528 1.908743e-176
## Fare_amount               7.04240211  8.21093790 1.745704e-187
## Trip_distance             2.21692072  2.64286050 5.252237e-205
## trip_distance_km          3.56778807  4.25327169 5.252237e-205
```

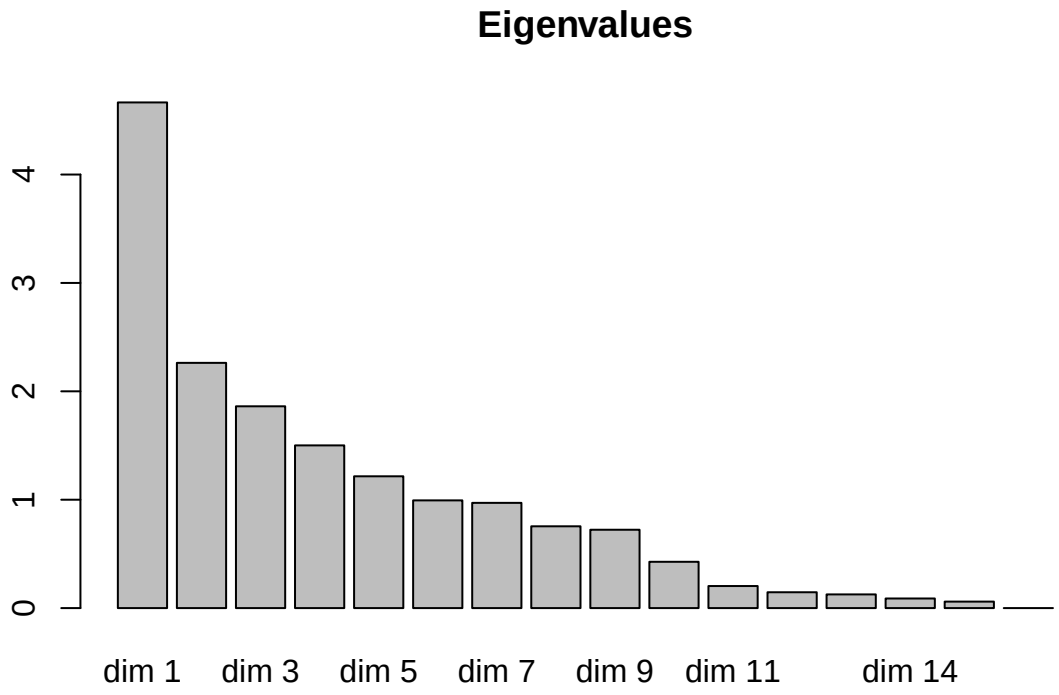
```
##
## $quanti$YesOut
##                            v.test Mean in category Overall mean
## Trip_distance           30.552414       10.0497343   2.73668393
## trip_distance_km        30.552414       16.1734795   4.40426587
## Fare_amount             29.203596       33.4261409  11.70873776
## trip_length             28.320943       15.4228496   4.57452370
## Total_amount            27.910198       37.9362185  14.13787376
## travel_time             21.783624       33.7299575  12.65794487
## Dropoff_longitude       11.349392      -73.8859544 -73.93547744
## Tip_amount              10.596633        3.1091597   1.19928788
## espeed                   8.513398       30.4121115  22.28520498
## Tolls_amount             8.418471        0.6682353   0.09350192
## Pickup_longitude         6.600138      -73.9114737 -73.93622358
## mis_ind                  6.073257        0.1596639   0.02629982
## Pickup_latitude          3.994578       40.7667140  40.74646566
## Dropoff_latitude         3.057580       40.7600959  40.74443156
## Extra                   -5.743620        0.1596639   0.34806797
## improvement_surcharge -32.647953        0.1764706   0.29481489
## MTA_tax               -33.194703        0.2899160   0.49089622
##                        sd in category  Overall sd       p.value
## Trip_distance              6.00807417  2.64286050 5.252237e-205
## trip_distance_km           9.66905811  4.25327169 5.252237e-205
## Fare_amount               17.51251556  8.21093790 1.745704e-187
## trip_length               10.29777621  4.22937528 1.908743e-176
## Total_amount              20.90944493  9.41466867 2.006558e-171
## travel_time               25.83355249 10.68063515 3.317601e-105
## Dropoff_longitude          0.07505847  0.04817875  7.468044e-30
## Tip_amount                 4.48908790  1.99002040  3.089027e-26
## espeed                    18.18980454 10.54007035  1.689085e-17
## Tolls_amount               2.03590859  0.75379701  3.814303e-17
## Pickup_longitude           0.05614219  0.04140391  4.107739e-11
## mis_ind                    0.48478000  0.24245901  1.253417e-09
## Pickup_latitude            0.07699582  0.05596793  6.480951e-05
## Dropoff_latitude           0.08623901  0.05656589  2.231319e-03
## Extra                      0.28939791  0.36218118  9.267368e-09
## improvement_surcharge      0.14764589  0.04002327 8.567615e-234
## MTA_tax                    0.24679286  0.06685067 1.283715e-241
##
##
## attr(,"class")
## [1] "catdes" "list "
```

```
#we execute Pca again with the too contributive individuals as suplementary
contr_ind = c(outliers, outliers2, outliers3, outliers4)
res.pca<-PCA(df[,vars_con_pca],ind.sup=contr_ind, quanti.sup = 13, quali.sup = 14, ncp = 6 ) # TotalAmo
```

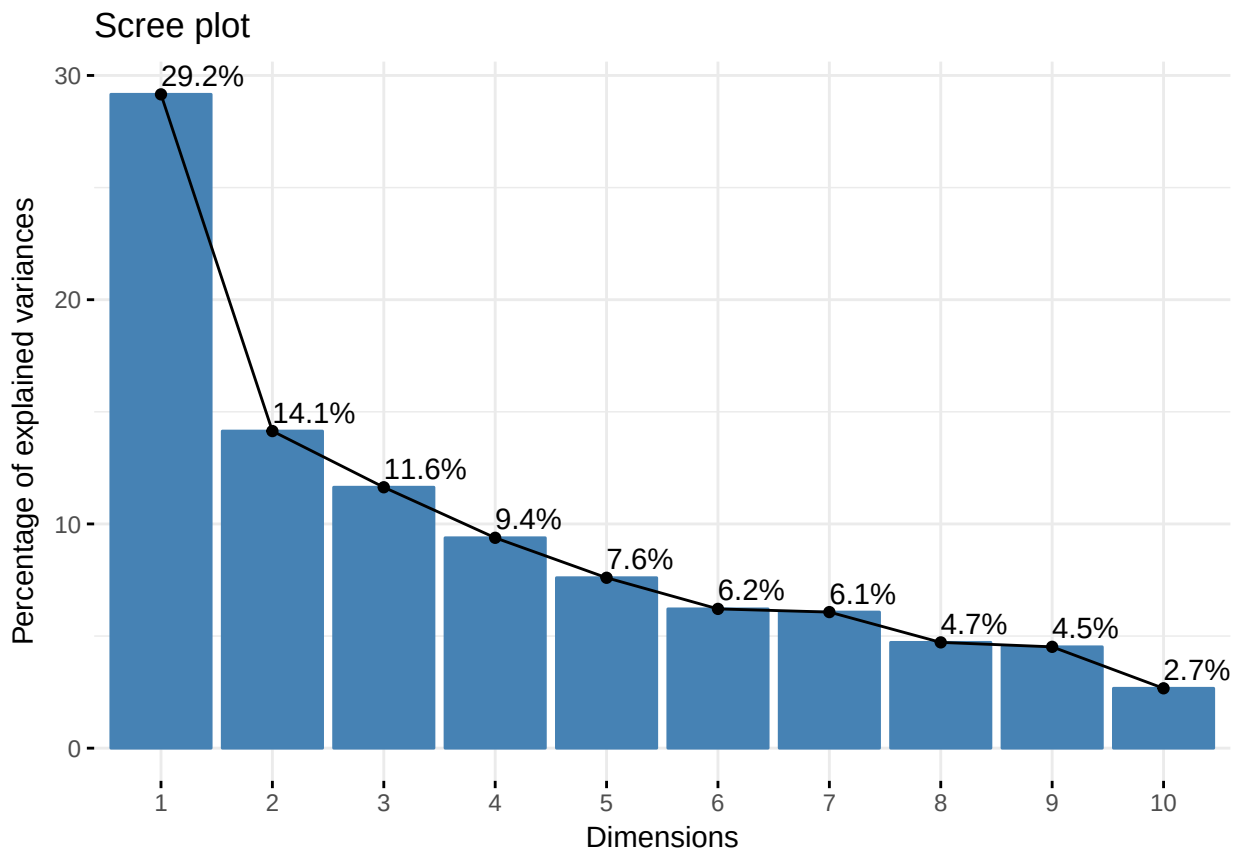

Individuals factor map (PCA)
Dim 2 (14.14%)
Dim 1 (29.16%)
AnyTip No
AnyTip Yes
1093535
207378
248487
−5
0
5
10
15
20
25



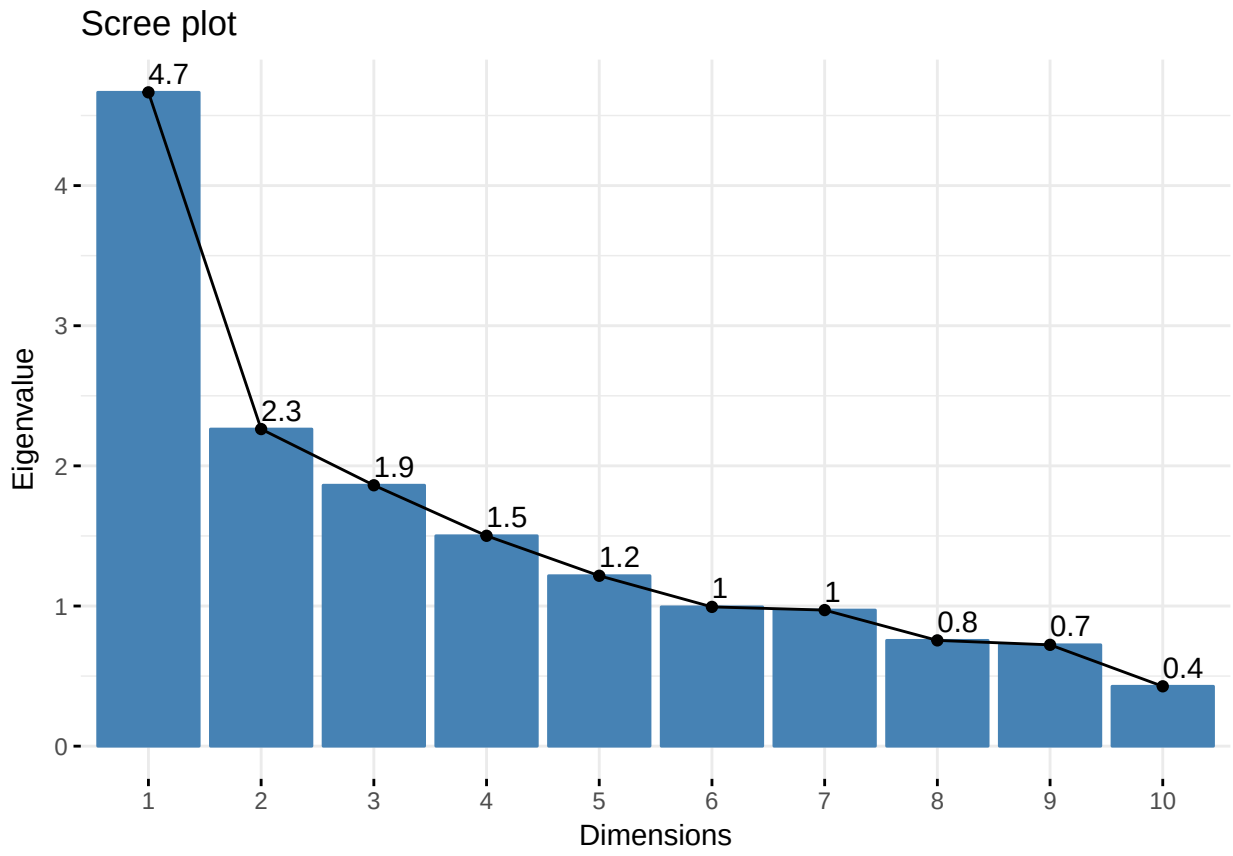

```r
barplot(res.pca$eig[,1], main="Eigenvalues", names.arg = paste("dim", 1:nrow(res.pca$eig)))
```




```
# With the PCA transformation the PC1 covers 31.6% of the variance, PC2 - 14.8%, PCA3 - 9.9%, PCA4 - 8.
kaiser <- res.pca$eig[1:7,1] #keep only EV >=1 ->first 7
#If we use the kaiser rule we have to keep all EV greater than 1, which results in saving the first 7 d
#facto extra
fviz_eig(res.pca, addlabels = TRUE)
```




```
fviz_eig(res.pca, choice = "eigenvalue",addlabels = TRUE)
```




```
#According to the elbow rule we have to take the first 4 dimentions as the slope of the graphic shows.
elbow <- res.pca$eig[1:4,1]
```

III Interpret axis

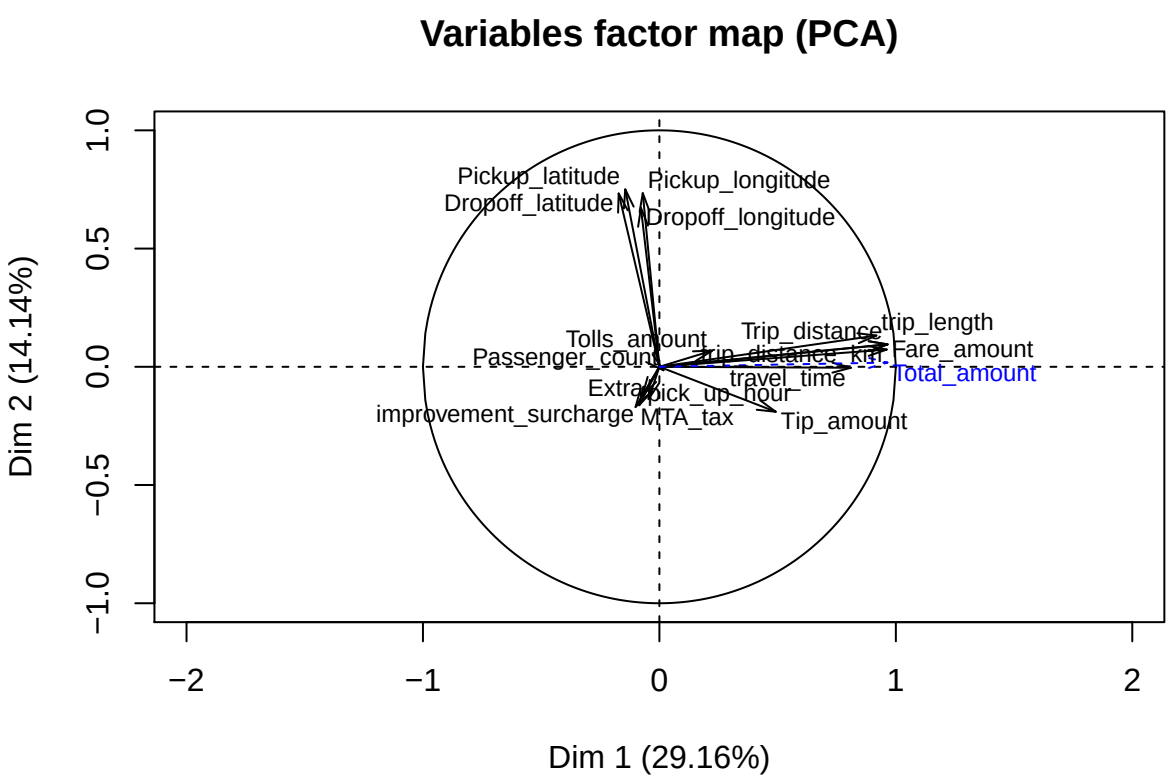
```
# Interential criteria

dimdesc (res.pca, axes=1:4)
```

```
## $Dim.1
## $Dim.1$quanti
##                         correlation       p.value
## trip_distance_km         0.96509559  0.000000e+00
## Trip_distance            0.96509559  0.000000e+00
## Total_amount             0.96454484  0.000000e+00
## Fare_amount              0.96172705  0.000000e+00
## trip_length              0.91784007  0.000000e+00
## travel_time              0.81022681  0.000000e+00
## Tip_amount               0.49226488 6.333255e-293
## Tolls_amount             0.22318147  1.625496e-55
## Extra                   -0.04568190  1.505238e-03
## Pickup_longitude        -0.07030646  1.019068e-06
## pick_up_hour            -0.07156043  6.513918e-07
## Dropoff_longitude       -0.07815542  5.473663e-08
## improvement_surcharge   -0.08575299  2.440014e-09
```

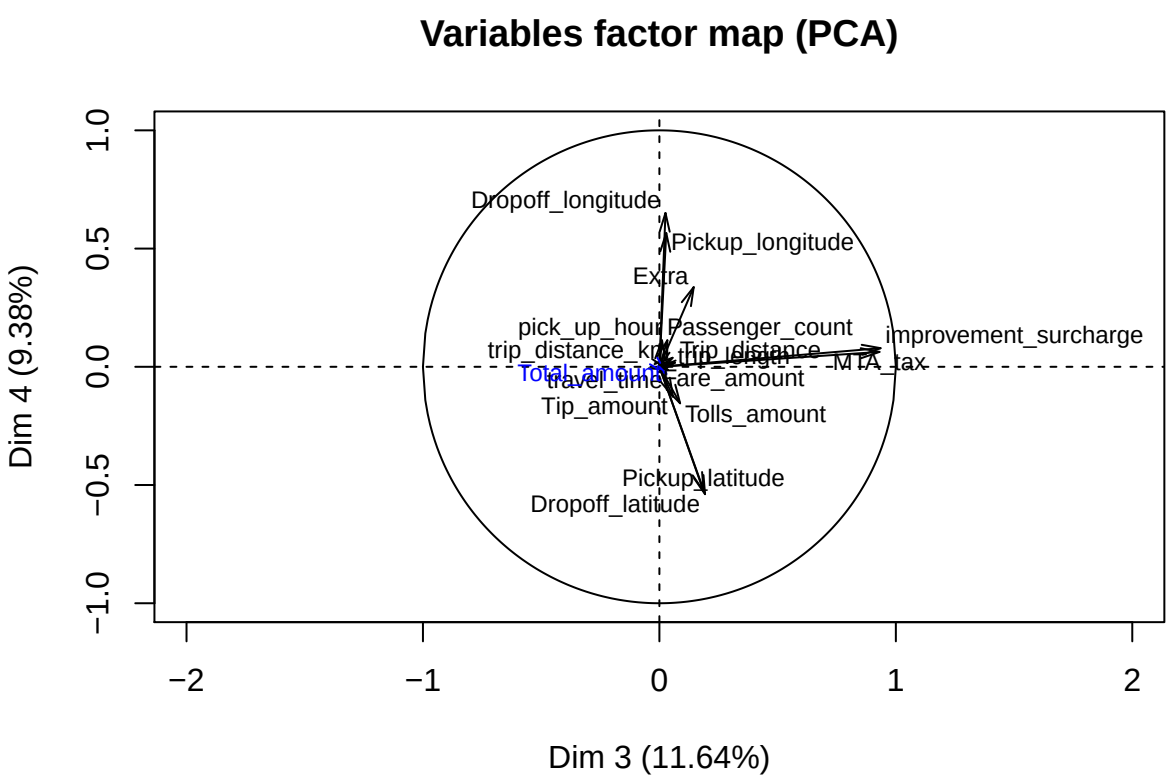
```
## MTA_tax               -0.10088958  2.165363e-12
## Pickup_latitude       -0.14431103  7.250616e-24
## Dropoff_latitude      -0.17219697  2.012981e-33
##
## $Dim.1$quali
##                R2      p.value
## AnyTip 0.04897507 1.362274e-54
##
## $Dim.1$category
##              Estimate      p.value
## AnyTip Yes  0.4840463 1.362274e-54
## AnyTip No  -0.4840463 1.362274e-54
##
##
## $Dim.2
## $Dim.2$quanti
##                         correlation      p.value
## Pickup_latitude          0.75045070 0.000000e+00
## Pickup_longitude         0.73446222 0.000000e+00
## Dropoff_latitude         0.73304058 0.000000e+00
## Dropoff_longitude        0.67231617 0.000000e+00
## trip_length              0.13216290 3.042306e-20
## Trip_distance            0.09489302 4.002228e-11
## trip_distance_km         0.09489302 4.002228e-11
## Fare_amount              0.07297183 3.901306e-07
## Tolls_amount             0.06828264 2.065743e-06
## pick_up_hour            -0.12077206 3.866547e-17
## Extra                   -0.13665526 1.517909e-21
## improvement_surcharge   -0.16279008 5.281567e-30
## MTA_tax                 -0.16997574 1.346573e-32
## Tip_amount              -0.19020752 1.563519e-40
##
## $Dim.2$quali
##                R2     p.value
## AnyTip 0.06254233 1.08729e-69
##
## $Dim.2$category
##              Estimate     p.value
## AnyTip No   0.3809245 1.08729e-69
## AnyTip Yes -0.3809245 1.08729e-69
##
##
## $Dim.3
## $Dim.3$quanti
##                         correlation      p.value
## improvement_surcharge    0.93553572 0.000000e+00
## MTA_tax                  0.93090316 0.000000e+00
## Dropoff_latitude         0.19312950 9.332083e-42
## Pickup_latitude          0.18697627 3.347236e-39
## Extra                    0.14512114 4.047844e-24
## Tolls_amount             0.08762732 1.085451e-09
## trip_distance_km         0.06478021 6.704044e-06
## Trip_distance            0.06478021 6.704044e-06
## Tip_amount               0.05746561 6.509376e-05
```

```
## trip_length            0.05596409 1.006468e-04
## travel_time            0.03592425 1.258584e-02
## pick_up_hour           0.03374261 1.909588e-02
## Total_amount           0.02999217 3.724797e-02
## Pickup_longitude       0.02899418 4.404177e-02
##
## $Dim.3$quali
##                 R2      p.value
## AnyTip 0.002861887 0.0002013579
##
## $Dim.3$category
##               Estimate      p.value
## AnyTip Yes  0.07391661 0.0002013579
## AnyTip No  -0.07391661 0.0002013579
##
##
## $Dim.4
## $Dim.4$quanti
##                       correlation       p.value
## Dropoff_longitude      0.65049262  0.000000e+00
## Pickup_longitude       0.56582547  0.000000e+00
## Extra                  0.33685426 2.630864e-128
## pick_up_hour           0.11215661  5.607321e-15
## Passenger_count        0.11186214  6.604082e-15
## improvement_surcharge  0.07871419  4.395574e-08
## MTA_tax                0.06292403  1.222190e-05
## trip_length            0.03817211  8.012940e-03
## Total_amount          -0.03066770  3.317370e-02
## Tip_amount            -0.12635852  1.259431e-18
## Tolls_amount          -0.15443404  3.912992e-27
## Pickup_latitude       -0.52601666  0.000000e+00
## Dropoff_latitude      -0.53874029  0.000000e+00
##
## $Dim.4$quali
##                 R2    p.value
## AnyTip 0.02219539 2.4047e-25
##
## $Dim.4$category
##              Estimate    p.value
## AnyTip No   0.1848331 2.4047e-25
## AnyTip Yes -0.1848331 2.4047e-25
```

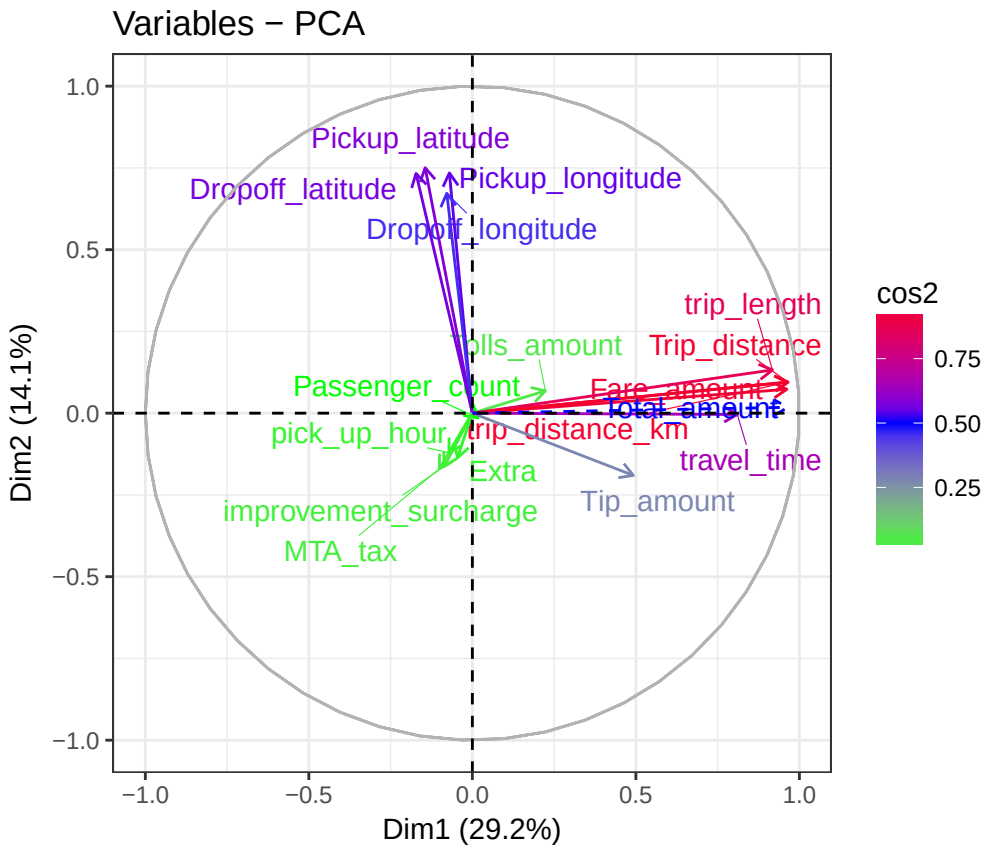
```
#The first Dimention is best described by the quantative variables Total_amount, trip_distance and Fare
plot(res.pca,choix="var", cex = 0.75)
```



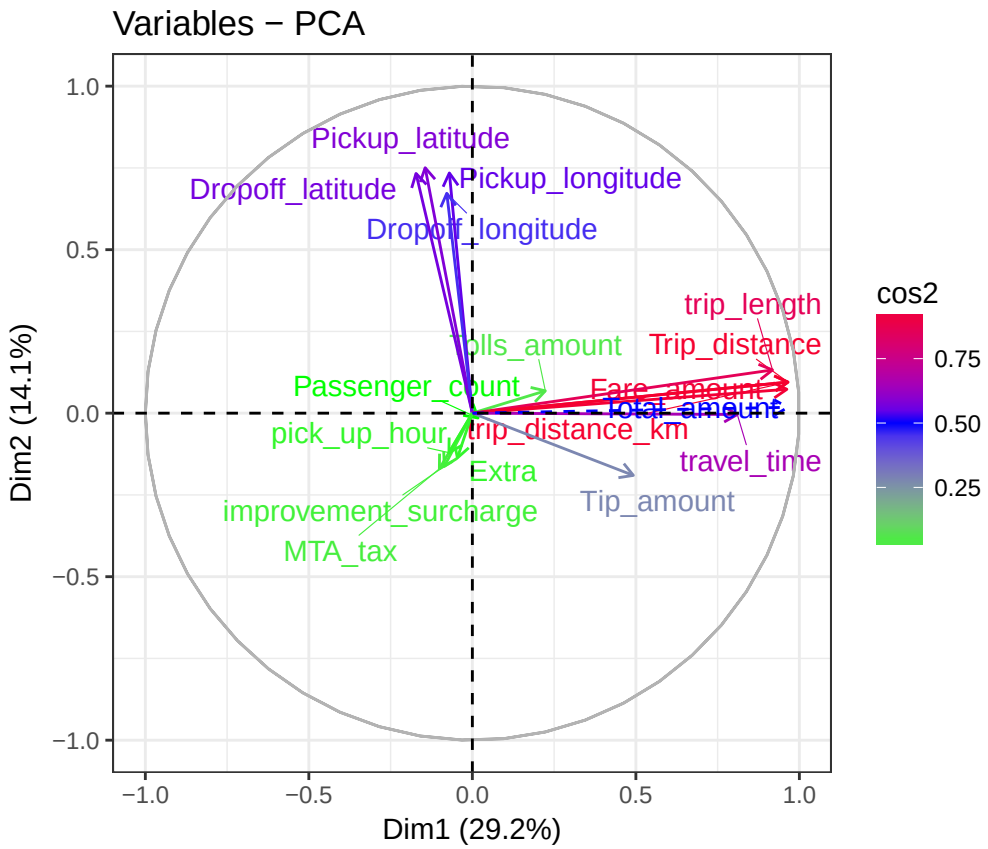

```
plot(res.pca,choix="var", cex = 0.75, axes = (3:4))# 3rd and 4th PCA
```




```
#modern factoextra
# Components with higher cos^2 are more importaint to interpret both active and supplementary observati
fviz_pca_var(res.pca,col.var="cos2", repel=TRUE)+scale_color_gradient2(low="green", mid="blue", high="r
```




```
fviz_pca_var(res.pca,col.var="cos2", repel=TRUE)+scale_color_gradient2(low="green", mid="blue", high="r
```



#Leveling For better understanding of the results we will name the levels of the factorial variables

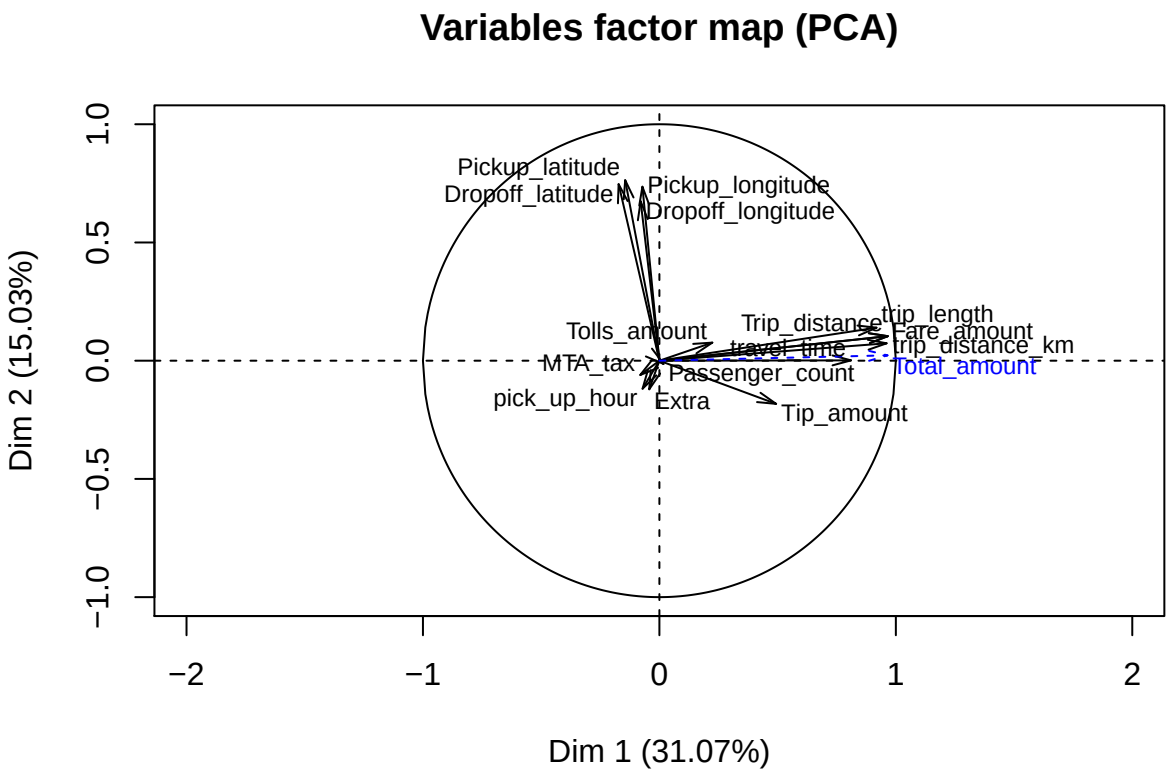
```
levels(df$f.fare_amount)
```

```
## [1] "(0,6]"     "(6,9]"     "(9,14]"    "(14,60.5]"
```

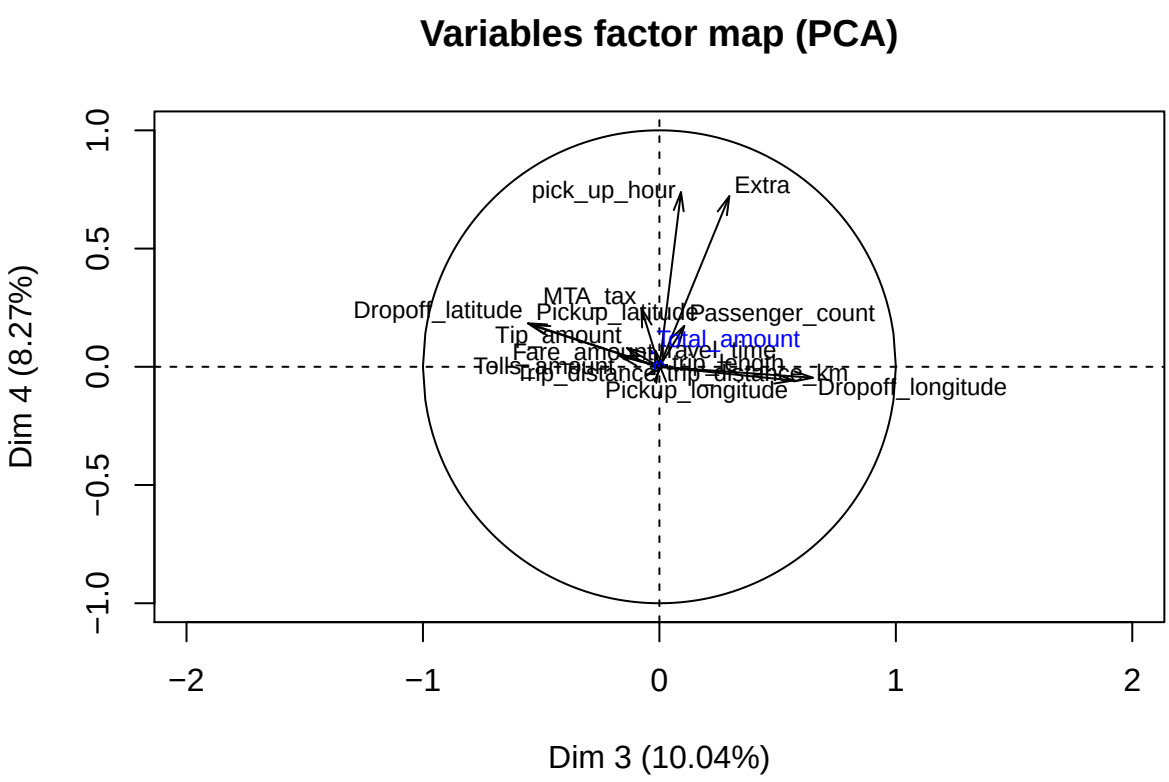
```
levels(df$f.passenger) <- c("onePassager","multiPassagers")
levels(df$f.pickup_longitude) <- c("p.Y1","p.Y2","p.Y3","p.Y4") #pickup->p., longtitude->Y
levels(df$f.pickup_latitude) <- c("p.X1","p.X2","p.X3","p.X4") #pickup->p., latitude->X
levels(df$f.distance) <- c("Dist1","Dist2", "Dist3", "Dist4")#distance 1-4, 1 shortest
levels(df$f.dropoff_longitude) <- c("d.Y1","d.Y2", "d.Y3", "d.Y4")
levels(df$f.dropoff_latitude) <- c("d.X1","d.X2", "d.X3", "d.X4")
levels(df$f.fare_amount) <- c("FAmount1","FAmount2", "FAmount3", "FAmount4")
levels(df$f.extra) <- c("smallExtra","highExtra")
levels(df$f.MTA_tax) <- c("smallMTA","highMTA")
levels(df$f.Improvement_surcharge) <- c("smallSurcharge","highSurcharge")
levels(df$f.tip_amount) <- c("smallTip","highTip")
levels(df$f.toll) <- c("smallToll","highToll")
levels(df$f.total) <- c("CheapestTrip","CheapTrip","MediumTrip","ExpensiveTrip")
levels(df$f.ttime) <- c("Time1","Time2", "Time3", "Time4")
levels(df$f.espeed) <- c("Speed1","Speed2", "Speed3","Speed4")
levels(df$f.outlierPCAd1) <- c("Normald1", "Outlierd1")
levels(df$f.outlierPCAd2) <- c("Normald2", "Outlierd2")
levels(df$f.outlierPCAd3) <- c("Normald3", "Outlierd3")
levels(df$f.outlierPCAd4) <- c("Normald4", "Outlierd4")
levels(df$f.outlierPCA) <- c("Normal", "Outlier")
```

IV PCA execution with supplementary individuals

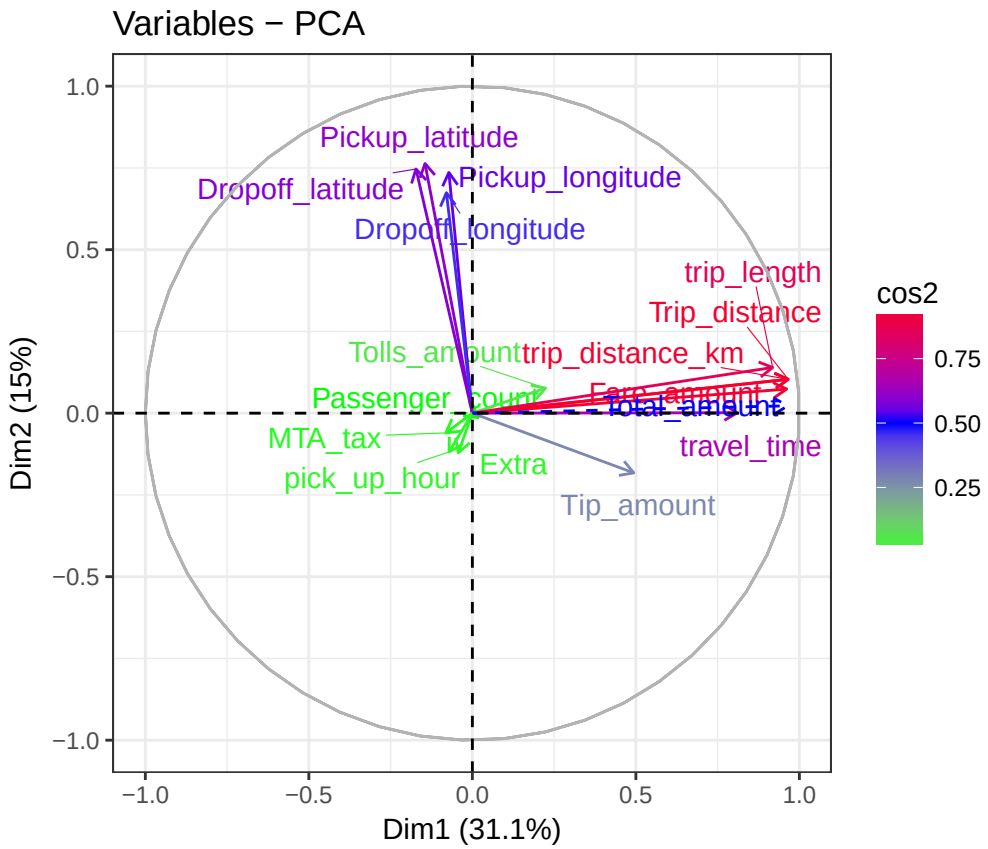
```
vars_con_pca <- names(df)[c(6:16,18,23:26)]
# We do a PCA analysis using the factorial variables Fare amount, total and the pickup perio in order t

res.pca<-PCA(df[,c(vars_con_pca, "f.fare_amount", "f.total", "pick_up_period", "f.passenger", "f.pickup_

plot(res.pca,choix="var", cex = 0.75)
```



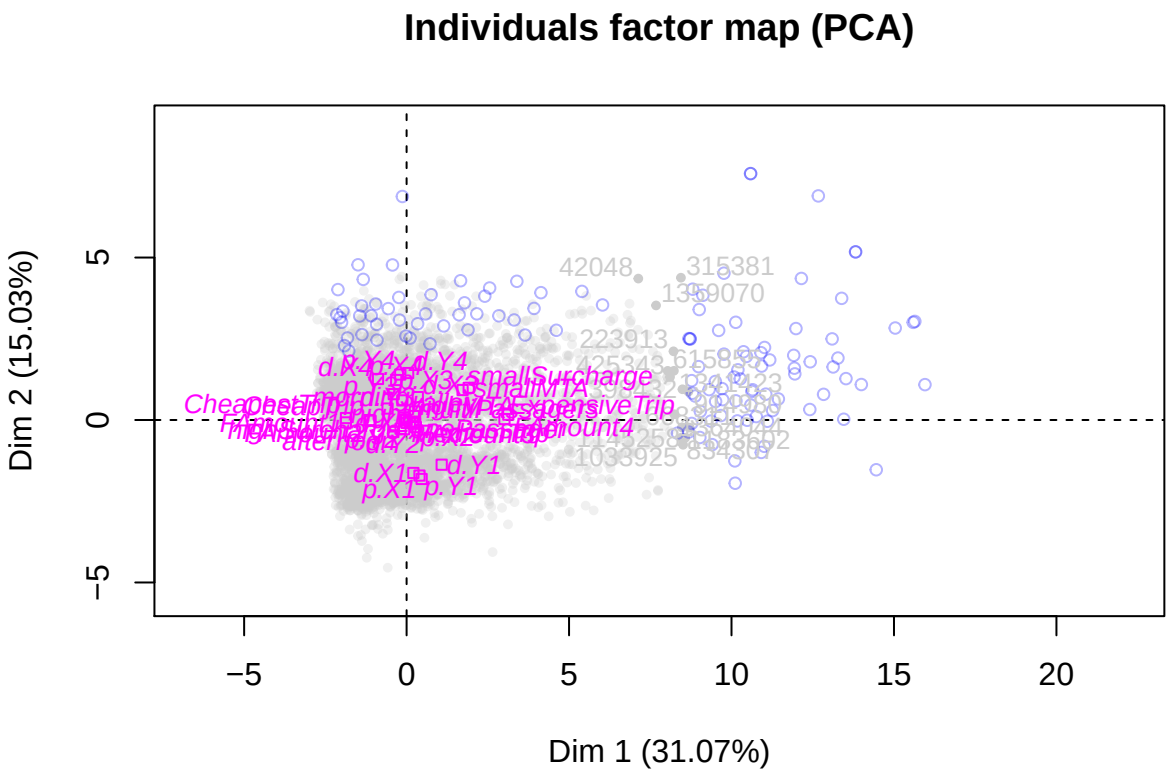

```
plot(res.pca,choix="var", cex = 0.75, axes = (3:4))# 3rd and 4th PCA
```




```
fviz_pca_var(res.pca,col.var="cos2", repel=TRUE)+scale_color_gradient2(low="green", mid="blue", high="r
```



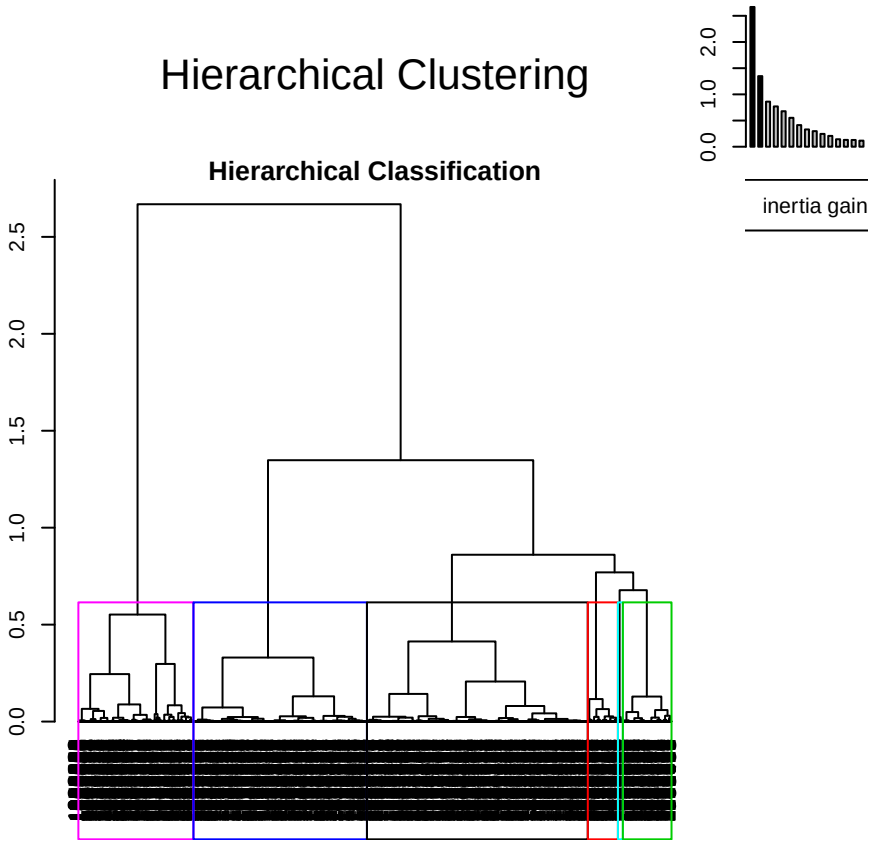

```
#We can see that trips in the afternoon tend to be longer and thus also more expensive than the ones du
# we see That the qualitative varibles are close to the center, so they are not of a big importaince fo
plot.PCA(res.pca, choix=c("ind"),cex=0.8,col.ind="grey80",select="contrib15",axes=c(1,2))
```



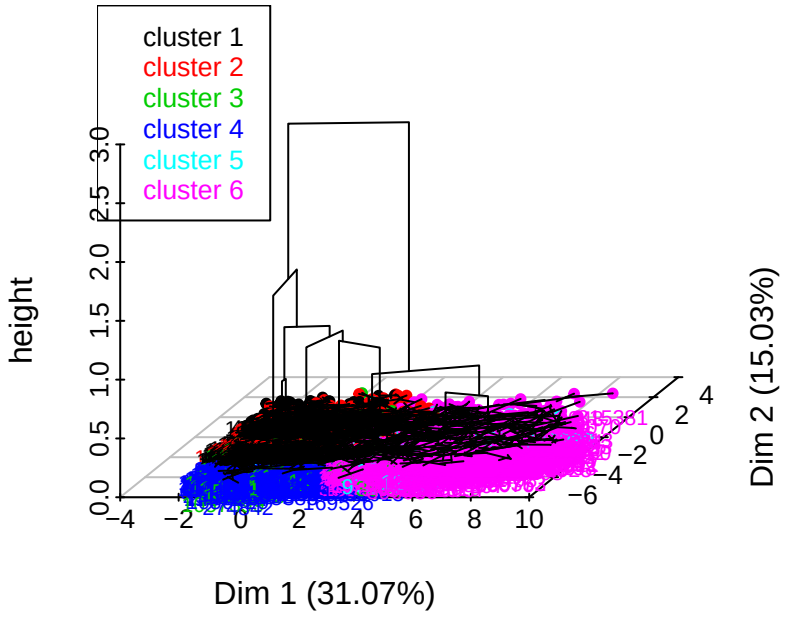
Hierarchical clustering

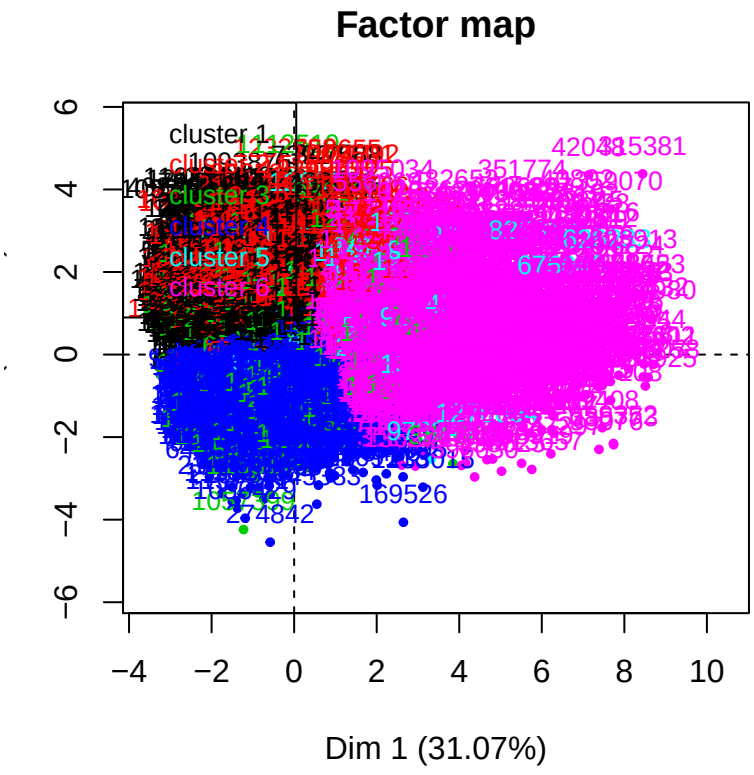
We generate 6 clusters using the hierarchical method, taking the projection obtained by the PCA as a source dataset. The resulting table is showing the distribution taken between the different clusters (excluding the multivariant outliers)

```
library(FactoMineR)
res.hcpc <-HCPC(res.pca, nb.clust = 6, order=TRUE)
```


Hierarchical Clustering
Hierarchical Classification
0.0
0.5
1.0
1.5
2.0
2.5
0.0
1.0
2.0
inertia gain

Hierarchical clustering on the factor map






```
table (res.hcpc$data.clust$clust)
```

```
##
##    1    2    3    4    5    6
## 1546  770  242 1520   40  706
```

Variable description

Below is listed the categorical description for each cluster and our explanation as it follows.

```
#Block A descripcion per variables
res.hcpc$desc.var
```

```
## $test.chi2
##                                 p.value df
## f.fare_amount             0.000000e+00 15
## f.total                   0.000000e+00 15
## f.pickup_longitude        0.000000e+00 15
## f.pickup_latitude         0.000000e+00 15
## f.dropoff_longitude       0.000000e+00 15
## f.dropoff_latitude        0.000000e+00 15
## f.Improvement_surcharge   0.000000e+00  5
## f.MTA_tax                 0.000000e+00  5
## f.passenger              1.049986e-310  5
## pick_up_period            1.689687e-37 15
##
```

```
## $category
## $category$`1`
##                                            Cla/Mod      Mod/Cla     Global
## f.dropoff_latitude=d.X4                  86.802480  63.38939198 23.4038143
## f.pickup_latitude=p.X4                   78.764143  58.53816300 23.8184080
## f.pickup_longitude=p.Y3                  60.921366  49.61190168 26.0986733
## f.dropoff_longitude=d.Y3                 56.783920  43.85510996 24.7512438
## f.pickup_latitude=p.X3                   47.304480  40.29754204 27.3009950
## f.total=CheapestTrip                     47.638110  38.48641656 25.8913765
## f.dropoff_longitude=d.Y2                 44.645441  39.90944373 28.6484245
## f.dropoff_latitude=d.X3                  46.685083  32.79430789 22.5124378
## f.fare_amount=FAmount1                   45.280000  36.61060802 25.9121061
## f.passenger=onePassager                  34.841076  92.17335058 84.7844113
## f.fare_amount=FAmount2                   41.995173  33.76455369 25.7669983
## f.pickup_longitude=p.Y2                  41.089965  30.72445019 23.9635158
## pick_up_period=morning                   40.599174  25.42043984 20.0663350
## f.total=CheapTrip                        39.536878  29.81888745 24.1708126
## f.MTA_tax=highMTA                        32.316054 100.00000000 99.1708126
## pick_up_period=valley                    37.490775  32.85899094 28.0887231
## f.Improvement_surcharge=highSurcharge    32.281655  99.93531695 99.2122720
## pick_up_period=afternoon                 29.159420  32.53557568 35.7587065
## f.Improvement_surcharge=smallSurcharge    2.631579   0.06468305  0.7877280
## f.MTA_tax=smallMTA                        0.000000   0.00000000  0.8291874
## pick_up_period=night                     18.298969   9.18499353 16.0862355
## f.passenger=multiPassagers               16.485014   7.82664942 15.2155887
## f.pickup_longitude=p.Y4                  18.197136  13.97153946 24.6061360
## f.dropoff_longitude=d.Y4                 13.043478   9.50840880 23.3623549
## f.dropoff_longitude=d.Y1                  9.277431   6.72703752 23.2379768
## f.fare_amount=FAmount4                    7.565217   5.62742561 23.8391376
## f.total=ExpensiveTrip                     7.471264   5.88615783 25.2487562
## f.pickup_longitude=p.Y1                   7.201309   5.69210867 25.3316750
## f.dropoff_latitude=d.X2                   4.412865   3.81630013 27.7155887
## f.pickup_latitude=p.X2                    1.581722   1.16429495 23.5903814
## f.pickup_latitude=p.X1                    0.000000   0.00000000 25.2902156
## f.dropoff_latitude=d.X1                   0.000000   0.00000000 26.3681592
##                                             p.value     v.test
## f.dropoff_latitude=d.X4                  0.000000e+00        Inf
## f.pickup_latitude=p.X4                  1.957686e-319  38.209510
## f.pickup_longitude=p.Y3                 1.904821e-137  24.954572
## f.dropoff_longitude=d.Y3                 2.598830e-94  20.602640
## f.pickup_latitude=p.X3                   1.576731e-42  13.667998
## f.total=CheapestTrip                     2.926549e-41  13.453748
## f.dropoff_longitude=d.Y2                 1.234387e-31  11.702716
## f.dropoff_latitude=d.X3                  1.314774e-30  11.500287
## f.fare_amount=FAmount1                   2.466878e-30  11.445845
## f.passenger=onePassager                  7.167529e-25  10.298357
## f.fare_amount=FAmount2                   8.107353e-18   8.598062
## f.pickup_longitude=p.Y2                  9.410087e-14   7.448929
## pick_up_period=morning                   3.339184e-10   6.282152
## f.total=CheapTrip                        5.056945e-10   6.217327
## f.MTA_tax=highMTA                        1.797721e-07   5.219123
## pick_up_period=valley                    5.124877e-07   5.021578
## f.Improvement_surcharge=highSurcharge    7.895675e-06   4.467993
## pick_up_period=afternoon                 1.294194e-03  -3.217264
```

```
## f.Improvement_surcharge=smallSurcharge  7.895675e-06  -4.467993
## f.MTA_tax=smallMTA                      1.797721e-07  -5.219123
## pick_up_period=night                    1.101049e-20  -9.325843
## f.passenger=multiPassagers              7.167529e-25 -10.298357
## f.pickup_longitude=p.Y4                 3.119354e-34 -12.199709
## f.dropoff_longitude=d.Y4                1.629942e-61 -16.548922
## f.dropoff_longitude=d.Y1                1.454075e-90 -20.180453
## f.fare_amount=FAmount4                 3.490642e-110 -22.302546
## f.total=ExpensiveTrip                  2.496279e-119 -23.225844
## f.pickup_longitude=p.Y1                6.819551e-123 -23.575851
## f.dropoff_latitude=d.X2                1.012764e-177 -28.424307
## f.pickup_latitude=p.X2                 1.685521e-189 -29.361875
## f.pickup_latitude=p.X1                 1.068745e-245 -33.476241
## f.dropoff_latitude=d.X1                1.793052e-258 -34.342976
##
## $category$`2`
##                                          Cla/Mod     Mod/Cla     Global
## f.dropoff_longitude=d.Y4               63.7976930  93.3766234 23.3623549
## f.pickup_longitude=p.Y4                61.8365628  95.3246753 24.6061360
## f.dropoff_latitude=d.X2                32.0119671  55.5844156 27.7155887
## f.pickup_latitude=p.X3                 31.3591496  53.6363636 27.3009950
## f.pickup_latitude=p.X2                 27.6801406  40.9090909 23.5903814
## f.dropoff_latitude=d.X3                22.2836096  31.4285714 22.5124378
## f.fare_amount=FAmount3                 20.9144793  32.0779221 24.4817579
## pick_up_period=night                   21.9072165  22.0779221 16.0862355
## f.passenger=onePassager                16.8459658  89.4805195 84.7844113
## f.total=MediumTrip                     19.5633921  30.2597403 24.6890547
## f.total=CheapTrip                      19.3825043  29.3506494 24.1708126
## f.MTA_tax=highMTA                      16.0953177 100.0000000 99.1708126
## f.Improvement_surcharge=highSurcharge  16.0885917 100.0000000 99.2122720
## f.fare_amount=FAmount2                 18.0209171  29.0909091 25.7669983
## f.Improvement_surcharge=smallSurcharge  0.0000000   0.0000000  0.7877280
## f.MTA_tax=smallMTA                      0.0000000   0.0000000  0.8291874
## pick_up_period=morning                 11.8801653  14.9350649 20.0663350
## f.passenger=multiPassagers             11.0354223  10.5194805 15.2155887
## f.fare_amount=FAmount4                  9.4782609  14.1558442 23.8391376
## f.total=ExpensiveTrip                   8.5385878  13.5064935 25.2487562
## f.dropoff_latitude=d.X1                 6.1320755  10.1298701 26.3681592
## f.dropoff_longitude=d.Y3                4.1876047   6.4935065 24.7512438
## f.pickup_longitude=p.Y3                 2.8594122   4.6753247 26.0986733
## f.dropoff_latitude=d.X4                 1.9486271   2.8571429 23.4038143
## f.pickup_latitude=p.X4                  1.8276762   2.7272727 23.8184080
## f.pickup_latitude=p.X1                  1.7213115   2.7272727 25.2902156
## f.dropoff_longitude=d.Y1                0.0000000   0.0000000 23.2379768
## f.pickup_longitude=p.Y2                 0.0000000   0.0000000 23.9635158
## f.pickup_longitude=p.Y1                 0.0000000   0.0000000 25.3316750
## f.dropoff_longitude=d.Y2                0.0723589   0.1298701 28.6484245
##                                            p.value     v.test
## f.dropoff_longitude=d.Y4               0.000000e+00        Inf
## f.pickup_longitude=p.Y4                0.000000e+00        Inf
## f.dropoff_latitude=d.X2                7.178227e-72  17.927624
## f.pickup_latitude=p.X3                 6.129233e-65  17.017140
## f.pickup_latitude=p.X2                 7.403488e-32  11.746007
## f.dropoff_latitude=d.X3                3.874251e-10   6.259012
```

```
## f.fare_amount=FAmount3                   1.803442e-07   5.218535
## pick_up_period=night                     1.861887e-06   4.767865
## f.passenger=onePassager                  4.115799e-05   4.100881
## f.total=MediumTrip                       1.227384e-04   3.840591
## f.total=CheapTrip                        3.190940e-04   3.599284
## f.MTA_tax=highMTA                        9.239318e-04   3.312721
## f.Improvement_surcharge=highSurcharge    1.312246e-03   3.213288
## f.fare_amount=FAmount2                   2.268824e-02   2.278644
## f.Improvement_surcharge=smallSurcharge   1.312246e-03  -3.213288
## f.MTA_tax=smallMTA                       9.239318e-04  -3.312721
## pick_up_period=morning                   6.857330e-05  -3.981182
## f.passenger=multiPassagers               4.115799e-05  -4.100881
## f.fare_amount=FAmount4                   6.144948e-13  -7.197227
## f.total=ExpensiveTrip                    6.365231e-18  -8.625789
## f.dropoff_latitude=d.X1                  1.744146e-33 -12.058748
## f.dropoff_longitude=d.Y3                 3.538772e-46 -14.266504
## f.pickup_longitude=p.Y3                  2.309089e-63 -16.803281
## f.dropoff_latitude=d.X4                  5.010687e-66 -17.163165
## f.pickup_latitude=p.X4                   5.971590e-69 -17.549775
## f.pickup_latitude=p.X1                   1.653393e-75 -18.387531
## f.dropoff_longitude=d.Y1                 2.212389e-98 -21.051568
## f.pickup_longitude=p.Y2                 6.066721e-102 -21.436781
## f.pickup_longitude=p.Y1                 9.134978e-109 -22.155981
## f.dropoff_longitude=d.Y2                2.625940e-123 -23.616224
##
## $category$`3`
##                               Cla/Mod   Mod/Cla   Global       p.value
## f.passenger=multiPassagers 32.970027 100.00000 15.21559 1.767383e-215
## f.pickup_latitude=p.X3      7.820805  42.56198 27.30100  1.624205e-07
## pick_up_period=night        6.829897  21.90083 16.08624  1.492579e-02
## pick_up_period=afternoon    6.028986  42.97521 35.75871  1.758346e-02
## f.pickup_longitude=p.Y4     6.318450  30.99174 24.60614  2.069560e-02
## f.dropoff_latitude=d.X3     6.353591  28.51240 22.51244  2.520208e-02
## f.pickup_latitude=p.X1      3.524590  17.76860 25.29022  4.482445e-03
## f.pickup_latitude=p.X4      3.394256  16.11570 23.81841  2.844901e-03
## f.fare_amount=FAmount4      3.391304  16.11570 23.83914  2.776892e-03
## f.total=ExpensiveTrip       3.366174  16.94215 25.24876  1.599873e-03
## pick_up_period=morning      2.789256  11.15702 20.06633  1.718326e-04
## f.passenger=onePassager     0.000000   0.00000 84.78441 1.767383e-215
##                               v.test
## f.passenger=multiPassagers  31.330971
## f.pickup_latitude=p.X3       5.237893
## pick_up_period=night         2.434175
## pick_up_period=afternoon     2.374275
## f.pickup_longitude=p.Y4      2.313492
## f.dropoff_latitude=d.X3      2.238291
## f.pickup_latitude=p.X1      -2.842050
## f.pickup_latitude=p.X4      -2.984018
## f.fare_amount=FAmount4      -2.991413
## f.total=ExpensiveTrip       -3.155930
## pick_up_period=morning      -3.757189
## f.passenger=onePassager    -31.330971
##
## $category$`4`
```

```
##                                          Cla/Mod      Mod/Cla     Global
## f.dropoff_latitude=d.X1                78.3805031  65.59210526 26.3681592
## f.pickup_latitude=p.X1                 77.8688525  62.50000000 25.2902156
## f.pickup_longitude=p.Y1                71.8494272  57.76315789 25.3316750
## f.dropoff_longitude=d.Y1               55.9322034  41.25000000 23.2379768
## f.pickup_latitude=p.X2                 49.1212654  36.77631579 23.5903814
## f.dropoff_longitude=d.Y2               41.2445731  37.50000000 28.6484245
## f.total=MediumTrip                     40.6381192  31.84210526 24.6890547
## f.fare_amount=FAmount3                 40.4741744  31.44736842 24.4817579
## f.dropoff_latitude=d.X2                38.6686612  34.01315789 27.7155887
## f.pickup_longitude=p.Y2                39.0138408  29.67105263 23.9635158
## pick_up_period=afternoon               36.8695652  41.84210526 35.7587065
## f.passenger=onePassager                33.0073350  88.81578947 84.7844113
## f.MTA_tax=highMTA                      31.7725753 100.00000000 99.1708126
## f.Improvement_surcharge=highSurcharge  31.7384037  99.93421053 99.2122720
## f.total=CheapTrip                      35.0771870  26.90789474 24.1708126
## pick_up_period=night                   35.8247423  18.28947368 16.0862355
## f.fare_amount=FAmount2                 34.2719228  28.02631579 25.7669983
## pick_up_period=valley                  27.9704797  24.93421053 28.0887231
## f.Improvement_surcharge=smallSurcharge  2.6315789   0.06578947  0.7877280
## f.MTA_tax=smallMTA                      0.0000000   0.00000000  0.8291874
## f.passenger=multiPassagers             23.1607629  11.18421053 15.2155887
## f.dropoff_longitude=d.Y3               25.0418760  19.67105263 24.7512438
## pick_up_period=morning                 23.4504132  14.93421053 20.0663350
## f.total=ExpensiveTrip                  20.8538588  16.71052632 25.2487562
## f.fare_amount=FAmount4                 16.9565217  12.82894737 23.8391376
## f.pickup_longitude=p.Y3                13.5027800  11.18421053 26.0986733
## f.dropoff_longitude=d.Y4                2.1295475   1.57894737 23.3623549
## f.pickup_longitude=p.Y4                 1.7691660   1.38157895 24.6061360
## f.dropoff_latitude=d.X3                 0.5524862   0.39473684 22.5124378
## f.dropoff_latitude=d.X4                 0.0000000   0.00000000 23.4038143
## f.pickup_latitude=p.X4                  0.0000000   0.00000000 23.8184080
## f.pickup_latitude=p.X3                  0.8352316   0.72368421 27.3009950
##                                            p.value     v.test
## f.dropoff_latitude=d.X1                 0.000000e+00        Inf
## f.pickup_latitude=p.X1                  0.000000e+00        Inf
## f.pickup_longitude=p.Y1                6.693401e-259  34.371632
## f.dropoff_longitude=d.Y1                4.443693e-85  19.546205
## f.pickup_latitude=p.X2                  2.450253e-46  14.292123
## f.dropoff_longitude=d.Y2                8.863747e-20   9.102056
## f.total=MediumTrip                      1.335307e-14   7.702403
## f.fare_amount=FAmount3                  5.363019e-14   7.522764
## f.dropoff_latitude=d.X2                 5.493028e-11   6.556920
## f.pickup_longitude=p.Y2                 4.975712e-10   6.219869
## pick_up_period=afternoon                2.777327e-09   5.944246
## f.passenger=onePassager                 6.557298e-08   5.402897
## f.MTA_tax=highMTA                       2.470504e-07   5.159924
## f.Improvement_surcharge=highSurcharge   1.044148e-05   4.407823
## f.total=CheapTrip                       2.763672e-03   2.992870
## pick_up_period=night                    5.106376e-03   2.800245
## f.fare_amount=FAmount2                  1.541761e-02   2.422418
## pick_up_period=valley                   8.845761e-04  -3.324877
## f.Improvement_surcharge=smallSurcharge  1.044148e-05  -4.407823
## f.MTA_tax=smallMTA                      2.470504e-07  -5.159924
```

```
## f.passenger=multiPassagers               6.557298e-08  -5.402897
## f.dropoff_longitude=d.Y3                 1.909868e-08  -5.619973
## pick_up_period=morning                   7.533134e-10  -6.154465
## f.total=ExpensiveTrip                    2.194209e-21  -9.495375
## f.fare_amount=FAmount4                   8.816550e-37 -12.668698
## f.pickup_longitude=p.Y3                  8.234914e-64 -16.864317
## f.dropoff_longitude=d.Y4                3.138703e-173 -28.058584
## f.pickup_longitude=p.Y4                 5.374037e-191 -29.478859
## f.dropoff_latitude=d.X3                 5.385427e-195 -29.789237
## f.dropoff_latitude=d.X4                 3.514308e-219 -31.601560
## f.pickup_latitude=p.X4                  8.453861e-224 -31.935996
## f.pickup_latitude=p.X3                  1.441938e-238 -32.982619
##
## $category$`5`
##                                            Cla/Mod Mod/Cla     Global
## f.MTA_tax=smallMTA                      100.0000000   100.0  0.8291874
## f.Improvement_surcharge=smallSurcharge   94.7368421    90.0  0.7877280
## f.fare_amount=FAmount4                    2.5217391    72.5 23.8391376
## f.total=ExpensiveTrip                     2.1346470    65.0 25.2487562
## f.pickup_longitude=p.Y4                   1.5164280    45.0 24.6061360
## f.dropoff_longitude=d.Y2                  0.4341534    15.0 28.6484245
## f.total=CheapestTrip                      0.3202562    10.0 25.8913765
## f.total=CheapTrip                         0.2572899     7.5 24.1708126
## f.pickup_longitude=p.Y1                   0.1636661     5.0 25.3316750
## f.fare_amount=FAmount1                    0.0000000     0.0 25.9121061
## f.Improvement_surcharge=highSurcharge     0.0835771    10.0 99.2122720
## f.MTA_tax=highMTA                         0.0000000     0.0 99.1708126
##                                             p.value     v.test
## f.MTA_tax=smallMTA                     4.423513e-100  21.236189
## f.Improvement_surcharge=smallSurcharge  6.790423e-84  19.406576
## f.fare_amount=FAmount4                  1.023586e-10   6.463426
## f.total=ExpensiveTrip                   1.396358e-07   5.265727
## f.pickup_longitude=p.Y4                 5.073992e-03   2.802298
## f.dropoff_longitude=d.Y2                4.857554e-02  -1.972299
## f.total=CheapestTrip                    1.475913e-02  -2.438236
## f.total=CheapTrip                       7.690669e-03  -2.665357
## f.pickup_longitude=p.Y1                 9.651084e-04  -3.300505
## f.fare_amount=FAmount1                  5.823330e-06  -4.532704
## f.Improvement_surcharge=highSurcharge   6.790423e-84 -19.406576
## f.MTA_tax=highMTA                      4.423513e-100 -21.236189
##
## $category$`6`
##                                            Cla/Mod     Mod/Cla     Global
## f.total=ExpensiveTrip                   57.63546798  99.4334278 25.2487562
## f.fare_amount=FAmount4                  60.08695652  97.8753541 23.8391376
## f.dropoff_longitude=d.Y1                29.79482605  47.3087819 23.2379768
## f.dropoff_latitude=d.X3                 23.48066298  36.1189802 22.5124378
## pick_up_period=morning                  20.24793388  27.7620397 20.0663350
## f.dropoff_latitude=d.X2                 18.47419596  34.9858357 27.7155887
## f.MTA_tax=highMTA                       14.75752508 100.0000000 99.1708126
## f.Improvement_surcharge=highSurcharge   14.75135813 100.0000000 99.2122720
## f.pickup_longitude=p.Y3                 16.75933280  29.8866856 26.0986733
## f.Improvement_surcharge=smallSurcharge   0.00000000   0.0000000  0.7877280
## f.MTA_tax=smallMTA                       0.00000000   0.0000000  0.8291874
```

```
## f.pickup_latitude=p.X3                 11.99696279  22.3796034 27.3009950
## f.pickup_longitude=p.Y4                10.36225779  17.4220963 24.6061360
## pick_up_period=afternoon               11.24637681  27.4787535 35.7587065
## f.dropoff_latitude=d.X1                10.37735849  18.6968839 26.3681592
## f.dropoff_longitude=d.Y2                9.11722142  17.8470255 28.6484245
## f.dropoff_longitude=d.Y3                7.37018425  12.4645892 24.7512438
## f.dropoff_latitude=d.X4                 6.37732507  10.1983003 23.4038143
## f.fare_amount=FAmount3                  0.93141406   1.5580737 24.4817579
## f.total=CheapTrip                       0.08576329   0.1416431 24.1708126
## f.total=MediumTrip                      0.08396306   0.1416431 24.6890547
## f.fare_amount=FAmount1                  0.24000000   0.4249292 25.9121061
## f.total=CheapestTrip                    0.16012810   0.2832861 25.8913765
## f.fare_amount=FAmount2                  0.08045052   0.1416431 25.7669983
##                                             p.value     v.test
## f.total=ExpensiveTrip                  0.000000e+00        Inf
## f.fare_amount=FAmount4                 0.000000e+00        Inf
## f.dropoff_longitude=d.Y1               3.269879e-53  15.355183
## f.dropoff_latitude=d.X3                3.660507e-19   8.946766
## pick_up_period=morning                 9.151655e-08   5.342810
## f.dropoff_latitude=d.X2                4.622248e-06   4.581241
## f.MTA_tax=highMTA                      1.734007e-03   3.132371
## f.Improvement_surcharge=highSurcharge  2.386117e-03   3.037421
## f.pickup_longitude=p.Y3                1.415835e-02   2.453221
## f.Improvement_surcharge=smallSurcharge 2.386117e-03  -3.037421
## f.MTA_tax=smallMTA                     1.734007e-03  -3.132371
## f.pickup_latitude=p.X3                 1.268004e-03  -3.223125
## f.pickup_longitude=p.Y4                7.839962e-07  -4.939310
## pick_up_period=afternoon               4.517265e-07  -5.045758
## f.dropoff_latitude=d.X1                2.611956e-07  -5.149490
## f.dropoff_longitude=d.Y2               9.822562e-13  -7.132970
## f.dropoff_longitude=d.Y3               3.657158e-18  -8.688978
## f.dropoff_latitude=d.X4                4.856303e-22  -9.651244
## f.fare_amount=FAmount3                 1.121359e-75 -18.408574
## f.total=CheapTrip                      3.500075e-91 -20.250731
## f.total=MediumTrip                     1.660151e-93 -20.512645
## f.fare_amount=FAmount1                 7.299244e-95 -20.664041
## f.total=CheapestTrip                   8.838030e-97 -20.876064
## f.fare_amount=FAmount2                 2.127404e-98 -21.053424
##
##
## $quanti.var
##                           Eta2       P-value
## Pickup_longitude  0.610440929  0.000000e+00
## Pickup_latitude   0.623453710  0.000000e+00
## Dropoff_longitude 0.500783414  0.000000e+00
## Dropoff_latitude  0.602928048  0.000000e+00
## Passenger_count   0.775173386  0.000000e+00
## Trip_distance     0.616167072  0.000000e+00
## Fare_amount       0.606755661  0.000000e+00
## MTA_tax           1.000000000  0.000000e+00
## Total_amount      0.623052638  0.000000e+00
## trip_length       0.541934033  0.000000e+00
## trip_distance_km  0.616167072  0.000000e+00
## travel_time       0.419516323  0.000000e+00
```

```
## Tip_amount        0.189455156 1.290567e-216
## Tolls_amount      0.056288266  2.935040e-58
## Extra             0.019008107  2.012484e-18
## pick_up_hour      0.006267129  1.287593e-05
##
## $quanti
## $quanti$`1`
##                      v.test Mean in category Overall mean sd in category
## Pickup_latitude    48.057783      40.80164287  40.74596616     0.02844591
## Dropoff_latitude   47.703730      40.79963558  40.74404515     0.02984922
## MTA_tax             4.361073       0.50000000   0.49585406     0.00000000
## pick_up_hour        2.061754      13.79107374  13.49689055     5.94745844
## Dropoff_longitude  -3.909936     -73.94052459 -73.93669909     0.02199554
## Tolls_amount       -4.513445       0.01433376   0.07932421     0.28143134
## Extra              -6.236868       0.30530401   0.35271559     0.37412814
## Passenger_count   -12.149685       1.09055627   1.34929519     0.32700978
## Tip_amount        -12.405915       0.66772962   1.15217454     1.13394882
## travel_time       -18.292232       8.51571626  12.13813362     4.94022771
## trip_length       -20.145655       2.79934676   4.30691367     1.72583727
## Fare_amount       -21.663863       7.97412678  11.17300580     3.32178558
## Trip_distance     -21.746967       1.54542570   2.55628323     0.93713486
## trip_distance_km  -21.746967       2.48712158   4.11393908     1.50817236
## Total_amount      -22.293994       9.76130013  13.55080846     3.74387340
##                   Overall sd       p.value
## Pickup_latitude   0.05525463  0.000000e+00
## Dropoff_latitude  0.05557847  0.000000e+00
## MTA_tax           0.04534071  1.294264e-05
## pick_up_hour      6.80518339  3.923113e-02
## Dropoff_longitude 0.04666352  9.232040e-05
## Tolls_amount      0.68675249  6.378289e-06
## Extra             0.36255737  4.464176e-10
## Passenger_count   1.01567703  5.758692e-34
## Tip_amount        1.86240658  2.427233e-35
## travel_time       9.44475500  9.542732e-75
## trip_length       3.56906208  2.937974e-90
## Fare_amount       7.04240211 4.498821e-104
## Trip_distance     2.21692072 7.380104e-105
## trip_distance_km  3.56778807 7.380104e-105
## Total_amount      8.10688270 4.225627e-110
##
## $quanti$`2`
##                      v.test Mean in category Overall mean sd in category
## Pickup_longitude   49.517280      -73.8701097 -73.93683412     0.03031557
## Dropoff_longitude  46.394953      -73.8651694 -73.93669909     0.03510277
## Extra               3.513664        0.3948052   0.35271559     0.35305572
## MTA_tax             2.767557        0.5000000   0.49585406     0.00000000
## pick_up_hour       -1.996239       13.0480519  13.49689055     7.30941592
## Dropoff_latitude   -2.656398       40.7391672  40.74404515     0.03127258
## Pickup_latitude    -2.776506       40.7408974  40.74596616     0.02611218
## Tolls_amount       -3.495970        0.0000000   0.07932421     0.00000000
## trip_length        -5.119580        3.7032066   4.30691367     2.39469985
## Passenger_count    -6.422618        1.1337662   1.34929519     0.42521511
## Trip_distance      -6.703250        2.0652923   2.55628323     1.31346106
## trip_distance_km   -6.703250        3.3237657   4.11393908     2.11381068
```

```
## travel_time        -6.999270        9.9539895  12.13813362     5.26204849
## Fare_amount        -7.206908        9.4961039  11.17300580     4.30546139
## Total_amount       -8.854258       11.1791948  13.55080846     4.55686611
## Tip_amount        -10.789045        0.4882857   1.15217454     1.03679309
##                   Overall sd      p.value
## Pickup_longitude  0.04078401 0.000000e+00
## Dropoff_longitude 0.04666352 0.000000e+00
## Extra             0.36255737 4.419713e-04
## MTA_tax           0.04534071 5.647822e-03
## pick_up_hour      6.80518339 4.590794e-02
## Dropoff_latitude  0.05557847 7.898034e-03
## Pickup_latitude   0.05525463 5.494670e-03
## Tolls_amount      0.68675249 4.723425e-04
## trip_length       3.56906208 3.062176e-07
## Passenger_count   1.01567703 1.339500e-10
## Trip_distance     2.21692072 2.038337e-11
## trip_distance_km  3.56778807 2.038337e-11
## travel_time       9.44475500 2.573002e-12
## Fare_amount       7.04240211 5.723676e-13
## Total_amount      8.10688270 8.424230e-19
## Tip_amount        1.86240658 3.878003e-27
##
## $quanti$`3`
##                     v.test Mean in category Overall mean sd in category
## Passenger_count  61.002966        5.2314050    1.3492952      0.6133546
## Extra             2.754569        0.4152893    0.3527156      0.3462517
## trip_length      -2.988341        3.6386537    4.3069137      2.5377094
## Trip_distance    -3.167429        2.1163180    2.5562832      1.4696737
## trip_distance_km -3.167429        3.4058837    4.1139391      2.3652106
## travel_time      -3.268230       10.2040975   12.1381336      6.1479750
## Total_amount     -3.415945       11.8157025   13.5508085      5.5722554
## Fare_amount      -3.590193        9.5888430   11.1730058      4.6371835
##                   Overall sd      p.value
## Passenger_count   1.0156770 0.0000000000
## Extra             0.3625574 0.0058769516
## trip_length       3.5690621 0.0028049681
## Trip_distance     2.2169207 0.0015379310
## trip_distance_km  3.5677881 0.0015379310
## travel_time       9.4447550 0.0010822223
## Total_amount      8.1068827 0.0006356105
## Fare_amount       7.0424021 0.0003304335
##
## $quanti$`4`
##                      v.test Mean in category Overall mean sd in category
## Extra              5.160334      0.392434211   0.35271559     0.34808026
## MTA_tax            4.307198      0.500000000   0.49585406     0.00000000
## pick_up_hour       2.794807     13.900657895  13.49689055     7.22577620
## Tolls_amount      -4.940855      0.007289474   0.07932421     0.20082467
## travel_time       -8.911360     10.351340554  12.13813362     5.67319321
## Passenger_count   -9.761451      1.138815789   1.34929519     0.41652531
## Total_amount     -11.154141     11.631125000  13.55080846     4.59886764
## Fare_amount      -12.615745      9.286868421  11.17300580     4.07798611
## trip_distance_km -13.012030      3.128377632   4.11393908     1.86692872
## Trip_distance    -13.012030      1.943883739   2.55628323     1.16005573
```

```
## trip_length       -13.441420      3.288465697   4.30691367     1.95135451
## Dropoff_longitude -28.270126    -73.964704691 -73.93669909     0.02418498
## Pickup_longitude  -35.332859    -73.967426154 -73.93683412     0.01983254
## Dropoff_latitude  -44.420367     40.691633435  40.74404515     0.02577484
## Pickup_latitude   -45.702890     40.692355404  40.74596616     0.02225790
##                   Overall sd       p.value
## Extra             0.36255737  2.465093e-07
## MTA_tax           0.04534071  1.653358e-05
## pick_up_hour      6.80518339  5.193076e-03
## Tolls_amount      0.68675249  7.778056e-07
## travel_time       9.44475500  5.041055e-19
## Passenger_count   1.01567703  1.647807e-22
## Total_amount      8.10688270  6.834956e-29
## Fare_amount       7.04240211  1.729142e-36
## trip_distance_km  3.56778807  1.045276e-38
## Trip_distance     2.21692072  1.045276e-38
## trip_length       3.56906208  3.457379e-41
## Dropoff_longitude 0.04666352 8.053496e-176
## Pickup_longitude  0.04078401 1.838320e-273
## Dropoff_latitude  0.05557847  0.000000e+00
## Pickup_latitude   0.05525463  0.000000e+00
##
## $quanti$`5`
##                      v.test Mean in category Overall mean sd in category
## Fare_amount         8.309805        20.388500   11.1730058    10.35192049
## Total_amount        6.451768        21.787250   13.5508085    11.58155214
## trip_length         3.966920         6.536450    4.3069137     4.55463007
## Trip_distance       3.348801         3.725369    2.5562832     2.49525892
## trip_distance_km    3.348801         5.995401    4.1139391     4.01572996
## Pickup_longitude    3.110563       -73.916857  -73.9368341     0.03850257
## travel_time         2.912984        16.470605   12.1381336    10.71238890
## Dropoff_longitude   2.326190       -73.919606  -73.9366991     0.04937623
## Extra              -4.864257         0.075000    0.3527156     0.32691742
## MTA_tax           -69.447822         0.000000    0.4958541     0.00000000
##                   Overall sd      p.value
## Fare_amount       7.04240211 9.585916e-17
## Total_amount      8.10688270 1.105524e-10
## trip_length       3.56906208 7.280727e-05
## Trip_distance     2.21692072 8.116202e-04
## trip_distance_km  3.56778807 8.116202e-04
## Pickup_longitude  0.04078401 1.867308e-03
## travel_time       9.44475500 3.579930e-03
## Dropoff_longitude 0.04666352 2.000841e-02
## Extra             0.36255737 1.148873e-06
## MTA_tax           0.04534071 0.000000e+00
##
## $quanti$`6`
##                      v.test Mean in category Overall mean sd in category
## trip_distance_km  53.856564       10.7961454   4.11393908     3.63248760
## Trip_distance     53.856564        6.7084137   2.55628323     2.25712315
## Total_amount      53.723216       28.6968130  13.55080846     7.22191602
## Fare_amount       52.736211       24.0885269  11.17300580     6.29567920
## trip_length       50.396570       10.5620571   4.30691367     3.93848475
## travel_time       44.386634       26.7170452  12.13813362    13.46916511
```

```
## Tip_amount        28.800252        3.0174929   1.15217454     3.01019346
## Tolls_amount      16.416267        0.4713881   0.07932421     1.62853018
## MTA_tax            2.629373        0.5000000   0.49585406     0.00000000
## Passenger_count   -2.069319        1.2762040   1.34929519     0.78146030
## Extra             -2.641935        0.3194051   0.35271559     0.36142649
## Pickup_longitude  -4.440938      -73.9431328 -73.93683412     0.03624119
## pick_up_hour      -4.571467       12.4150142  13.49689055     6.68847481
## Dropoff_longitude -7.023627      -73.9480969 -73.93669909     0.05458784
##                   Overall sd       p.value
## trip_distance_km  3.56778807  0.000000e+00
## Trip_distance     2.21692072  0.000000e+00
## Total_amount      8.10688270  0.000000e+00
## Fare_amount       7.04240211  0.000000e+00
## trip_length       3.56906208  0.000000e+00
## travel_time       9.44475500  0.000000e+00
## Tip_amount        1.86240658 2.129452e-182
## Tolls_amount      0.68675249  1.462942e-60
## MTA_tax           0.04534071  8.554244e-03
## Passenger_count   1.01567703  3.851617e-02
## Extra             0.36255737  8.243394e-03
## Pickup_longitude  0.04078401  8.956776e-06
## pick_up_hour      6.80518339  4.843209e-06
## Dropoff_longitude 0.04666352  2.161816e-12
##
##
## attr(,"class")
## [1] "catdes" "list "
```

So, first we can assume that all the null hipothesis of independence for the qualitative variables taken can be denied by the chisquare test. It means that all of them have been used somehow to calculate the clustering distances and their splittings, as expected (because we took them from the axis interpretation analysis so we knew they were significative for PCA projections).

Diving inside each category, we can determine the following characterization: ### Category 1: It is defined by individuals contained between this coordinates rangs: - pickup_latitude=(40.8,40.91] - pickup_longitude=(-73.95,-73.92]

- dropoff_latitude=(40.8,40.91]
- dropoff_longitude=(-73.95,-73.91]

It also have a significative representation (almost 48 in Cla/Mod) of rows which total amount are contained in (-1,7.8] rang.

Category 2:

Very similar to previous case, defined also by coordinates values, but this times the rangs are the nexts: - f.pickup_latitude=(40.75,40.8] - f.pickup_longitude=(-73.92,-73.79]

- f.pickup_latitude=(40.75,40.8]
- f.dropoff_longitude=(-73.91,-73.75]

Category 3

As category 1 and 2, it seems to be characterized depending on the coordinates points where the client has been picked up and dropped off. This time, this rangs are: - f.pickup_latitude=(40.58,40.69] - f.pickup_longitude=(-

74.03,-73.96] - f.dropoff_longitude=(-74.03,-73.97] - f.dropoff_latitude=(40.58,40.69] However, we can appreciate that in any of the different clusters the rangs that defines the cluster itself are being overlapped (which makes totally sense). Furthermore, it also have a significative representation (almost 41%) of the rows which, this time, its total amount rang is: (11,16.6].

Category 4

This category is determined, with a huge difference between its 2 first v.test values, for passenger variable. Concretely, 100% of their individuals are included in (1,6] rang for f.passenger.

Category 5

This category is gathering all the rows with MTA_tax = 0 and also the 90% of the individuals which its improvement surcharge is 0 are in this category. So MTA_tax and Improvement_surcharge explains the behaviour of category 5 rows, and 63% of individuals of this category it has its total_amount value compressed between (16.6,46].

Category 6

Definitely, category 6 is defined by those rows which their total_amount is in the rang: (16.6,46] and their fare_amount between (14,42.5] (so , the most expensive one). We can appreciate how 100% of this individuals have: - MTA_tax = 0.5 - Improvement surcharge != 0

$quanti section

We can observe how all the peculiarities pointed at the cluster analysis are proved by the quantitative variables output. - For quanti 1, 2 and 3 the most significative quantitative variables are latitudes and longitudes - For quanti 4 we have passenger_count at the top - quanti 5 has a huge negative value for MTA_tax (which make sense with the description above) - And in quanti 6 Total_amount, trip_distance and Fare_amount are distinguished as more correlated.

Axes description

At this point, this output does not help anymore to detail our clusters. But we can also see how the interpretation of axis made in a past section corresponds to the characterization of the clusters, being the variables that explain the most each specific dimension the ones that are also distinguished for each cluster who has a higher v.test value for the dimension in question.

```
#Block B descripcion per eixos
res.hcpc$desc.axes
```

```
## $quanti.var
##            Eta2       P-value
## Dim.1 0.6607600  0.000000e+00
## Dim.2 0.6494640  0.000000e+00
## Dim.3 0.5419910  0.000000e+00
## Dim.5 0.5850848  0.000000e+00
## Dim.6 0.7729213  0.000000e+00
## Dim.4 0.1062099 1.040754e-114
##
## $quanti
```

```
## $quanti$`1`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.2  27.525274        0.8665425  2.891171e-12      0.7228325  1.5014716
## Dim.4   8.109728        0.1894092 -7.565656e-13      1.0423741  1.1139193
## Dim.5  -7.291932       -0.1541230  2.627846e-15      0.3882088  1.0080539
## Dim.6  -7.773557       -0.1621727 -2.016049e-13      0.3320971  0.9949857
## Dim.1 -25.432188       -1.1510970 -1.447302e-13      0.9706699  2.1586742
## Dim.3 -38.167268       -0.9818854  4.601893e-12      0.6473058  1.2269549
##             p.value
## Dim.2 8.750703e-167
## Dim.4  5.073328e-16
## Dim.5  3.055419e-13
## Dim.6  7.631245e-15
## Dim.1 1.111431e-142
## Dim.3  0.000000e+00
##
## $quanti$`2`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.3  41.823449        1.6954552  4.601893e-12      0.9726455  1.2269549
## Dim.2  27.682906        1.3733040  2.891171e-12      0.7055715  1.5014716
## Dim.5   6.722108        0.2238861  2.627846e-15      0.3830502  1.0080539
## Dim.4  -5.784506       -0.2128913 -7.565656e-13      1.0642113  1.1139193
## Dim.1  -9.245365       -0.6594001 -1.447302e-13      1.2291884  2.1586742
## Dim.6 -11.581202       -0.3807223 -2.016049e-13      0.4280619  0.9949857
##             p.value
## Dim.3  0.000000e+00
## Dim.2 1.121576e-168
## Dim.5  1.791138e-11
## Dim.4  7.272585e-09
## Dim.1  2.344394e-20
## Dim.6  5.132122e-31
##
## $quanti$`3`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.6 57.942132        3.6122061 -2.016049e-13      0.5935830  0.9949857
## Dim.5 14.307223        0.9036502  2.627846e-15      0.6393131  1.0080539
## Dim.4 10.377365        0.7242726 -7.565656e-13      1.0645016  1.1139193
## Dim.3  5.455195        0.4193727  4.601893e-12      1.0673606  1.2269549
## Dim.1 -3.302942       -0.4467342 -1.447302e-13      1.4689699  2.1586742
##            p.value
## Dim.6 0.000000e+00
## Dim.5 1.972319e-46
## Dim.4 3.143314e-25
## Dim.3 4.891899e-08
## Dim.1 9.567625e-04
##
## $quanti$`4`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.3   7.904502        0.2058935  4.601893e-12      0.6144627  1.2269549
## Dim.4  -7.395685       -0.1748927 -7.565656e-13      1.0497738  1.1139193
## Dim.1  -8.566018       -0.3925597 -1.447302e-13      1.1653594  2.1586742
## Dim.6 -10.442573       -0.2205789 -2.016049e-13      0.4125492  0.9949857
## Dim.2 -53.253956       -1.6974954  2.891171e-12      0.6580396  1.5014716
##            p.value
```

```
## Dim.3 2.690070e-15
## Dim.4 1.406815e-13
## Dim.1 1.071251e-17
## Dim.6 1.584559e-25
## Dim.2 0.000000e+00
##
## $quanti$`5`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.6  17.254477        2.7034950 -2.016049e-13      0.8460908  0.9949857
## Dim.1   5.621445        1.9109178 -1.447302e-13      2.0760185  2.1586742
## Dim.3   5.279964        1.0201559  4.601893e-12      1.4920056  1.2269549
## Dim.2   4.190979        0.9909223  2.891171e-12      1.3469019  1.5014716
## Dim.4 -17.159463       -3.0099851 -7.565656e-13      0.9840208  1.1139193
## Dim.5 -50.448777       -8.0083192  2.627846e-15      0.5886606  1.0080539
##            p.value
## Dim.6 1.035555e-66
## Dim.1 1.893669e-08
## Dim.3 1.292091e-07
## Dim.2 2.777532e-05
## Dim.4 5.340548e-66
## Dim.5 0.000000e+00
##
## $quanti$`6`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1 55.013403        4.1298834 -1.447302e-13      1.7840028  2.1586742
## Dim.5  5.828992        0.2043429  2.627846e-15      1.3791586  1.0080539
## Dim.2  4.002171        0.2089753  2.891171e-12      1.4213296  1.5014716
## Dim.4  3.000554        0.1162351 -7.565656e-13      1.0711674  1.1139193
## Dim.6 -4.221994       -0.1460887 -2.016049e-13      0.7529646  0.9949857
## Dim.3 -8.058584       -0.3438504  4.601893e-12      1.1968156  1.2269549
##            p.value
## Dim.1 0.000000e+00
## Dim.5 5.576305e-09
## Dim.2 6.276396e-05
## Dim.4 2.694888e-03
## Dim.6 2.421510e-05
## Dim.3 7.718345e-16
##
##
## attr(,"class")
## [1] "catdes" "list "
```

Invidual analysis

Again, this command can help us now to confirm the conclusions made until now. As an example, we will look a paragon of C6, and how its total_amount is served in some middle-point of the last rang (16.6,46] for this variable (total_amount = 26.3), and also look at how a distinguished C6 row has one of the possible highest values for total_amount (= 44.8).

If we keep tracking for the rest of the clusters, we can assume that the conclusions made below are concordant.

```
#Block C individus
res.hcpc$desc.ind
```

```
## $para
## Cluster: 1
##    253799    419422     92598    809989     87900
## 0.3811201 0.3904390 0.4366720 0.4591189 0.4703781
## ------------------------------------------------------------
## Cluster: 2
##    746656   1372589     90225   1362378    142290
## 0.5807398 0.6118779 0.6426381 0.6479828 0.6695885
## ------------------------------------------------------------
## Cluster: 3
##    473235    745377    415886   1370940   1361206
## 0.7053162 0.7638502 1.0287560 1.0542710 1.0656542
## ------------------------------------------------------------
## Cluster: 4
##    473230   1369263   1076497    474605    749722
## 0.5302566 0.5873180 0.6251593 0.6394274 0.6753597
## ------------------------------------------------------------
## Cluster: 5
##   1343967    829507    396850   1322755    469848
## 0.9410112 1.0839339 1.2499154 1.2841703 1.4562540
## ------------------------------------------------------------
## Cluster: 6
##    678042   1059592   1406084   1229822    116640
## 0.9553348 1.0369012 1.1123124 1.1123404 1.1718410
##
## $dist
## Cluster: 1
##  1178619   572868   915921   955214   203359
## 5.684318 5.556909 5.475445 5.388099 4.372158
## ------------------------------------------------------------
## Cluster: 2
##   532659   576477  1404537  1426703   657301
## 6.224754 6.156786 6.120060 6.018727 5.811136
## ------------------------------------------------------------
## Cluster: 3
##  1137082   329313  1112510  1394101  1123289
## 7.204914 6.975908 6.082006 5.786417 5.667288
## ------------------------------------------------------------
## Cluster: 4
##   274842  1033483   645383  1157271   697017
## 5.777649 5.668662 5.268899 5.252580 5.189690
## ------------------------------------------------------------
## Cluster: 5
##  1325016   825818   626233   675043   978944
## 10.36407 10.28404 10.28256 10.14405 10.11998
## ------------------------------------------------------------
## Cluster: 6
##    285458    868718    317072    425343    206032
## 12.448150 11.787798 10.438274 10.053165  9.876675
```

```
#parangon C6
df["678042",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 678042 VeriFone Inc.  2016-01-15 13:24:58   2016-01-15 14:00:17
```

```
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 678042     Store_and_fwd Standard rate        -73.90332        40.74579
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 678042         -73.98235          40.7681               1          4.95
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 678042          25     0     0.5          3            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 678042                   0.3         28.8  Credit card Street-hail       0
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 678042 AnyTip Yes    11.26821         7.966253    35.31667           13
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 678042         valley 19.14372 onePassager      Dist4               p.Y4
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 678042              p.X3                d.Y1               d.X3
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 678042      FAmount4 smallExtra   highMTA           highSurcharge
##        f.tip_amount    f.toll       f.total f.ttime f.espeed
## 678042      highTip smallToll ExpensiveTrip   Time4   Speed2
##        f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4
## 678042       Normald1       Normald2       Normald3       Normald4
##        f.outlierPCA
## 678042       Normal
```

```
#distinguished C6
df["285458",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 285458 VeriFone Inc.  2016-01-07 01:26:54   2016-01-07 01:41:13
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 285458     Store_and_fwd Standard rate        -73.94707        40.81071
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 285458         -74.01742         40.85104               1         10.47
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 285458          29   0.5     0.5          4         10.5
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 285458                   0.3         44.8  Credit card Street-hail       0
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 285458 AnyTip Yes    12.30616         16.84983    14.31667            1
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 285458          night 51.57415 onePassager      Dist4               p.Y2
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 285458              p.X4                d.Y1               d.X4
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 285458      FAmount4 smallExtra   highMTA           highSurcharge
##        f.tip_amount   f.toll       f.total f.ttime f.espeed f.outlierPCAd1
## 285458      highTip highToll ExpensiveTrip   Time3   Speed4       Normald1
##        f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4 f.outlierPCA
## 285458       Normald2       Normald3       Normald4       Normal
```

Assigning clusters groups

Now we assign to each row the cluster group decided by HPC method and we also consider as group 7 the outliers (which they haven't been taking into consideration until now).

```
#Donar-li una classe (the last one) a tots els outliers multidimensionals (sup.)
df$claHP<-7
df[row.names(res.hcpc$data.clust),"claHP"]<-res.hcpc$data.clust$clust
table(df$claHP)
```

```
##
##    1    2    3    4    5    6    7
## 1546  770  242 1520   40  706  119
```

K-Means Classification

We execute kmeans command defining 6 clusters in order to get the same number of groups as in the hierarchical process.

```
ppcc<-res.pca$ind$coord[,1:4]
dim(ppcc)
```

```
## [1] 4824    4
```

```
kc<-kmeans(ppcc,6,iter.max = 30, trace=T)
```

```
## KMNS(*, k=6): iter=  1, indx=0
##  QTRAN(): istep=4824, icoun=11
##  QTRAN(): istep=9648, icoun=26
##  QTRAN(): istep=14472, icoun=517
##  QTRAN(): istep=19296, icoun=517
##  QTRAN(): istep=24120, icoun=2185
##  QTRAN(): istep=28944, icoun=2477
## KMNS(*, k=6): iter=  2, indx=0
##  QTRAN(): istep=4824, icoun=71
##  QTRAN(): istep=9648, icoun=36
##  QTRAN(): istep=14472, icoun=682
##  QTRAN(): istep=19296, icoun=36
##  QTRAN(): istep=24120, icoun=351
##  QTRAN(): istep=28944, icoun=391
##  QTRAN(): istep=33768, icoun=2025
## KMNS(*, k=6): iter=  3, indx=43
##  QTRAN(): istep=4824, icoun=129
##  QTRAN(): istep=9648, icoun=985
## KMNS(*, k=6): iter=  4, indx=4824
```

```
table(kc$cluster)
```

```
##
##    1    2    3    4    5    6
##  659  764  726 1503  359  813
```

```
#plotcluster(ppcc, kc$cluster)
```

Assigning clusters groups

As we also did before in HPC, we assign the clusters in a way that group 7 is taken by the outliers.

```
df$claKM<-7
df[names(kc$cluster),"claKM"]<-kc$cluster
kc$betweenss/kc$totss
```

```
## [1] 0.6600474
```

```
table(df$claKM)
```

```
##
##    1    2    3    4    5    6    7
##  659  764  726 1503  359  813  119
```

Characterization of Kmeans clustering

As we didn't manage to execute catdes command, we've tried to be a bit creative and search for internet. At last, we found interesting "$centers" and we realized we could try to give it a try in order to get some notion about wether the clustering done by Kmeans was similar or not to HCPC.

```
kc$centers
```

```
##        Dim.1      Dim.2      Dim.3        Dim.4
## 1 -0.7632688 -1.7985079  0.4924921  0.989103028
## 2 -0.7106933 -1.5883491  0.0265183 -1.095376530
## 3  2.1040351 -0.4625913 -0.2866859 -0.007695426
## 4 -1.2875272  0.9003796 -0.9707284  0.202141937
## 5  5.5494307  0.6722143 -0.1869361  0.104489772
## 6 -0.6625558  1.4021663  1.7090240 -0.185357327
```

```
#catdes(df,47)
#veure si s'han posat d'acord o no
table(df$claHP,df$claKM)
```

```
##
##       1    2    3    4    5    6    7
##   1  21   16   83 1422    0    4    0
##   2   4    6   16    6    0  738    0
##   3  50   29   31   70    3   59    0
##   4 584  707  229    0    0    0    0
##   5   0    6   12    5    7   10    0
##   6   0    0  355    0  349    2    0
##   7   0    0    0    0    0    0  119
```

The output it's a bit messy, but we can certainly appreciate how cluster1 it's centred by Dimension 1 (which is the axis that increases with total_amount prices) and, at least, also check how Dimension 2 and cluster4 is very correlated. These two clusters, as we can observe in the next section, are precisely associated with cluster6 and cluster4 (in this order) by the labels done in HCPC. So, even though we haven't been able to interpret the categorical description, we can predict that both methods (kmeans and hierarchical) are actually generating a very similar groups of individuals.

Re-labeling

After all, we generate a new label for Kmeans groups so the cluster numbers are referring to the same group and avoid further confusions.

To finalize, we check the diagonal summatory of the number of invidiuals from the contingency table generated, to have an idea of the bias taken by kmeans respect to HCPC.

```
df$claHP<-factor(df$claHP,labels=paste("kHP-",1:7))
df$claKM<-factor(df$claKM,levels=c(3,4,1,5,6,2,7),labels=c("kKM-3","kKM-4","kKM-1","kKM-5","kKM-6","kKM
tt<-table(df$claHP,df$claKM)
tt
```

```
##
##          kKM-3 kKM-4 kKM-1 kKM-5 kKM-6 kKM-2 kKM-7
##   kHP- 1    83  1422    21     0     4    16     0
##   kHP- 2    16     6     4     0   738     6     0
##   kHP- 3    31    70    50     3    59    29     0
##   kHP- 4   229     0   584     0     0   707     0
##   kHP- 5    12     5     0     7    10     6     0
##   kHP- 6   355     0     0   349     2     0     0
##   kHP- 7     0     0     0     0     0     0   119
```

```
sum(diag(tt)/sum(tt))
```

```
## [1] 0.05421809
```

Load Required Packages: to be increased over the course

```
rm(list=ls())
requiredPackages <- c("effects","FactoMineR","car", "factoextra","RColorBrewer","ggplot2","missMDA","mv
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]

if(length(missingPackages)) install.packages(missingPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
## [1] TRUE
##
## [[8]]
## [1] TRUE
##
## [[9]]
```

```
## [1] TRUE
##
## [[10]]
## [1] TRUE
##
## [[11]]
## [1] TRUE
##
## [[12]]
## [1] TRUE
##
## [[13]]
## [1] TRUE
##
## [[14]]
## [1] TRUE
```

```
# Useful function and Data
#setwd("C://Users/Sergi/Desktop/Sergi/ADEI") #Change
load("DataPCA2.RData")
calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q2=s.x[3],        q3=s.x[5], ma
```

Correspondence Analysis: f.cost (discretization of Total_amount) vs f.hour and f.tt and period

In this section we will interprete two Correspondence Analysis executed on our discretizated target variable (f.total) with 2 different factors of our data: pick_up_period and f.espeed.

Total_amount vs pick_up_period

We generate the contingency table for this two variables and execute the chisq test.

```
tt<-table(df[,c("f.total","pick_up_period")])
tt
```

```
##                pick_up_period
## f.total         night morning valley afternoon
##   CheapestTrip    203     275    363       412
##   CheapTrip       177     218    331       461
##   MediumTrip      185     204    340       474
##   ExpensiveTrip   235     300    357       408
```

```
prop.table(tt,1)
```

```
##                pick_up_period
## f.total             night   morning    valley afternoon
##   CheapestTrip  0.1620112 0.2194733 0.2897047 0.3288109
##   CheapTrip     0.1491154 0.1836563 0.2788543 0.3883741
##   MediumTrip    0.1537822 0.1695761 0.2826268 0.3940150
##   ExpensiveTrip 0.1807692 0.2307692 0.2746154 0.3138462
```

```
prop.table(table(df$pick_up_period))
```

```
##
##     night   morning    valley afternoon
## 0.1618450 0.2016994 0.2814081 0.3550475
```

Before getting into the chisq test, taking a look to the contingency table we can actually say very little, because the categories of pick_up_period are not well balanced and so "afternoon" is always taking a higher number of individuals for each category.

For the previous reason, we look at their marginal tables to better extrapolate our first impressions. The marginal table relative to f.total it is just a confirmation for what we already said (not very helpfull): we have an unbalanced distribution: most of the trips were made during the afternoon in every range of price contended.

```
prop.table(tt,2)
```

```
##                pick_up_period
## f.total             night   morning    valley afternoon
##   CheapestTrip  0.2537500 0.2758275 0.2609633 0.2347578
##   CheapTrip     0.2212500 0.2186560 0.2379583 0.2626781
##   MediumTrip    0.2312500 0.2046138 0.2444285 0.2700855
##   ExpensiveTrip 0.2937500 0.3009027 0.2566499 0.2324786
```

```
prop.table(table(df$f.total))
```

```
##
##  CheapestTrip     CheapTrip    MediumTrip ExpensiveTrip
##     0.2534898     0.2401376     0.2433745     0.2629982
```

Then, moving on to the second marginal table (now, relative to pick_up_period variable) things get more even. we can appreciate that the cheapest trips take place, in a greater proportion, during the morning, as me can also say that during the night occur the most expensive trips. On the other hand, the medium price categories are more associated to the afternoon.

```
chisq.test(tt)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tt
## X-squared = 38.085, df = 9, p-value = 1.683e-05
```

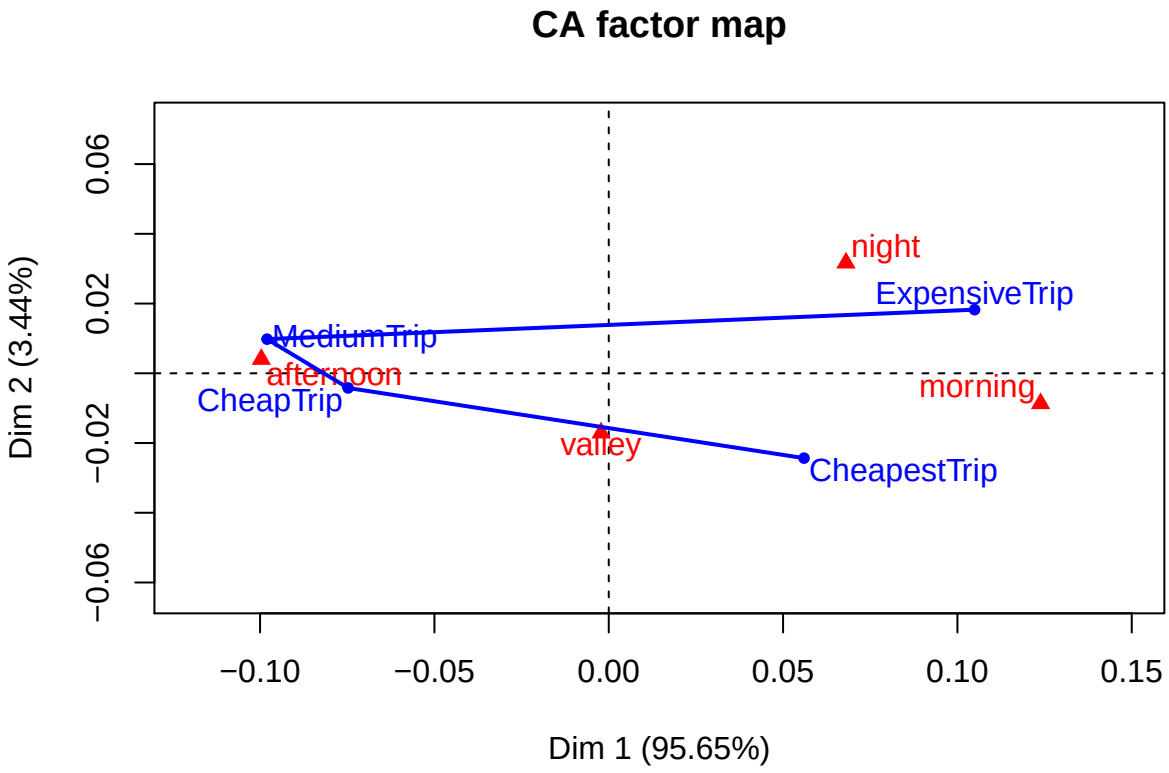
From the chi-square test we see that the p.value is less than 0.05 which meand that we can reject the Nul hypotesis thus the two factors are dependent. The df is 9 which means that the chi value of the threshold is 16,9 and we see that we have X-squared = 36,42 which is bigger than 16,9. => We can deffinetly reject the hypotesis. Factors are dependent and so we can affirm that the conclusions made above are statistically significant.

CA plot interpretation

The plot visually represents all the information we could have interpreted from the tables:

- At night and morning the trips tend to be more expensive.
- In the afternoon we find middle prices
- Cheap trips stay at valley.

```
res.ca<-CA(tt)
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="blue")
```

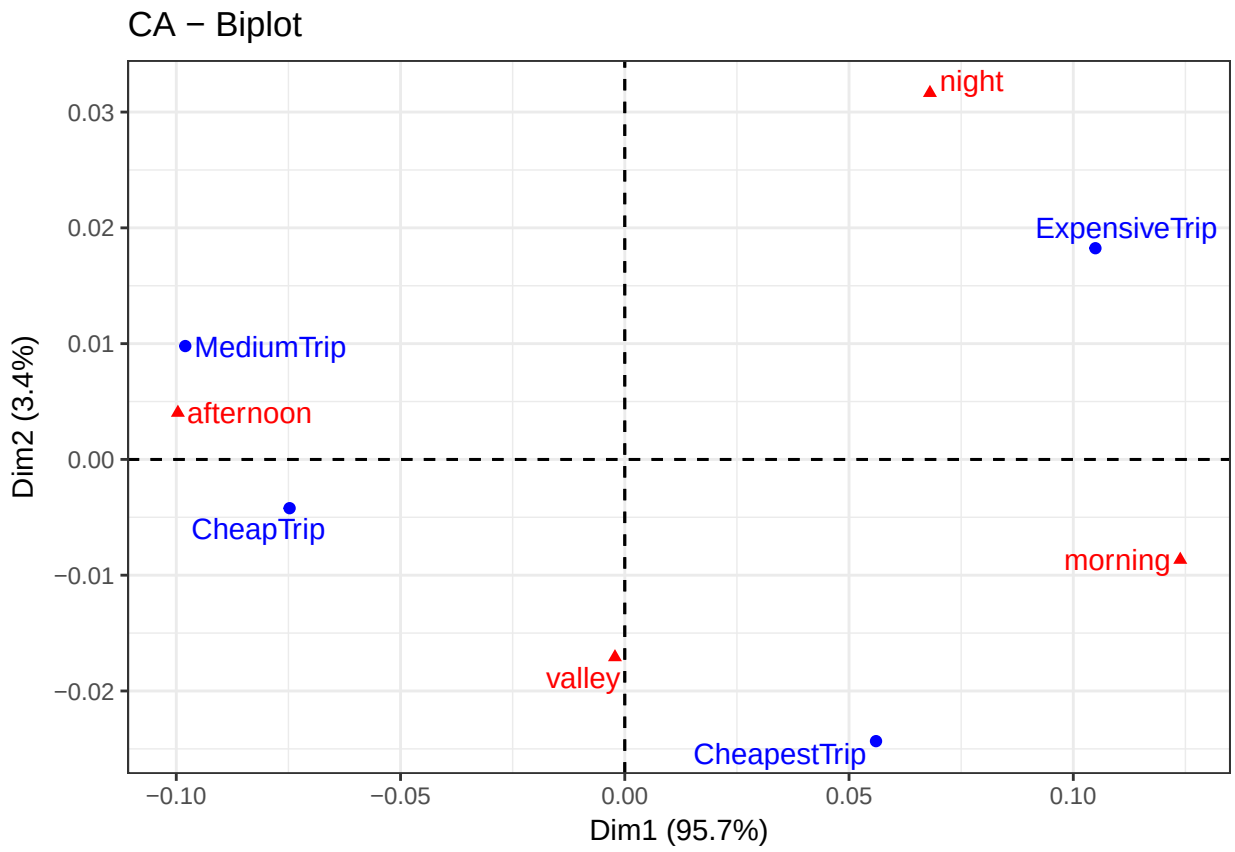



```
summary(res.ca,dig=2)
```

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 38.08459 (p-value =  1.683268e-
##
## Eigenvalues
##                         Dim.1   Dim.2   Dim.3
## Variance                0.007   0.000   0.000
## % of var.              95.652   3.442   0.906
## Cumulative % of var.   95.652  99.094 100.000
##
## Rows
##                  Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## CheapestTrip  |     0.951 |  0.056 10.793  0.837 | -0.024 56.598  0.158 |
## CheapTrip     |     1.385 | -0.075 18.200  0.969 | -0.004  1.608  0.003 |
## MediumTrip    |     2.385 | -0.098 31.713  0.980 |  0.010  8.797  0.010 |
## ExpensiveTrip |     2.984 |  0.105 39.294  0.970 |  0.018 32.997  0.029 |
##                 Dim.3    ctr   cos2
## CheapestTrip    0.004  7.260  0.005 |
## CheapTrip      -0.013 56.178  0.028 |
```

```
## MediumTrip     0.010 35.153  0.010 |
## ExpensiveTrip -0.002  1.408  0.000 |
##
## Columns
##                 Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## night         |     0.920 |  0.068 10.167  0.814 |  0.032 61.153  0.176 |
## morning       |     3.132 |  0.124 41.983  0.988 | -0.009  5.712  0.005 |
## valley        |     0.112 | -0.002  0.018  0.012 | -0.017 30.959  0.733 |
## afternoon     |     3.541 | -0.100 47.832  0.996 |  0.004  2.176  0.002 |
##                  Dim.3    ctr   cos2
## night          0.007 12.496  0.009 |
## morning       -0.011 32.135  0.007 |
## valley         0.010 40.881  0.255 |
## afternoon     -0.005 14.487  0.003 |
```

```
fviz_ca_biplot(res.ca,repel=TRUE)+theme_bw()
```



Eigenvalues and dominant axes analysis. How many axes we have to consider

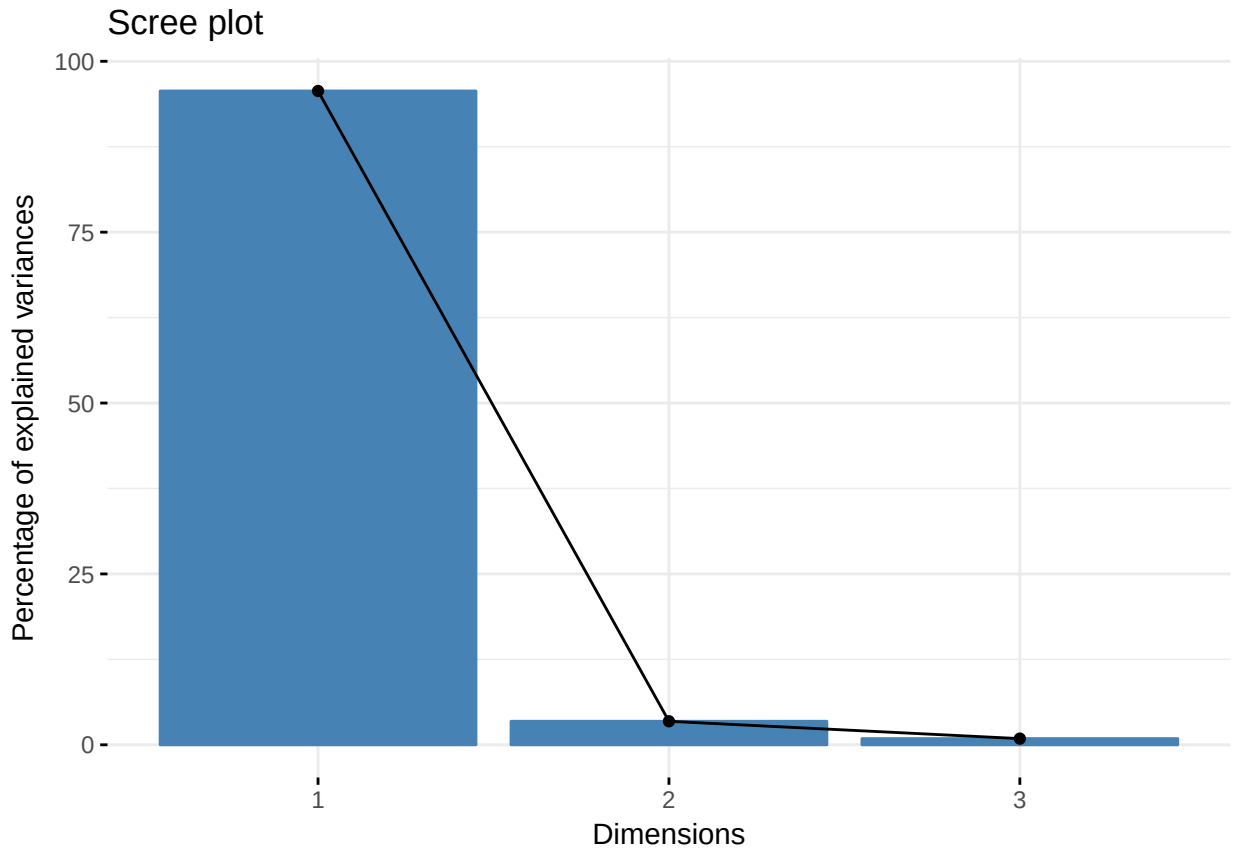
The two first dimensions explain 99% of the variance, so there is no need of further exploration. (Actually, only the first dim has more than 95% of the variance). We can be confident that the patterns we see in the CA plot represent the patterns that we would see if we could peer into n-dimensional space.

```
res.ca$eig
```

```
##         eigenvalue percentage of variance
## dim 1 0.0073697821             95.6524349
```

```
## dim 2 0.0002651851              3.4418382
## dim 3 0.0000697840              0.9057269
##       cumulative percentage of variance
## dim 1                           95.65243
## dim 2                           99.09427
## dim 3                          100.00000
```

```
fviz_eig(res.ca)
```



Individuals

For the next output we deduce that the more extreme individuals are actually the most contributive ones for the specific axe of their coordinates.

```
res.ca$row$contrib[,1:2]
```

```
##                  Dim 1     Dim 2
## CheapestTrip  10.79251 56.598295
## CheapTrip     18.20028  1.607563
## MediumTrip    31.71276  8.796824
## ExpensiveTrip 39.29445 32.997318
```

```
res.ca$row$coord[,1:2]
```

```
##                     Dim 1        Dim 2
## CheapestTrip   0.05601550 -0.024333018
## CheapTrip     -0.07473711 -0.004213359
```

```
## MediumTrip    -0.09799566  0.009790390
## ExpensiveTrip  0.10493408  0.018240531
```

Total_amount vs f.ttime

Now we will look at the contingency table for this two variables and execute the chisq test.

```
tt<-table(df[,c("f.total","f.ttime")])
tt
```

```
##                f.ttime
## f.total         Time1 Time2 Time3 Time4
##   CheapestTrip   1011   233     5     4
##   CheapTrip       207   740   232     8
##   MediumTrip        3   225   747   228
##   ExpensiveTrip     6    10   232  1052
```

```
prop.table(tt,1)
```

```
##                f.ttime
## f.total               Time1       Time2       Time3       Time4
##   CheapestTrip  0.806863528 0.185953711 0.003990423 0.003192338
##   CheapTrip     0.174389217 0.623420388 0.195450716 0.006739680
##   MediumTrip    0.002493766 0.187032419 0.620947631 0.189526185
##   ExpensiveTrip 0.004615385 0.007692308 0.178461538 0.809230769
```

```
prop.table(table(df$f.ttime))
```

```
##
##     Time1     Time2     Time3     Time4
## 0.2482298 0.2443860 0.2460045 0.2613797
```

```
prop.table(tt,2)
```

```
##                f.ttime
## f.total               Time1       Time2       Time3       Time4
##   CheapestTrip  0.823960880 0.192880795 0.004111842 0.003095975
##   CheapTrip     0.168704156 0.612582781 0.190789474 0.006191950
##   MediumTrip    0.002444988 0.186258278 0.614309211 0.176470588
##   ExpensiveTrip 0.004889976 0.008278146 0.190789474 0.814241486
```

```
prop.table(table(df$f.total))
```

```
##
##  CheapestTrip     CheapTrip    MediumTrip ExpensiveTrip
##     0.2534898     0.2401376     0.2433745     0.2629982
```

```
chisq.test(tt)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tt
## X-squared = 6387.1, df = 9, p-value < 2.2e-16
```

We can appreciate a directionality between the increment of f.total and the f.ttime values. The diagonal of the contingency table has most of individuals contended and this is proved by the marginal tables too

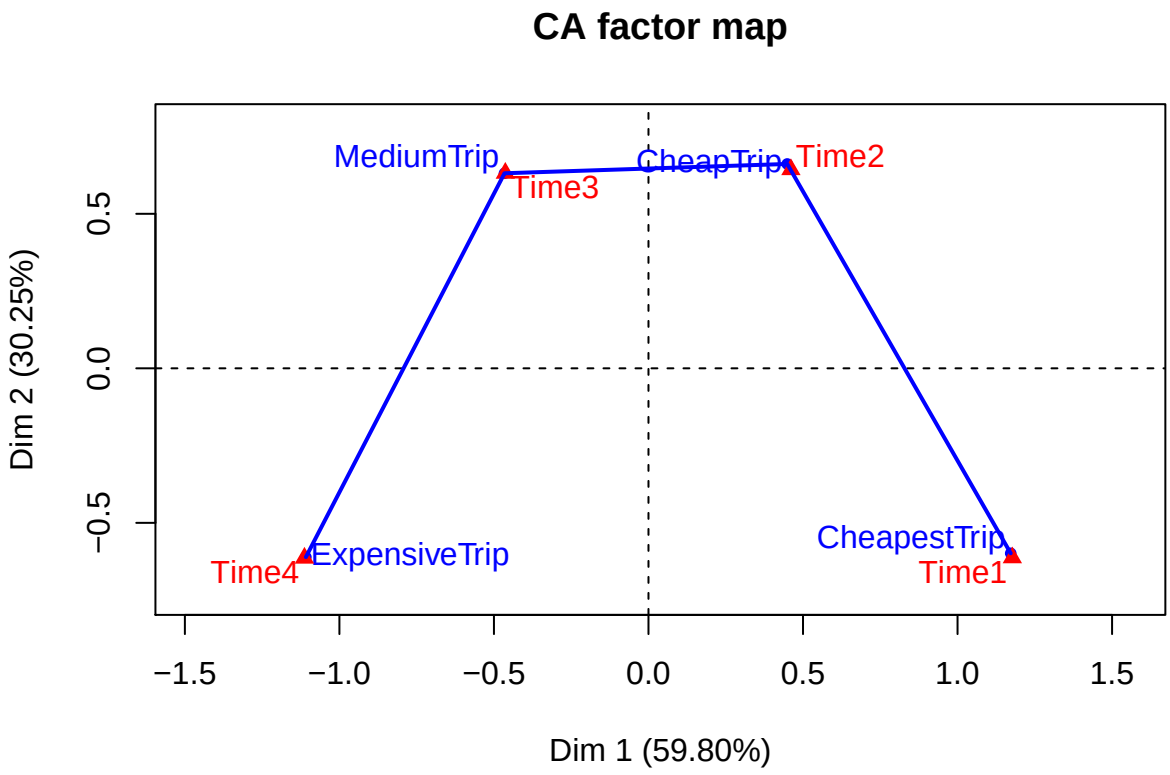
(no matter which one you look at). Lastly, they have a really good balanced distribution for each of their categories (almost equitative in each of them).

The chisq test numerically validates what we have already seen: these two variable are completetly dependent.

CA plot interpretation

The plot shows a visual representation of the conclusion above: the longer the trips the more they cost.

```
res.ca<-CA(tt)
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="blue")
```



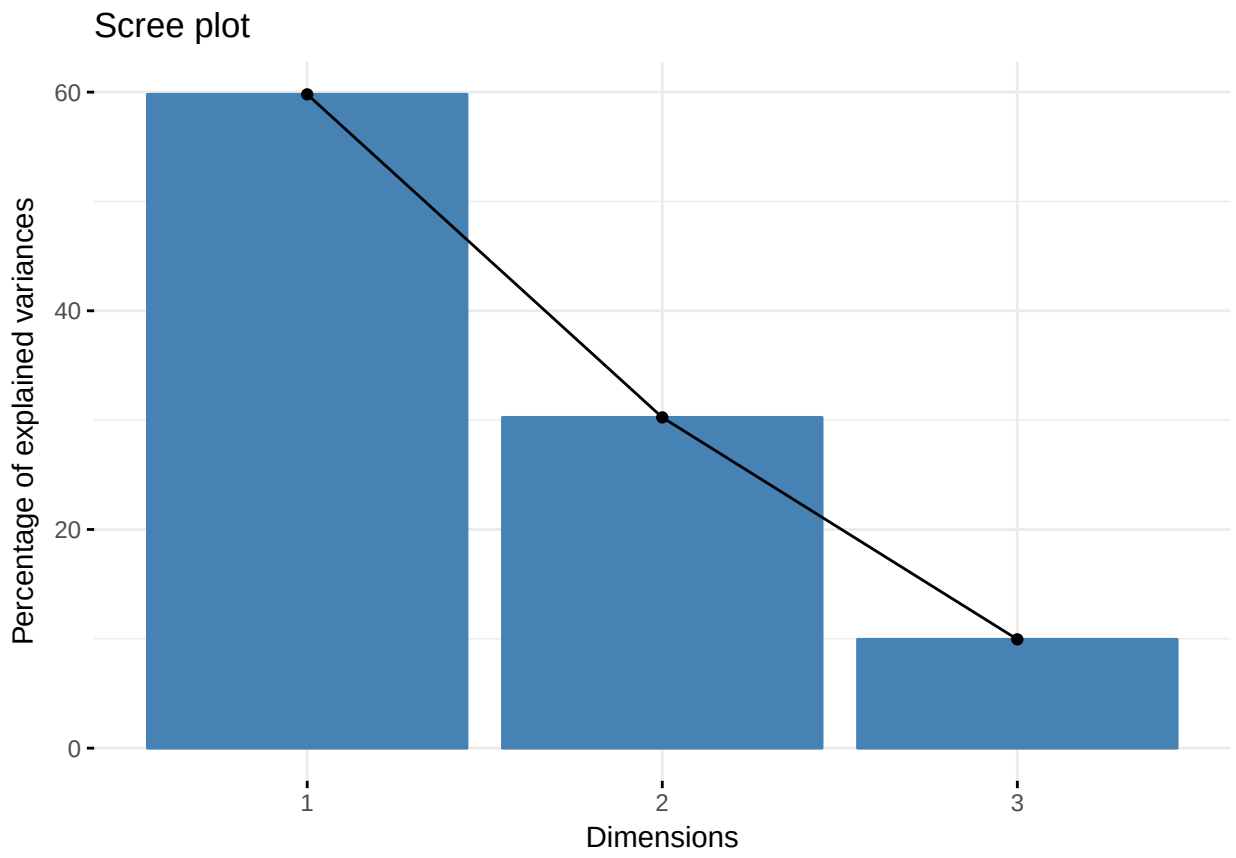
Eigenvalues and dominant axes analysis. How many axes we have to consider

We need to consider the first 2 dimentions which together explains 90% of the data.

```
res.ca$eig
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.7726698              59.797140                           59.79714
## dim 2  0.3908832              30.250561                           90.04770
## dim 3  0.1285988               9.952299                          100.00000
```

```
fviz_eig(res.ca)
```




```
summary(res.ca)
```

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 6387.106 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.773   0.391   0.129
## % of var.             59.797  30.251   9.952
## Cumulative % of var.  59.797  90.048 100.000
##
## Rows
##                 Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## CheapestTrip  |   447.227 |   1.172  45.071   0.779 |  -0.599  23.233
## CheapTrip     |   208.511 |   0.449   6.264   0.232 |   0.661  26.860
## MediumTrip    |   206.368 |  -0.466   6.835   0.256 |   0.631  24.799
## ExpensiveTrip |   430.046 |  -1.109  41.830   0.752 |  -0.611  25.108
##                  cos2     Dim.3     ctr    cos2
## CheapestTrip    0.203 |   0.179   6.347   0.018 |
## CheapTrip       0.504 |  -0.479  42.862   0.264 |
## MediumTrip      0.470 |   0.482  44.028   0.274 |
## ExpensiveTrip   0.228 |  -0.182   6.763   0.020 |
##
```

```
## Columns
##                 Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## Time1         |   446.043 |   1.177  44.535   0.771 |  -0.613  23.840
## Time2         |   208.278 |   0.461   6.719   0.249 |   0.643  25.861
## Time3         |   206.804 |  -0.463   6.832   0.255 |   0.632  25.102
## Time4         |   431.027 |  -1.113  41.914   0.751 |  -0.614  25.197
##                  cos2     Dim.3     ctr    cos2
## Time1           0.209 |   0.188   6.802   0.020 |
## Time2           0.485 |  -0.476  42.981   0.265 |
## Time3           0.474 |   0.477  43.466   0.270 |
## Time4           0.229 |  -0.182   6.751   0.020 |
```

```
res.ca$row$contrib[,1:2]
```

```
##                   Dim 1    Dim 2
## CheapestTrip  45.071354 23.23262
## CheapTrip      6.263752 26.86034
## MediumTrip     6.834878 24.79942
## ExpensiveTrip 41.830016 25.10762
```

```
res.ca$row$coord[,1:2]
```

```
##                    Dim 1      Dim 2
## CheapestTrip   1.1721064 -0.5985389
## CheapTrip      0.4489358  0.6612249
## MediumTrip    -0.4658272  0.6311124
## ExpensiveTrip -1.1085738 -0.6108715
```

```
res.ca$row$cos2[,1:2]
```

```
##                   Dim 1     Dim 2
## CheapestTrip  0.7786928 0.2030565
## CheapTrip     0.2321132 0.5035351
## MediumTrip    0.2559076 0.4697286
## ExpensiveTrip 0.7515660 0.2282116
```

```
sum(res.ca$eig[1:2,1])/sum(res.ca$eig[1:3,1])
```

```
## [1] 0.900477
```

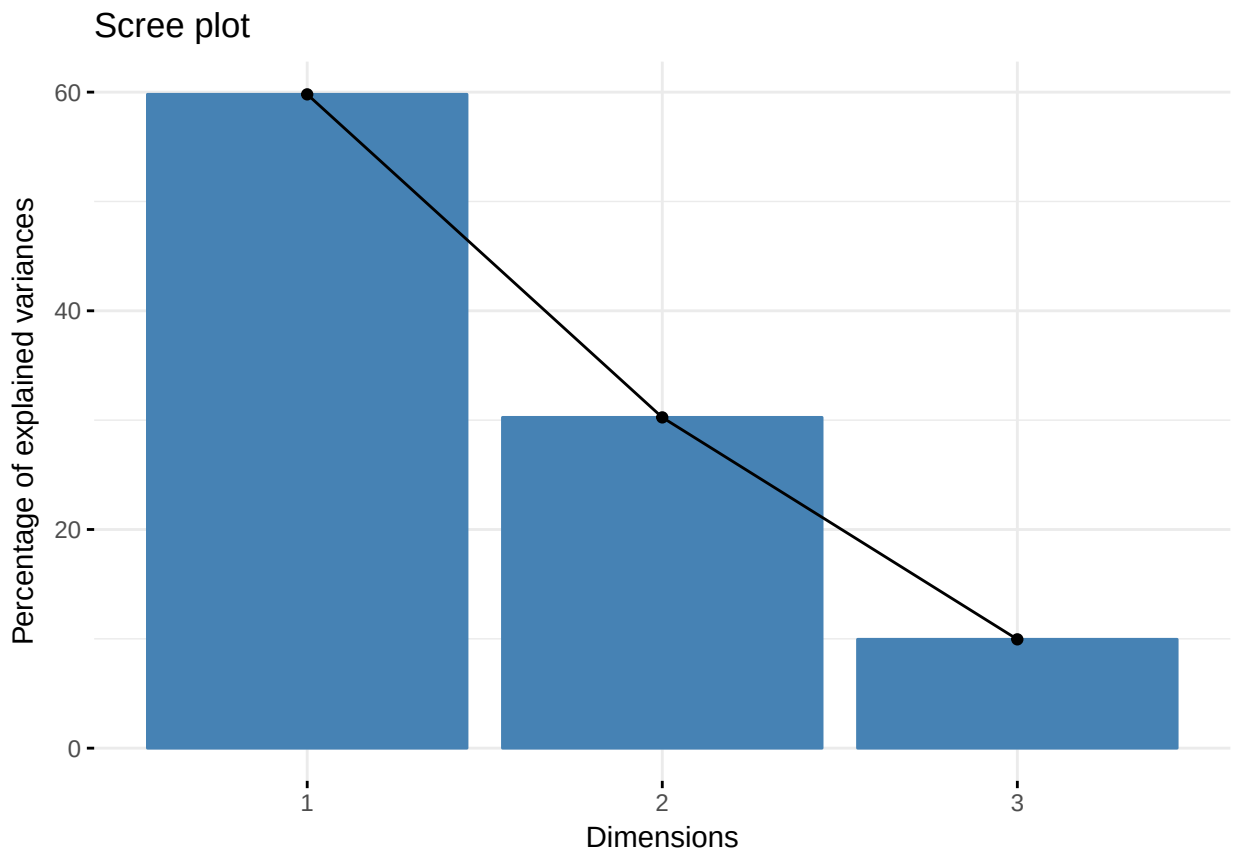
```
sum(res.ca$eig[1:3,1])
```

```
## [1] 1.292152
```

```
res.ca$call$marge.col
```

```
##     Time1     Time2     Time3     Time4
## 0.2482298 0.2443860 0.2460045 0.2613797
```

```
fviz_eig(res.ca)
```

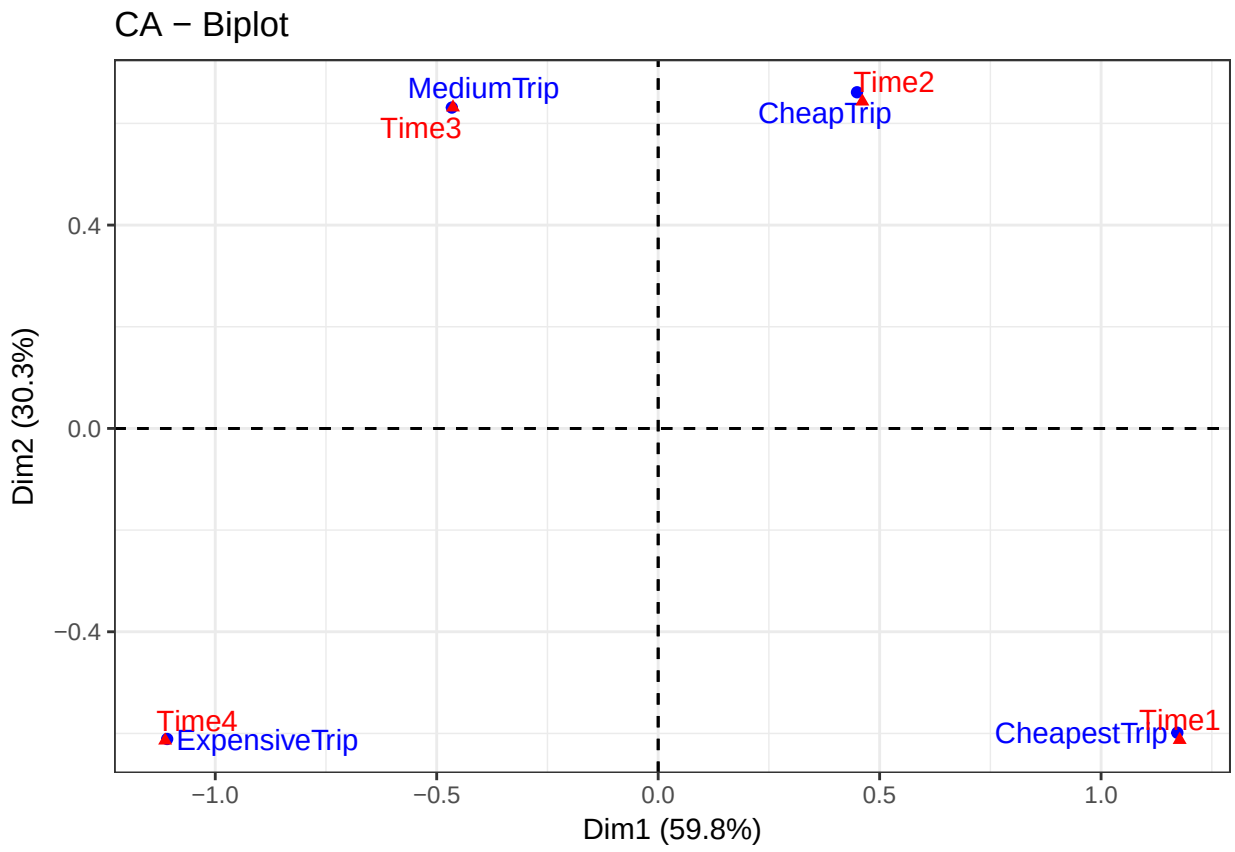



```
summary(res.ca,dig=2)
```

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 6387.106 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.773   0.391   0.129
## % of var.             59.797  30.251   9.952
## Cumulative % of var.  59.797  90.048 100.000
##
## Rows
##                 Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## CheapestTrip  |   447.227 |   1.172  45.071   0.779 |  -0.599  23.233
## CheapTrip     |   208.511 |   0.449   6.264   0.232 |   0.661  26.860
## MediumTrip    |   206.368 |  -0.466   6.835   0.256 |   0.631  24.799
## ExpensiveTrip |   430.046 |  -1.109  41.830   0.752 |  -0.611  25.108
##                  cos2     Dim.3     ctr    cos2
## CheapestTrip    0.203 |   0.179   6.347   0.018 |
## CheapTrip       0.504 |  -0.479  42.862   0.264 |
## MediumTrip      0.470 |   0.482  44.028   0.274 |
## ExpensiveTrip   0.228 |  -0.182   6.763   0.020 |
##
```

```
## Columns
##                 Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## Time1         |   446.043 |   1.177  44.535   0.771 |  -0.613  23.840
## Time2         |   208.278 |   0.461   6.719   0.249 |   0.643  25.861
## Time3         |   206.804 |  -0.463   6.832   0.255 |   0.632  25.102
## Time4         |   431.027 |  -1.113  41.914   0.751 |  -0.614  25.197
##                  cos2     Dim.3     ctr    cos2
## Time1           0.209 |   0.188   6.802   0.020 |
## Time2           0.485 |  -0.476  42.981   0.265 |
## Time3           0.474 |   0.477  43.466   0.270 |
## Time4           0.229 |  -0.182   6.751   0.020 |
```

```
fviz_ca_biplot(res.ca,repel=TRUE)+theme_bw()
```



Multiple Correspondence Analysis

In order to compute MCA we gather all the categorical variables that we want to use as an active ones in vars_cat vector. We are not taking all distributed nor factors in it because we want to avoid using some extremely unbalanced variables such as Store_and_fwd_flag.

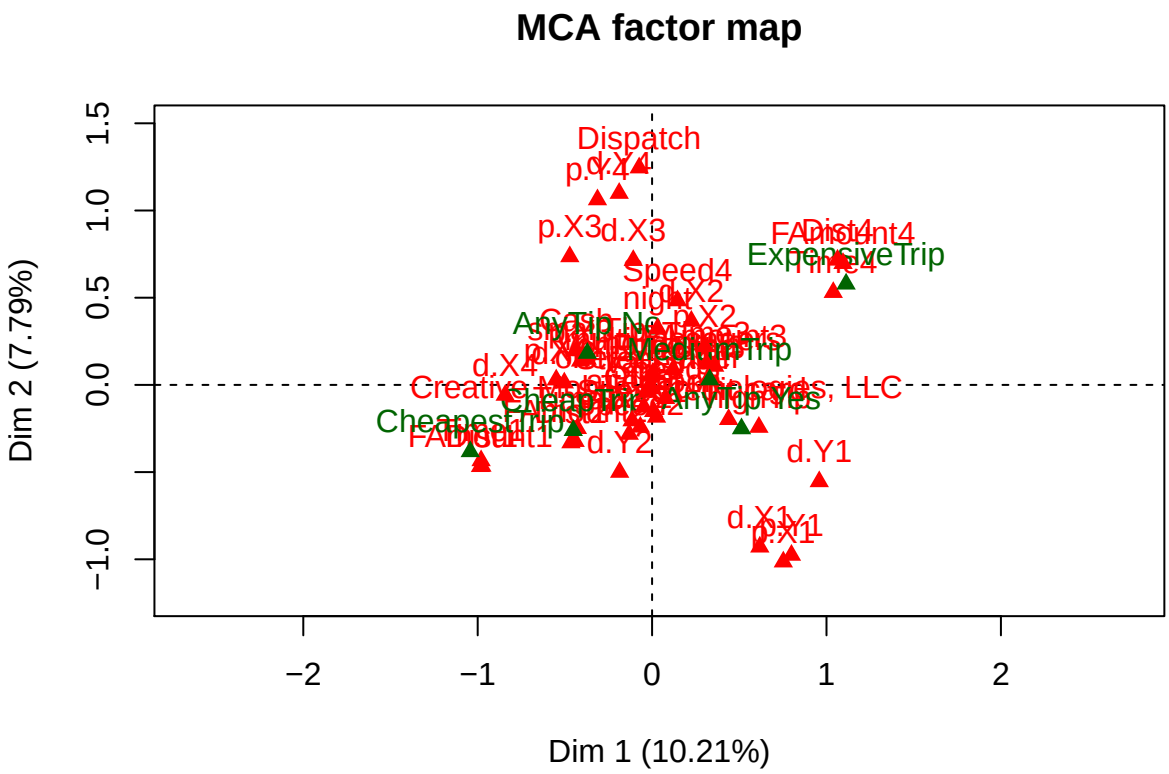
As a suplementary we will use the targets itselfs and the factor of the total_amount too.

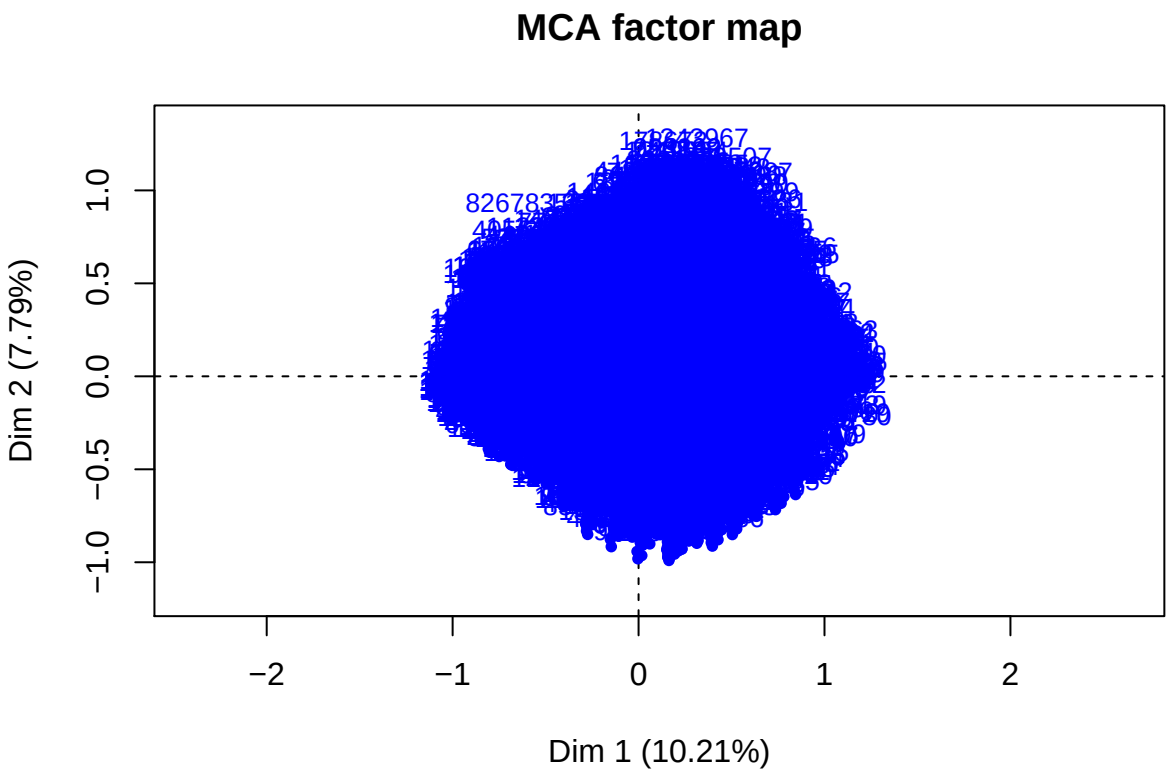
```
names(df)
```

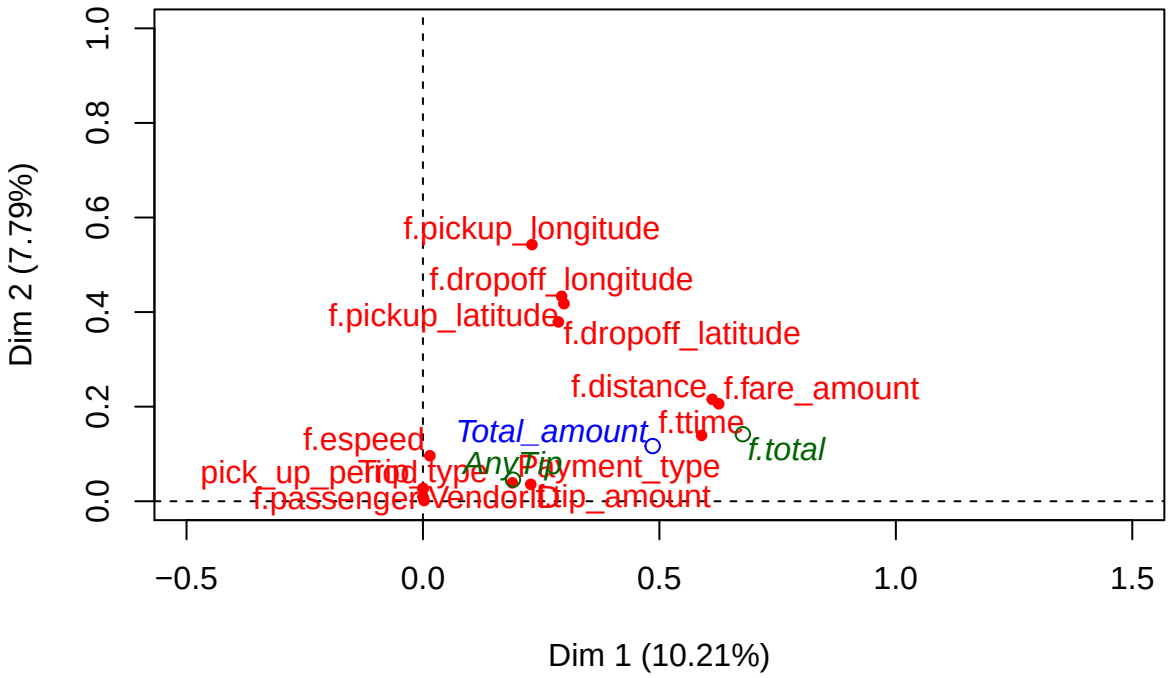
```
##  [1] "VendorID"               "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"  "Store_and_fwd_flag"
```

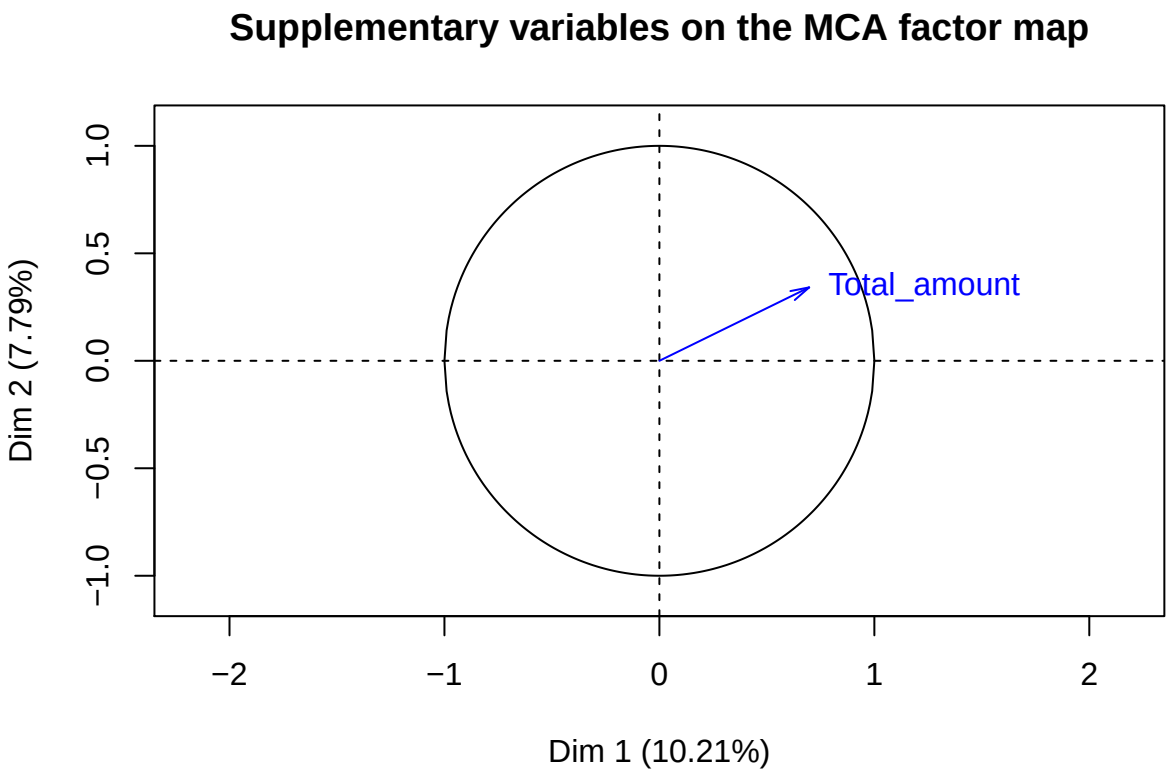
```
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"          "f.outlierPCAd3"
## [47] "f.outlierPCAd4"          "f.outlierPCA"
## [49] "claHP"
```

```
vars_cat <- names(df)[c(1,19,20,22,27,29:35,39,41,42,43)]
#add f.totalamount as quali.sup
res.mca<-MCA(df[,c(vars_cat,"Total_amount")],quali.sup=c(4,14),quanti.sup=17)
```


MCA factor map
Dispatch
p.X4
d.X4
p.X3
d.X3
Speed4
ExpensiveTrip
Medium
Creative
Cheapest
Trip
d.X4
d.Y2
d.Y1
d.X1
p.X1
Dim 2 (7.79%)
1.5
1.0
0.5
0.0
−1.0
−2
−1
0
1
2
Dim 1 (10.21%)


MCA factor map
Dim 2 (7.79%)
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Dim 1 (10.21%)

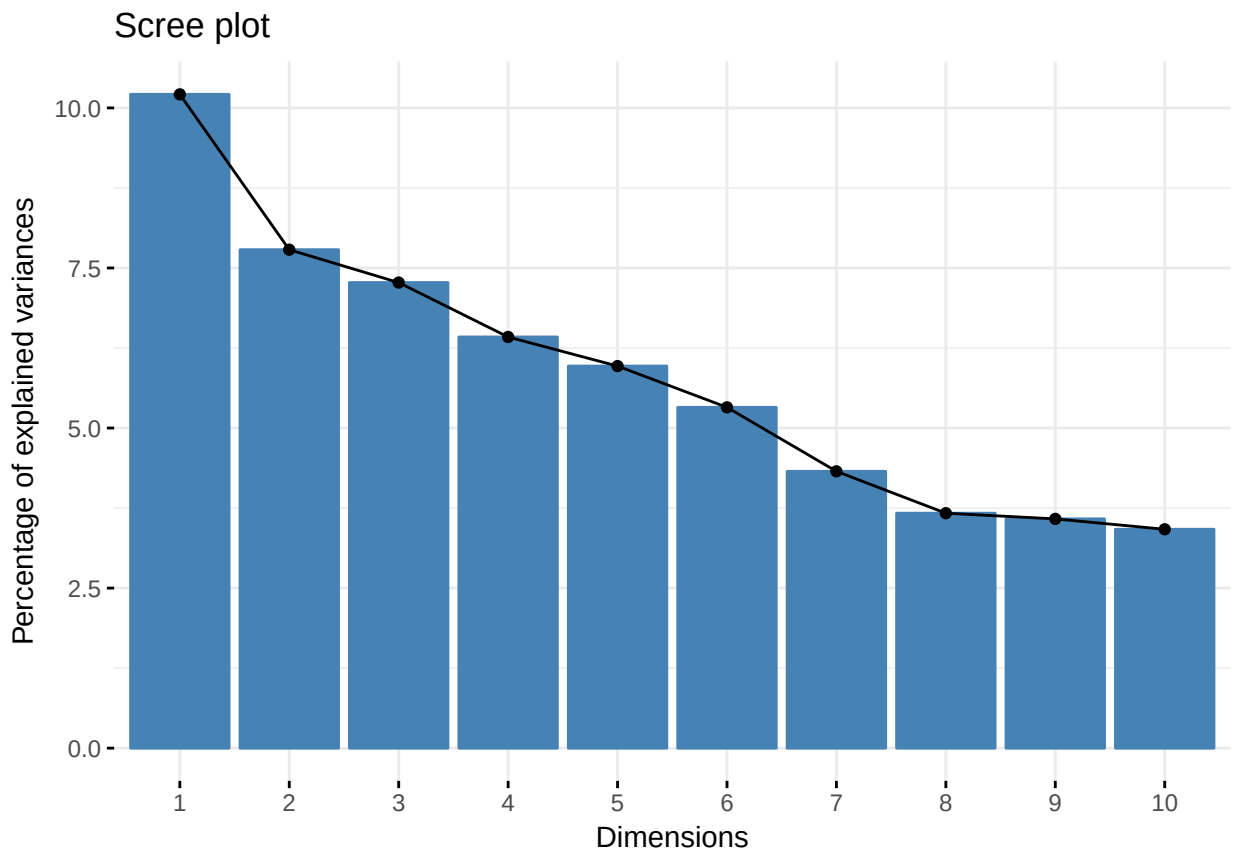

f.pickup_longitude
f.dropoff_longitude
f.pickup_latitude
f.dropoff_latitude
f.distance
f.fare_amount
f.ttime
Total_amount
f.total
f.espeed
AnyTip
pick_up_period
Trip_type
Payment_type
f.passengerVendorID
Tip_amount
Dim 1 (10.21%)
Dim 2 (7.79%)
−0.5
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0



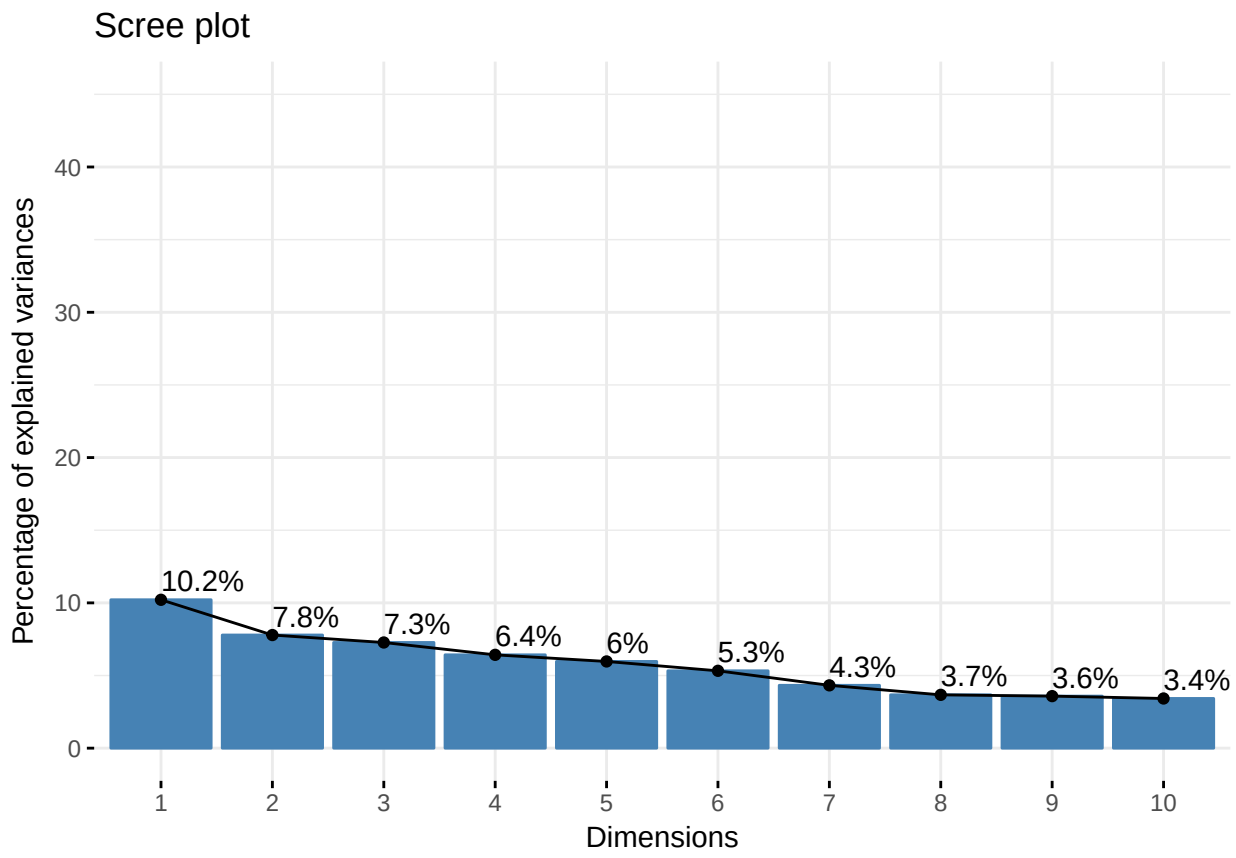
Axes analysis

Below are shown the histograms of Eigen values and explained variances -in this order- as the numerical table for each dimension. In order to get the number of dimensions which we will take into account from now on, we keep only the dimensions associated to eigenvalues greater than 1/(nbvar), which is until the 17th dimension and we got almost 80% of the variance explained.

```
fviz_eig(res.mca)
```




```
fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0, 45))
```




```
get_eig(res.mca)
```

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.240712106       10.2120287                     10.21203
## Dim.2  0.183512638        7.7853847                     17.99741
## Dim.3  0.171414292        7.2721215                     25.26953
## Dim.4  0.151381923        6.4222634                     31.69180
## Dim.5  0.140668693        5.9677627                     37.65956
## Dim.6  0.125469246        5.3229377                     42.98250
## Dim.7  0.101914121        4.3236294                     47.30613
## Dim.8  0.086505324        3.6699228                     50.97605
## Dim.9  0.084409284        3.5809999                     54.55705
## Dim.10 0.080580908        3.4185840                     57.97563
## Dim.11 0.077611033        3.2925893                     61.26822
## Dim.12 0.072364247        3.0699984                     64.33822
## Dim.13 0.069827289        2.9623698                     67.30059
## Dim.14 0.069413144        2.9448000                     70.24539
## Dim.15 0.068521682        2.9069804                     73.15237
## Dim.16 0.067756161        2.8745038                     76.02688
## Dim.17 0.064303229        2.7280158                     78.75489
## Dim.18 0.060008695        2.5458234                     81.30072
## Dim.19 0.058779581        2.4936792                     83.79439
## Dim.20 0.053542367        2.2714944                     86.06589
## Dim.21 0.051284356        2.1757000                     88.24159
## Dim.22 0.048359678        2.0516227                     90.29321
## Dim.23 0.043514113        1.8460533                     92.13927
```

```
## Dim.24 0.036361339        1.5426023                   93.68187
## Dim.25 0.033126222        1.4053549                   95.08722
## Dim.26 0.025385632        1.0769662                   96.16419
## Dim.27 0.022023139        0.9343150                   97.09850
## Dim.28 0.018816707        0.7982845                   97.89679
## Dim.29 0.014610133        0.6198238                   98.51661
## Dim.30 0.013222868        0.5609702                   99.07758
## Dim.31 0.010033031        0.4256437                   99.50323
## Dim.32 0.007343640        0.3115484                   99.81477
## Dim.33 0.004366037        0.1852258                  100.00000
```

```
length <-length(which(res.mca$eig[,1]>=(1/length(df[,c(vars_cat)]))));length
```

```
## [1] 17
```

```
res.mca$eig[1:length,1]
```

```
##      dim 1      dim 2      dim 3      dim 4      dim 5      dim 6
## 0.24071211 0.18351264 0.17141429 0.15138192 0.14066869 0.12546925
##      dim 7      dim 8      dim 9     dim 10     dim 11     dim 12
## 0.10191412 0.08650532 0.08440928 0.08058091 0.07761103 0.07236425
##     dim 13     dim 14     dim 15     dim 16     dim 17
## 0.06982729 0.06941314 0.06852168 0.06775616 0.06430323
```

```
summary(res.mca, dig=2,nbelements=50, ncp=4)
```

```
##
## Call:
## MCA(X = df[, c(vars_cat, "Total_amount")], quanti.sup = 17, quali.sup = c(4,
##      14))
##
##
## Eigenvalues
##                         Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance                0.241   0.184   0.171   0.151   0.141   0.125
## % of var.              10.212   7.785   7.272   6.422   5.968   5.323
## Cumulative % of var.   10.212  17.997  25.270  31.692  37.660  42.982
##                         Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance                0.102   0.087   0.084   0.081   0.078   0.072
## % of var.               4.324   3.670   3.581   3.419   3.293   3.070
## Cumulative % of var.   47.306  50.976  54.557  57.976  61.268  64.338
##                        Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18
## Variance                0.070   0.069   0.069   0.068   0.064   0.060
## % of var.               2.962   2.945   2.907   2.875   2.728   2.546
## Cumulative % of var.   67.301  70.245  73.152  76.027  78.755  81.301
##                        Dim.19  Dim.20  Dim.21  Dim.22  Dim.23  Dim.24
## Variance                0.059   0.054   0.051   0.048   0.044   0.036
## % of var.               2.494   2.271   2.176   2.052   1.846   1.543
## Cumulative % of var.   83.794  86.066  88.242  90.293  92.139  93.682
##                        Dim.25  Dim.26  Dim.27  Dim.28  Dim.29  Dim.30
## Variance                0.033   0.025   0.022   0.019   0.015   0.013
## % of var.               1.405   1.077   0.934   0.798   0.620   0.561
## Cumulative % of var.   95.087  96.164  97.099  97.897  98.517  99.078
##                        Dim.31  Dim.32  Dim.33
## Variance                0.010   0.007   0.004
## % of var.               0.426   0.312   0.185
```

```
## Cumulative % of var.  99.503  99.815 100.000
##
## Individuals (the 50 first)
##                                      Dim.1    ctr   cos2    Dim.2    ctr
## 285                                |  0.375  0.012  0.065 |  0.302  0.010
## 307                                |  0.097  0.001  0.004 |  0.692  0.053
## 401                                |  0.150  0.002  0.010 |  0.411  0.019
## 593                                | -0.660  0.037  0.167 |  0.585  0.038
## 636                                |  0.639  0.034  0.173 | -0.442  0.022
## 886                                |  0.345  0.010  0.051 | -0.791  0.069
## 904                                | -0.217  0.004  0.021 |  0.339  0.013
## 978                                |  0.326  0.009  0.040 |  0.560  0.035
## 1135                               | -0.358  0.011  0.059 | -0.182  0.004
## 1282                               | -0.481  0.019  0.099 |  0.008  0.000
## 1409                               | -0.539  0.024  0.129 |  0.115  0.001
## 1475                               |  0.225  0.004  0.021 | -0.341  0.013
## 1495                               |  0.891  0.067  0.355 | -0.298  0.010
## 1905                               | -0.122  0.001  0.007 |  0.671  0.050
## 2126                               | -0.079  0.001  0.002 |  0.673  0.050
## 2151                               | -0.663  0.037  0.195 | -0.052  0.000
## 2201                               |  0.862  0.062  0.332 | -0.061  0.000
## 2271                               | -0.215  0.004  0.020 |  0.400  0.018
## 2747                               | -0.617  0.032  0.167 |  0.497  0.027
## 3065                               |  0.912  0.070  0.368 |  0.146  0.002
## 3089                               |  0.566  0.027  0.142 |  0.390  0.017
## 3130                               | -0.705  0.042  0.220 |  0.320  0.011
## 3221                               | -0.614  0.032  0.165 | -0.115  0.001
## 3420                               | -0.763  0.049  0.264 | -0.162  0.003
## 3679                               |  0.123  0.001  0.006 |  0.487  0.026
## 4310                               |  0.034  0.000  0.000 |  0.065  0.000
## 4754                               | -0.405  0.014  0.071 | -0.531  0.031
## 5241                               |  0.573  0.028  0.129 | -0.629  0.044
## 5277                               | -0.448  0.017  0.081 | -0.312  0.011
## 5649                               |  0.239  0.005  0.020 | -0.324  0.012
## 6353                               | -0.215  0.004  0.016 |  0.202  0.005
## 6364                               |  0.031  0.000  0.000 | -0.613  0.041
## 6755                               | -0.288  0.007  0.037 | -0.097  0.001
## 6869                               | -0.475  0.019  0.095 | -0.279  0.009
## 7079                               | -0.437  0.016  0.087 |  0.472  0.025
## 7211                               | -0.327  0.009  0.047 |  0.374  0.015
## 7342                               | -0.174  0.003  0.013 | -0.535  0.032
## 7802                               | -0.039  0.000  0.001 | -0.642  0.045
## 8138                               |  0.304  0.008  0.039 |  0.074  0.001
## 8443                               |  0.309  0.008  0.043 | -0.185  0.004
## 8619                               |  0.261  0.006  0.033 |  0.795  0.070
## 8891                               | -0.230  0.004  0.024 |  0.429  0.020
## 8960                               | -0.376  0.012  0.065 | -0.042  0.000
## 9207                               |  0.218  0.004  0.021 |  0.594  0.039
## 9503                               |  0.158  0.002  0.011 |  0.838  0.077
## 9747                               |  0.112  0.001  0.006 |  0.657  0.048
## 9765                               |  0.200  0.003  0.006 |  0.925  0.094
## 9984                               |  0.619  0.032  0.143 | -0.538  0.032
## 10034                              |  0.759  0.048  0.251 |  0.300  0.010
## 10199                              | -0.536  0.024  0.126 |  0.528  0.031
```

```
##                                cos2    Dim.3    ctr   cos2    Dim.4
## 285                           0.042 |  0.104  0.001  0.005 | -0.072
## 307                           0.215 |  0.037  0.000  0.001 | -0.127
## 401                           0.076 | -0.136  0.002  0.008 | -0.104
## 593                           0.131 |  0.517  0.032  0.102 | -0.698
## 636                           0.083 | -0.377  0.017  0.060 | -0.105
## 886                           0.269 | -0.069  0.001  0.002 | -0.376
## 904                           0.050 | -0.177  0.004  0.014 | -0.618
## 978                           0.117 |  0.329  0.013  0.040 |  0.032
## 1135                          0.015 |  0.594  0.042  0.163 | -0.560
## 1282                          0.000 |  0.532  0.033  0.121 | -0.047
## 1409                          0.006 | -0.520  0.032  0.120 |  0.288
## 1475                          0.049 | -0.402  0.019  0.067 |  0.377
## 1495                          0.040 |  0.389  0.018  0.068 |  0.080
## 1905                          0.205 | -0.162  0.003  0.012 |  0.181
## 2126                          0.173 | -0.519  0.032  0.103 | -0.584
## 2151                          0.001 | -0.455  0.024  0.092 |  0.549
## 2201                          0.002 |  0.377  0.017  0.063 | -0.015
## 2271                          0.068 | -0.523  0.032  0.117 | -0.361
## 2747                          0.109 |  0.140  0.002  0.009 | -0.660
## 3065                          0.009 |  0.468  0.026  0.097 |  0.070
## 3089                          0.067 |  0.378  0.017  0.063 | -0.165
## 3130                          0.045 |  0.502  0.030  0.111 | -0.396
## 3221                          0.006 | -0.457  0.025  0.091 |  0.445
## 3420                          0.012 |  0.604  0.043  0.165 |  0.262
## 3679                          0.089 | -0.135  0.002  0.007 | -0.024
## 4310                          0.002 | -0.375  0.017  0.060 |  0.671
## 4754                          0.122 |  0.592  0.041  0.152 |  0.253
## 5241                          0.156 | -0.425  0.021  0.071 | -0.089
## 5277                          0.039 |  0.493  0.029  0.098 | -0.574
## 5649                          0.036 | -0.442  0.023  0.067 | -0.302
## 6353                          0.015 | -0.129  0.002  0.006 |  0.588
## 6364                          0.133 |  0.331  0.013  0.039 | -0.566
## 6755                          0.004 | -0.456  0.025  0.094 | -0.277
## 6869                          0.033 |  0.241  0.007  0.025 |  0.390
## 7079                          0.101 | -0.534  0.034  0.129 | -0.616
## 7211                          0.062 | -0.552  0.036  0.135 | -0.692
## 7342                          0.126 |  0.565  0.038  0.140 | -0.627
## 7802                          0.188 | -0.445  0.023  0.091 | -0.299
## 8138                          0.002 | -0.437  0.023  0.081 | -0.087
## 8443                          0.015 | -0.195  0.005  0.017 |  0.160
## 8619                          0.302 |  0.336  0.013  0.054 | -0.147
## 8891                          0.083 | -0.507  0.030  0.116 | -0.244
## 8960                          0.001 |  0.602  0.043  0.165 | -0.145
## 9207                          0.159 | -0.203  0.005  0.019 | -0.488
## 9503                          0.314 |  0.294  0.010  0.039 |  0.071
## 9747                          0.193 |  0.306  0.011  0.042 |  0.298
## 9765                          0.135 |  0.322  0.012  0.016 |  0.421
## 9984                          0.108 | -0.407  0.020  0.062 | -0.186
## 10034                         0.039 |  0.418  0.021  0.076 |  0.111
## 10199                         0.122 | -0.202  0.005  0.018 | -0.606
##                                 ctr   cos2
## 285                           0.001  0.002 |
## 307                           0.002  0.007 |
```

```
## 401                                0.001  0.005 |
## 593                                0.065  0.187 |
## 636                                0.001  0.005 |
## 886                                0.019  0.061 |
## 904                                0.051  0.167 |
## 978                                0.000  0.000 |
## 1135                               0.042  0.145 |
## 1282                               0.000  0.001 |
## 1409                               0.011  0.037 |
## 1475                               0.019  0.059 |
## 1495                               0.001  0.003 |
## 1905                               0.004  0.015 |
## 2126                               0.046  0.130 |
## 2151                               0.040  0.134 |
## 2201                               0.000  0.000 |
## 2271                               0.017  0.056 |
## 2747                               0.058  0.192 |
## 3065                               0.001  0.002 |
## 3089                               0.004  0.012 |
## 3130                               0.021  0.069 |
## 3221                               0.026  0.087 |
## 3420                               0.009  0.031 |
## 3679                               0.000  0.000 |
## 4310                               0.060  0.192 |
## 4754                               0.009  0.028 |
## 5241                               0.001  0.003 |
## 5277                               0.044  0.133 |
## 5649                               0.012  0.031 |
## 6353                               0.046  0.123 |
## 6364                               0.043  0.113 |
## 6755                               0.010  0.034 |
## 6869                               0.020  0.064 |
## 7079                               0.051  0.172 |
## 7211                               0.064  0.212 |
## 7342                               0.053  0.173 |
## 7802                               0.012  0.041 |
## 8138                               0.001  0.003 |
## 8443                               0.003  0.012 |
## 8619                               0.003  0.010 |
## 8891                               0.008  0.027 |
## 8960                               0.003  0.010 |
## 9207                               0.032  0.108 |
## 9503                               0.001  0.002 |
## 9747                               0.012  0.040 |
## 9765                               0.024  0.028 |
## 9984                               0.005  0.013 |
## 10034                              0.002  0.005 |
## 10199                              0.049  0.161 |
##
## Categories
##                                       Dim.1     ctr    cos2  v.test
## Creative Mobile Technologies, LLC |   0.025   0.004   0.000   0.931 |
## VeriFone Inc.                     |  -0.007   0.001   0.000  -0.931 |
## Credit card                       |   0.439   2.823   0.188  30.481 |
```

```
## Cash                                |  -0.434   2.787   0.187 -30.432 |
## Other                               |  -0.047   0.000   0.000  -0.281 |
## Street-hail                         |   0.001   0.000   0.000   0.689 |
## Dispatch                            |  -0.074   0.003   0.000  -0.689 |
## night                               |   0.031   0.005   0.000   0.972 |
## morning                             |   0.004   0.000   0.000   0.151 |
## valley                              |  -0.033   0.009   0.000  -1.430 |
## afternoon                           |   0.009   0.001   0.000   0.470 |
## onePassager                         |  -0.022   0.012   0.003  -3.580 |
## multiPassagers                      |   0.120   0.065   0.003   3.580 |
## Dist1                               |  -0.976   7.022   0.315 -39.447 |
## Dist2                               |  -0.465   1.588   0.071 -18.740 |
## Dist3                               |   0.326   0.766   0.034  12.979 |
## Dist4                               |   1.063   8.778   0.401  44.501 |
## p.Y1                                |   0.799   4.776   0.215  32.612 |
## p.Y2                                |  -0.067   0.032   0.001  -2.637 |
## p.Y3                                |  -0.410   1.298   0.059 -17.092 |
## p.Y4                                |  -0.313   0.733   0.033 -12.770 |
## p.X1                                |   0.752   4.252   0.192  30.790 |
## p.X2                                |   0.302   0.632   0.028  11.725 |
## p.X3                                |  -0.472   1.780   0.082 -20.146 |
## p.X4                                |  -0.550   2.183   0.097 -21.926 |
## d.Y1                                |   0.958   6.308   0.277  36.973 |
## d.Y2                                |  -0.186   0.288   0.013  -8.165 |
## d.Y3                                |  -0.503   1.853   0.083 -20.234 |
## d.Y4                                |  -0.189   0.255   0.011  -7.489 |
## d.X1                                |   0.618   2.990   0.137  26.015 |
## d.X2                                |   0.225   0.412   0.019   9.730 |
## d.X3                                |  -0.108   0.077   0.003  -4.064 |
## d.X4                                |  -0.842   5.020   0.222 -33.136 |
## FAmount1                            |  -0.984   7.272   0.328 -40.267 |
## FAmount2                            |  -0.441   1.465   0.066 -18.084 |
## FAmount3                            |   0.359   0.929   0.041  14.294 |
## FAmount4                            |   1.095   8.893   0.400  44.439 |
## smallTip                            |  -0.372   2.558   0.228 -33.568 |
## highTip                             |   0.612   4.208   0.228  33.568 |
## Time1                               |  -0.980   7.081   0.317 -39.606 |
## Time2                               |  -0.428   1.331   0.059 -17.125 |
## Time3                               |   0.311   0.707   0.032  12.497 |
## Time4                               |   1.039   8.368   0.382  43.438 |
## Speed1                              |  -0.115   0.096   0.004  -4.615 |
## Speed2                              |  -0.131   0.120   0.005  -5.123 |
## Speed3                              |   0.077   0.042   0.002   3.023 |
## Speed4                              |   0.146   0.177   0.008   6.407 |
##                                      Dim.2     ctr    cos2  v.test
## Creative Mobile Technologies, LLC  -0.184   0.284   0.009  -6.780 |
## VeriFone Inc.                       0.050   0.078   0.009   6.780 |
## Credit card                        -0.198   0.752   0.038 -13.739 |
## Cash                                0.198   0.759   0.039  13.864 |
## Other                              -0.127   0.004   0.000  -0.752 |
## Street-hail                        -0.022   0.018   0.027 -11.577 |
## Dispatch                            1.245   1.037   0.027  11.577 |
## night                               0.323   0.657   0.020   9.977 |
## morning                             0.067   0.035   0.001   2.368 |
```

```
## valley                             -0.042   0.019   0.001  -1.849 |
## afternoon                          -0.152   0.319   0.013  -7.927 |
## onePassager                        -0.016   0.008   0.001  -2.632 |
## multiPassagers                      0.088   0.046   0.001   2.632 |
## Dist1                              -0.466   2.098   0.072 -18.824 |
## Dist2                              -0.331   1.055   0.036 -13.338 |
## Dist3                               0.042   0.016   0.001   1.652 |
## Dist4                               0.716   5.225   0.182  29.977 |
## p.Y1                               -0.976   9.348   0.321 -39.835 |
## p.Y2                               -0.241   0.534   0.018  -9.422 |
## p.Y3                                0.137   0.190   0.007   5.715 |
## p.Y4                                1.062  11.050   0.379  43.299 |
## p.X1                               -1.015  10.140   0.349 -41.517 |
## p.X2                                0.220   0.441   0.015   8.545 |
## p.X3                                0.736   5.675   0.200  31.407 |
## p.X4                                0.029   0.008   0.000   1.163 |
## d.Y1                               -0.554   2.768   0.093 -21.384 |
## d.Y2                               -0.500   2.733   0.098 -21.960 |
## d.Y3                                0.012   0.001   0.000   0.477 |
## d.Y4                                1.099  11.365   0.385  43.624 |
## d.X1                               -0.929   8.877   0.310 -39.139 |
## d.X2                                0.366   1.430   0.051  15.818 |
## d.X3                                0.714   4.420   0.146  26.872 |
## d.X4                               -0.059   0.032   0.001  -2.307 |
## FAmount1                           -0.467   2.152   0.074 -19.127 |
## FAmount2                           -0.324   1.034   0.036 -13.265 |
## FAmount3                            0.108   0.110   0.004   4.287 |
## FAmount4                            0.697   4.726   0.162  28.286 |
## smallTip                            0.147   0.526   0.036  13.295 |
## highTip                            -0.243   0.866   0.036 -13.295 |
## Time1                              -0.434   1.820   0.062 -17.530 |
## Time2                              -0.249   0.592   0.020  -9.971 |
## Time3                               0.122   0.142   0.005   4.893 |
## Time4                               0.531   2.865   0.100  22.191 |
## Speed1                             -0.205   0.404   0.014  -8.251 |
## Speed2                             -0.283   0.734   0.025 -11.050 |
## Speed3                             -0.078   0.055   0.002  -3.036 |
## Speed4                              0.483   2.550   0.091  21.221 |
##                                     Dim.3     ctr    cos2  v.test
## Creative Mobile Technologies, LLC  -0.017   0.002   0.000  -0.608 |
## VeriFone Inc.                       0.005   0.001   0.000   0.608 |
## Credit card                         0.035   0.025   0.001   2.400 |
## Cash                               -0.040   0.033   0.002  -2.777 |
## Other                               0.379   0.042   0.001   2.252 |
## Street-hail                        -0.003   0.000   0.000  -1.475 |
## Dispatch                            0.159   0.018   0.000   1.475 |
## night                               0.083   0.046   0.001   2.564 |
## morning                             0.142   0.169   0.005   5.010 |
## valley                             -0.012   0.002   0.000  -0.547 |
## afternoon                          -0.109   0.174   0.006  -5.660 |
## onePassager                        -0.007   0.002   0.000  -1.198 |
## multiPassagers                      0.040   0.010   0.000   1.198 |
## Dist1                               1.009  10.539   0.337  40.781 |
## Dist2                              -0.829   7.079   0.226 -33.392 |
```

```
## Dist3                              -0.894   8.086   0.256 -35.587 |
## Dist4                               0.654   4.667   0.152  27.383 |
## p.Y1                                0.065   0.045   0.001   2.657 |
## p.Y2                                0.062   0.038   0.001   2.429 |
## p.Y3                                0.069   0.051   0.002   2.867 |
## p.Y4                               -0.195   0.397   0.013  -7.934 |
## p.X1                                0.083   0.072   0.002   3.379 |
## p.X2                               -0.037   0.013   0.000  -1.444 |
## p.X3                               -0.093   0.098   0.003  -3.986 |
## p.X4                                0.053   0.029   0.001   2.122 |
## d.Y1                                0.224   0.485   0.015   8.649 |
## d.Y2                               -0.004   0.000   0.000  -0.186 |
## d.Y3                               -0.067   0.046   0.001  -2.703 |
## d.Y4                               -0.141   0.201   0.006  -5.605 |
## d.X1                               -0.028   0.008   0.000  -1.163 |
## d.X2                                0.031   0.011   0.000   1.350 |
## d.X3                               -0.009   0.001   0.000  -0.322 |
## d.X4                                0.003   0.000   0.000   0.105 |
## FAmount1                            1.093  12.590   0.405  44.711 |
## FAmount2                           -1.001  10.598   0.341 -41.039 |
## FAmount3                           -0.853   7.364   0.234 -33.971 |
## FAmount4                            0.739   5.691   0.182  30.000 |
## smallTip                           -0.010   0.003   0.000  -0.909 |
## highTip                             0.017   0.004   0.000   0.909 |
## Time1                               1.063  11.696   0.373  42.954 |
## Time2                              -0.914   8.509   0.270 -36.546 |
## Time3                              -0.798   6.533   0.208 -32.055 |
## Time4                               0.596   3.871   0.126  24.931 |
## Speed1                              0.020   0.004   0.000   0.784 |
## Speed2                             -0.141   0.196   0.006  -5.517 |
## Speed3                             -0.108   0.115   0.004  -4.228 |
## Speed4                              0.192   0.434   0.014   8.459 |
##                                    Dim.4     ctr    cos2  v.test
## Creative Mobile Technologies, LLC   0.007   0.000   0.000   0.254 |
## VeriFone Inc.                      -0.002   0.000   0.000  -0.254 |
## Credit card                         0.182   0.776   0.033  12.675 |
## Cash                               -0.193   0.879   0.037 -13.558 |
## Other                               0.888   0.264   0.006   5.272 |
## Street-hail                        -0.012   0.007   0.008  -6.411 |
## Dispatch                            0.689   0.386   0.008   6.411 |
## night                              -0.380   1.103   0.028 -11.740 |
## morning                             0.370   1.306   0.035  13.091 |
## valley                              0.125   0.208   0.006   5.507 |
## afternoon                          -0.136   0.312   0.010  -7.116 |
## onePassager                         0.045   0.080   0.011   7.423 |
## multiPassagers                     -0.249   0.446   0.011  -7.423 |
## Dist1                              -0.345   1.394   0.039 -13.938 |
## Dist2                              -0.117   0.159   0.004  -4.708 |
## Dist3                               0.095   0.103   0.003   3.772 |
## Dist4                               0.350   1.511   0.043  14.640 |
## p.Y1                               -0.368   1.612   0.046 -15.026 |
## p.Y2                                0.298   0.987   0.027  11.631 |
## p.Y3                                0.863   9.152   0.262  35.995 |
## p.Y4                               -0.803   7.657   0.217 -32.737 |
```

```
## p.X1                              -0.350   1.462   0.041 -14.319 |
## p.X2                              -0.649   4.645   0.129 -25.201 |
## p.X3                              -0.283   1.022   0.030 -12.102 |
## p.X4                               1.300  19.427   0.544  51.865 |
## d.Y1                              -0.159   0.275   0.008  -6.125 |
## d.Y2                               0.359   1.704   0.050  15.751 |
## d.Y3                               0.482   2.700   0.076  19.371 |
## d.Y4                              -0.755   6.510   0.182 -29.988 |
## d.X1                              -0.425   2.256   0.065 -17.922 |
## d.X2                              -0.671   5.837   0.171 -29.030 |
## d.X3                               0.058   0.035   0.001   2.165 |
## d.X4                               1.190  15.942   0.444  46.827 |
## FAmount1                          -0.433   2.237   0.063 -17.711 |
## FAmount2                          -0.032   0.012   0.000  -1.314 |
## FAmount3                           0.075   0.064   0.002   2.986 |
## FAmount4                           0.398   1.866   0.053  16.144 |
## smallTip                          -0.109   0.349   0.020  -9.834 |
## highTip                            0.179   0.574   0.020   9.834 |
## Time1                             -0.401   1.887   0.053 -16.213 |
## Time2                             -0.107   0.133   0.004  -4.290 |
## Time3                              0.092   0.097   0.003   3.680 |
## Time4                              0.395   1.927   0.055  16.529 |
## Speed1                             0.063   0.046   0.001   2.533 |
## Speed2                             0.014   0.002   0.000   0.564 |
## Speed3                            -0.209   0.489   0.014  -8.187 |
## Speed4                             0.109   0.157   0.005   4.779 |
##
## Categorical variables (eta2)
##                                    Dim.1 Dim.2 Dim.3 Dim.4
## VendorID                         | 0.000 0.009 0.000 0.000 |
## Payment_type                     | 0.189 0.039 0.002 0.041 |
## Trip_type                        | 0.000 0.027 0.000 0.008 |
## pick_up_period                   | 0.000 0.026 0.009 0.062 |
## f.passenger                      | 0.003 0.001 0.000 0.011 |
## f.distance                       | 0.612 0.216 0.729 0.067 |
## f.pickup_longitude               | 0.230 0.543 0.013 0.411 |
## f.pickup_latitude                | 0.298 0.418 0.005 0.563 |
## f.dropoff_longitude              | 0.293 0.433 0.018 0.237 |
## f.dropoff_latitude               | 0.286 0.379 0.000 0.510 |
## f.fare_amount                    | 0.625 0.206 0.870 0.089 |
## f.tip_amount                     | 0.228 0.036 0.000 0.020 |
## f.ttime                          | 0.589 0.139 0.735 0.086 |
## f.espeed                         | 0.015 0.096 0.018 0.015 |
##
## Supplementary categories
##                                       Dim.1    cos2  v.test     Dim.2
## AnyTip No                          |  -0.371   0.190 -30.665 |   0.182
## AnyTip Yes                         |   0.513   0.190  30.665 |  -0.251
## CheapestTrip                       |  -1.043   0.369 -42.708 |  -0.382
## CheapTrip                          |  -0.452   0.065 -17.863 |  -0.261
## MediumTrip                         |   0.330   0.035  13.152 |   0.030
## ExpensiveTrip                      |   1.112   0.442  46.711 |   0.579
##                                       cos2  v.test     Dim.3    cos2
## AnyTip No                           0.046  15.013 |  -0.018   0.000
```

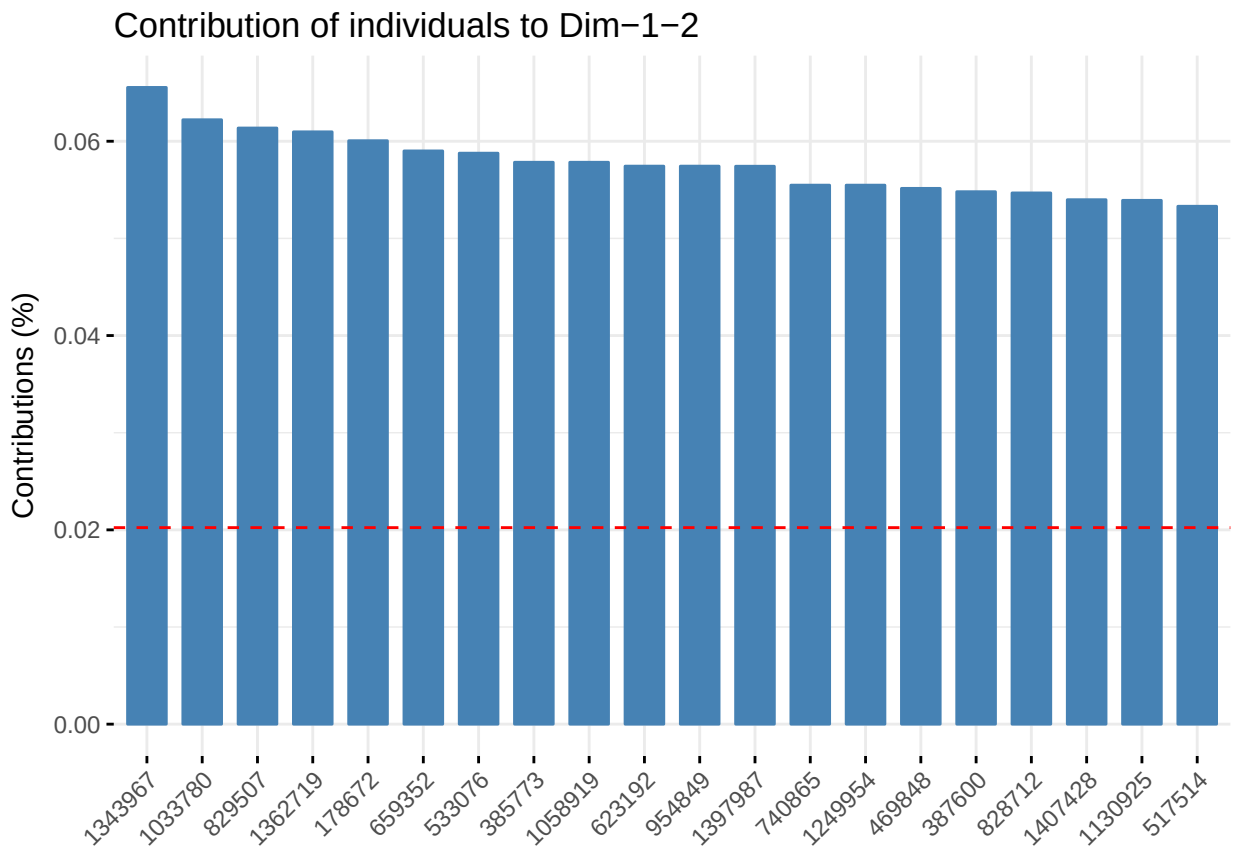
```
## AnyTip Yes                          0.046 -15.013 |   0.025   0.000
## CheapestTrip                        0.050 -15.656 |   0.836   0.237
## CheapTrip                           0.022 -10.311 |  -0.730   0.168
## MediumTrip                          0.000   1.196 |  -0.817   0.215
## ExpensiveTrip                       0.120  24.308 |   0.617   0.136
##                                    v.test     Dim.4    cos2  v.test
## AnyTip No                          -1.517 |  -0.120   0.020  -9.885 |
## AnyTip Yes                          1.517 |   0.165   0.020   9.885 |
## CheapestTrip                       34.234 |  -0.354   0.043 -14.506 |
## CheapTrip                         -28.848 |  -0.138   0.006  -5.447 |
## MediumTrip                        -32.592 |   0.108   0.004   4.297 |
## ExpensiveTrip                      25.931 |   0.367   0.048  15.430 |
##
## Supplementary categorical variables (eta2)
##                                     Dim.1 Dim.2 Dim.3 Dim.4
## AnyTip                            | 0.190 0.046 0.000 0.020 |
## f.total                           | 0.676 0.142 0.568 0.075 |
##
## Supplementary continuous variable
##                                     Dim.1   Dim.2   Dim.3   Dim.4
## Total_amount                      | 0.697 | 0.342 | 0.193 | 0.243 |
```

Individuals

We have 4866 observation so each individual should have $100/4866 \sim 0.02$ contribution.So the expected avaradge contribution is $0.02*Dim1 + 0.02*$Dim2 $= 0.02*0.1+0.02*0.78 = 0.035$.So as we see in the two plots there are some too contributive individuals.

Too contributive groups would be: - Individuals with longest travel time, high total price, high fare amount and longest distance - Individulas with shortest travel time, lowest Fare Amount and lowest total amount - Individuals with a dispach trip type and the biggest pickup and dropoff longtitude - Individuals with smallest pickup and droppoff longitude and latitude.

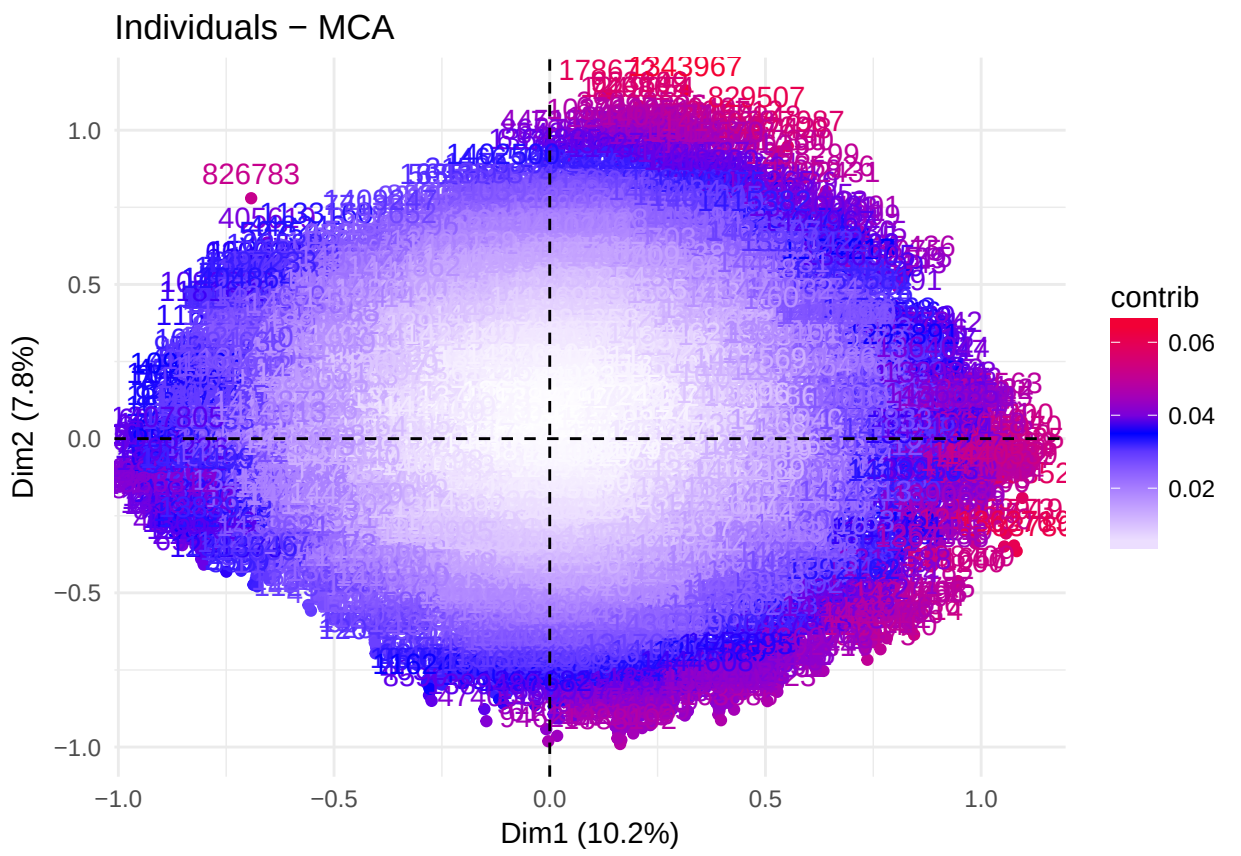
```
fviz_contrib(res.mca, choice = "ind", axes = 1:2, top = 20)
```



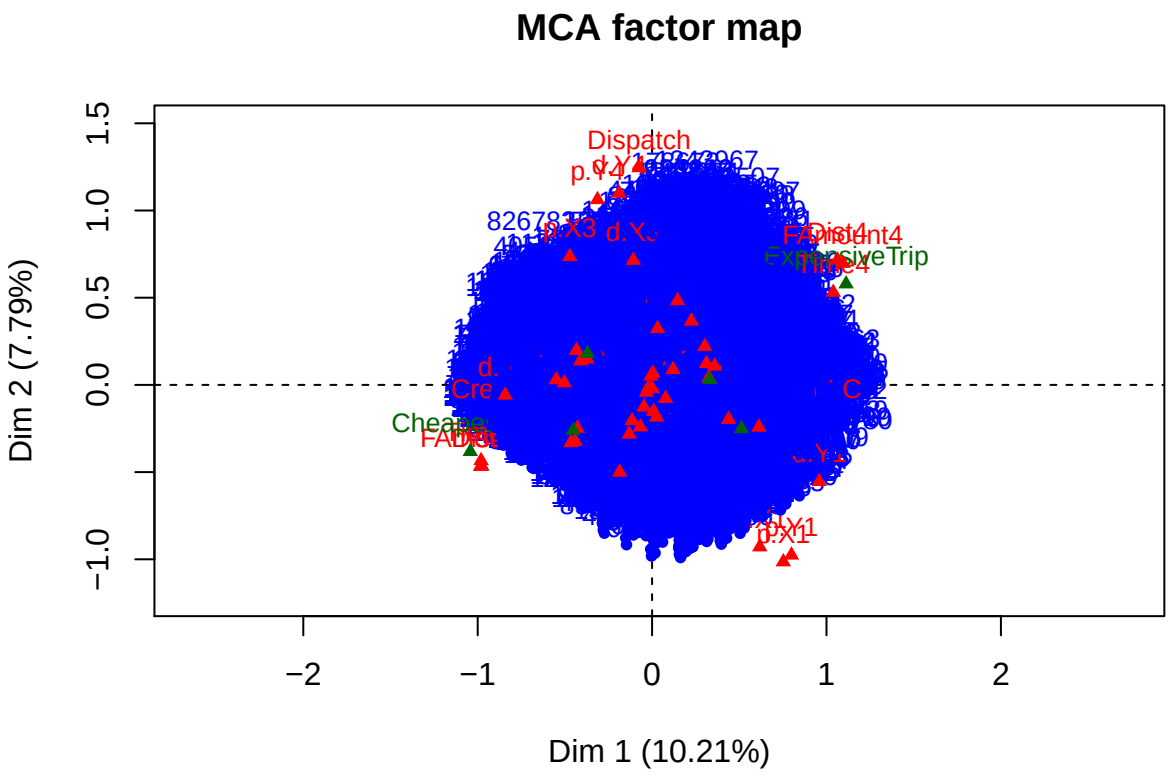

```
head(res.mca$ind$contrib)
```

```
##            Dim 1      Dim 2        Dim 3        Dim 4       Dim 5
## 285 0.0118166324 0.01004614 0.0012672814 0.0006861235 0.005730337
## 307 0.0007881181 0.05276895 0.0001606660 0.0021552330 0.001544272
## 401 0.0019001724 0.01863540 0.0021780373 0.0014488616 0.023048032
## 593 0.0366643489 0.03779164 0.0315087675 0.0650350774 0.004841817
## 636 0.0343240250 0.02152825 0.0167396766 0.0014726896 0.062537228
## 886 0.0100317628 0.06900959 0.0005687912 0.0189143315 0.014698087
```

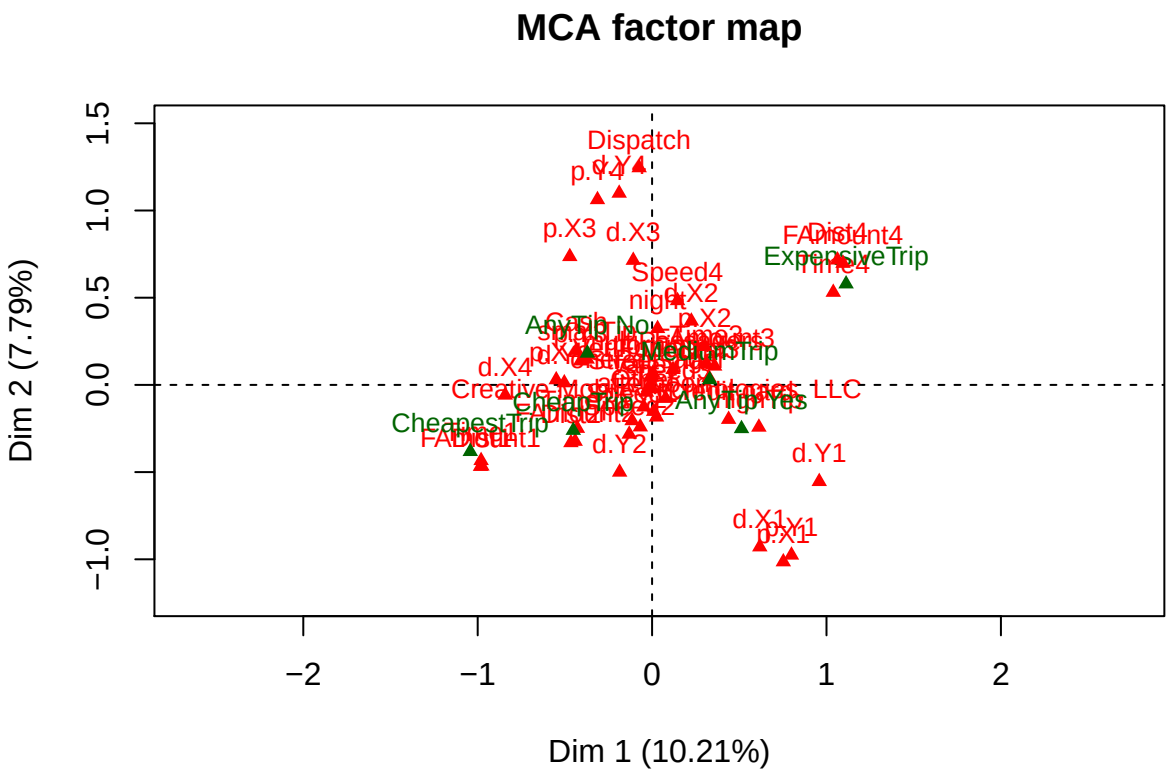
```
fviz_mca_ind(res.mca, col.ind="contrib")+
scale_color_gradient2(low="white", mid="blue",
                      high="red", midpoint=0.035)+theme_minimal()
```




```
plot.MCA(res.mca,choix=c("ind"),cex=0.8)
```



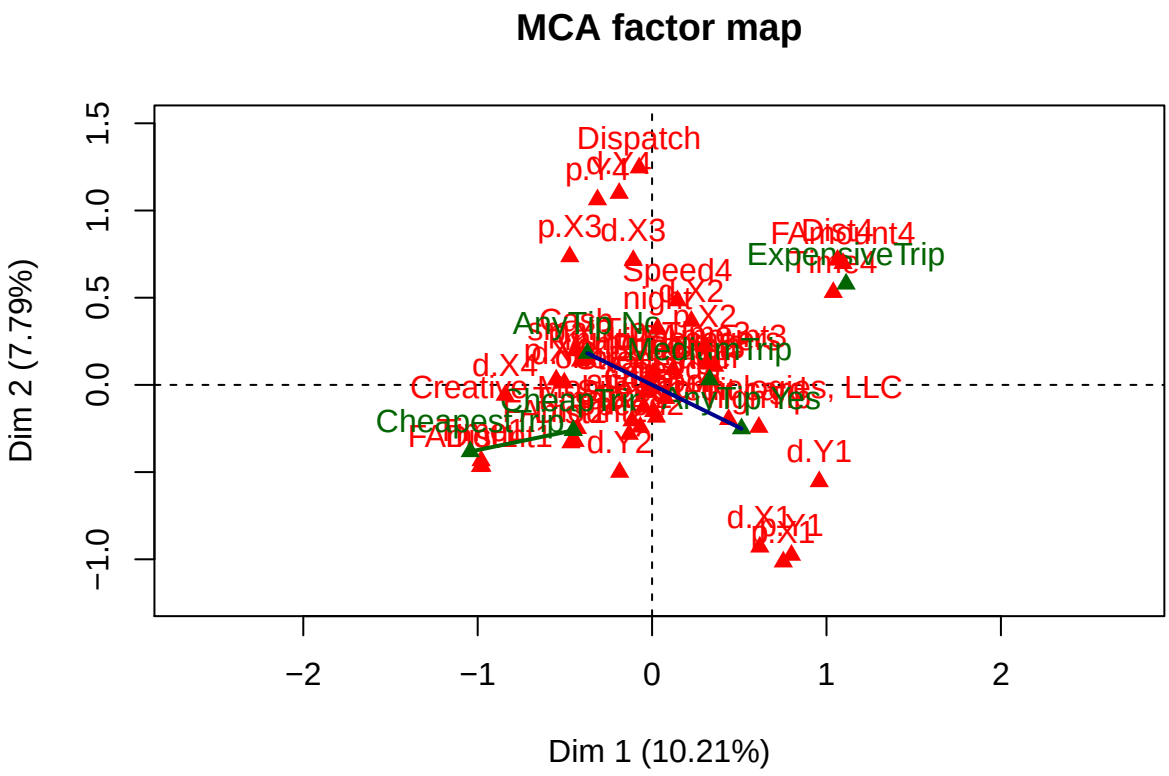

```
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),cex=0.8)
```



Representation of categories

We see that the pickup and dropoff coordinates and fare amount are best correlated to the first 2 dimentions. Rare categories can be found near the border such as Dispach.

```
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),axes=c(1,2))
lines(res.mca$quali.sup$coord[3:4,1],res.mca$quali.sup$coord[3:4,2],lwd=2,col="darkgreen") # Trencada d
lines(res.mca$quali.sup$coord[1:2,1],res.mca$quali.sup$coord[1:2,2],lwd=2,col="darkblue") # Trencada An
```




```
res.mca$var
```

```
## $coord
##                                        Dim 1       Dim 2        Dim 3
## Creative Mobile Technologies, LLC  0.025309318 -0.18438062 -0.016536519
## VeriFone Inc.                     -0.006925662  0.05045406  0.004525067
## Credit card                        0.438795770 -0.19778694  0.034548701
## Cash                              -0.433854349  0.19765896 -0.039597277
## Other                             -0.047269927 -0.12660869  0.379398784
## Street-hail                        0.001297067 -0.02178261 -0.002775562
## Dispatch                          -0.074131195  1.24494010  0.158631556
## night                              0.031460579  0.32297465  0.083006427
## morning                            0.004264005  0.06700790  0.141775953
## valley                            -0.032511692 -0.04202793 -0.012439509
## afternoon                          0.009005178 -0.15198048 -0.108519892
## onePassager                       -0.021605135 -0.01588207 -0.007230660
## multiPassagers                     0.120031713  0.08823608  0.040171397
## Dist1                             -0.975987936 -0.46574062  1.008985844
## Dist2                             -0.465425412 -0.33126297 -0.829321185
## Dist3                              0.326058940  0.04150950 -0.894045935
## Dist4                              1.063003775  0.71606659  0.654112576
## p.Y1                               0.799094299 -0.97607306  0.065114954
## p.Y2                              -0.067479376 -0.24108391  0.062147778
## p.Y3                              -0.409998535  0.13709811  0.068781376
## p.Y4                              -0.313235490  1.06207700 -0.194624227
## p.X1                               0.752411211 -1.01456843  0.082561162
```

```
## p.X2                                0.301880601  0.21999669 -0.037178236
## p.X3                               -0.471835233  0.73559969 -0.093352817
## p.X4                               -0.549631121  0.02914464  0.053189352
## d.Y1                                0.958411179 -0.55431722  0.224192051
## d.Y2                               -0.186057760 -0.50041906 -0.004230181
## d.Y3                               -0.503349662  0.01186255 -0.067229422
## d.Y4                               -0.188674850  1.09897165 -0.141194342
## d.X1                                0.617543059 -0.92908845 -0.027602807
## d.X2                                0.225002784  0.36578200  0.031209187
## d.X3                               -0.107990813  0.71404635 -0.008569177
## d.X4                               -0.842190099 -0.05864436  0.002657391
## FAmount1                           -0.984020860 -0.46741057  1.092614958
## FAmount2                           -0.441206049 -0.32364546 -1.001270292
## FAmount3                            0.358720495  0.10757805 -0.852512468
## FAmount4                            1.094741462  0.69682299  0.739038528
## smallTip                           -0.372332268  0.14746967 -0.010086859
## highTip                             0.612385977 -0.24254775  0.016590158
## Time1                              -0.980442581 -0.43396518  1.063339550
## Time2                              -0.428350103 -0.24940178 -0.914105494
## Time3                               0.311207565  0.12184332 -0.798287257
## Time4                               1.038716387  0.53064330  0.596160304
## Speed1                             -0.114865694 -0.20536380  0.019524070
## Speed2                             -0.131013873 -0.28259118 -0.141086429
## Speed3                              0.077272151 -0.07759419 -0.108067223
## Speed4                              0.145777256  0.48286670  0.192482894
##                                          Dim 4        Dim 5
## Creative Mobile Technologies, LLC  0.006909478  0.034393189
## VeriFone Inc.                     -0.001890715 -0.009411380
## Credit card                        0.182460968 -0.072698279
## Cash                              -0.193290505  0.072098089
## Other                              0.888134282 -0.007562576
## Street-hail                       -0.012063259 -0.002517008
## Dispatch                           0.689450726  0.143854434
## night                             -0.380039256  0.092519093
## morning                            0.370456934 -0.167289622
## valley                             0.125176179 -0.080338739
## afternoon                         -0.136429756  0.116537701
## onePassager                        0.044797197 -0.003675250
## multiPassagers                    -0.248879920  0.020418595
## Dist1                             -0.344853864  0.252264341
## Dist2                             -0.116920269 -0.984381582
## Dist3                              0.094751862  1.118422003
## Dist4                              0.349720215 -0.347725737
## p.Y1                              -0.368177388  0.007877618
## p.Y2                               0.297599544 -0.053224957
## p.Y3                               0.863451609  0.086185355
## p.Y4                              -0.803013193 -0.047054947
## p.X1                              -0.349925818  0.020552847
## p.X2                              -0.648824565 -0.043671749
## p.X3                              -0.283444570 -0.029311793
## p.X4                               1.300121730  0.053003520
## d.Y1                              -0.158769560 -0.121747224
## d.Y2                               0.358928384  0.013650530
## d.Y3                               0.481893043  0.158979101
```

```
## d.Y4                                -0.755471205 -0.061319963
## d.X1                                -0.425436986  0.106635559
## d.X2                                -0.671301209 -0.155157184
## d.X3                                 0.057534763  0.002901756
## d.X4                                 1.190187168  0.057750148
## FAmount1                            -0.432794880  0.290259768
## FAmount2                            -0.032061931 -1.088199879
## FAmount3                             0.074926068  1.298417251
## FAmount4                             0.397711102 -0.452435160
## smallTip                            -0.109073912  0.062704656
## highTip                              0.179397113 -0.103132216
## Time1                               -0.401358187  0.287576605
## Time2                               -0.107302749 -0.979676997
## Time3                                0.091633287  1.000376462
## Time4                                0.395249334 -0.298656702
## Speed1                               0.063044751 -0.207118928
## Speed2                               0.014418020 -0.014768902
## Speed3                              -0.209252060  0.191985687
## Speed4                               0.108747264  0.032312848
##
## $contrib
##                                          Dim 1        Dim 2        Dim 3
## Creative Mobile Technologies, LLC 4.083841e-03  0.284295794 2.448201e-03
## VeriFone Inc.                     1.117506e-03  0.077794933 6.699277e-04
## Credit card                       2.822630e+00  0.752239698 2.457217e-02
## Cash                              2.786534e+00  0.758650030 3.259553e-02
## Other                             4.694846e-04  0.004417844 4.247117e-02
## Street-hail                       4.906432e-05  0.018150669 3.154960e-04
## Dispatch                          2.804170e-03  1.037364096 1.803152e-02
## night                             4.753427e-03  0.657116285 4.646733e-02
## morning                           1.088213e-04  0.035250282 1.689408e-01
## valley                            8.826524e-03  0.019347220 1.814546e-03
## afternoon                         8.543684e-04  0.319203853 1.742329e-01
## onePassager                       1.173837e-02  0.008320321 1.846293e-03
## multiPassagers                    6.521489e-02  0.046225232 1.025745e-02
## Dist1                             7.022159e+00  2.097500062 1.053908e+01
## Dist2                             1.587813e+00  1.055059188 7.079385e+00
## Dist3                             7.658732e-01  0.016281398 8.086028e+00
## Dist4                             8.777833e+00  5.224641246 4.667385e+00
## p.Y1                              4.776365e+00  9.347555805 4.453627e-02
## p.Y2                              3.190042e-02  0.534100140 3.799762e-02
## p.Y3                              1.297744e+00  0.190335278 5.128815e-02
## p.Y4                              7.327330e-01 11.049630601 3.972359e-01
## p.X1                              4.251589e+00 10.139939732 7.188582e-02
## p.X2                              6.324288e-01  0.440560455 1.347007e-02
## p.X3                              1.780198e+00  5.675480967 9.785734e-02
## p.X4                              2.183497e+00  0.008053029 2.871509e-02
## d.Y1                              6.308312e+00  2.767949319 4.847310e-01
## d.Y2                              2.880330e-01  2.733041906 2.090815e-04
## d.Y3                              1.852551e+00  0.001349643 4.640880e-02
## d.Y4                              2.553756e-01 11.364655066 2.008338e-01
## d.X1                              2.989930e+00  8.877131842 8.388494e-03
## d.X2                              4.124194e-01  1.429683451 1.114241e-02
## d.X3                              7.708047e-02  4.420338259 6.815534e-04
```

```
## d.X4                                5.020159e+00  0.031928735 7.018736e-05
## FAmount1                            7.271924e+00  2.152136150 1.259001e+01
## FAmount2                            1.465426e+00  1.034312939 1.059826e+01
## FAmount3                            9.285393e-01  0.109538596 7.364449e+00
## FAmount4                            8.892524e+00  4.725837558 5.690979e+00
## smallTip                            2.558283e+00  0.526410604 2.636636e-03
## highTip                             4.207685e+00  0.865803209 4.336553e-03
## Time1                               7.080637e+00  1.819573936 1.169561e+01
## Time2                               1.330602e+00  0.591672535 8.509301e+00
## Time3                               7.069965e-01  0.142151819 6.532603e+00
## Time4                               8.368351e+00  2.864727142 3.870998e+00
## Speed1                              9.639504e-02  0.404160365 3.910800e-03
## Speed2                              1.203541e-01  0.734472317 1.959962e-01
## Speed3                              4.190289e-02  0.055422783 1.150898e-01
## Speed4                              1.772006e-01  2.550187668 4.338313e-01
##                                            Dim 4        Dim 5
## Creative Mobile Technologies, LLC 4.839743e-04 1.290486e-02
## VeriFone Inc.                     1.324351e-04 3.531297e-03
## Credit card                       7.760558e-01 1.325799e-01
## Cash                              8.794708e-01 1.316814e-01
## Other                             2.635316e-01 2.056323e-05
## Street-hail                       6.748296e-03 3.161628e-04
## Dispatch                          3.856850e-01 1.806963e-02
## night                             1.102946e+00 7.034555e-02
## morning                           1.306105e+00 2.866270e-01
## valley                            2.080549e-01 9.222763e-02
## afternoon                         3.118191e-01 2.448465e-01
## onePassager                       8.024526e-02 5.812566e-04
## multiPassagers                    4.458188e-01 3.229289e-03
## Dist1                             1.394042e+00 8.027753e-01
## Dist2                             1.593319e-01 1.215420e+01
## Dist3                             1.028405e-01 1.541972e+01
## Dist4                             1.510717e+00 1.607281e+00
## p.Y1                              1.612277e+00 7.943131e-04
## p.Y2                              9.866036e-01 3.396138e-02
## p.Y3                              9.152175e+00 9.812775e-02
## p.Y4                              7.657254e+00 2.829534e-02
## p.X1                              1.462233e+00 5.428567e-03
## p.X2                              4.645364e+00 2.264868e-02
## p.X3                              1.021522e+00 1.175636e-02
## p.X4                              1.942683e+01 3.474720e-02
## d.Y1                              2.752757e-01 1.741919e-01
## d.Y2                              1.704457e+00 2.653054e-03
## d.Y3                              2.699951e+00 3.162360e-01
## d.Y4                              6.510451e+00 4.615891e-02
## d.X1                              2.256426e+00 1.525567e-01
## d.X2                              5.837435e+00 3.355889e-01
## d.X3                              3.479007e-02 9.523422e-05
## d.X4                              1.594233e+01 4.039280e-02
## FAmount1                          2.236809e+00 1.082717e+00
## FAmount2                          1.230507e-02 1.525453e+01
## FAmount3                          6.441362e-02 2.081694e+01
## FAmount4                          1.866213e+00 2.599053e+00
## smallTip                          3.491025e-01 1.241616e-01
```

```
## highTip                             5.741793e-01 2.042123e-01
## Time1                               1.886758e+00 1.042403e+00
## Time2                               1.327688e-01 1.191013e+01
## Time3                               9.746472e-02 1.250099e+01
## Time4                               1.926691e+00 1.183834e+00
## Speed1                              4.617382e-02 5.363077e-01
## Speed2                              2.317722e-03 2.617117e-03
## Speed3                              4.886089e-01 4.426252e-01
## Speed4                              1.568001e-01 1.489830e-02
##
## $cos2
##                                          Dim 1        Dim 2        Dim 3
## Creative Mobile Technologies, LLC 1.752838e-04 9.302751e-03 7.482885e-05
## VeriFone Inc.                     1.752838e-04 9.302751e-03 7.482885e-05
## Credit card                       1.879996e-01 3.819682e-02 1.165455e-03
## Cash                              1.873937e-01 3.889557e-02 1.560981e-03
## Other                             1.593431e-05 1.143116e-04 1.026492e-03
## Street-hail                       9.615313e-05 2.711804e-02 4.402918e-04
## Dispatch                          9.615313e-05 2.711804e-02 4.402918e-04
## night                             1.911210e-04 2.014243e-02 1.330450e-03
## morning                           4.593815e-06 1.134462e-03 5.078591e-03
## valley                            4.139361e-04 6.917198e-04 6.059832e-05
## afternoon                         4.464198e-05 1.271553e-02 6.483022e-03
## onePassager                       2.593301e-03 1.401372e-03 2.904657e-04
## multiPassagers                    2.593301e-03 1.401372e-03 2.904657e-04
## Dist1                             3.148679e-01 7.170143e-02 3.365191e-01
## Dist2                             7.106234e-02 3.599856e-02 2.256238e-01
## Dist3                             3.408424e-02 5.524036e-04 2.562602e-01
## Dist4                             4.007099e-01 1.818307e-01 1.517279e-01
## p.Y1                              2.152111e-01 3.210948e-01 1.428993e-03
## p.Y2                              1.407282e-03 1.796285e-02 1.193686e-03
## p.Y3                              5.911268e-02 6.609657e-03 1.663634e-03
## p.Y4                              3.299727e-02 3.793570e-01 1.273884e-02
## p.X1                              1.918254e-01 3.487854e-01 2.309656e-03
## p.X2                              2.781845e-02 1.477389e-02 4.219297e-04
## p.X3                              8.212161e-02 1.995997e-01 3.214634e-03
## p.X4                              9.727778e-02 2.735196e-04 9.110043e-04
## d.Y1                              2.766053e-01 9.252806e-02 1.513551e-02
## d.Y2                              1.348885e-02 9.757691e-02 6.972641e-06
## d.Y3                              8.284391e-02 4.601267e-05 1.477882e-03
## d.Y4                              1.135001e-02 3.850714e-01 6.356278e-03
## d.X1                              1.369413e-01 3.099660e-01 2.735939e-04
## d.X2                              1.915779e-02 5.063073e-02 3.685823e-04
## d.X3                              3.341978e-03 1.461110e-01 2.104303e-05
## d.X4                              2.221695e-01 1.077251e-03 2.211948e-06
## FAmount1                          3.280985e-01 7.402743e-02 4.045106e-01
## FAmount2                          6.617162e-02 3.560639e-02 3.407937e-01
## FAmount3                          4.134558e-02 3.718468e-03 2.335169e-01
## FAmount4                          3.995941e-01 1.618978e-01 1.821084e-01
## smallTip                          2.280111e-01 3.576844e-02 1.673426e-04
## highTip                           2.280111e-01 3.576844e-02 1.673426e-04
## Time1                             3.174046e-01 6.218400e-02 3.733471e-01
## Time2                             5.934363e-02 2.011757e-02 2.702520e-01
## Time3                             3.159908e-02 4.843705e-03 2.079182e-01
```

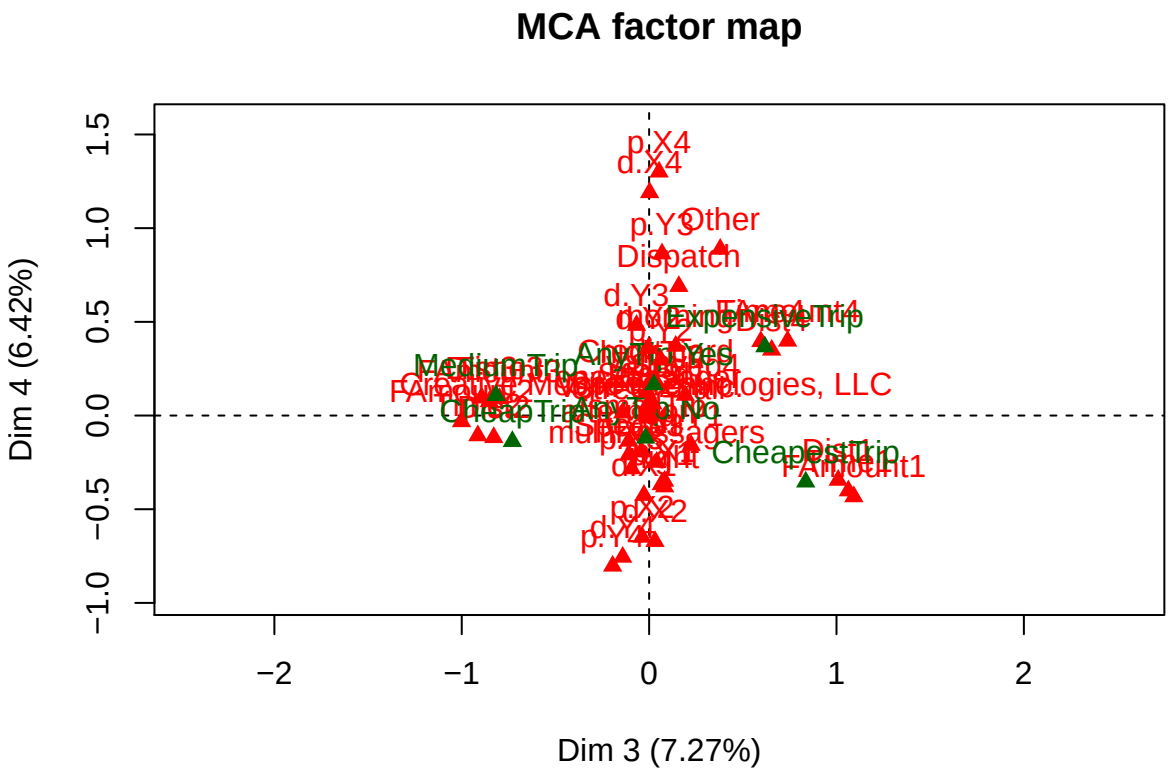
```
## Time4                                3.818077e-01 9.964512e-02 1.257699e-01
## Speed1                               4.309515e-03 1.377512e-02 1.245055e-04
## Speed2                               5.310806e-03 2.470831e-02 6.158804e-03
## Speed3                               1.849518e-03 1.864967e-03 3.617434e-03
## Speed4                               8.305473e-03 9.112538e-02 1.448002e-02
##                                             Dim 4        Dim 5
## Creative Mobile Technologies, LLC 1.306385e-05 3.236874e-04
## VeriFone Inc.                     1.306385e-05 3.236874e-04
## Credit card                       3.250663e-02 5.160363e-03
## Cash                              3.719530e-02 5.175050e-03
## Other                             5.624977e-03 4.078523e-07
## Street-hail                       8.317023e-03 3.620828e-04
## Dispatch                          8.317023e-03 3.620828e-04
## night                             2.788894e-02 1.652867e-03
## morning                           3.467477e-02 7.070923e-03
## valley                            6.136172e-03 2.527576e-03
## afternoon                         1.024653e-02 7.476386e-03
## onePassager                       1.114912e-02 7.504343e-05
## multiPassagers                    1.114912e-02 7.504343e-05
## Dist1                             3.931061e-02 2.103542e-02
## Dist2                             4.484550e-03 3.178822e-01
## Dist3                             2.878306e-03 4.010263e-01
## Dist4                             4.337124e-02 4.287796e-02
## p.Y1                              4.568597e-02 2.091502e-05
## p.Y2                              2.737180e-02 8.755269e-04
## p.Y3                              2.621754e-01 2.612059e-03
## p.Y4                              2.168610e-01 7.446405e-04
## p.X1                              4.149040e-02 1.431329e-04
## p.X2                              1.285041e-01 5.821886e-04
## p.X3                              2.963555e-02 3.169287e-04
## p.X4                              5.443009e-01 9.046497e-04
## d.Y1                              7.590864e-03 4.463493e-03
## d.Y2                              5.019899e-02 7.260698e-05
## d.Y3                              7.593156e-02 8.264205e-03
## d.Y4                              1.819718e-01 1.198870e-03
## d.X1                              6.499357e-02 4.083231e-03
## d.X2                              1.705314e-01 9.109894e-03
## d.X3                              9.486164e-04 2.412969e-06
## d.X4                              4.437054e-01 1.044649e-03
## FAmount1                          6.346874e-02 2.854758e-02
## FAmount2                          3.494365e-04 4.025374e-01
## FAmount3                          1.803776e-03 5.416831e-01
## FAmount4                          5.273893e-02 6.825093e-02
## smallTip                          1.956754e-02 6.466870e-03
## highTip                           1.956754e-02 6.466870e-03
## Time1                             5.319038e-02 2.730712e-02
## Time2                             3.723900e-03 3.104146e-01
## Time3                             2.739559e-03 3.265135e-01
## Time4                             5.528317e-02 3.156423e-02
## Speed1                            1.298212e-03 1.401159e-02
## Speed2                            6.431868e-05 6.748734e-05
## Speed3                            1.356289e-02 1.141696e-02
## Speed4                            4.621910e-03 4.080706e-04
##
```

```
## $v.test
##                                      Dim 1       Dim 2       Dim 3
## Creative Mobile Technologies, LLC   0.9307269  -6.7804275  -0.6081153
## VeriFone Inc.                      -0.9307269   6.7804275   0.6081153
## Credit card                        30.4810404 -13.7393112   2.3999328
## Cash                              -30.4318853  13.8644109  -2.7774754
## Other                              -0.2806196  -0.7516170   2.2523146
## Street-hail                         0.6893394 -11.5765868  -1.4751007
## Dispatch                           -0.6893394  11.5765868   1.4751007
## night                               0.9718643   9.9771691   2.5641925
## morning                             0.1506739   2.3678078   5.0098299
## valley                             -1.4302700  -1.8489131  -0.5472448
## afternoon                           0.4697027  -7.9271776  -5.6603088
## onePassager                        -3.5799575  -2.6316495  -1.1981158
## multiPassagers                      3.5799575   2.6316495   1.1981158
## Dist1                             -39.4471466 -18.8241453  40.7808448
## Dist2                             -18.7400657 -13.3380982 -33.3921035
## Dist3                              12.9786103   1.6522647 -35.5870438
## Dist4                              44.5006530  29.9767806  27.3831923
## p.Y1                               32.6124704 -39.8352907   2.6574579
## p.Y2                               -2.6371927  -9.4219118   2.4288260
## p.Y3                              -17.0919526   5.7153239   2.8673469
## p.Y4                              -12.7699841  43.2987538  -7.9344403
## p.X1                               30.7896277 -41.5174359   3.3785082
## p.X2                               11.7251347   8.5447387  -1.4440140
## p.X3                              -20.1455953  31.4073486  -3.9858153
## p.X4                              -21.9259387   1.1626408   2.1218349
## d.Y1                               36.9727374 -21.3839585   8.6486823
## d.Y2                               -8.1646738 -21.9596238  -0.1856308
## d.Y3                              -20.2339959   0.4768591  -2.7025345
## d.Y4                               -7.4894434  43.6236521  -5.6047059
## d.X1                               26.0146839 -39.1388777  -1.1627988
## d.X2                                9.7302510  15.8182518   1.3496421
## d.X3                               -4.0639950  26.8715523  -0.3224820
## d.X4                              -33.1355053  -2.3073301   0.1045536
## FAmount1                          -40.2673903 -19.1270375  44.7111995
## FAmount2                          -18.0836990 -13.2652468 -41.0390352
## FAmount3                           14.2944003   4.2868019 -33.9711688
## FAmount4                           44.4386523  28.2860162  29.9996641
## smallTip                          -33.5682984  13.2953981  -0.9093993
## highTip                            33.5682984 -13.2953981   0.9093993
## Time1                             -39.6057247 -17.5303539  42.9544108
## Time2                             -17.1253087  -9.9710085 -36.5456637
## Time3                              12.4965061   4.8926055 -32.0551386
## Time4                              43.4383874  22.1911290  24.9310039
## Speed1                             -4.6149349  -8.2508581   0.7844145
## Speed2                             -5.1230854 -11.0502708  -5.5169565
## Speed3                              3.0232961  -3.0358961  -4.2281625
## Speed4                              6.4066877  21.2212539   8.4593291
##                                      Dim 4       Dim 5
## Creative Mobile Technologies, LLC   0.2540897   1.2647779
## VeriFone Inc.                      -0.2540897  -1.2647779
## Credit card                        12.6746895  -5.0500012
## Cash                              -13.5579936   5.0571828
```

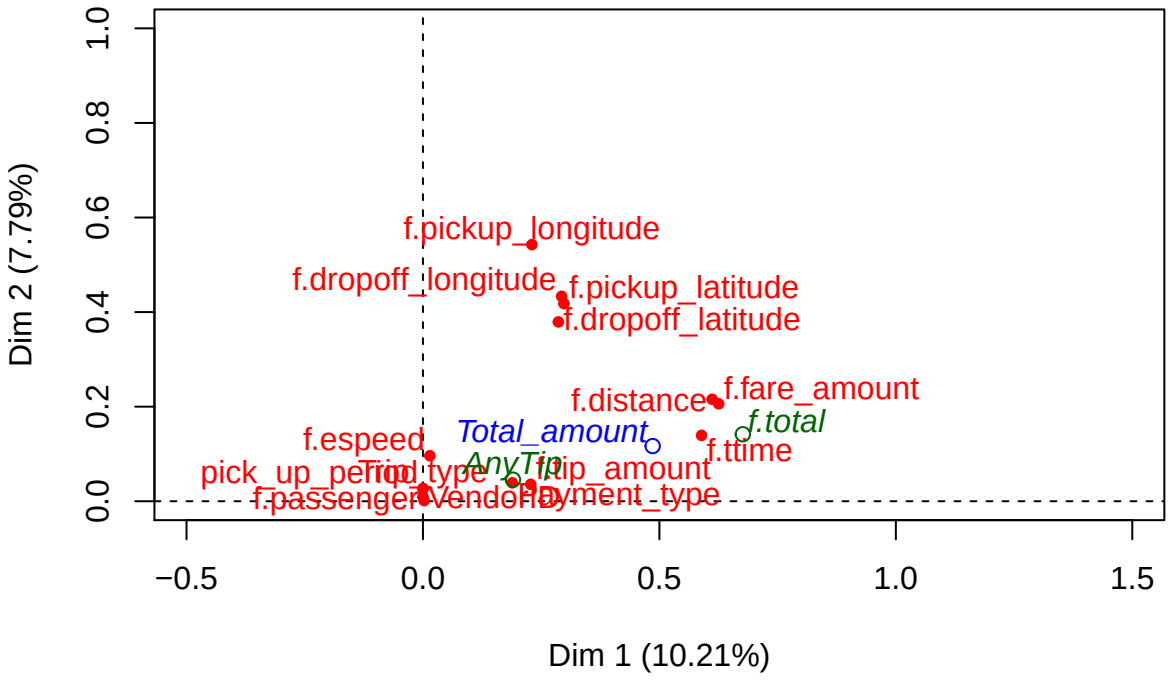
```
## Other                              5.2724413  -0.0448955
## Street-hail                       -6.4111408  -1.3376895
## Dispatch                           6.4111408   1.3376895
## night                            -11.7399800   2.8580529
## morning                           13.0905573  -5.9113872
## valley                             5.5068107  -3.5343005
## afternoon                         -7.1160645   6.0785112
## onePassager                        7.4228677  -0.6089866
## multiPassagers                    -7.4228677   0.6089866
## Dist1                            -13.9381855  10.1959339
## Dist2                             -4.7077222 -39.6355143
## Dist3                              3.7715497  44.5182191
## Dist4                             14.6403788 -14.5568837
## p.Y1                             -15.0259790   0.3214997
## p.Y2                              11.6306254  -2.0801091
## p.Y3                              35.9954310   3.5928812
## p.Y4                             -32.7372410  -1.9183360
## p.X1                             -14.3194114   0.8410487
## p.X2                             -25.2005442  -1.6962240
## p.X3                             -12.1020204  -1.2515037
## p.X4                              51.8645840   2.1144216
## d.Y1                              -6.1248714  -4.6966566
## d.Y2                              15.7506636   0.5990189
## d.Y3                              19.3714680   6.3907512
## d.Y4                             -29.9884106  -2.4340944
## d.X1                             -17.9220033   4.4921408
## d.X2                             -29.0304375  -6.7097763
## d.X3                               2.1651933   0.1092012
## d.X4                              46.8272581   2.2721478
## FAmount1                         -17.7105192  11.8778004
## FAmount2                          -1.3141214 -44.6020155
## FAmount3                           2.9856761  51.7397144
## FAmount4                          16.1442185 -18.3656228
## smallTip                          -9.8337585   5.6532532
## highTip                            9.8337585  -5.6532532
## Time1                            -16.2131696  11.6168759
## Time2                             -4.2899317 -39.1671928
## Time3                              3.6795247  40.1700087
## Time4                             16.5290486 -12.4896129
## Speed1                             2.5329357  -8.3213737
## Speed2                             0.5637933  -0.5775140
## Speed3                            -8.1870496   7.5114976
## Speed4                             4.7792761   1.4201003
##
## $eta2
##                            Dim 1       Dim 2        Dim 3        Dim 4
## VendorID             1.752838e-04 0.009302751 7.482885e-05 1.306385e-05
## Payment_type         1.890429e-01 0.038930933 2.391134e-03 4.067150e-02
## Trip_type            9.615313e-05 0.027118042 4.402918e-04 8.317023e-03
## pick_up_period       4.900994e-04 0.026486098 9.394152e-03 6.207407e-02
## f.passenger          2.593301e-03 0.001401372 2.904657e-04 1.114912e-02
## f.distance           6.117734e-01 0.215643401 7.288644e-01 6.711826e-02
## f.pickup_longitude   2.304635e-01 0.542651837 1.274433e-02 4.113294e-01
## f.pickup_latitude    2.981652e-01 0.417851816 5.085856e-03 5.628126e-01
```

```
## f.dropoff_longitude 2.933313e-01 0.433342970 1.757092e-02 2.371578e-01
## f.dropoff_latitude  2.864336e-01 0.379186938 4.867429e-04 5.101475e-01
## f.fare_amount       6.254129e-01 0.206094884 8.697762e-01 8.858320e-02
## f.tip_amount        2.280111e-01 0.035768436 1.673426e-04 1.956754e-02
## f.ttime             5.892926e-01 0.139201229 7.345431e-01 8.569966e-02
## f.espeed            1.468810e-02 0.096196231 1.797038e-02 1.470616e-02
##                            Dim 5
## VendorID            3.236874e-04
## Payment_type        5.204667e-03
## Trip_type           3.620828e-04
## pick_up_period      1.366829e-02
## f.passenger         7.504343e-05
## f.distance          5.904931e-01
## f.pickup_longitude  3.174193e-03
## f.pickup_latitude   1.468766e-03
## f.dropoff_longitude 1.061958e-02
## f.dropoff_latitude  1.041071e-02
## f.fare_amount       7.828851e-01
## f.tip_amount        6.466870e-03
## f.ttime             5.245860e-01
## f.espeed            1.962367e-02
```

```
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),axes=c(3,4))
```

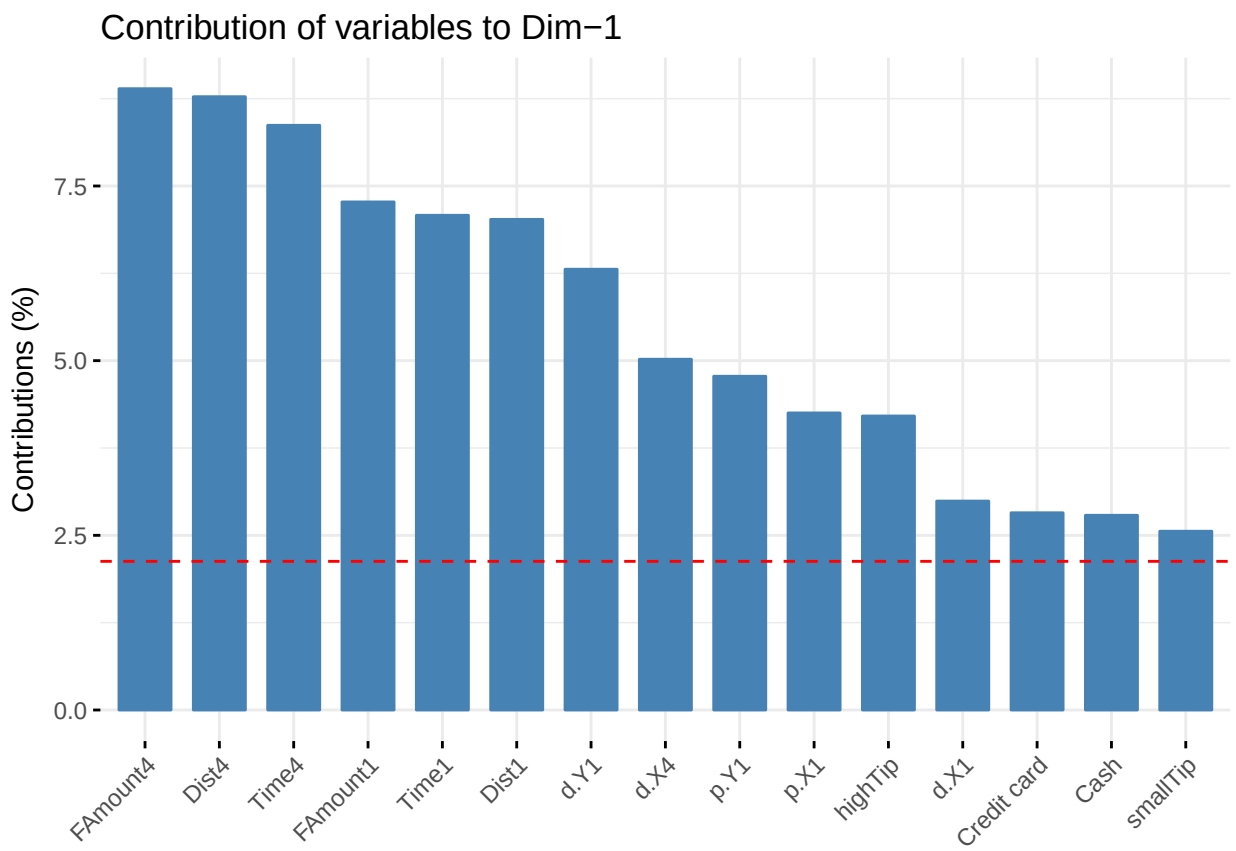



```
plot.MCA(res.mca,choix=c("var"))
```

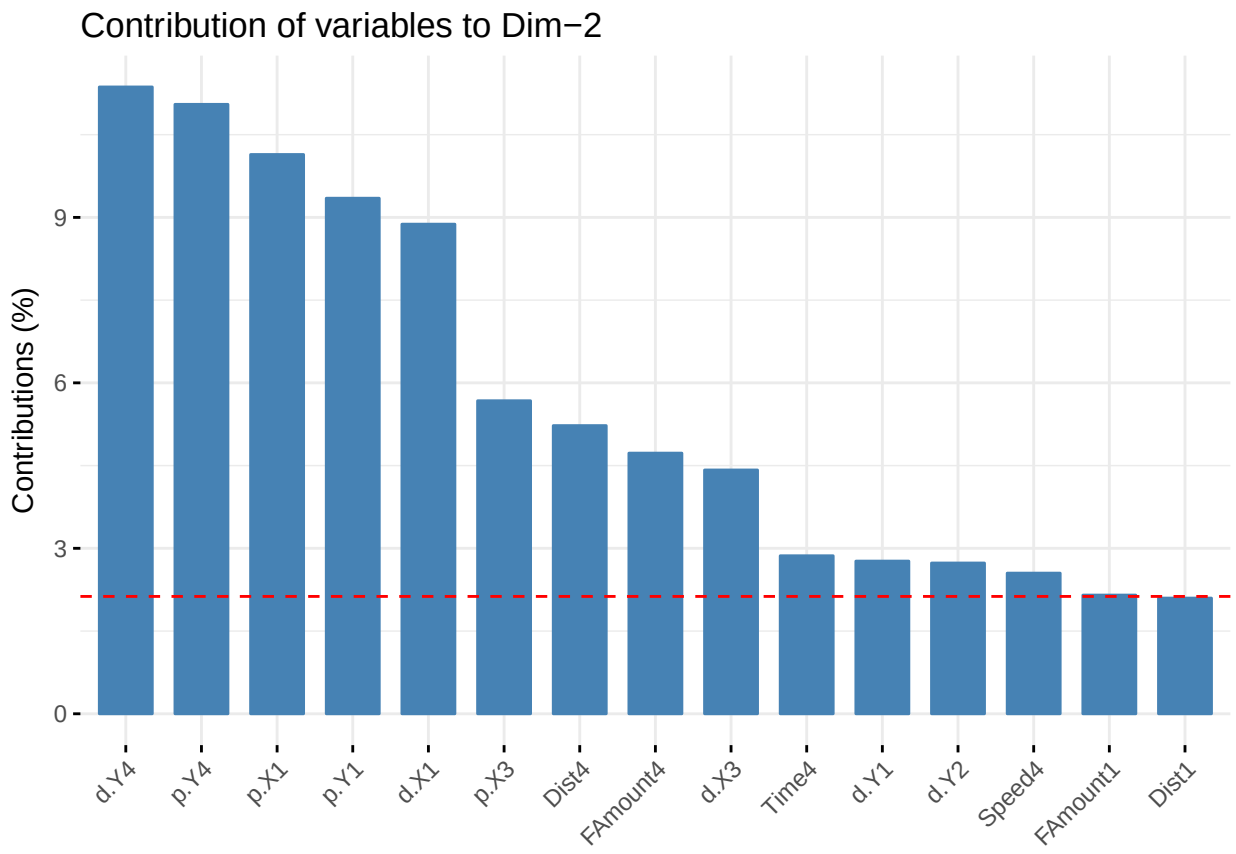


We see that the the provider of the record,categories of the rate code are most contributive and the the type of the payment. Also it makes difference when it is night.

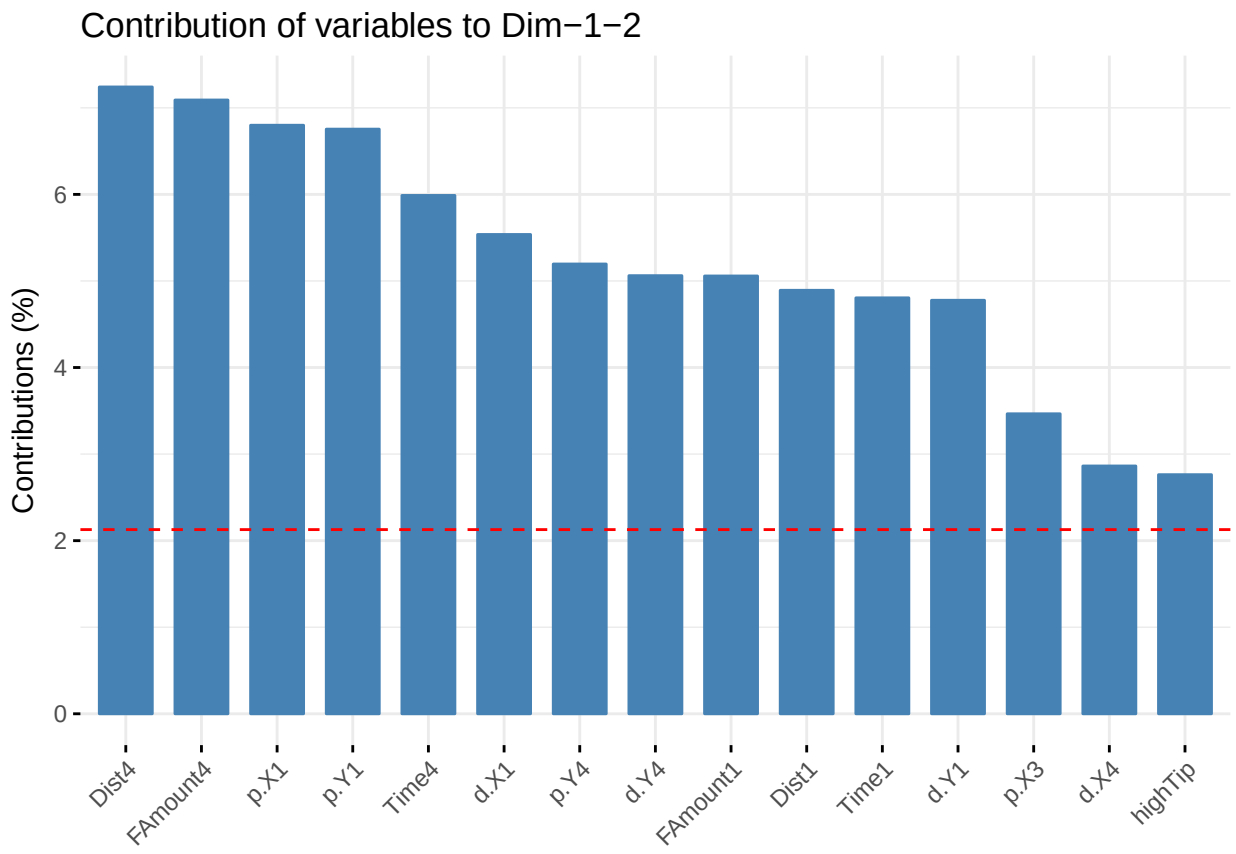
```
# Contributions of rows to dimension 1
fviz_contrib(res.mca, choice = "var", axes = 1, top = 15)
```




```
# Contributions of rows to dimension 2
fviz_contrib(res.mca, choice = "var", axes = 2, top = 15)
```



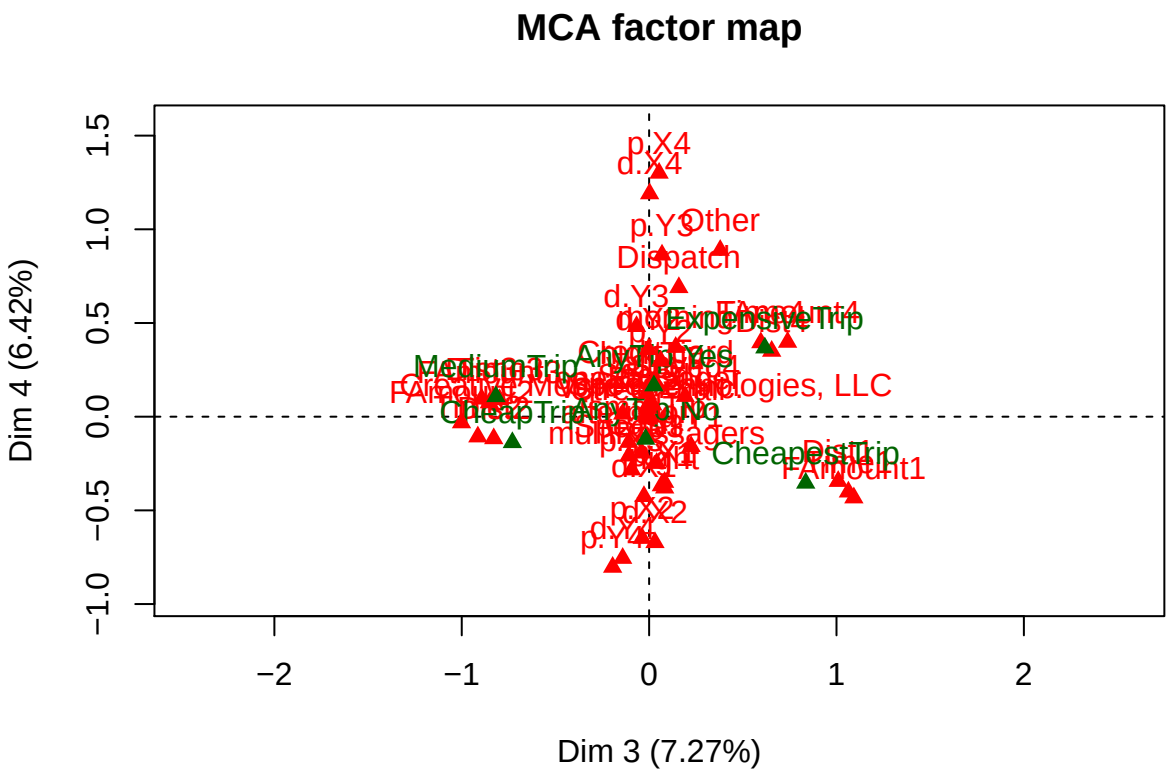

```
# Total contribution to dimension 1 and 2
fviz_contrib(res.mca, choice = "var", axes = 1:2, top = 15)
```



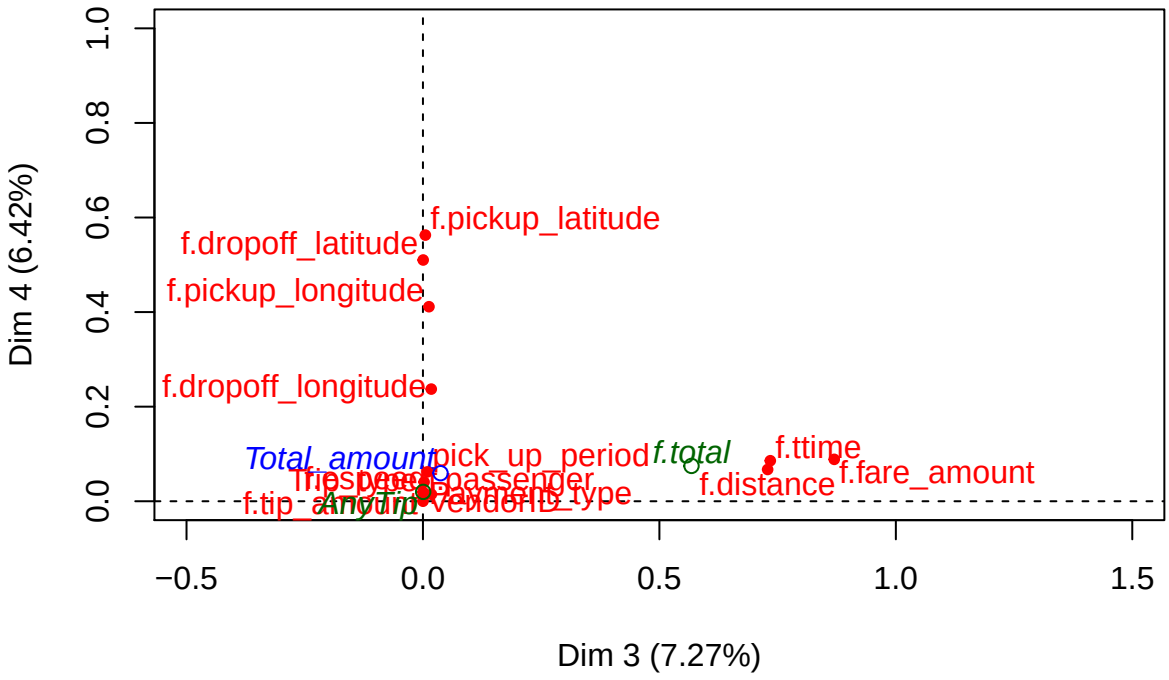

```
head(round(res.mca$var$contrib,2), 10)
```

```
##                                 Dim 1 Dim 2 Dim 3 Dim 4 Dim 5
## Creative Mobile Technologies, LLC  0.00  0.28  0.00  0.00  0.01
## VeriFone Inc.                      0.00  0.08  0.00  0.00  0.00
## Credit card                        2.82  0.75  0.02  0.78  0.13
## Cash                               2.79  0.76  0.03  0.88  0.13
## Other                              0.00  0.00  0.04  0.26  0.00
## Street-hail                        0.00  0.02  0.00  0.01  0.00
## Dispatch                           0.00  1.04  0.02  0.39  0.02
## night                              0.00  0.66  0.05  1.10  0.07
## morning                            0.00  0.04  0.17  1.31  0.29
## valley                             0.01  0.02  0.00  0.21  0.09
```

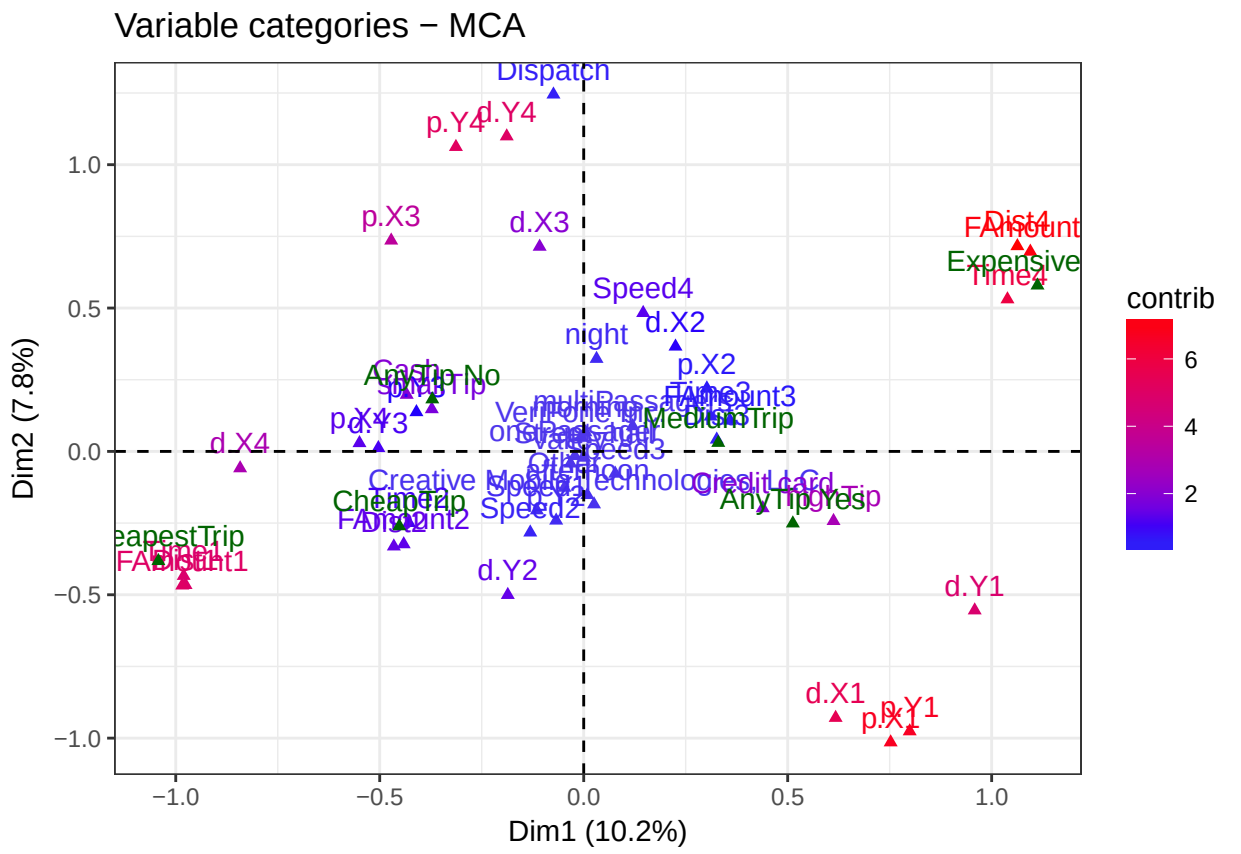
```
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),axes=c(3,4))
```



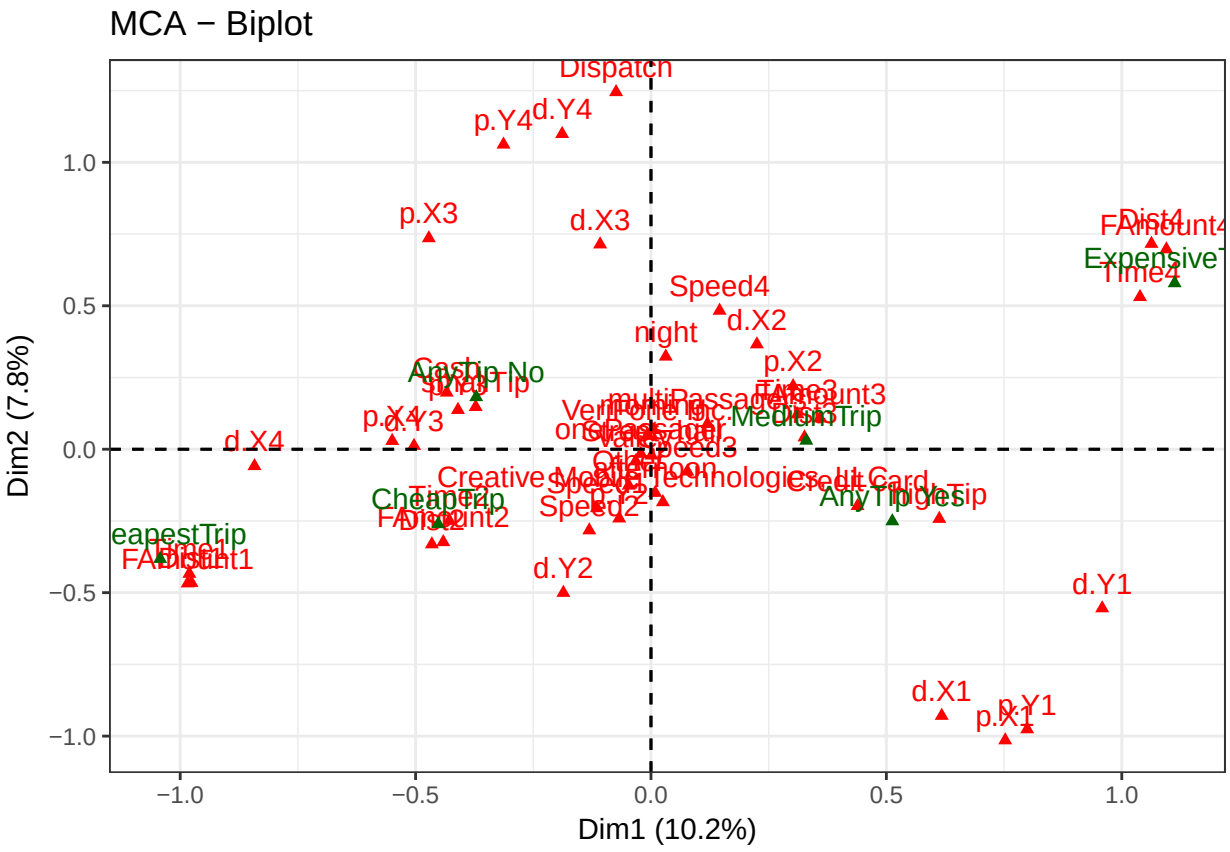

```
plot.MCA(res.mca,choix=c("var"),axes=c(3,4))
```




```
# Use modern ggplot facilities
fviz_mca_var(res.mca, col.var="contrib")+
    scale_color_gradient2(low="green", mid="blue",
    high="red", midpoint=0.75)+theme_bw()
```




```
fviz_mca_biplot(res.mca, invisible="ind",axes=1:2,repel=FALSE)+theme_bw()
```

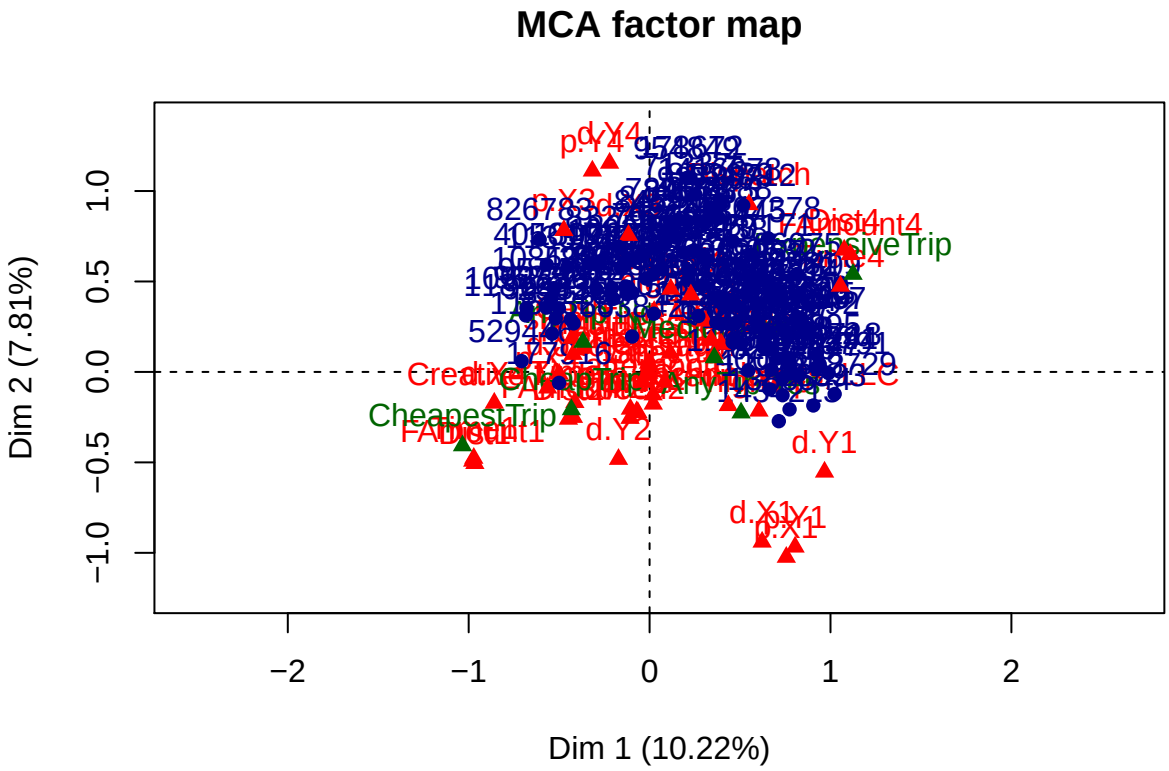


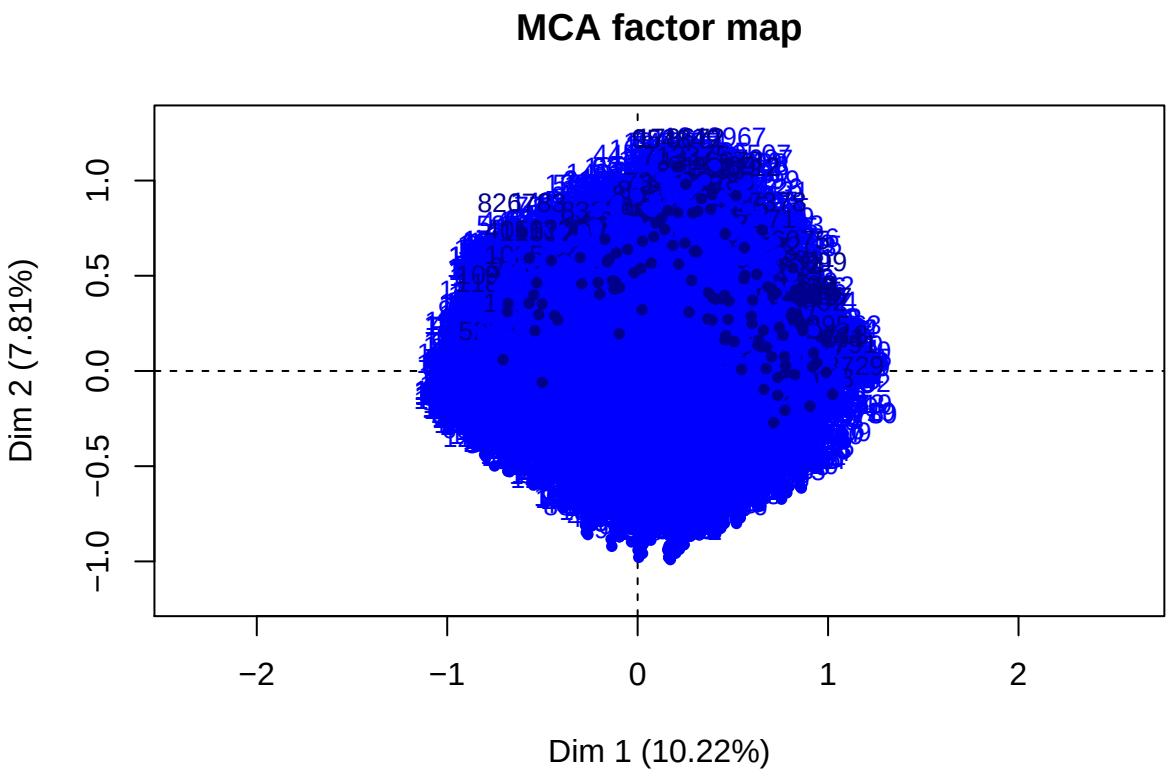
MCA with all supplementary variables

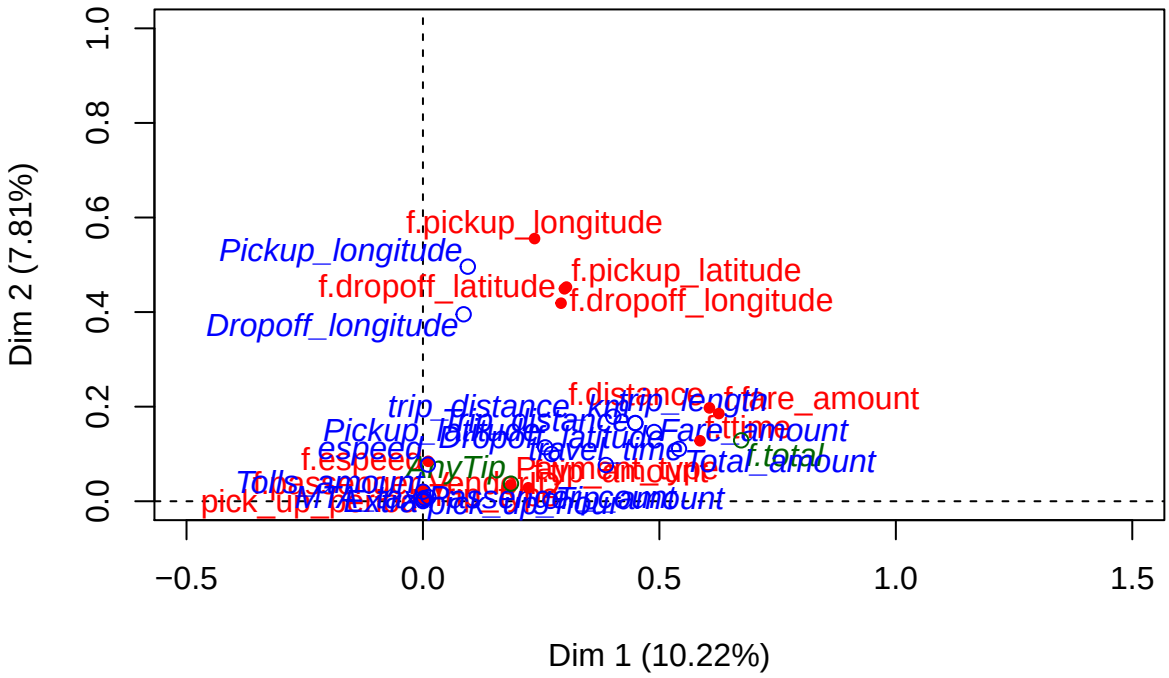
```
vars_con_pca <- names(df)[c(6:16,18,23:26,28)]
vec_out <- which(df$f.outlierPCA == "Outlier"); length(vec_out)
```

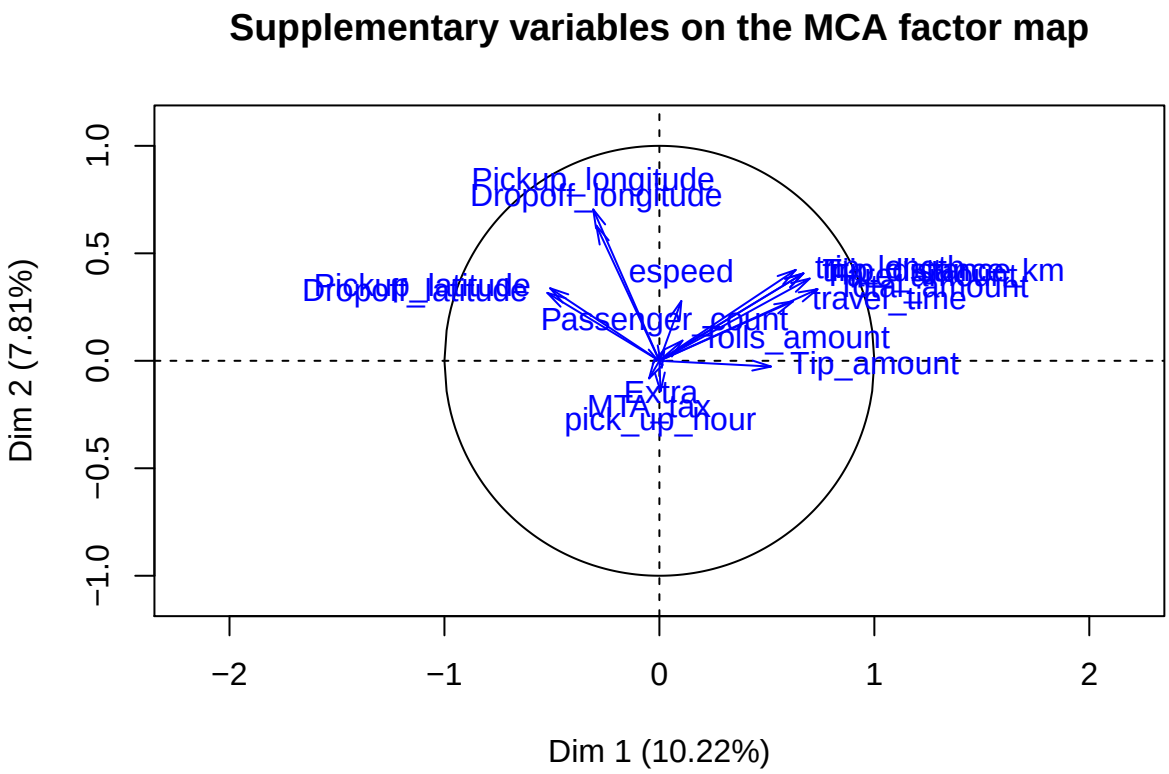
```
## [1] 119
```

```
res.mca<-MCA(df[,c(vars_cat,vars_con_pca)], ind.sup = vec_out, quali.sup=c(4,14),quanti.sup=17:33,ncp=1
```


MCA factor map
Dim 2 (7.81%)
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Dim 1 (10.22%)
d.Y4
p.Y4
CheapestTrip
d.Y2
d.Y1
d.X1y1
d.X1


MCA factor map
Dim 2 (7.81%)
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Dim 1 (10.22%)


Dim 2 (7.81%)
1.0
0.8
0.6
0.4
0.2
0.0
f.pickup_longitude
Pickup_longitude
f.pickup_latitude
f.dropoff_latitude
f.dropoff_longitude
Dropoff_longitude
f.distance_km
trip_length
f.fare_amount
f.ttime
f.total
f.speed
Total_amount
−0.5
0.0
0.5
1.0
1.5
Dim 1 (10.22%)

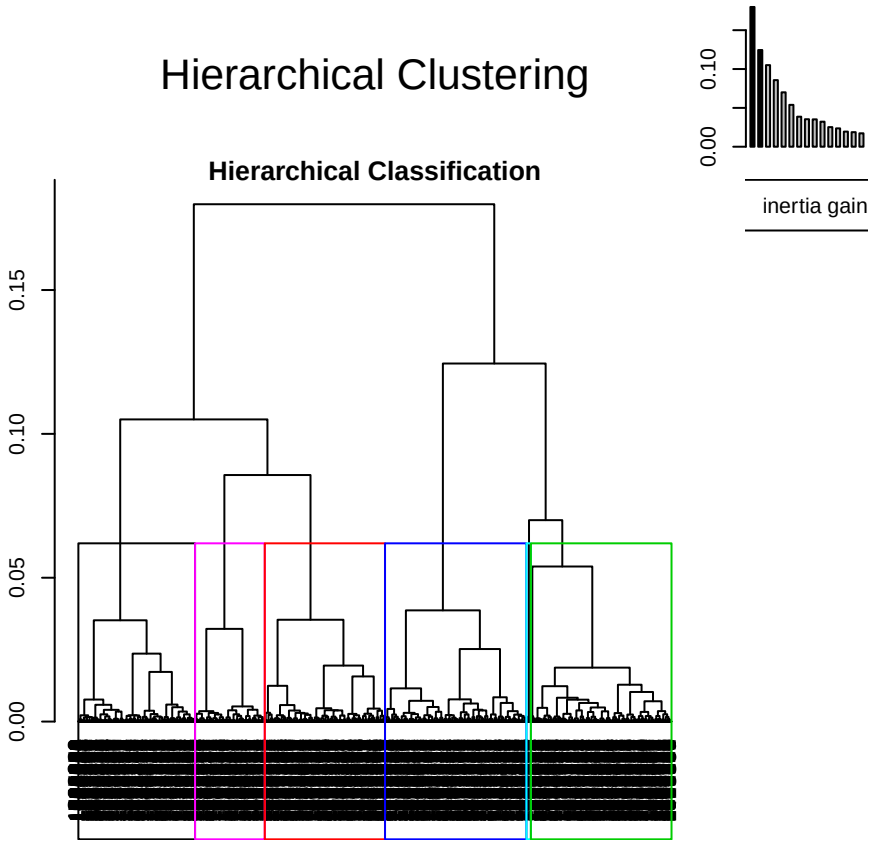


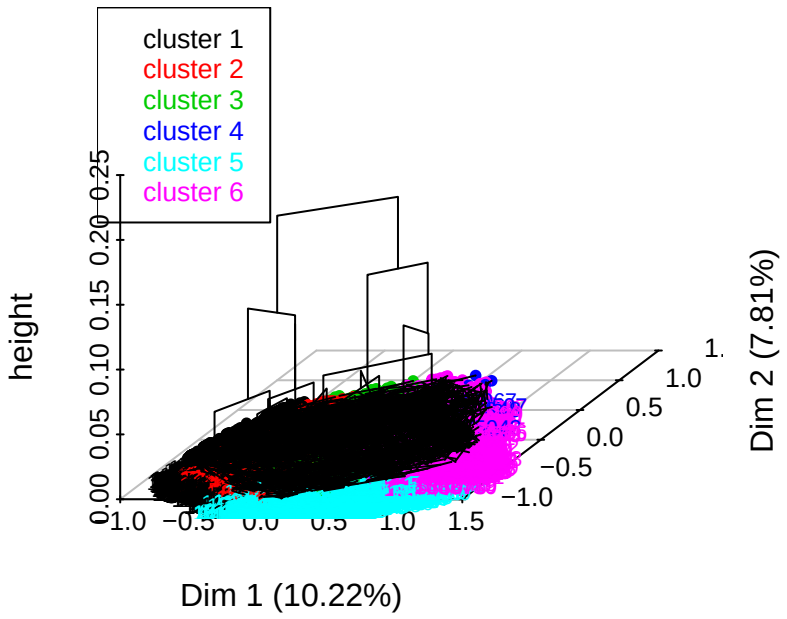
Parangons and class-specific individuals.

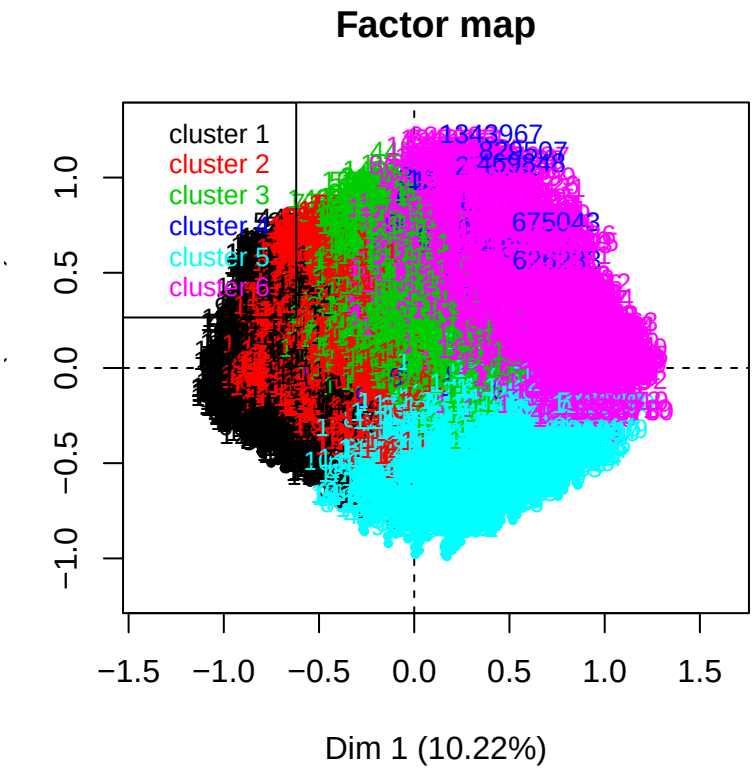
Synthesis through HCPC: Hierarchical Clustering

With the last MCA computation with 17 axes taken into account, we generate the clusters through hierarchical technique.

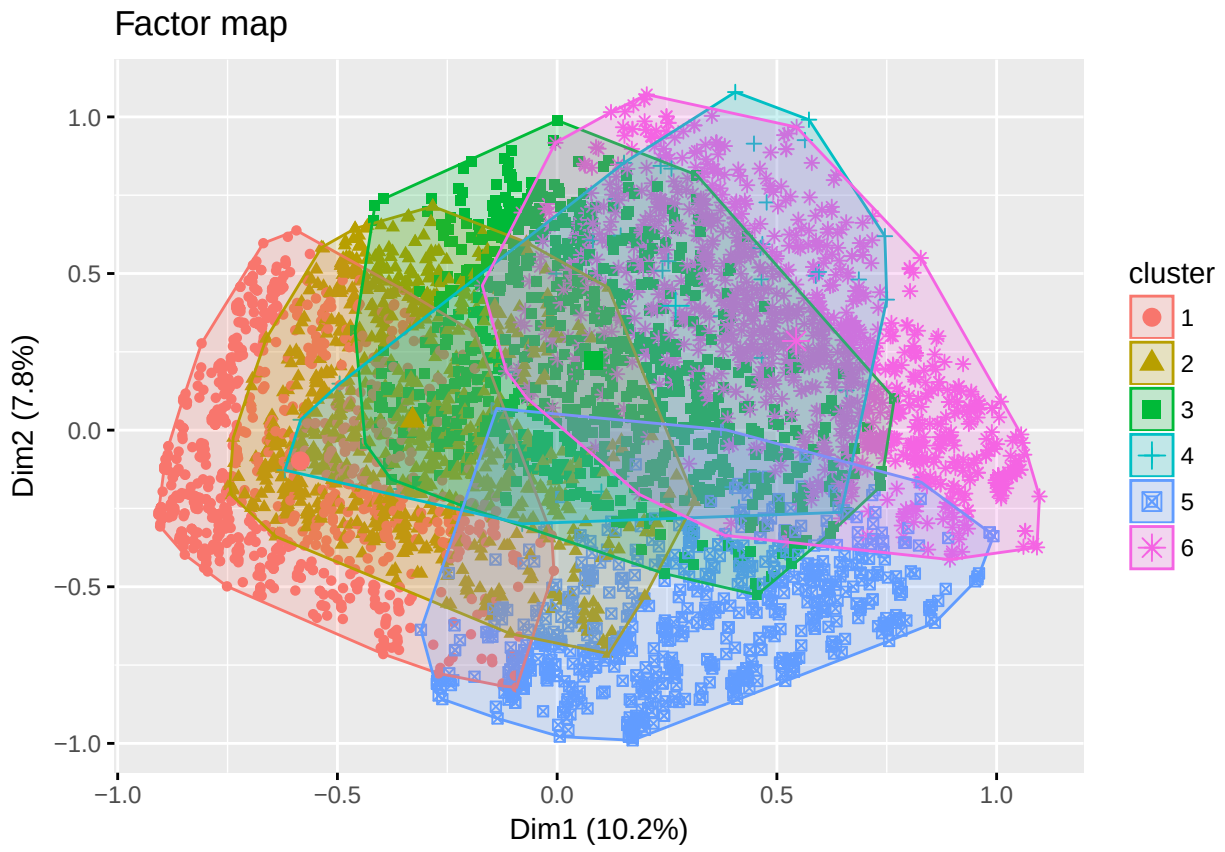
```
res.hcpc<-HCPC(res.mca, nb.clust = 6,order=TRUE)
```


Hierarchical Clustering
Hierarchical Classification
0.00
0.05
0.10
0.15
0.00
0.10
inertia gain


Hierarchical clustering on the factor map
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
height
0.00
0.05
0.10
0.15
0.20
0.25
−1.0
−0.5
0.0
0.5
1.0
1.5
Dim 1 (10.22%)
−1.0
−0.5
0.0
0.5
1.0
Dim 2 (7.81%)




```
fviz_cluster(res.hcpc, geom = "point", main = "Factor map")
```




```
df$MCAhp<-7
df[row.names(res.hcpc$data.clust),"MCAhp"]<-res.hcpc$data.clust$clust
table(df$MCAhp)
```

```
##
##    1    2    3    4    5    6    7
## 1021  981  911   36  761 1114  119
```

Description of clusters

```
res.hcpc$desc.var
```

```
## $test.chi2
##                            p.value df
## Trip_type            0.000000e+00  5
## f.distance           0.000000e+00 15
## f.pickup_longitude   0.000000e+00 15
## f.pickup_latitude    0.000000e+00 15
## f.dropoff_longitude  0.000000e+00 15
## f.dropoff_latitude   0.000000e+00 15
## f.fare_amount        0.000000e+00 15
## f.total              0.000000e+00 15
## f.ttime              0.000000e+00 15
## f.tip_amount         1.163483e-47  5
## f.espeed             2.187891e-34 15
```

```
## AnyTip              7.368541e-32  5
## Payment_type        1.444697e-28 10
## pick_up_period      3.187248e-11 15
##
## $category
## $category$`1`
##                               Cla/Mod      Mod/Cla     Global       p.value
## f.ttime=Time1              77.24649629  91.77277179 25.1451078  0.000000e+00
## f.total=CheapestTrip       70.53642914  86.28795299 25.8913765  0.000000e+00
## f.fare_amount=FAmount1     81.44000000  99.70617042 25.9121061  0.000000e+00
## f.distance=Dist1           72.51028807  86.28795299 25.1865672  0.000000e+00
## f.dropoff_latitude=d.X4    34.80956599  38.49167483 23.4038143  7.193155e-35
## f.tip_amount=smallTip      26.02602603  76.39569050 62.1268657  1.875968e-27
## f.dropoff_longitude=d.Y3   31.07202680  36.33692458 24.7512438  7.135812e-21
## Payment_type=Cash          26.34854772  62.19392752 49.9585406  9.910637e-19
## AnyTip=AnyTip No           25.39398281  69.44172380 57.8772803  1.394752e-17
## f.pickup_latitude=p.X3     29.38496583  37.90401567 27.3009950  6.191670e-17
## f.dropoff_longitude=d.Y2   28.21997106  38.19784525 28.6484245  9.406004e-14
## f.pickup_latitude=p.X4     28.19843342  31.73359452 23.8184080  6.558617e-11
## f.pickup_longitude=p.Y2    27.94117647  31.63565132 23.9635158  2.581913e-10
## f.pickup_longitude=p.Y3    27.40270056  33.79040157 26.0986733  6.480060e-10
## Trip_type=Street-hail      21.32414369 100.00000000 99.2537313  1.847450e-04
## f.pickup_longitude=p.Y4    24.85256950  28.89324192 24.6061360  4.054324e-04
## pick_up_period=night       24.61340206  18.70714985 16.0862355  1.132902e-02
## f.dropoff_latitude=d.X2    23.41062079  30.65621939 27.7155887  1.886825e-02
## pick_up_period=valley      19.18819188  25.46523017 28.0887231  3.476492e-02
## Trip_type=Dispatch          0.00000000   0.00000000  0.7462687  1.847450e-04
## AnyTip=AnyTip Yes          15.35433071  30.55827620 42.1227197  1.394752e-17
## Payment_type=Credit card   16.00840336  37.31635651 49.3366501  3.701997e-18
## f.total=CheapTrip          11.66380789  13.32027424 24.1708126  1.109779e-21
## f.distance=Dist2           11.51315789  13.71204701 25.2072968  1.634311e-23
## f.tip_amount=highTip       13.19102354  23.60430950 37.8731343  1.875968e-27
## f.ttime=Time2               6.72199170   7.93339863 24.9792703  5.565559e-54
## f.dropoff_latitude=d.X1     6.52515723   8.12928501 26.3681592  1.940118e-59
## f.pickup_latitude=p.X1      5.65573770   6.75808031 25.2902156  2.523494e-64
## f.pickup_longitude=p.Y1     4.74631751   5.68070519 25.3316750  1.472556e-73
## f.dropoff_longitude=d.Y1    0.98126673   1.07737512 23.2379768 1.018710e-113
## f.fare_amount=FAmount4      0.08695652   0.09794319 23.8391376 1.891320e-136
## f.fare_amount=FAmount3      0.08467401   0.09794319 24.4817579 7.737359e-141
## f.ttime=Time4               0.16666667   0.19588639 24.8756219 3.543207e-141
## f.total=ExpensiveTrip       0.24630542   0.29382958 25.2487562 1.510242e-141
## f.total=MediumTrip          0.08396306   0.09794319 24.6890547 2.907585e-142
## f.ttime=Time3               0.08291874   0.09794319 25.0000000 2.075531e-144
## f.distance=Dist3            0.00000000   0.00000000 24.6683250 8.652017e-145
## f.distance=Dist4            0.00000000   0.00000000 24.9378109 1.178347e-146
## f.fare_amount=FAmount2      0.08045052   0.09794319 25.7669983 9.480759e-150
##                                v.test
## f.ttime=Time1                     Inf
## f.total=CheapestTrip              Inf
## f.fare_amount=FAmount1            Inf
## f.distance=Dist1                  Inf
## f.dropoff_latitude=d.X4    12.318601
## f.tip_amount=smallTip      10.855587
## f.dropoff_longitude=d.Y3    9.371723
```

```
## Payment_type=Cash          8.836113
## AnyTip=AnyTip No           8.535562
## f.pickup_latitude=p.X3     8.361517
## f.dropoff_longitude=d.Y2   7.448986
## f.pickup_latitude=p.X4     6.530419
## f.pickup_longitude=p.Y2    6.322005
## f.pickup_longitude=p.Y3    6.178288
## Trip_type=Street-hail      3.739012
## f.pickup_longitude=p.Y4    3.536522
## pick_up_period=night       2.532383
## f.dropoff_latitude=d.X2    2.348124
## pick_up_period=valley     -2.111086
## Trip_type=Dispatch        -3.739012
## AnyTip=AnyTip Yes         -8.535562
## Payment_type=Credit card  -8.687593
## f.total=CheapTrip         -9.566134
## f.distance=Dist2          -9.993075
## f.tip_amount=highTip     -10.855587
## f.ttime=Time2            -15.469596
## f.dropoff_latitude=d.X1  -16.258635
## f.pickup_latitude=p.X1   -16.934062
## f.pickup_longitude=p.Y1  -18.142472
## f.dropoff_longitude=d.Y1 -22.663851
## f.fare_amount=FAmount4   -24.862564
## f.fare_amount=FAmount3   -25.265061
## f.ttime=Time4            -25.295908
## f.total=ExpensiveTrip    -25.329545
## f.total=MediumTrip       -25.394405
## f.ttime=Time3            -25.587989
## f.distance=Dist3         -25.622111
## f.distance=Dist4         -25.788993
## f.fare_amount=FAmount2   -26.063411
##
## $category$`2`
##                              Cla/Mod     Mod/Cla     Global       p.value
## f.ttime=Time2            64.06639004  78.6952090 24.9792703  0.000000e+00
## f.fare_amount=FAmount2   76.83024940  97.3496432 25.7669983  0.000000e+00
## f.distance=Dist2         62.08881579  76.9622834 25.2072968  0.000000e+00
## f.total=CheapTrip        59.60548885  70.8460754 24.1708126 1.016952e-282
## f.dropoff_latitude=d.X4  32.59521701  37.5127421 23.4038143  2.493138e-29
## f.pickup_latitude=p.X3   27.86636295  37.4108053 27.3009950  7.412860e-15
## f.pickup_latitude=p.X4   28.28546562  33.1294597 23.8184080  8.594903e-14
## f.dropoff_longitude=d.Y3 27.13567839  33.0275229 24.7512438  4.953286e-11
## f.pickup_longitude=p.Y4  26.79022746  32.4159021 24.6061360  4.973478e-10
## f.pickup_longitude=p.Y3  25.49642573  32.7217125 26.0986733  2.009652e-07
## f.dropoff_longitude=d.Y2 24.96382055  35.1681957 28.6484245  6.177232e-07
## f.tip_amount=smallTip    22.52252252  68.8073394 62.1268657  1.078292e-06
## f.dropoff_longitude=d.Y4 25.11091393  28.8481142 23.3623549  7.870050e-06
## AnyTip=AnyTip No         22.38538682  63.7104995 57.8772803  3.108231e-05
## Payment_type=Cash        22.73858921  55.8613660 49.9585406  3.417051e-05
## f.espeed=Speed2          24.41760138  28.8481142 24.0257048  9.567659e-05
## f.pickup_longitude=p.Y2  24.22145329  28.5423038 23.9635158  2.071880e-04
## Trip_type=Street-hail    20.48872180 100.0000000 99.2537313  2.697194e-04
## f.espeed=Speed1          23.90939597  29.0519878 24.7097844  4.887177e-04
```

```
## Trip_type=Dispatch        0.00000000   0.0000000  0.7462687  2.697194e-04
## Payment_type=Credit card 17.89915966  43.4250765 49.3366501  3.297663e-05
## AnyTip=AnyTip Yes        17.51968504  36.2895005 42.1227197  3.108231e-05
## f.tip_amount=highTip     16.74876847  31.1926606 37.8731343  1.078292e-06
## f.espeed=Speed4          14.95468278  20.1834862 27.4461028  5.184878e-09
## f.total=CheapestTrip     13.29063251  16.9215087 25.8913765  1.110453e-13
## f.ttime=Time3            11.35986733  13.9653415 25.0000000  7.417449e-21
## f.total=MediumTrip        9.99160369  12.1304791 24.6890547  2.795774e-27
## f.distance=Dist3          9.57983193  11.6207951 24.6683250  1.633952e-29
## f.distance=Dist1          9.13580247  11.3149847 25.1865672  6.776588e-33
## f.dropoff_latitude=d.X1   6.76100629   8.7665647 26.3681592  1.976501e-52
## f.ttime=Time1             5.35861500   6.6258919 25.1451078  1.122377e-61
## f.pickup_longitude=p.Y1   5.07364975   6.3200815 25.3316750  5.427107e-65
## f.pickup_latitude=p.X1    4.83606557   6.0142712 25.2902156  3.117872e-67
## f.dropoff_longitude=d.Y1  2.58697591   2.9561672 23.2379768  1.080668e-84
## f.fare_amount=FAmount1    1.68000000   2.1406728 25.9121061 2.709405e-110
## f.ttime=Time4             0.58333333   0.7135576 24.8756219 7.533018e-125
## f.fare_amount=FAmount3    0.33869602   0.4077472 24.4817579 6.581358e-128
## f.fare_amount=FAmount4    0.08695652   0.1019368 23.8391376 2.939956e-130
## f.distance=Dist4          0.08312552   0.1019368 24.9378109 1.934397e-137
## f.total=ExpensiveTrip     0.08210181   0.1019368 25.2487562 1.703346e-139
##                              v.test
## f.ttime=Time2                   Inf
## f.fare_amount=FAmount2          Inf
## f.distance=Dist2                Inf
## f.total=CheapTrip         35.930601
## f.dropoff_latitude=d.X4   11.243495
## f.pickup_latitude=p.X3     7.777232
## f.pickup_latitude=p.X4     7.460876
## f.dropoff_longitude=d.Y3   6.572333
## f.pickup_longitude=p.Y4    6.219939
## f.pickup_longitude=p.Y3    5.198443
## f.dropoff_longitude=d.Y2   4.985592
## f.tip_amount=smallTip      4.876784
## f.dropoff_longitude=d.Y4   4.468689
## AnyTip=AnyTip No           4.165388
## Payment_type=Cash          4.143728
## f.espeed=Speed2            3.901302
## f.pickup_longitude=p.Y2    3.710087
## Trip_type=Street-hail      3.642770
## f.espeed=Speed1            3.486865
## Trip_type=Dispatch        -3.642770
## Payment_type=Credit card  -4.151872
## AnyTip=AnyTip Yes         -4.165388
## f.tip_amount=highTip      -4.876784
## f.espeed=Speed4           -5.841127
## f.total=CheapestTrip      -7.427051
## f.ttime=Time3             -9.367637
## f.total=MediumTrip       -10.819077
## f.distance=Dist3         -11.280724
## f.distance=Dist1         -11.946441
## f.dropoff_latitude=d.X1  -15.238061
## f.ttime=Time1            -16.571371
## f.pickup_longitude=p.Y1  -17.024263
```

```
## f.pickup_latitude=p.X1   -17.323676
## f.dropoff_longitude=d.Y1 -19.500806
## f.fare_amount=FAmount1   -22.313881
## f.ttime=Time4            -23.765859
## f.fare_amount=FAmount3   -24.059867
## f.fare_amount=FAmount4   -24.283344
## f.distance=Dist4         -24.953955
## f.total=ExpensiveTrip    -25.142586
##
## $category$`3`
##                              Cla/Mod     Mod/Cla     Global       p.value
## f.total=MediumTrip       59.36188077  77.6070252 24.6890547  0.000000e+00
## f.fare_amount=FAmount3   75.02116850  97.2557629 24.4817579  0.000000e+00
## f.distance=Dist3         61.34453782  80.1317234 24.6683250  0.000000e+00
## f.ttime=Time3            53.89718076  71.3501647 25.0000000 1.395886e-248
## f.dropoff_latitude=d.X3  26.61141805  31.7233809 22.5124378  7.721087e-13
## f.pickup_longitude=p.Y4  25.94776748  33.8090011 24.6061360  3.171329e-12
## f.dropoff_longitude=d.Y4 25.64330080  31.7233809 23.3623549  1.148799e-10
## f.pickup_latitude=p.X3   24.29764617  35.1262349 27.3009950  7.858908e-09
## f.pickup_latitude=p.X2   24.78031634  30.9549945 23.5903814  1.374181e-08
## Trip_type=Street-hail    19.02673350 100.0000000 99.2537313  5.181009e-04
## f.pickup_longitude=p.Y3  21.28673550  29.4182217 26.0986733  1.206818e-02
## f.pickup_latitude=p.X4   21.14882507  26.6739846 23.8184080  2.595573e-02
## f.espeed=Speed4          20.92145015  30.4061471 27.4461028  2.731642e-02
## f.dropoff_latitude=d.X2  20.86761406  30.6256861 27.7155887  3.048080e-02
## f.dropoff_longitude=d.Y3 21.02177554  27.5521405 24.7512438  3.096445e-02
## f.espeed=Speed3          20.88772846  26.3446762 23.8184080  4.852664e-02
## Trip_type=Dispatch        0.00000000   0.0000000  0.7462687  5.181009e-04
## f.espeed=Speed1          15.01677852  19.6487377 24.7097844  6.285743e-05
## f.ttime=Time4            11.58333333  15.2579583 24.8756219  8.551161e-15
## f.dropoff_longitude=d.Y1 10.43710972  12.8430296 23.2379768  4.267936e-18
## f.dropoff_latitude=d.X1   9.74842767  13.6114160 26.3681592  1.537775e-24
## f.distance=Dist4          9.39318371  12.4039517 24.9378109  8.892606e-25
## f.ttime=Time2             9.37759336  12.4039517 24.9792703  6.567686e-25
## f.total=CheapTrip         8.91938250  11.4160263 24.1708126  2.342093e-26
## f.total=ExpensiveTrip     8.12807882  10.8671789 25.2487562  1.624624e-32
## f.pickup_longitude=p.Y1   8.01963993  10.7574094 25.3316750  2.452847e-33
## f.pickup_latitude=p.X1    5.40983607   7.2447859 25.2902156  4.552233e-53
## f.distance=Dist2          5.01644737   6.6959385 25.2072968  1.977632e-56
## f.fare_amount=FAmount2    1.68946098   2.3051592 25.7669983  1.440336e-98
## f.ttime=Time1             0.74196208   0.9879254 25.1451078 3.917679e-112
## f.fare_amount=FAmount4    0.26086957   0.3293085 23.8391376 3.149542e-115
## f.distance=Dist1          0.57613169   0.7683864 25.1865672 9.815453e-116
## f.total=CheapestTrip      0.08006405   0.1097695 25.8913765 5.293697e-132
## f.fare_amount=FAmount1    0.08000000   0.1097695 25.9121061 3.949348e-132
##                              v.test
## f.total=MediumTrip              Inf
## f.fare_amount=FAmount3          Inf
## f.distance=Dist3                Inf
## f.ttime=Time3             33.673853
## f.dropoff_latitude=d.X3    7.166018
## f.pickup_longitude=p.Y4    6.969915
## f.dropoff_longitude=d.Y4   6.445948
## f.pickup_latitude=p.X3     5.771457
```

```
## f.pickup_latitude=p.X2     5.676575
## Trip_type=Street-hail      3.471219
## f.pickup_longitude=p.Y3    2.510144
## f.pickup_latitude=p.X4     2.226873
## f.espeed=Speed4            2.206966
## f.dropoff_latitude=d.X2    2.163786
## f.dropoff_longitude=d.Y3   2.157529
## f.espeed=Speed3            1.972727
## Trip_type=Dispatch        -3.471219
## f.espeed=Speed1           -4.001819
## f.ttime=Time4             -7.759133
## f.dropoff_longitude=d.Y1  -8.671412
## f.dropoff_latitude=d.X1  -10.224653
## f.distance=Dist4         -10.277587
## f.ttime=Time2            -10.306762
## f.total=CheapTrip        -10.622498
## f.total=ExpensiveTrip    -11.873531
## f.pickup_longitude=p.Y1  -12.030629
## f.pickup_latitude=p.X1   -15.333710
## f.distance=Dist2         -15.828538
## f.fare_amount=FAmount2   -21.071900
## f.ttime=Time1            -22.502562
## f.fare_amount=FAmount4   -22.816435
## f.distance=Dist1         -22.867379
## f.total=CheapestTrip     -24.447933
## f.fare_amount=FAmount1   -24.459893
##
## $category$`4`
##                             Cla/Mod    Mod/Cla     Global      p.value
## Trip_type=Dispatch      100.0000000 100.000000  0.7462687 1.058594e-91
## f.fare_amount=FAmount4    2.1739130  69.444444 23.8391376 9.072251e-09
## f.total=ExpensiveTrip     1.8062397  61.111111 25.2487562 6.251877e-06
## f.distance=Dist4          1.6625104  55.555556 24.9378109 9.899774e-05
## AnyTip=AnyTip No          1.0744986  83.333333 57.8772803 1.311850e-03
## f.tip_amount=smallTip     1.0343677  86.111111 62.1268657 1.808003e-03
## f.ttime=Time4             1.3333333  44.444444 24.8756219 1.085058e-02
## f.pickup_longitude=p.Y4   1.2636900  41.666667 24.6061360 2.489373e-02
## f.espeed=Speed3           1.2184508  38.888889 23.8184080 4.426811e-02
## f.total=CheapestTrip      0.3202562  11.111111 25.8913765 3.387680e-02
## f.espeed=Speed2           0.2588438   8.333333 24.0257048 1.867676e-02
## f.total=CheapTrip         0.2572899   8.333333 24.1708126 1.778362e-02
## f.ttime=Time1             0.2473207   8.333333 25.1451078 1.273495e-02
## f.pickup_longitude=p.Y1   0.1636661   5.555556 25.3316750 2.605130e-03
## f.tip_amount=highTip      0.2736727  13.888889 37.8731343 1.808003e-03
## AnyTip=AnyTip Yes         0.2952756  16.666667 42.1227197 1.311850e-03
## f.fare_amount=FAmount1    0.0000000   0.000000 25.9121061 1.954110e-05
## Trip_type=Street-hail     0.0000000   0.000000 99.2537313 1.058594e-91
##                             v.test
## Trip_type=Dispatch       20.309555
## f.fare_amount=FAmount4    5.747220
## f.total=ExpensiveTrip     4.517687
## f.distance=Dist4          3.893035
## AnyTip=AnyTip No          3.213375
## f.tip_amount=smallTip     3.120083
```

```
## f.ttime=Time4             2.547474
## f.pickup_longitude=p.Y4   2.243048
## f.espeed=Speed3           2.011543
## f.total=CheapestTrip     -2.121535
## f.espeed=Speed2          -2.351921
## f.total=CheapTrip        -2.370093
## f.ttime=Time1            -2.491097
## f.pickup_longitude=p.Y1  -3.010855
## f.tip_amount=highTip     -3.120083
## AnyTip=AnyTip Yes        -3.213375
## f.fare_amount=FAmount1   -4.270071
## Trip_type=Street-hail   -20.309555
##
## $category$`5`
##                                          Cla/Mod     Mod/Cla
## f.dropoff_latitude=d.X1                55.4245283  92.6412615
## f.pickup_latitude=p.X1                 57.7049180  92.5098555
## f.pickup_longitude=p.Y1                56.5466448  90.8015769
## f.dropoff_longitude=d.Y1               45.5842997  67.1484888
## f.total=MediumTrip                     23.0058774  36.0052562
## f.fare_amount=FAmount3                 22.6926334  35.2168200
## f.distance=Dist3                       22.3529412  34.9540079
## f.fare_amount=FAmount2                 20.9171360  34.1655716
## f.ttime=Time3                          20.8126036  32.9829172
## f.distance=Dist2                       20.7236842  33.1143233
## f.tip_amount=highTip                   18.9381500  45.4664915
## AnyTip=AnyTip Yes                      18.6515748  49.8028909
## pick_up_period=afternoon               18.6666667  42.3127464
## f.total=CheapTrip                      19.5540309  29.9605782
## f.espeed=Speed1                        18.8758389  29.5663601
## Payment_type=Credit card               17.5210084  54.7963206
## f.ttime=Time2                          18.7551867  29.6977661
## Trip_type=Street-hail                  15.8939014 100.0000000
## pick_up_period=valley                  18.3025830  32.5886991
## VendorID=Creative Mobile Technologies, LLC 18.5934489  25.3613666
## f.espeed=Speed3                        17.7545692  26.8068331
## pick_up_period=morning                 13.2231405  16.8199737
## VendorID=VeriFone Inc.                 15.0026413  74.6386334
## Trip_type=Dispatch                      0.0000000   0.0000000
## Payment_type=Cash                      14.1078838  44.6780552
## f.dropoff_longitude=d.Y2               13.0969609  23.7844941
## AnyTip=AnyTip No                       13.6819484  50.1971091
## f.tip_amount=smallTip                  13.8471805  54.5335085
## pick_up_period=night                    8.1185567   8.2785808
## f.espeed=Speed4                         9.8187311  17.0827858
## f.ttime=Time4                           7.1666667  11.3009198
## f.dropoff_longitude=d.Y3                5.6113903   8.8042050
## f.pickup_latitude=p.X2                  4.8330404   7.2273325
## f.total=ExpensiveTrip                   5.1724138   8.2785808
## f.pickup_longitude=p.Y2                 4.1522491   6.3074901
## f.dropoff_latitude=d.X2                 4.0388930   7.0959264
## f.distance=Dist4                        2.7431421   4.3363995
## f.fare_amount=FAmount4                  2.1739130   3.2851511
## f.pickup_longitude=p.Y3                 1.5885624   2.6281209
```

```
## f.dropoff_longitude=d.Y4                    0.1774623   0.2628121
## f.dropoff_latitude=d.X4                     0.1771479   0.2628121
## f.dropoff_latitude=d.X3                     0.0000000   0.0000000
## f.pickup_latitude=p.X4                      0.1740644   0.2628121
## f.pickup_longitude=p.Y4                     0.1684920   0.2628121
## f.pickup_latitude=p.X3                      0.0000000   0.0000000
##                                                Global       p.value
## f.dropoff_latitude=d.X1                    26.3681592  0.000000e+00
## f.pickup_latitude=p.X1                     25.2902156  0.000000e+00
## f.pickup_longitude=p.Y1                    25.3316750  0.000000e+00
## f.dropoff_longitude=d.Y1                   23.2379768 2.661148e-182
## f.total=MediumTrip                         24.6890547  2.430975e-14
## f.fare_amount=FAmount3                     24.4817579  3.825788e-13
## f.distance=Dist3                           24.6683250  3.643755e-12
## f.fare_amount=FAmount2                     25.7669983  1.783095e-08
## f.ttime=Time3                              25.0000000  6.443236e-08
## f.distance=Dist2                           25.2072968  9.150667e-08
## f.tip_amount=highTip                       37.8731343  3.116705e-06
## AnyTip=AnyTip Yes                          42.1227197  3.321410e-06
## pick_up_period=afternoon                   35.7587065  4.685752e-05
## f.total=CheapTrip                          24.1708126  6.695338e-05
## f.espeed=Speed1                            24.7097844  8.633219e-04
## Payment_type=Credit card                   49.3366501  1.033292e-03
## f.ttime=Time2                              24.9792703  1.248539e-03
## Trip_type=Street-hail                      99.2537313  2.019028e-03
## pick_up_period=valley                      28.0887231  2.931728e-03
## VendorID=Creative Mobile Technologies, LLC 21.5174129  5.615800e-03
## f.espeed=Speed3                            23.8184080  3.671993e-02
## pick_up_period=morning                     20.0663350  1.351902e-02
## VendorID=VeriFone Inc.                     78.4825871  5.615800e-03
## Trip_type=Dispatch                          0.7462687  2.019028e-03
## Payment_type=Cash                          49.9585406  1.501683e-03
## f.dropoff_longitude=d.Y2                   28.6484245  1.054906e-03
## AnyTip=AnyTip No                           57.8772803  3.321410e-06
## f.tip_amount=smallTip                      62.1268657  3.116705e-06
## pick_up_period=night                       16.0862355  9.310057e-12
## f.espeed=Speed4                            27.4461028  4.104734e-13
## f.ttime=Time4                              24.8756219  5.940037e-24
## f.dropoff_longitude=d.Y3                   24.7512438  9.447013e-34
## f.pickup_latitude=p.X2                     23.5903814  1.757231e-37
## f.total=ExpensiveTrip                      25.2487562  6.763147e-38
## f.pickup_longitude=p.Y2                    23.9635158  9.478976e-44
## f.dropoff_latitude=d.X2                    27.7155887  2.606825e-53
## f.distance=Dist4                           24.9378109  3.655377e-60
## f.fare_amount=FAmount4                     23.8391376  1.557871e-63
## f.pickup_longitude=p.Y3                    26.0986733  4.543789e-79
## f.dropoff_longitude=d.Y4                   23.3623549  4.394042e-93
## f.dropoff_latitude=d.X4                    23.4038143  2.784612e-93
## f.dropoff_latitude=d.X3                    22.5124378  1.219210e-93
## f.pickup_latitude=p.X4                     23.8184080  2.863632e-95
## f.pickup_longitude=p.Y4                    24.6061360  4.418634e-99
## f.pickup_latitude=p.X3                     27.3009950 2.640776e-117
##                                               v.test
## f.dropoff_latitude=d.X1                          Inf
```

```
## f.pickup_latitude=p.X1                              Inf
## f.pickup_longitude=p.Y1                             Inf
## f.dropoff_longitude=d.Y1                       28.792521
## f.total=MediumTrip                              7.625496
## f.fare_amount=FAmount3                          7.261586
## f.distance=Dist3                                6.950352
## f.fare_amount=FAmount2                          5.631828
## f.ttime=Time3                                   5.406043
## f.distance=Dist2                                5.342830
## f.tip_amount=highTip                            4.662975
## AnyTip=AnyTip Yes                               4.649872
## pick_up_period=afternoon                        4.070772
## f.total=CheapTrip                               3.986860
## f.espeed=Speed1                                 3.331652
## Payment_type=Credit card                        3.281302
## f.ttime=Time2                                   3.227553
## Trip_type=Street-hail                           3.087419
## pick_up_period=valley                           2.974807
## VendorID=Creative Mobile Technologies, LLC      2.769410
## f.espeed=Speed3                                 2.088864
## pick_up_period=morning                         -2.469796
## VendorID=VeriFone Inc.                         -2.769410
## Trip_type=Dispatch                             -3.087419
## Payment_type=Cash                              -3.174358
## f.dropoff_longitude=d.Y2                       -3.275458
## AnyTip=AnyTip No                               -4.649872
## f.tip_amount=smallTip                          -4.662975
## pick_up_period=night                           -6.816785
## f.espeed=Speed4                                -7.252062
## f.ttime=Time4                                 -10.092881
## f.dropoff_longitude=d.Y3                      -12.109148
## f.pickup_latitude=p.X2                        -12.794614
## f.total=ExpensiveTrip                         -12.868583
## f.pickup_longitude=p.Y2                       -13.871116
## f.dropoff_latitude=d.X2                       -15.369872
## f.distance=Dist4                              -16.360595
## f.fare_amount=FAmount4                        -16.826603
## f.pickup_longitude=p.Y3                       -18.826926
## f.dropoff_longitude=d.Y4                      -20.465252
## f.dropoff_latitude=d.X4                       -20.487476
## f.dropoff_latitude=d.X3                       -20.527654
## f.pickup_latitude=p.X4                        -20.709167
## f.pickup_longitude=p.Y4                       -21.127777
## f.pickup_latitude=p.X3                        -23.024645
##
## $category$`6`
##                                Cla/Mod      Mod/Cla     Global
## f.ttime=Time4              79.16666667  85.27827648 24.8756219
## f.total=ExpensiveTrip      84.56486043  92.45960503 25.2487562
## f.fare_amount=FAmount4     95.21739130  98.29443447 23.8391376
## f.distance=Dist4           86.11803824  92.99820467 24.9378109
## f.dropoff_longitude=d.Y1   39.96431757  40.21543986 23.2379768
## f.tip_amount=highTip       31.74603175  52.06463196 37.8731343
## f.espeed=Speed4            33.91238671  40.30520646 27.4461028
```

```
## Payment_type=Credit card   28.69747899  61.31059246 49.3366501
## AnyTip=AnyTip Yes          29.28149606  53.41113106 42.1227197
## f.dropoff_latitude=d.X2    29.16978310  35.00897666 27.7155887
## f.dropoff_latitude=d.X3    29.74217311  28.99461400 22.5124378
## f.pickup_latitude=p.X2     28.29525483  28.90484740 23.5903814
## Trip_type=Street-hail      23.26649958 100.00000000 99.2537313
## pick_up_period=morning     26.75619835  23.24955117 20.0663350
## f.dropoff_longitude=d.Y4   25.90949423  26.21184919 23.3623549
## f.pickup_longitude=p.Y1    25.45008183  27.91741472 25.3316750
## f.pickup_latitude=p.X1     25.40983607  27.82764811 25.2902156
## f.passenger=multiPassagers 26.02179837  17.14542190 15.2155887
## Payment_type=Other         38.23529412   1.16696589  0.7048093
## f.pickup_longitude=p.Y4    20.97725358  22.35188510 24.6061360
## f.passenger=onePassager    22.56723716  82.85457810 84.7844113
## f.espeed=Speed3            20.62663185  21.27468582 23.8184080
## f.dropoff_latitude=d.X1    20.59748428  23.51885099 26.3681592
## f.espeed=Speed1            19.21140940  20.55655296 24.7097844
## Trip_type=Dispatch          0.00000000   0.00000000  0.7462687
## pick_up_period=afternoon   19.65217391  30.43087971 35.7587065
## f.pickup_latitude=p.X3     17.84358390  21.09515260 27.3009950
## f.espeed=Speed2            17.16997412  17.86355476 24.0257048
## AnyTip=AnyTip No           18.58882521  46.58886894 57.8772803
## f.dropoff_longitude=d.Y2   14.90593343  18.49192101 28.6484245
## f.dropoff_longitude=d.Y3   14.07035176  15.08078995 24.7512438
## Payment_type=Cash          17.34439834  37.52244165 49.9585406
## f.ttime=Time3              12.85240464  13.91382406 25.0000000
## f.dropoff_latitude=d.X4    12.31178034  12.47755835 23.4038143
## f.tip_amount=smallTip      17.81781782  47.93536804 62.1268657
## f.total=MediumTrip          6.96893367   7.45062837 24.6890547
## f.distance=Dist3            6.21848739   6.64272890 24.6683250
## f.fare_amount=FAmount3      1.43945809   1.52603232 24.4817579
## f.ttime=Time2               0.66390041   0.71813285 24.9792703
## f.total=CheapTrip           0.00000000   0.00000000 24.1708126
## f.distance=Dist2            0.24671053   0.26929982 25.2072968
## f.ttime=Time1               0.08244023   0.08976661 25.1451078
## f.distance=Dist1            0.08230453   0.08976661 25.1865672
## f.fare_amount=FAmount1      0.16000000   0.17953321 25.9121061
## f.total=CheapestTrip        0.08006405   0.08976661 25.8913765
## f.fare_amount=FAmount2      0.00000000   0.00000000 25.7669983
##                                  p.value     v.test
## f.ttime=Time4               0.000000e+00        Inf
## f.total=ExpensiveTrip       0.000000e+00        Inf
## f.fare_amount=FAmount4      0.000000e+00        Inf
## f.distance=Dist4            0.000000e+00        Inf
## f.dropoff_longitude=d.Y1    8.844202e-49  14.678565
## f.tip_amount=highTip        3.485324e-28  11.008296
## f.espeed=Speed4             1.199800e-26  10.684744
## Payment_type=Credit card    6.382490e-20   9.137649
## AnyTip=AnyTip Yes           4.987263e-18   8.653662
## f.dropoff_latitude=d.X2     1.054323e-09   6.100962
## f.dropoff_latitude=d.X3     7.204388e-09   5.786089
## f.pickup_latitude=p.X2      2.759375e-06   4.687963
## Trip_type=Street-hail       7.545199e-05   3.958405
## pick_up_period=morning      2.805509e-03   2.988282
```

```
## f.dropoff_longitude=d.Y4    1.105229e-02   2.541041
## f.pickup_longitude=p.Y1     2.459623e-02   2.247686
## f.pickup_latitude=p.X1      2.730006e-02   2.207200
## f.passenger=multiPassagers  4.285203e-02   2.025149
## Payment_type=Other          4.778214e-02   1.979301
## f.pickup_longitude=p.Y4     4.531693e-02  -2.001701
## f.passenger=onePassager     4.285203e-02  -2.025149
## f.espeed=Speed3             2.215835e-02  -2.287642
## f.dropoff_latitude=d.X1     1.324766e-02  -2.477041
## f.espeed=Speed1             2.060586e-04  -3.711471
## Trip_type=Dispatch          7.545199e-05  -3.958405
## pick_up_period=afternoon    2.014635e-05  -4.263263
## f.pickup_latitude=p.X3      6.800507e-08  -5.396364
## f.espeed=Speed2             1.991903e-08  -5.612703
## AnyTip=AnyTip No            4.987263e-18  -8.653662
## f.dropoff_longitude=d.Y2    1.214253e-18  -8.813383
## f.dropoff_longitude=d.Y3    8.273068e-19  -8.856278
## Payment_type=Cash           2.098396e-21  -9.500025
## f.ttime=Time3               2.240079e-24 -10.188140
## f.dropoff_latitude=d.X4     5.818455e-25 -10.318400
## f.tip_amount=smallTip       3.485324e-28 -11.008296
## f.total=MediumTrip          4.236652e-62 -16.629848
## f.distance=Dist3            9.465360e-69 -17.523593
## f.fare_amount=FAmount3     3.195171e-127 -23.994221
## f.ttime=Time2              2.370468e-145 -25.672516
## f.total=CheapTrip          2.249970e-156 -26.641422
## f.distance=Dist2           6.754472e-157 -26.686487
## f.ttime=Time1              4.139958e-161 -27.047025
## f.distance=Dist1           1.984193e-161 -27.074167
## f.fare_amount=FAmount1     1.333430e-164 -27.342298
## f.total=CheapestTrip       6.828641e-167 -27.534271
## f.fare_amount=FAmount2     1.128428e-168 -27.682687
##
##
## $quanti.var
##                          Eta2       P-value
## Pickup_latitude   0.285870927  0.000000e+00
## Dropoff_latitude  0.290669712  0.000000e+00
## Trip_distance     0.698841376  0.000000e+00
## Fare_amount       0.731526050  0.000000e+00
## MTA_tax           0.899306759  0.000000e+00
## Total_amount      0.700345544  0.000000e+00
## trip_length       0.628154655  0.000000e+00
## trip_distance_km  0.698841376  0.000000e+00
## travel_time       0.529263476  0.000000e+00
## Pickup_longitude  0.219067898 1.847862e-255
## Dropoff_longitude 0.149135427 5.561266e-166
## Tip_amount        0.140830764 7.409396e-156
## espeed            0.047204505  2.377617e-48
## Tolls_amount      0.023731395  2.488938e-23
## pick_up_hour      0.010859403  4.019144e-10
## Extra             0.008369962  1.179017e-07
##
## $quanti
```

```
## $quanti$`1`
##                       v.test Mean in category Overall mean sd in category
## Dropoff_latitude   14.262158       40.7660736  40.74404515     0.04762073
## Pickup_latitude    13.122650       40.7661165  40.74596616     0.04776969
## Pickup_longitude    8.288745      -73.9274397 -73.93683412     0.03393086
## Dropoff_longitude   7.018313      -73.9275978 -73.93669909     0.03467279
## MTA_tax             2.901700        0.4995103   0.49585406     0.01564027
## espeed             -2.129339       21.4812550  22.08472780     9.66780744
## Tolls_amount       -4.156367        0.0000000   0.07932421     0.00000000
## Tip_amount        -13.423594        0.4574143   1.15217454     1.05888118
## trip_length       -29.235982        1.4071444   4.30691367     0.57428450
## trip_distance_km  -29.472500        1.1917542   4.11393908     0.42033837
## Trip_distance     -29.472500        0.7405218   2.55628323     0.26118615
## travel_time       -29.535055        4.3860198  12.13813362     6.30155037
## Total_amount      -31.016529        6.5630362  13.55080846     1.85728504
## Fare_amount       -31.839204        4.9417630  11.17300580     1.47497865
##                     Overall sd       p.value
## Dropoff_latitude   0.05557847  3.766226e-46
## Pickup_latitude    0.05525463  2.442427e-39
## Pickup_longitude   0.04078401  1.144497e-16
## Dropoff_longitude  0.04666352  2.245630e-12
## MTA_tax            0.04534071  3.711439e-03
## espeed            10.19814341  3.322619e-02
## Tolls_amount       0.68675249  3.233486e-05
## Tip_amount         1.86240658  4.398546e-41
## trip_length        3.56906208 6.768762e-188
## trip_distance_km   3.56778807 6.483325e-191
## Trip_distance      2.21692072 6.483325e-191
## travel_time        9.44475500 1.021732e-191
## Total_amount       8.10688270 3.227093e-211
## Fare_amount        7.04240211 1.857021e-222
##
## $quanti$`2`
##                       v.test Mean in category Overall mean sd in category
## Dropoff_latitude   14.448878       40.7669318  40.74404515     0.04753400
## Pickup_latitude    13.787337       40.7676777  40.74596616     0.04732869
## Pickup_longitude   10.233289      -73.9249396 -73.93683412     0.03593739
## Dropoff_longitude   9.038661      -73.9246786 -73.93669909     0.03805218
## MTA_tax             3.208428        0.5000000   0.49585406     0.00000000
## Tolls_amount       -3.475807        0.0112946   0.07932421     0.24988898
## espeed             -7.386894       19.9377624  22.08472780     7.72225077
## Tip_amount         -9.702815        0.6371662   1.15217454     0.92621990
## travel_time       -14.174582        8.3227078  12.13813362     6.30667243
## trip_length       -16.902775        2.5876037   4.30691367     0.84010836
## Fare_amount       -17.934643        7.5733945  11.17300580     0.95740879
## trip_distance_km  -18.018190        2.2818261   4.11393908     0.54803121
## Trip_distance     -18.018190        1.4178610   2.55628323     0.34053081
## Total_amount      -18.083459        9.3727217  13.55080846     1.41628199
##                     Overall sd      p.value
## Dropoff_latitude   0.05557847 2.548038e-47
## Pickup_latitude    0.05525463 3.037688e-43
## Pickup_longitude   0.04078401 1.406589e-24
## Dropoff_longitude  0.04666352 1.586031e-19
## MTA_tax            0.04534071 1.334625e-03
```

```
## Tolls_amount       0.68675249 5.093178e-04
## espeed            10.19814341 1.502980e-13
## Tip_amount         1.86240658 2.932932e-22
## travel_time        9.44475500 1.316248e-45
## trip_length        3.56906208 4.292203e-64
## Fare_amount        7.04240211 6.326897e-72
## trip_distance_km   3.56778807 1.402589e-72
## Trip_distance      2.21692072 1.402589e-72
## Total_amount       8.10688270 4.302271e-73
##
## $quanti$`3`
##                     v.test Mean in category Overall mean sd in category
## Pickup_longitude  10.750641      -73.9237495  -73.9368341     0.04092909
## Dropoff_longitude  9.413416      -73.9235903  -73.9366991     0.04763743
## Pickup_latitude    8.359555       40.7597506   40.7459662     0.04789780
## Dropoff_latitude   7.804163       40.7569892   40.7440451     0.04830627
## MTA_tax            2.658434        0.4994512    0.4958541     0.01655664
## travel_time        2.176414       12.7515697   12.1381336     2.94762140
## trip_length        2.161218        4.5371054    4.3069137     1.41927100
## espeed             2.112101       22.7275234   22.0847278     9.59746796
##                    Overall sd      p.value
## Pickup_longitude   0.04078401 5.885114e-27
## Dropoff_longitude  0.04666352 4.802695e-21
## Pickup_latitude    0.05525463 6.295549e-17
## Dropoff_latitude   0.05557847 5.989792e-15
## MTA_tax            0.04534071 7.850476e-03
## travel_time        9.44475500 2.952432e-02
## trip_length        3.56906208 3.067850e-02
## espeed            10.19814341 3.467779e-02
##
## $quanti$`4`
##                       v.test Mean in category Overall mean sd in category
## Fare_amount         7.305602        19.716667   11.1730058    10.26272922
## Total_amount        5.122326        20.446667   13.5508085    10.90353154
## trip_length         3.707645         6.504366    4.3069137     4.69531078
## Trip_distance       3.294198         3.769021    2.5562832     2.55203635
## trip_distance_km    3.294198         6.065652    4.1139391     4.10710439
## travel_time         3.269089        17.265386   12.1381336    10.68095735
## Pickup_longitude    2.780605       -73.918002  -73.9368341     0.04002803
## Dropoff_longitude   2.046625       -73.920840  -73.9366991     0.05172427
## Extra              -5.858423         0.000000    0.3527156     0.00000000
## MTA_tax           -65.856463         0.000000    0.4958541     0.00000000
##                    Overall sd      p.value
## Fare_amount        7.04240211 2.760287e-13
## Total_amount       8.10688270 3.017890e-07
## trip_length        3.56906208 2.091953e-04
## Trip_distance      2.21692072 9.870293e-04
## trip_distance_km   3.56778807 9.870293e-04
## travel_time        9.44475500 1.078943e-03
## Pickup_longitude   0.04078401 5.425779e-03
## Dropoff_longitude  0.04666352 4.069493e-02
## Extra              0.36255737 4.672833e-09
## MTA_tax            0.04534071 0.000000e+00
##
```

```
## $quanti$`5`
##                       v.test Mean in category Overall mean sd in category
## pick_up_hour        6.899793      15.05913272  13.49689055     5.99806572
## MTA_tax             2.748286       0.50000000   0.49585406     0.00000000
## Tolls_amount       -2.834417       0.01455979   0.07932421     0.28363577
## espeed             -7.153894      19.65735728  22.08472780     7.90003927
## travel_time        -7.746569       9.70384226  12.13813362     4.92319364
## Total_amount       -9.667956      10.94308804  13.55080846     4.23763258
## Fare_amount       -10.872343       8.62549277  11.17300580     3.55513462
## trip_length       -10.887594       3.01403286   4.30691367     1.70544652
## Trip_distance     -11.691908       1.69388473   2.55628323     0.94067392
## trip_distance_km  -11.691908       2.72604323   4.11393908     1.51386793
## Dropoff_longitude -25.159729     -73.97576122 -73.93669909     0.02174930
## Pickup_longitude  -31.125436     -73.97906964 -73.93683412     0.01593457
## Dropoff_latitude  -34.585794      40.68008981  40.74404515     0.01900779
## Pickup_latitude   -34.900248      40.68180538  40.74596616     0.01556289
##                     Overall sd       p.value
## pick_up_hour       6.80518339  5.207831e-12
## MTA_tax            0.04534071  5.990769e-03
## Tolls_amount       0.68675249  4.590933e-03
## espeed            10.19814341  8.435015e-13
## travel_time        9.44475500  9.440852e-15
## Total_amount       8.10688270  4.125330e-22
## Fare_amount        7.04240211  1.561388e-27
## trip_length        3.56906208  1.320831e-27
## Trip_distance      2.21692072  1.402016e-31
## trip_distance_km   3.56778807  1.402016e-31
## Dropoff_longitude  0.04666352 1.105986e-139
## Pickup_longitude   0.04078401 1.090705e-212
## Dropoff_latitude   0.05557847 4.131595e-262
## Pickup_latitude    0.05525463 7.370235e-267
##
## $quanti$`6`
##                       v.test Mean in category Overall mean sd in category
## Fare_amount        54.915317       21.3354937  11.17300580     6.28226824
## Trip_distance      54.697688        5.7427168   2.55628323     2.25987539
## trip_distance_km   54.697688        9.2420068   4.11393908     3.63691690
## Total_amount       54.134845       25.0831239  13.55080846     7.62422748
## trip_length        51.133968        9.1025841   4.30691367     3.90395764
## travel_time        46.176856       23.5985603  12.13813362     9.11412762
## Tip_amount         24.395809        2.3460952   1.15217454     2.67402505
## espeed             13.150365       25.6087988  22.08472780    12.96228336
## Tolls_amount       10.434654        0.2676302   0.07932421     1.22454709
## MTA_tax             2.726329        0.4991023   0.49585406     0.02116665
## Dropoff_longitude  -2.834038      -73.9401742 -73.93669909     0.05814368
## Pickup_latitude    -3.404320       40.7410232  40.74596616     0.05671848
## pick_up_hour       -3.556733       12.8608618  13.49689055     6.80106085
## Dropoff_latitude   -4.760720       40.7370923  40.74404515     0.05615605
##                     Overall sd       p.value
## Fare_amount        7.04240211  0.000000e+00
## Trip_distance      2.21692072  0.000000e+00
## trip_distance_km   3.56778807  0.000000e+00
## Total_amount       8.10688270  0.000000e+00
## trip_length        3.56906208  0.000000e+00
```

```
## travel_time        9.44475500  0.000000e+00
## Tip_amount         1.86240658 1.894633e-131
## espeed            10.19814341  1.693546e-39
## Tolls_amount       0.68675249  1.722393e-25
## MTA_tax            0.04534071  6.404317e-03
## Dropoff_longitude  0.04666352  4.596392e-03
## Pickup_latitude    0.05525463  6.632898e-04
## pick_up_hour       6.80518339  3.754961e-04
## Dropoff_latitude   0.05557847  1.929034e-06
##
##
## attr(,"class")
## [1] "catdes" "list "
```

Cluster 1

Category

86% of its individuals have the CheapestTrip category for f.total. Dist1, FAmount1 and Time1 for their respective variables (f.distance, f.fare_amount and f.ttime) are also predominant (more than 85% in Mod/Cla for each one). This means that Cluster 1 is defined by the cheapest and shortest trips with a small tips. #### Quanti As a prove of the conclusions made before, we can see how the most remarkable quantitatives variables are trip_length, Fare_amount, Total_amount and travel_time. The mean of all of this variables inside the cluster is clearly above the overall mean in the dataset.

Cluster 2

Category

We can see that here we still have a predominance of cheap trips (with a 71% of individuals of cluster 2 have cheaptrip) but now f.distance, f.fare_amount and f.ttime get their second level (category) of values. So we got cheap trips with a bigger distances than in cluster 1 here. #### Quanti The same as the case before. The numerical variables agreed with the conclusions made. The most significance variables are still trip_length, Fare_amount, Total_amount and travel_time and their mean is a bit higher than the cluster1 but still above the overall mean.

Cluster 3

Category

In cluster 3 seems to be still splitting by its cost and lentgh of the journey, now with middle trips which contains medium costs (77% of the rows) and the 3rd level of f.fare_amount, f.distance and time, which also means longer trips than before. #### Quanti Travel time and trip_length in this cluster are in the mean values, which is logic for what we just said, but it is curious how the variables with biggest v.test values are actually longitudes and latitudes.

Cluster 4

Category

Clearly defined by one parameter: the type of the trip "Dispatch". 100% of its individuals are of this type and 100% of this category is included in this cluster. #### Quanti MTA_tax is the most remarkable variables here and it doesn't say too much except of the cluster except we only have rows without taxes nor extra's.

Cluster 5

Category

It has been build from the longitutds and latitudes of the first level/category of the pick_up and dropoff variables. #### Quanti As we already pointed out, mostly longitudes and latitudes involved here.

Cluster 6

Category

Here we find the most expensive trips with all the elements that it follows from it: large trips (Dist4) that take a lot of time (Time4) with a big Fare amount (FAmount4). All of this classes have more than 85% of representation inside this cluster. #### Quanti Fare_amount, trip_length, travel_time and Total_amount clearly over the overall mean and with a lot of significance here.

Individuals

Again, this command can help us now to confirm the conclusions made until now.

```
res.hcpc$desc.ind
```

```
## $para
## Cluster: 1
##    411232    445076    461517    187874    842676
## 0.4115936 0.6264737 0.6647595 0.6689452 0.6792066
## ------------------------------------------------------------
## Cluster: 2
##    346461    300660    632552   1294374    186363
## 0.6751316 0.7371433 0.7371433 0.7606227 0.7822721
## ------------------------------------------------------------
## Cluster: 3
##   1210493     80326   1321174    151963    649075
## 0.5356644 0.6296142 0.7167908 0.7208800 0.7412111
## ------------------------------------------------------------
## Cluster: 4
##    424236    278068    984283    794973    825818
## 0.8271495 0.8458822 0.9157033 0.9635802 0.9649285
## ------------------------------------------------------------
## Cluster: 5
##   1267046   1369263    322743    483381    371250
## 0.7245503 0.7287315 0.7375132 0.7422188 0.7461320
## ------------------------------------------------------------
## Cluster: 6
##    307788    803251   1245227    808547   1035060
## 0.5641748 0.5952830 0.5952830 0.6475194 0.6544842
##
## $dist
## Cluster: 1
##  1241879   203359   529632  1282516  1383646
## 3.140528 3.112007 3.112007 3.109630 3.080040
## ------------------------------------------------------------
## Cluster: 2
##   259005   840165   390598   163666   840014
```

```
## 3.145964 3.138056 3.107597 2.997960 2.994313
## ------------------------------------------------------------
## Cluster: 3
##  1371481  1322779   665911   353521    36933
## 3.166014 3.016792 2.947542 2.944879 2.917540
## ------------------------------------------------------------
## Cluster: 4
##   444409   169710   161512   431494   272451
## 3.451944 3.385436 3.368163 3.359398 3.359352
## ------------------------------------------------------------
## Cluster: 5
##  1367682   839935  1123562    73710   980167
## 3.143346 2.930257 2.899628 2.866396 1.646619
## ------------------------------------------------------------
## Cluster: 6
##   844434    18596   301650   504126  1000815
## 3.290353 3.219564 3.176062 3.081069 3.063079
```

Cluster 1

For a parangon of C1 we expect some row with a cheapTrip and short distance. We can check Total_amount (8.3) is below the mean (11.0) as the rest of significant quantitatives variables explained for Cluster1. On the contrary, for the distinguished, we have a Total_amount over the mean (20.34).

```
#parangon C1
df["419422",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 419422 VeriFone Inc.  2016-01-09 19:52:50   2016-01-09 20:00:05
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 419422      Store_and_fwd Standard rate        -73.93687        40.80215
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 419422         -73.94244         40.78634               1          1.57
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 419422         7.5     0     0.5          0            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 419422                   0.3          8.3         Cash Street-hail       0
##           AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 419422 AnyTip No    2.377077          2.52667        7.25           19
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 419422      afternoon 19.67236 onePassager      Dist2               p.Y3
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 419422              p.X4                d.Y3               d.X3
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 419422      FAmount2 smallExtra   highMTA           highSurcharge
##        f.tip_amount    f.toll   f.total f.ttime f.espeed f.outlierPCAd1
## 419422     smallTip smallToll CheapTrip   Time2   Speed2       Normald1
##        f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4 f.outlierPCA claHP
## 419422       Normald2       Normald3       Normald4       Normal     1
##        MCAhp
## 419422     2
```

```
#distinguished C1
df["572868",]
```

```
##                               VendorID lpep_pickup_datetime
## 572868 Creative Mobile Technologies, LLC  2016-01-13 07:30:41
##        Lpep_dropoff_datetime Store_and_fwd_flag    RateCodeID
## 572868   2016-01-13 07:43:33      Store_and_fwd Standard rate
##        Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
## 572868          -73.957        40.74821         -73.97182         40.75252
##        Passenger_count Trip_distance Fare_amount Extra MTA_tax Tip_amount
## 572868               1             3          12     0     0.5          2
##        Tolls_amount improvement_surcharge Total_amount Payment_type
## 572868         5.54                   0.3        20.34  Credit card
##          Trip_type mis_ind     AnyTip trip_length trip_distance_km
## 572868 Street-hail       0 AnyTip Yes    2.127662         4.828032
##        travel_time pick_up_hour pick_up_period   espeed f.passenger
## 572868    12.86667            7        morning 9.921739 onePassager
##        f.distance f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 572868      Dist3               p.Y2              p.X3                d.Y1
##        f.dropoff_latitude f.fare_amount    f.extra f.MTA_tax
## 572868               d.X3      FAmount3 smallExtra   highMTA
##        f.Improvement_surcharge f.tip_amount   f.toll       f.total f.ttime
## 572868           highSurcharge      highTip highToll ExpensiveTrip   Time3
##        f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 572868   Speed1       Normald1       Normald2       Normald3
##        f.outlierPCAd4 f.outlierPCA claHP MCAhp
## 572868       Normald4       Normal     1     3
```

Cluster 2

Cheap trips (total_amount below the mean) with bigger distances than cluster 1 for parangons as we predicted.

```
#parangon
df["746656",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 746656 VeriFone Inc.  2016-01-16 17:14:24   2016-01-16 17:25:34
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 746656      Store_and_fwd Standard rate        -73.88785        40.74715
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 746656         -73.85871         40.73974               1          2.05
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 746656         9.5     0     0.5          0            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 746656                   0.3         10.3         Cash Street-hail       0
##           AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 746656 AnyTip No    4.064021         3.299155    11.16667           17
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 746656         valley 21.83653 onePassager      Dist3               p.Y4
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 746656              p.X3                d.Y4               d.X2
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 746656      FAmount3 smallExtra   highMTA           highSurcharge
##        f.tip_amount    f.toll   f.total f.ttime f.espeed f.outlierPCAd1
## 746656     smallTip smallToll CheapTrip   Time3   Speed3       Normald1
##        f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4 f.outlierPCA claHP
```

```
## 746656       Normald2       Normald3       Normald4       Normal     2
##        MCAhp
## 746656     3
```

```
#distinguished
df["532659",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 532659 VeriFone Inc.  2016-01-12 08:45:02   2016-01-12 08:58:58
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 532659      Store_and_fwd Standard rate        -73.80746        40.70039
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 532659         -73.75623          40.6506               3          5.85
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 532659          19     0     0.5          0            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 532659                   0.3         19.8         Cash Street-hail       0
##           AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 532659 AnyTip No    11.23258         9.414662    13.93333            8
##        pick_up_period   espeed    f.passenger f.distance
## 532659        morning 48.36995 multiPassagers      Dist4
##        f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 532659               p.Y4              p.X2                d.Y4
##        f.dropoff_latitude f.fare_amount    f.extra f.MTA_tax
## 532659               d.X1      FAmount4 smallExtra   highMTA
##        f.Improvement_surcharge f.tip_amount    f.toll       f.total
## 532659           highSurcharge     smallTip smallToll ExpensiveTrip
##        f.ttime f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 532659   Time3   Speed4       Normald1       Normald2       Normald3
##        f.outlierPCAd4 f.outlierPCA claHP MCAhp
## 532659       Normald4       Normal     2     6
```

Cluster 3

Parangon with medium costs -> total_amount in the mean. Distinguished with high espeed and low travel_time

```
#parangon
df["473230",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 473230 VeriFone Inc.  2016-01-10 18:49:26   2016-01-10 19:01:23
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 473230      Store_and_fwd Standard rate          -73.976        40.68373
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 473230         -73.94895         40.68416               1          2.14
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 473230        10.5     0     0.5          0            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 473230                   0.3         11.3  Credit card Street-hail       0
##           AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 473230 AnyTip No    3.054484         3.443996       11.95           18
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 473230      afternoon 15.33632 onePassager      Dist3               p.Y1
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
```

```
## 473230              p.X1                d.Y2               d.X1
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 473230      FAmount3 smallExtra   highMTA           highSurcharge
##        f.tip_amount    f.toll    f.total f.ttime f.espeed f.outlierPCAd1
## 473230     smallTip smallToll MediumTrip   Time3   Speed2       Normald1
##        f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4 f.outlierPCA claHP
## 473230       Normald2       Normald3       Normald4       Normal     4
##        MCAhp
## 473230     5
```

```
#distinguished
df["274842",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 274842 VeriFone Inc.  2016-01-06 20:02:39   2016-01-06 20:09:14
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 274842      Store_and_fwd Standard rate        -74.02743        40.62211
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 274842        -74.01118         40.60867               1          1.46
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 274842           7     1     0.5       1.76            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 274842                   0.3        10.56  Credit card Street-hail       0
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 274842 AnyTip Yes    3.302202         2.349642    6.583333           20
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 274842      afternoon 30.09602 onePassager      Dist2               p.Y1
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 274842              p.X1                d.Y1               d.X1
##        f.fare_amount   f.extra f.MTA_tax f.Improvement_surcharge
## 274842      FAmount2 highExtra   highMTA           highSurcharge
##        f.tip_amount    f.toll   f.total f.ttime f.espeed f.outlierPCAd1
## 274842      highTip smallToll CheapTrip   Time2   Speed4       Normald1
##        f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4 f.outlierPCA claHP
## 274842       Normald2       Normald3       Normald4       Normal     4
##        MCAhp
## 274842     5
```

Cluster 4

```
#parangon
df["473235",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 473235 VeriFone Inc.  2016-01-10 18:50:53   2016-01-10 19:01:54
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 473235      Store_and_fwd Standard rate        -73.90318        40.74598
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 473235        -73.92684         40.76157               5          1.93
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 473235         9.5     0     0.5       2.06            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 473235                   0.3        12.36  Credit card Street-hail       0
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
```

```
## 473235 AnyTip Yes    4.363912         3.106034    11.01667           18
##        pick_up_period   espeed    f.passenger f.distance
## 473235      afternoon 23.76715 multiPassagers      Dist3
##        f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 473235               p.Y4              p.X3               d.Y3
##        f.dropoff_latitude f.fare_amount    f.extra f.MTA_tax
## 473235               d.X3      FAmount3 smallExtra   highMTA
##        f.Improvement_surcharge f.tip_amount    f.toll    f.total f.ttime
## 473235           highSurcharge      highTip smallToll MediumTrip   Time3
##        f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 473235   Speed3       Normald1       Normald2       Normald3
##        f.outlierPCAd4 f.outlierPCA claHP MCAhp
## 473235       Normald4       Normal     3     3
```

```
#distinguished
df["1137082",]
```

```
##              VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 1137082 VeriFone Inc.  2016-01-26 07:44:44   2016-01-26 07:59:34
##         Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 1137082      Store_and_fwd Standard rate        -73.91051        40.76951
##         Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 1137082         -73.93724         40.82439               6          6.13
##         Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 1137082        18.5     0     0.5       4.97         5.54
##         improvement_surcharge Total_amount Payment_type   Trip_type
## 1137082                   0.3        29.81  Credit card Street-hail
##         mis_ind     AnyTip trip_length trip_distance_km travel_time
## 1137082       0 AnyTip Yes    9.074375         9.865279    14.83333
##         pick_up_hour pick_up_period   espeed    f.passenger f.distance
## 1137082            7        morning 36.70534 multiPassagers      Dist4
##         f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 1137082               p.Y4              p.X3               d.Y3
##         f.dropoff_latitude f.fare_amount    f.extra f.MTA_tax
## 1137082               d.X4      FAmount4 smallExtra   highMTA
##         f.Improvement_surcharge f.tip_amount   f.toll       f.total
## 1137082           highSurcharge      highTip highToll ExpensiveTrip
##         f.ttime f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 1137082   Time3   Speed4       Normald1       Normald2       Normald3
##         f.outlierPCAd4 f.outlierPCA claHP MCAhp
## 1137082       Normald4       Normal     3     6
```

Cluster 5

```
#parangon C1
df["272451",]
```

```
##              VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 272451 VeriFone Inc.  2016-01-06 19:26:39   2016-01-06 19:26:43
##        Store_and_fwd_flag   RateCodeID Pickup_longitude Pickup_latitude
## 272451      Store_and_fwd Special rate        -73.93873        40.67981
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 272451         -73.93874         40.67983               2      4.061167
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
```

```
## 272451          20     0       0          0            0
##        improvement_surcharge Total_amount Payment_type Trip_type mis_ind
## 272451                     0           20  Credit card  Dispatch       1
##           AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 272451 AnyTip No    6.578058         6.535815    17.56467           19
##        pick_up_period  espeed    f.passenger f.distance f.pickup_longitude
## 272451      afternoon 22.4703 multiPassagers      Dist4               p.Y3
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 272451              p.X1                d.Y3               d.X1
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 272451      FAmount4 smallExtra  smallMTA          smallSurcharge
##        f.tip_amount    f.toll       f.total f.ttime f.espeed
## 272451     smallTip smallToll ExpensiveTrip   Time4   Speed3
##        f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4
## 272451       Normald1       Normald2       Normald3       Normald4
##        f.outlierPCA claHP MCAhp
## 272451       Normal     5     4
```

```
#distinguished C1
df["675043",]
```

```
##                              VendorID lpep_pickup_datetime
## 675043 Creative Mobile Technologies, LLC  2016-01-15 11:30:15
##        Lpep_dropoff_datetime Store_and_fwd_flag   RateCodeID
## 675043   2016-01-15 12:10:05      Store_and_fwd Special rate
##        Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
## 675043       -73.86209        40.73549         -73.98948         40.71623
##        Passenger_count Trip_distance Fare_amount Extra MTA_tax Tip_amount
## 675043               1           8.6          30     0       0          0
##        Tolls_amount improvement_surcharge Total_amount Payment_type
## 675043            0                     0           30  Credit card
##        Trip_type mis_ind    AnyTip trip_length trip_distance_km
## 675043  Dispatch       0 AnyTip No    16.30613         13.84036
##        travel_time pick_up_hour pick_up_period   espeed f.passenger
## 675043    39.83333           11        morning 24.56154 onePassager
##        f.distance f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 675043      Dist4               p.Y4              p.X2               d.Y1
##        f.dropoff_latitude f.fare_amount    f.extra f.MTA_tax
## 675043               d.X2      FAmount4 smallExtra  smallMTA
##        f.Improvement_surcharge f.tip_amount    f.toll       f.total
## 675043          smallSurcharge     smallTip smallToll ExpensiveTrip
##        f.ttime f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 675043   Time4   Speed3       Normald1       Normald2       Normald3
##        f.outlierPCAd4 f.outlierPCA claHP MCAhp
## 675043       Normald4       Normal     5     4
```

Cluster 6

```
#parangon C1
df["678042",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 678042 VeriFone Inc.  2016-01-15 13:24:58   2016-01-15 14:00:17
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
```

```
## 678042      Store_and_fwd Standard rate        -73.90332        40.74579
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 678042         -73.98235          40.7681               1          4.95
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 678042          25     0     0.5          3            0
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 678042                   0.3         28.8  Credit card Street-hail       0
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 678042 AnyTip Yes    11.26821         7.966253    35.31667           13
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 678042         valley 19.14372 onePassager      Dist4               p.Y4
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 678042              p.X3                d.Y1               d.X3
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 678042      FAmount4 smallExtra   highMTA           highSurcharge
##        f.tip_amount    f.toll       f.total f.ttime f.espeed
## 678042      highTip smallToll ExpensiveTrip   Time4   Speed2
##        f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4
## 678042       Normald1       Normald2       Normald3       Normald4
##        f.outlierPCA claHP MCAhp
## 678042       Normal     6     6
```

```
#distinguished C1
df["285458",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 285458 VeriFone Inc.  2016-01-07 01:26:54   2016-01-07 01:41:13
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 285458      Store_and_fwd Standard rate        -73.94707        40.81071
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 285458         -74.01742         40.85104               1         10.47
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 285458          29   0.5     0.5          4         10.5
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 285458                   0.3         44.8  Credit card Street-hail       0
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 285458 AnyTip Yes    12.30616         16.84983    14.31667            1
##        pick_up_period   espeed f.passenger f.distance f.pickup_longitude
## 285458          night 51.57415 onePassager      Dist4               p.Y2
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 285458              p.X4                d.Y1               d.X4
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 285458      FAmount4 smallExtra   highMTA           highSurcharge
##        f.tip_amount   f.toll       f.total f.ttime f.espeed f.outlierPCAd1
## 285458      highTip highToll ExpensiveTrip   Time3   Speed4       Normald1
##        f.outlierPCAd2 f.outlierPCAd3 f.outlierPCAd4 f.outlierPCA claHP
## 285458       Normald2       Normald3       Normald4       Normal     6
##        MCAhp
## 285458     6
```

Comparison with PCA clustering

As it fails everytime we re-execute this commands, we commented it in order to show the proper results.

```
tt<-table(df$claHP,df$MCAhp); tt
```

```
##
##       1   2   3   4   5   6   7
##   1 561 525 369   0   2  89   0
##   2 184 220 254   0   2 110   0
##   3  59  52  64   0  30  37   0
##   4 215 183 216   0 720 186   0
##   5   1   0   1  36   0   2   0
##   6   1   1   7   0   7 690   0
##   7   0   0   0   0   0   0 119
```

```
df$MCAhp<-factor(df$MCAhp,levels=c(1,3,5,2,4,6,7),labels=c("mcaC-1","mcaC-3","mcaC-5","mcaC-2","mcaC-4"
tt<-table(df$claHP,df$MCAhp)
tt
```

```
##
##     mcaC-1 mcaC-3 mcaC-5 mcaC-2 mcaC-4 mcaC-6 mcaC-7
##   1    561    369      2    525      0     89      0
##   2    184    254      2    220      0    110      0
##   3     59     64     30     52      0     37      0
##   4    215    216    720    183      0    186      0
##   5      1      1      0      0     36      2      0
##   6      1      7      7      1      0    690      0
##   7      0      0      0      0      0      0    119
```

```
sum(diag(tt)/sum(tt))
```

```
## [1] 0.3789197
```

We can see, thou, that only with the table of contingencies is enough to appreciate how different are the clusters generated by both techniques. The diagonal summatory it's about 0.5 which means only a 50% of matching in the cluster structure.

To conclude, for our target total_amount, this clustering with MCA is taking a better approach to split by its categories (low, medium, and high prices) than the clusterings resulting from numerical PCA method.

MODELING

A statistical model is an expression that attempts to explain patterns in the observed values of a response variable by relating the response variable to a set of predictor variables and parameters. Consider the following familiar statistical model:

y = mx+c

This simple statistical model relates a response variable (y) to a single predictor variable (x) as a straight line according to the values of two constant parameters: m - the degree to which y changes per unit of change in x (gradient of line) c - the value of y when x = 0 (yintercept).

In complex systems, variables are typically the result of many influential and interacting factors and therefore simple models usually fail to fully explain a response variable. Consequently, the statistical model also has an error component that represents the portion of the response variable that the model fails to explain. Hence, statistical models are of the form:

response variable = model + error

Linear models in R

Les variables que utilitzarem per a comen?ar a fer el model s?n les variables cont?nues m?s significatives per a nosaltres. Per comen?ar hem de veure quines s?n les variables m?s relacionades amb la nostra variable resposta Total_amount.

```
load("Taxi5000_raw_DataDefinitivev1.RData")

# Quines variables s?n les m?s adientes per crear el nostre model? La resposta ?s
#sencilla, les variables m?s realcionades amb la nostra variable resposta s?n les
#seleccionades com a possibles candidates per a formar part del nostre model
condes(df[,vars_con],12)
```

```
## $quanti
##                   correlation      p.value
## Fare_amount        0.97065430 0.000000e+00
## trip_distance_km   0.91823220 0.000000e+00
## Trip_distance      0.91823220 0.000000e+00
## trip_length        0.85397745 0.000000e+00
## travel_time        0.75157892 0.000000e+00
## Tip_amount         0.57255145 0.000000e+00
## Tolls_amount       0.26464968 8.715509e-79
## MTA_tax           -0.03690045 1.004520e-02
## pick_up_hour      -0.04876788 6.664099e-04
## Dropoff_longitude -0.04996180 4.894714e-04
## Pickup_longitude  -0.05145902 3.293237e-04
## Pickup_latitude   -0.09542770 2.553859e-11
## Dropoff_latitude  -0.11714458 2.449070e-16
```

```
# Les variables m?s relacionades amb Total_amount (variable resposta) s?n:
# Trip_distance,Fare_amount,trip_distance_km,Trip_distance,travel_time,trip_length

# Feature Selection
cor(df[,c("Total_amount",vars_con)],method="spearman") #coeficient no parametric
```

```
##                   Total_amount Pickup_longitude Pickup_latitude
## Total_amount        1.00000000     -0.084165191     -0.11737286
## Pickup_longitude   -0.08416519      1.000000000      0.43217937
## Pickup_latitude    -0.11737286      0.432179371      1.00000000
## Dropoff_longitude  -0.12342014      0.733240913      0.26156584
## Dropoff_latitude   -0.13938888      0.359710073      0.88330586
## Passenger_count     0.02725452     -0.004024973     -0.03681770
## Trip_distance       0.92930404     -0.037153427     -0.09179998
## Fare_amount         0.97044508     -0.043014730     -0.08446424
## Extra               0.04318150     -0.020041909     -0.10407521
## MTA_tax            -0.04693595     -0.127040566     -0.08146209
## Tip_amount          0.44816381     -0.222141780     -0.14098034
## Tolls_amount        0.17772789      0.005501400      0.03115867
## Total_amount.1      1.00000000     -0.084165191     -0.11737286
## trip_length         0.86151458     -0.018482240     -0.07035942
## trip_distance_km    0.92930404     -0.037153427     -0.09179998
## travel_time         0.92665798     -0.073455598     -0.11369354
## pick_up_hour       -0.01124713     -0.057228330     -0.05332343
## espeed              0.04446030      0.093007240      0.05094887
##                   Dropoff_longitude Dropoff_latitude Passenger_count
```

```
## Total_amount          -0.123420137      -0.13938888     0.027254518
## Pickup_longitude       0.733240913       0.35971007    -0.004024973
## Pickup_latitude        0.261565840       0.88330586    -0.036817696
## Dropoff_longitude      1.000000000       0.26069037     0.008467316
## Dropoff_latitude       0.260690366       1.00000000    -0.041299727
## Passenger_count        0.008467316      -0.04129973     1.000000000
## Trip_distance         -0.076319646      -0.11476034     0.026752615
## Fare_amount           -0.084168706      -0.10635697     0.025310219
## Extra                  0.047705598      -0.11471654     0.063079504
## MTA_tax               -0.105938506      -0.07574832     0.004787270
## Tip_amount            -0.263440477      -0.13170792     0.006445604
## Tolls_amount          -0.018713024       0.05294082     0.025246138
## Total_amount.1        -0.123420137      -0.13938888     0.027254518
## trip_length           -0.053867733      -0.10112188     0.023524240
## trip_distance_km      -0.076319646      -0.11476034     0.026752615
## travel_time           -0.108767509      -0.13167205     0.021172043
## pick_up_hour          -0.027431165      -0.03845777     0.035733261
## espeed                 0.092767616       0.02483334     0.007583965
##                    Trip_distance Fare_amount       Extra     MTA_tax
## Total_amount         0.92930404  0.97044508  0.04318150 -0.04693595
## Pickup_longitude    -0.03715343 -0.04301473 -0.02004191 -0.12704057
## Pickup_latitude     -0.09179998 -0.08446424 -0.10407521 -0.08146209
## Dropoff_longitude   -0.07631965 -0.08416871  0.04770560 -0.10593851
## Dropoff_latitude    -0.11476034 -0.10635697 -0.11471654 -0.07574832
## Passenger_count      0.02675262  0.02531022  0.06307950  0.00478727
## Trip_distance        1.00000000  0.95291398 -0.02505440 -0.04604451
## Fare_amount          0.95291398  1.00000000 -0.03641036 -0.08712929
## Extra               -0.02505440 -0.03641036  1.00000000  0.12768968
## MTA_tax             -0.04604451 -0.08712929  0.12768968  1.00000000
## Tip_amount           0.27736241  0.27945628  0.02476966  0.08162160
## Tolls_amount         0.15597281  0.14642683 -0.02246175  0.01556496
## Total_amount.1       0.92930404  0.97044508  0.04318150 -0.04693595
## trip_length          0.92583392  0.88628774 -0.02783478 -0.03321429
## trip_distance_km     1.00000000  0.95291398 -0.02505440 -0.04604451
## travel_time          0.88279086  0.95086954 -0.02992717 -0.05134989
## pick_up_hour        -0.03557468 -0.03816450  0.32783036  0.01063518
## espeed               0.21363466  0.05014010 -0.00595466  0.01047900
##                    Tip_amount Tolls_amount Total_amount.1 trip_length
## Total_amount       0.448163808   0.17772789     1.00000000  0.86151458
## Pickup_longitude  -0.222141780   0.00550140    -0.08416519 -0.01848224
## Pickup_latitude   -0.140980336   0.03115867    -0.11737286 -0.07035942
## Dropoff_longitude -0.263440477  -0.01871302    -0.12342014 -0.05386773
## Dropoff_latitude  -0.131707916   0.05294082    -0.13938888 -0.10112188
## Passenger_count    0.006445604   0.02524614     0.02725452  0.02352424
## Trip_distance      0.277362409   0.15597281     0.92930404  0.92583392
## Fare_amount        0.279456280   0.14642683     0.97044508  0.88628774
## Extra              0.024769664  -0.02246175     0.04318150 -0.02783478
## MTA_tax            0.081621605   0.01556496    -0.04693595 -0.03321429
## Tip_amount         1.000000000   0.10376169     0.44816381  0.25366674
## Tolls_amount       0.103761686   1.00000000     0.17772789  0.14446835
## Total_amount.1     0.448163808   0.17772789     1.00000000  0.86151458
## trip_length        0.253666741   0.14446835     0.86151458  1.00000000
## trip_distance_km   0.277362409   0.15597281     0.92930404  0.92583392
## travel_time        0.272968823   0.11525739     0.92665798  0.81307964
```

```
## pick_up_hour      -0.001712168  -0.03867120    -0.01124713 -0.04124642
## espeed             0.006728818   0.07538278     0.04446030  0.39789846
##                   trip_distance_km travel_time pick_up_hour       espeed
## Total_amount            0.92930404  0.92665798 -0.011247127  0.044460299
## Pickup_longitude       -0.03715343 -0.07345560 -0.057228330  0.093007240
## Pickup_latitude        -0.09179998 -0.11369354 -0.053323428  0.050948866
## Dropoff_longitude      -0.07631965 -0.10876751 -0.027431165  0.092767616
## Dropoff_latitude       -0.11476034 -0.13167205 -0.038457766  0.024833340
## Passenger_count         0.02675262  0.02117204  0.035733261  0.007583965
## Trip_distance           1.00000000  0.88279086 -0.035574680  0.213634656
## Fare_amount             0.95291398  0.95086954 -0.038164501  0.050140103
## Extra                  -0.02505440 -0.02992717  0.327830359 -0.005954660
## MTA_tax                -0.04604451 -0.05134989  0.010635179  0.010478999
## Tip_amount              0.27736241  0.27296882 -0.001712168  0.006728818
## Tolls_amount            0.15597281  0.11525739 -0.038671201  0.075382781
## Total_amount.1          0.92930404  0.92665798 -0.011247127  0.044460299
## trip_length             0.92583392  0.81307964 -0.041246425  0.397898464
## trip_distance_km        1.00000000  0.88279086 -0.035574680  0.213634656
## travel_time             0.88279086  1.00000000 -0.012941788 -0.141913503
## pick_up_hour           -0.03557468 -0.01294179  1.000000000 -0.051142362
## espeed                  0.21363466 -0.14191350 -0.051142362  1.000000000
```

```
# Quantes variables agafem? Hem d'intentar arribar al millor model que pugem representar.
```

Les variables m?s correlacionades amb la nostra variable resposta Total_amount s?n les que tenen una correlaci? propera a 1. En el nostre cas les variables m?s correlacionades s?n: Trip_distance, Fare_amount, trip_distance_km, Trip_distance, travel_time, trip_length, per? no totes s?n ?tils.

Multiple Linear Regression issues

Target numeric Total_amount

Creem un nou model m10, amb les variables que m?s es relacionen amb la nostra variable resposta.

```
# Creem un nou model amb les variables Fare_amount,
# trip_distance_km, travel_time, Tip_amount, espeed.
m10<-lm(Total_amount~Fare_amount+trip_distance_km
        +travel_time+Tip_amount+espeed,data=df)
summary(m10)
```

```
##
## Call:
## lm(formula = Total_amount ~ Fare_amount + trip_distance_km +
##     travel_time + Tip_amount + espeed, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8159 -0.3585 -0.0687  0.1727 11.8113
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.1175243  0.0407993  27.391  < 2e-16 ***
## Fare_amount       0.9864011  0.0053612 183.989  < 2e-16 ***
## trip_distance_km  0.0585972  0.0102656   5.708 1.21e-08 ***
```

```
## travel_time      -0.0041999  0.0019113  -2.197    0.028 *
## Tip_amount        1.0491685  0.0066837 156.974  < 2e-16 ***
## espeed            0.0003075  0.0014119   0.218    0.828
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7687 on 4860 degrees of freedom
## Multiple R-squared:  0.9907, Adjusted R-squared:  0.9907
## F-statistic: 1.032e+05 on 5 and 4860 DF,  p-value: < 2.2e-16
```

```
#Explicaci?:
# 1. L'estimaci? de B0 ?s 1.1175243 amb un error estandar 0.0407993.
# El contrast de H0: B0 = 0 davant H1: B0 != 0 llen?a un valor estadistic
# de 27.391 i un p-valor inferior a 2e-16.

# 2. L'estimaci? de B1 ?s 0.9864011 amb un error estandar 0.0053612.
# El contrast de H0: B1 = 0 davant H1: B1 != 0 llen?a un valor estadistic
# de 183.989 i un p-valor inferior a 2e-16.

# 3. L'estimaci? de B2 ?s 0.0585972 amb un error estandar 0.0102656.
# El contrast de H0: B2 = 0 davant H1: B2 != 0 llen?a un valor estadistic
# de 5.708 i un p-valor de 1.21e-08.

# 4. L'estimaci? de B3 ?s -0.0041999 amb un error estandar 0.0019113.
# El contrast de H0: B3 = 0 davant H1: B3 != 0 llen?a un valor estadistic
# de -2.197 i un p-valor de 0.028.

# 5. L'estimaci? de B4 ?s 1.0491685 amb un error estandar 0.0066837.
#. El contrast de H0: B4 = 0 davant H1: B4 != 0 llen?a un valor estadistic
# de 156.974 i un p-valor inferior a 2e-16.

# 6. L'estimaci? de B5 ?s 0.0003075 amb un error estandar 0.0014119.
# El contrast de H0: B5 = 0 davant H1: B5 != 0 llen?a un valor estadistic
# de 0.218 i un p-valor de 0.828.

# La recta ajustada apareix, per tant, especificada a trav?s dels cinc
#coeficients:
# Total_amount = 1.1175243+0.9864011*Fare_amount+0.0585972*trip_distance_km
# -0.0041999*travel_time+1.0491685*Tip_amount+0.0003075*espeed

# 7. L'error estandar de l'ajust te un valor de 0.7687.

# 8. El coeficient R^2 te el valor 0.9907, que indica que el 99.07%
# de tota la variabilitat que te el fenomen relatiu al preu pagat.

# Les variables m?s significatives s?n les que tenen un p-valor menor a 0.05,
# ja que no podem rebutjar la hipotesi nul?la. En aquest cas les m?s
# significatives s?n: Fare_amount, trip_distance_km, travel_time,Tip_amount.
# Observem que podem descartar la variable espeed, almenys de moment.
```

```
# Test net effects
# All net effects are significant?
Anova(m10)
```

```
## Anova Table (Type II tests)
```

```
##
## Response: Total_amount
##                   Sum Sq   Df    F value    Pr(>F)
## Fare_amount      20002.6    1 33851.8818 < 2.2e-16 ***
## trip_distance_km    19.3    1    32.5826  1.21e-08 ***
## travel_time          2.9    1     4.8284   0.02804 *
## Tip_amount       14559.9    1 24640.8544 < 2.2e-16 ***
## espeed               0.0    1     0.0474   0.82762
## Residuals         2871.7 4860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# No tots els efectes son significatius, ja que obtenim variables que
# tenen un p-valor major al nostre llindar 0.05
```

```
# Apliquem la comanda step al model anterior per determinar
# quina combinaci? es la m?s adient i poder treure variables
# que no s?n necessaries
m11<-step(m10)
```

```
## Start:  AIC=-2554.16
## Total_amount ~ Fare_amount + trip_distance_km + travel_time +
##     Tip_amount + espeed
##
##                    Df Sum of Sq     RSS     AIC
## - espeed            1       0.0  2871.7 -2556.1
## <none>                           2871.7 -2554.2
## - travel_time       1       2.9  2874.6 -2551.3
## - trip_distance_km  1      19.3  2891.0 -2523.6
## - Tip_amount        1   14559.9 17431.7  6219.1
## - Fare_amount       1   20002.6 22874.3  7541.3
##
## Step:  AIC=-2556.11
## Total_amount ~ Fare_amount + trip_distance_km + travel_time +
##     Tip_amount
##
##                    Df Sum of Sq     RSS     AIC
## <none>                           2871.7 -2556.1
## - travel_time       1       3.4  2875.1 -2552.4
## - trip_distance_km  1      23.7  2895.4 -2518.2
## - Tip_amount        1   14597.2 17469.0  6227.5
## - Fare_amount       1   20952.6 23824.3  7737.3
```

```
m11<-step(m10,k=log(nrow(df)))  # BIC
```

```
## Start:  AIC=-2515.22
## Total_amount ~ Fare_amount + trip_distance_km + travel_time +
##     Tip_amount + espeed
##
##                    Df Sum of Sq     RSS     AIC
## - espeed            1       0.0  2871.7 -2523.7
## - travel_time       1       2.9  2874.6 -2518.9
## <none>                           2871.7 -2515.2
## - trip_distance_km  1      19.3  2891.0 -2491.2
## - Tip_amount        1   14559.9 17431.7  6251.5
## - Fare_amount       1   20002.6 22874.3  7573.8
```

```
##
## Step:  AIC=-2523.66
## Total_amount ~ Fare_amount + trip_distance_km + travel_time +
##     Tip_amount
##
##                     Df Sum of Sq     RSS     AIC
## - travel_time        1       3.4  2875.1 -2526.5
## <none>                            2871.7 -2523.7
## - trip_distance_km   1      23.7  2895.4 -2492.2
## - Tip_amount         1   14597.2 17469.0  6253.5
## - Fare_amount        1   20952.6 23824.3  7763.3
##
## Step:  AIC=-2526.46
## Total_amount ~ Fare_amount + trip_distance_km + Tip_amount
##
##                     Df Sum of Sq     RSS     AIC
## <none>                            2875.1 -2526.5
## - trip_distance_km   1      28.2  2903.3 -2487.5
## - Tip_amount         1   14623.0 17498.1  6253.1
## - Fare_amount        1   28439.1 31314.2  9085.0
```

```
vif(m11)
```

```
##      Fare_amount trip_distance_km       Tip_amount
##         7.918040         7.972860         1.180566
```

```
# Observem que l'unica variable que compleix la nostra cota
# es Tip_amount, per tant podem agafar aquesta variable per
# al nostre model, i hem de continuar treballant per tdeterminar
# quina es la millor entre Fare_amount trip_distance_km
```

Apliquem la comanda AIC per determinar quin ?s el millor model fins al moment

```
AIC(m10,m11)
```

```
##     df      AIC
## m10  7 11256.95
## m11  5 11258.69
```

```
# Comparem tos el models que tenim fins el momnet per
# determinar quin ?s el millor.
# Segons el dit a classe el millor model es el que te el Valor
# de AIC m?s petit, en el nostre cas ?s m10, per? aquest mostra colinealitat.
```

Creem el model m20 i m21 amb totes les variables cont?nues.

```
vars_con_model <- vars_con[c(1:5,7:8,10:11,13:17)]
m20<-lm(Total_amount~.,data= df[,c("Total_amount",vars_con_model)])
summary(m20)
```

```
##
## Call:
## lm(formula = Total_amount ~ ., data = df[, c("Total_amount",
##     vars_con_model)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79231 -0.00442  0.01181  0.02682  0.17710
```

```
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       -1.587e+01  3.172e+00   -5.005 5.80e-07 ***
## Pickup_longitude  -2.428e-01  5.859e-02   -4.143 3.48e-05 ***
## Pickup_latitude   -6.843e-02  5.633e-02   -1.215  0.22455
## Dropoff_longitude -4.596e-02  5.044e-02   -0.911  0.36217
## Dropoff_latitude  -4.629e-02  5.517e-02   -0.839  0.40148
## Passenger_count    1.026e-03  1.414e-03    0.726  0.46811
## Fare_amount        9.974e-01  7.298e-04 1366.603  < 2e-16 ***
## Extra              1.034e+00  4.116e-03  251.243  < 2e-16 ***
## Tip_amount         1.004e+00  8.909e-04 1126.514  < 2e-16 ***
## Tolls_amount       1.003e+00  2.158e-03  464.670  < 2e-16 ***
## trip_length        4.104e-03  1.548e-03    2.651  0.00804 **
## trip_distance_km   1.324e-03  1.435e-03    0.923  0.35603
## travel_time       -8.956e-04  2.601e-04   -3.443  0.00058 ***
## pick_up_hour      -4.297e-04  2.188e-04   -1.964  0.04960 *
## espeed            -2.508e-04  2.343e-04   -1.070  0.28448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09978 on 4851 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 2.208e+06 on 14 and 4851 DF,  p-value: < 2.2e-16
```

```
vif(m20)
```

```
##  Pickup_longitude   Pickup_latitude Dropoff_longitude  Dropoff_latitude
##          2.843856          4.806349          2.781542          4.707439
##   Passenger_count       Fare_amount             Extra        Tip_amount
##          1.005299         12.554735          1.087191          1.252882
##      Tolls_amount       trip_length  trip_distance_km       travel_time
##          1.064394          9.101900         11.430521          3.130691
##      pick_up_hour            espeed
##          1.084849          2.237831
```

```
m21<-step(m20,k=log(nrow(df)))  # BIC
```

```
## Start:  AIC=-22318.35
## Total_amount ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Dropoff_latitude + Passenger_count + Fare_amount + Extra +
##     Tip_amount + Tolls_amount + trip_length + trip_distance_km +
##     travel_time + pick_up_hour + espeed
##
##                     Df Sum of Sq     RSS      AIC
## - Passenger_count    1       0.0    48.3 -22326.3
## - Dropoff_latitude   1       0.0    48.3 -22326.1
## - Dropoff_longitude  1       0.0    48.3 -22326.0
## - trip_distance_km   1       0.0    48.3 -22326.0
## - espeed             1       0.0    48.3 -22325.7
## - Pickup_latitude    1       0.0    48.3 -22325.4
## - pick_up_hour       1       0.0    48.3 -22323.0
## - trip_length        1       0.1    48.4 -22319.8
## <none>                               48.3 -22318.3
## - travel_time        1       0.1    48.4 -22315.0
```

```
## - Pickup_longitude   1       0.2    48.5 -22309.6
## - Extra              1     628.4   676.7  -9480.9
## - Tolls_amount       1    2149.5  2197.8  -3748.8
## - Tip_amount         1   12633.3 12681.6   4779.9
## - Fare_amount        1   18592.1 18640.4   6654.2
##
## Step:  AIC=-22326.31
## Total_amount ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Dropoff_latitude + Fare_amount + Extra + Tip_amount + Tolls_amount +
##     trip_length + trip_distance_km + travel_time + pick_up_hour +
##     espeed
##
##                      Df Sum of Sq     RSS      AIC
## - Dropoff_latitude   1       0.0    48.3 -22334.1
## - Dropoff_longitude  1       0.0    48.3 -22334.0
## - trip_distance_km   1       0.0    48.3 -22333.9
## - espeed             1       0.0    48.3 -22333.7
## - Pickup_latitude    1       0.0    48.3 -22333.3
## - pick_up_hour       1       0.0    48.3 -22331.0
## - trip_length        1       0.1    48.4 -22327.8
## <none>                               48.3 -22326.3
## - travel_time        1       0.1    48.4 -22323.1
## - Pickup_longitude   1       0.2    48.5 -22317.6
## - Extra              1     629.8   678.1  -9479.2
## - Tolls_amount       1    2149.5  2197.8  -3757.2
## - Tip_amount         1   12633.3 12681.6   4771.4
## - Fare_amount        1   18593.3 18641.6   6646.0
##
## Step:  AIC=-22334.08
## Total_amount ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Fare_amount + Extra + Tip_amount + Tolls_amount + trip_length +
##     trip_distance_km + travel_time + pick_up_hour + espeed
##
##                      Df Sum of Sq     RSS      AIC
## - trip_distance_km   1       0.0    48.3 -22341.7
## - Dropoff_longitude  1       0.0    48.3 -22341.5
## - espeed             1       0.0    48.3 -22341.5
## - pick_up_hour       1       0.0    48.3 -22338.7
## - trip_length        1       0.1    48.4 -22335.3
## <none>                               48.3 -22334.1
## - travel_time        1       0.1    48.4 -22330.8
## - Pickup_latitude    1       0.2    48.5 -22326.1
## - Pickup_longitude   1       0.2    48.5 -22325.9
## - Extra              1     631.2   679.5  -9477.6
## - Tolls_amount       1    2158.9  2207.2  -3744.9
## - Tip_amount         1   12655.3 12703.6   4771.4
## - Fare_amount        1   18594.4 18642.7   6637.8
##
## Step:  AIC=-22341.72
## Total_amount ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Fare_amount + Extra + Tip_amount + Tolls_amount + trip_length +
##     travel_time + pick_up_hour + espeed
##
##                     Df Sum of Sq   RSS      AIC
```

```
## - espeed             1         0    48 -22349.2
## - Dropoff_longitude  1         0    48 -22349.2
## - pick_up_hour       1         0    48 -22346.3
## <none>                             48 -22341.7
## - trip_length        1         0    48 -22339.7
## - travel_time        1         0    48 -22337.3
## - Pickup_latitude    1         0    48 -22333.7
## - Pickup_longitude   1         0    48 -22333.6
## - Extra              1       631   680  -9485.3
## - Tolls_amount       1      2175  2224  -3717.4
## - Tip_amount         1     12740 12789   4795.3
## - Fare_amount        1     33422 33471   9477.0
##
## Step:  AIC=-22349.23
## Total_amount ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Fare_amount + Extra + Tip_amount + Tolls_amount + trip_length +
##     travel_time + pick_up_hour
##
##                      Df Sum of Sq   RSS      AIC
## - Dropoff_longitude  1         0    48 -22356.6
## - pick_up_hour       1         0    48 -22354.1
## <none>                             48 -22349.2
## - travel_time        1         0    48 -22345.4
## - trip_length        1         0    48 -22345.3
## - Pickup_longitude   1         0    48 -22341.1
## - Pickup_latitude    1         0    48 -22340.7
## - Extra              1       632   680  -9491.1
## - Tolls_amount       1      2176  2224  -3725.0
## - Tip_amount         1     12777 12826   4800.9
## - Fare_amount        1     39804 39853  10317.7
##
## Step:  AIC=-22356.59
## Total_amount ~ Pickup_longitude + Pickup_latitude + Fare_amount +
##     Extra + Tip_amount + Tolls_amount + trip_length + travel_time +
##     pick_up_hour
##
##                     Df Sum of Sq   RSS      AIC
## - pick_up_hour       1         0    48 -22361.5
## <none>                             48 -22356.6
## - travel_time        1         0    48 -22353.2
## - trip_length        1         0    48 -22353.0
## - Pickup_latitude    1         0    48 -22348.7
## - Pickup_longitude   1         1    49 -22303.9
## - Extra              1       633   682  -9486.6
## - Tolls_amount       1      2177  2226  -3730.1
## - Tip_amount         1     12926 12974   4848.4
## - Fare_amount        1     39810 39858  10309.8
##
## Step:  AIC=-22361.49
## Total_amount ~ Pickup_longitude + Pickup_latitude + Fare_amount +
##     Extra + Tip_amount + Tolls_amount + trip_length + travel_time
##
##                     Df Sum of Sq   RSS      AIC
## <none>                             48 -22361.5
```

```
## - trip_length       1         0    48 -22357.8
## - travel_time       1         0    48 -22357.0
## - Pickup_latitude   1         0    49 -22353.4
## - Pickup_longitude  1         1    49 -22310.2
## - Extra             1       672   720  -9227.8
## - Tolls_amount      1      2178  2226  -3736.7
## - Tip_amount        1     12927 12975   4840.4
## - Fare_amount       1     39913 39961  10313.9
```

```
summary(m21)
```

```
##
## Call:
## lm(formula = Total_amount ~ Pickup_longitude + Pickup_latitude +
##     Fare_amount + Extra + Tip_amount + Tolls_amount + trip_length +
##     travel_time, data = df[, c("Total_amount", vars_con_model)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78790 -0.00395  0.01178  0.02665  0.16924
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      -1.560e+01  3.148e+00   -4.954 7.51e-07 ***
## Pickup_longitude -2.821e-01  3.641e-02   -7.746 1.14e-14 ***
## Pickup_latitude  -1.097e-01  2.696e-02   -4.068 4.82e-05 ***
## Fare_amount       9.981e-01  4.986e-04 2001.953  < 2e-16 ***
## Extra             1.032e+00  3.974e-03  259.756  < 2e-16 ***
## Tip_amount        1.004e+00  8.811e-04 1139.328  < 2e-16 ***
## Tolls_amount      1.003e+00  2.144e-03  467.663  < 2e-16 ***
## trip_length       3.656e-03  1.046e-03    3.496 0.000477 ***
## travel_time      -8.364e-04  2.326e-04   -3.596 0.000327 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09979 on 4857 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 3.863e+06 on 8 and 4857 DF,  p-value: < 2.2e-16
```

```
vif(m21)
```

```
## Pickup_longitude  Pickup_latitude      Fare_amount            Extra
##         1.097732         1.100133         5.856486         1.013072
##       Tip_amount     Tolls_amount      trip_length      travel_time
##         1.224887         1.050713         4.153853         2.502456
```

```
# Test Fisher: m20 - m21 H0 Equivalents
anova(m21,m20)
```

```
## Analysis of Variance Table
##
## Model 1: Total_amount ~ Pickup_longitude + Pickup_latitude + Fare_amount +
##     Extra + Tip_amount + Tolls_amount + trip_length + travel_time
## Model 2: Total_amount ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Dropoff_latitude + Passenger_count + Fare_amount + Extra +
##     Tip_amount + Tolls_amount + trip_length + trip_distance_km +
```

```
##     travel_time + pick_up_hour + espeed
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4857 48.369
## 2   4851 48.292  6  0.077418 1.2961 0.2552
```

```
# All net effects are significant?
Anova(m21)
```

```
## Anova Table (Type II tests)
##
## Response: Total_amount
##                   Sum Sq   Df    F value    Pr(>F)
## Pickup_longitude      1    1 6.0005e+01 1.144e-14 ***
## Pickup_latitude       0    1 1.6549e+01 4.816e-05 ***
## Fare_amount       39913    1 4.0078e+06 < 2.2e-16 ***
## Extra               672    1 6.7473e+04 < 2.2e-16 ***
## Tip_amount        12927    1 1.2981e+06 < 2.2e-16 ***
## Tolls_amount       2178    1 2.1871e+05 < 2.2e-16 ***
## trip_length           0    1 1.2219e+01 0.0004774 ***
## travel_time           0    1 1.2930e+01 0.0003266 ***
## Residuals            48 4857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(m21)  # Check association between explanatory vars
```

```
## Pickup_longitude  Pickup_latitude      Fare_amount            Extra
##         1.097732         1.100133         5.856486         1.013072
##       Tip_amount     Tolls_amount      trip_length      travel_time
##         1.224887         1.050713         4.153853         2.502456
```

```
round(cor(df[,vars_con],use="pairwise.complete.obs"),dig=2)
```

```
##                   Pickup_longitude Pickup_latitude Dropoff_longitude
## Pickup_longitude              1.00            0.26              0.79
## Pickup_latitude               0.26            1.00              0.15
## Dropoff_longitude             0.79            0.15              1.00
## Dropoff_latitude              0.22            0.88              0.16
## Passenger_count               0.00           -0.02              0.02
## Trip_distance                 0.00           -0.06              0.00
## Fare_amount                  -0.02           -0.07             -0.01
## Extra                        -0.01           -0.10              0.03
## MTA_tax                      -0.13           -0.10             -0.11
## Tip_amount                   -0.15           -0.13             -0.17
## Tolls_amount                  0.00            0.03             -0.02
## Total_amount                 -0.05           -0.10             -0.05
## trip_length                   0.01           -0.06              0.02
## trip_distance_km              0.00           -0.06              0.00
## travel_time                  -0.05           -0.09             -0.06
## pick_up_hour                 -0.05           -0.02             -0.04
## espeed                        0.10            0.07              0.12
##                   Dropoff_latitude Passenger_count Trip_distance
## Pickup_longitude              0.22            0.00          0.00
## Pickup_latitude               0.88           -0.02         -0.06
## Dropoff_longitude             0.16            0.02          0.00
## Dropoff_latitude              1.00           -0.02         -0.09
```

```
## Passenger_count              -0.02            1.00          0.02
## Trip_distance                -0.09            0.02          1.00
## Fare_amount                  -0.10            0.01          0.93
## Extra                        -0.10            0.05         -0.04
## MTA_tax                      -0.09            0.02         -0.05
## Tip_amount                   -0.12            0.01          0.39
## Tolls_amount                  0.05            0.00          0.20
## Total_amount                 -0.12            0.02          0.92
## trip_length                  -0.10            0.02          0.90
## trip_distance_km             -0.09            0.02          1.00
## travel_time                  -0.11            0.02          0.68
## pick_up_hour                 -0.01            0.02         -0.07
## espeed                        0.04            0.02          0.17
##                    Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## Pickup_longitude         -0.02 -0.01   -0.13      -0.15         0.00
## Pickup_latitude          -0.07 -0.10   -0.10      -0.13         0.03
## Dropoff_longitude        -0.01  0.03   -0.11      -0.17        -0.02
## Dropoff_latitude         -0.10 -0.10   -0.09      -0.12         0.05
## Passenger_count           0.01  0.05    0.02       0.01         0.00
## Trip_distance             0.93 -0.04   -0.05       0.39         0.20
## Fare_amount               1.00 -0.04   -0.08       0.38         0.17
## Extra                    -0.04  1.00    0.12       0.02        -0.02
## MTA_tax                  -0.08  0.12    1.00       0.06         0.02
## Tip_amount                0.38  0.02    0.06       1.00         0.15
## Tolls_amount              0.17 -0.02    0.02       0.15         1.00
## Total_amount              0.97  0.02   -0.04       0.57         0.26
## trip_length               0.87 -0.03   -0.04       0.36         0.18
## trip_distance_km          0.93 -0.04   -0.05       0.39         0.20
## travel_time               0.77 -0.02   -0.08       0.32         0.09
## pick_up_hour             -0.06  0.25    0.01      -0.02        -0.03
## espeed                    0.02 -0.04    0.02      -0.01         0.08
##                    Total_amount trip_length trip_distance_km travel_time
## Pickup_longitude          -0.05        0.01             0.00       -0.05
## Pickup_latitude           -0.10       -0.06            -0.06       -0.09
## Dropoff_longitude         -0.05        0.02             0.00       -0.06
## Dropoff_latitude          -0.12       -0.10            -0.09       -0.11
## Passenger_count            0.02        0.02             0.02        0.02
## Trip_distance              0.92        0.90             1.00        0.68
## Fare_amount                0.97        0.87             0.93        0.77
## Extra                      0.02       -0.03            -0.04       -0.02
## MTA_tax                   -0.04       -0.04            -0.05       -0.08
## Tip_amount                 0.57        0.36             0.39        0.32
## Tolls_amount               0.26        0.18             0.20        0.09
## Total_amount               1.00        0.85             0.92        0.75
## trip_length                0.85        1.00             0.90        0.65
## trip_distance_km           0.92        0.90             1.00        0.68
## travel_time                0.75        0.65             0.68        1.00
## pick_up_hour              -0.05       -0.06            -0.07        0.00
## espeed                     0.02        0.35             0.17       -0.21
##                    pick_up_hour espeed
## Pickup_longitude          -0.05   0.10
## Pickup_latitude           -0.02   0.07
## Dropoff_longitude         -0.04   0.12
## Dropoff_latitude          -0.01   0.04
```

```
## Passenger_count           0.02   0.02
## Trip_distance            -0.07   0.17
## Fare_amount              -0.06   0.02
## Extra                     0.25  -0.04
## MTA_tax                   0.01   0.02
## Tip_amount               -0.02  -0.01
## Tolls_amount             -0.03   0.08
## Total_amount             -0.05   0.02
## trip_length              -0.06   0.35
## trip_distance_km         -0.07   0.17
## travel_time               0.00  -0.21
## pick_up_hour              1.00  -0.09
## espeed                   -0.09   1.00
```

En el model m20 observem que hi ha variables que no s?n necess?ries, ja que hi ha moltes correlacions impl?cites entre les mateixes variables del model. El mateix passa amb el model m21. A l'hora de comparar l'equival?ncia entre els dos model observem que el p-valor ?s de 0.2552 per tant podem dir que no s?n equivalents.

```
# En aquest moment ja comencem a tenir una ida de quines variables
# continues necessitem per al nostre model
AIC(m20,m21,m10,m11)
```

```
##     df       AIC
## m20 16 -8604.588
## m21 10 -8608.794
## m10  7 11256.951
## m11  5 11258.687
```

Observem que de tots els models que hem vist fins al moment el millor ?s el m21.

Creem un nou model m23.

```
# Creem una nou model amb les variables trip_distance_km,
# Fare_amount, Extra, MTA_tax, Tip_amount, Tolls_amount
m23<-lm(Total_amount~trip_distance_km+Fare_amount+Extra+MTA_tax+Tip_amount+Tolls_amount,data=df)
summary(m23)
```

```
##
## Call:
## lm(formula = Total_amount ~ trip_distance_km + Fare_amount +
##     Extra + MTA_tax + Tip_amount + Tolls_amount, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29984 -0.00048  0.00024  0.00114  0.74780
##
## Coefficients:
##                    Estimate Std. Error   t value Pr(>|t|)
## (Intercept)       0.0139931  0.0015974     8.760  < 2e-16 ***
## trip_distance_km -0.0007437  0.0001667    -4.462 8.31e-06 ***
## Fare_amount       1.0004252  0.0000804 12442.730  < 2e-16 ***
## Extra             1.0027085  0.0005487  1827.251  < 2e-16 ***
## MTA_tax           1.5666741  0.0030942   506.322  < 2e-16 ***
## Tip_amount        0.9998823  0.0001202  8315.289  < 2e-16 ***
## Tolls_amount      1.0001088  0.0002961  3377.280  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01376 on 4859 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.709e+08 on 6 and 4859 DF,  p-value: < 2.2e-16
```

```
Anova(m23)
```

```
## Anova Table (Type II tests)
##
## Response: Total_amount
##                    Sum Sq   Df    F value    Pr(>F)
## trip_distance_km     0.0    1 1.9907e+01 8.315e-06 ***
## Fare_amount      29314.8    1 1.5482e+08 < 2.2e-16 ***
## Extra              632.2    1 3.3388e+06 < 2.2e-16 ***
## MTA_tax             48.5    1 2.5636e+05 < 2.2e-16 ***
## Tip_amount       13092.1    1 6.9144e+07 < 2.2e-16 ***
## Tolls_amount      2159.7    1 1.1406e+07 < 2.2e-16 ***
## Residuals            0.9 4859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Observem que totes les variables son significatives

vif(m23)
```

```
## trip_distance_km      Fare_amount            Extra          MTA_tax
##         8.111354         8.011136         1.015901         1.031321
##       Tip_amount     Tolls_amount
##         1.199979         1.053810
```

```
# Encara observem que hi ha variables que estan relacionades entre elles,
# ja que superen la cota de 3 en el valor resultant

AIC(m23)
```

```
## [1] -27892.93
```

Observem que aquest nou model m23 t? un AIC m?s petit que el model m22 per? no volem un model amb correlacions entre les pr?pies variables.

Creem el model m24.

```
# Creem un nou model amb les variables que creiem m?s significatives
# i que no estan ralacionades entre elles
m24<-lm(Total_amount~trip_distance_km
        +Extra+Tip_amount+Tolls_amount,data=df)
summary(m24)
```

```
##
## Call:
## lm(formula = Total_amount ~ trip_distance_km + Extra + Tip_amount +
##     Tolls_amount, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.936  -0.979  -0.385   0.498  35.378
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.05666    0.06615   61.33   <2e-16 ***
## trip_distance_km   1.91442    0.01150  166.53   <2e-16 ***
## Extra              1.04948    0.09741   10.77   <2e-16 ***
## Tip_amount         1.09362    0.02136   51.19   <2e-16 ***
## Tolls_amount       0.76123    0.05278   14.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.458 on 4861 degrees of freedom
## Multiple R-squared:  0.9046, Adjusted R-squared:  0.9045
## F-statistic: 1.152e+04 on 4 and 4861 DF,  p-value: < 2.2e-16
```

```
Anova(m24) # All Net effects are significant
```

```
## Anova Table (Type II tests)
##
## Response: Total_amount
##                   Sum Sq   Df  F value    Pr(>F)
## trip_distance_km 167590    1 27732.20 < 2.2e-16 ***
## Extra               701    1   116.07 < 2.2e-16 ***
## Tip_amount        15835    1  2620.37 < 2.2e-16 ***
## Tolls_amount       1257    1   207.98 < 2.2e-16 ***
## Residuals         29376 4861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# condicionat a tenir 3 variables en el model la quarta
# es significativa

vif(m24)
```

```
## trip_distance_km            Extra       Tip_amount     Tolls_amount
##         1.208775         1.003027         1.186837         1.049052
```

```
# Observem que no hi ha colinealitat entre les variables
# utilitzades per aquet model

AIC(m23,m24)
```

```
##     df       AIC
## m23  8 -27892.93
## m24  6  22569.69
```

```
# Agafem el model amb un AIC m?s petit
```

Diagnostics

L'homoscedasticitat ?s una propietat fonamental del model de regressi? lineal general i est? dins dels seus sup?sits cl?ssics b?sics. Es diu que hi ha homoscedasticitat quan la vari?ncia dels errors de la regressi? s?n els mateixos per a cada observaci? i (d'1 a n observacions).

Creem el model m25.

```r
# Heterocedastricitat (varian?a no constant)
# Creem un nou model amb les variables que hens han semblat
# les millors per a fer el model segons l'apartat anterior
# i apliquem el logaridme a la variable resposta.
m25 <-lm(log(Total_amount)~trip_distance_km
        +Extra+Tip_amount+Tolls_amount,data=df)
summary(m25)
```

```
##
## Call:
## lm(formula = log(Total_amount) ~ trip_distance_km + Extra + Tip_amount +
##     Tolls_amount, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5803 -0.1177  0.0230  0.1247  1.7199
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.839703   0.006091 302.054  < 2e-16 ***
## trip_distance_km 0.123742   0.001059 116.900  < 2e-16 ***
## Extra             0.110499   0.008970  12.319  < 2e-16 ***
## Tip_amount        0.065952   0.001967  33.526  < 2e-16 ***
## Tolls_amount      0.015307   0.004860   3.149  0.00165 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2264 on 4861 degrees of freedom
## Multiple R-squared:  0.8173, Adjusted R-squared:  0.8171
## F-statistic:  5435 on 4 and 4861 DF,  p-value: < 2.2e-16
```

```r
# Explicaci?:
# 1. L'estimaci? de B0 ?s 1.839703 amb un error estandar 0.006091.

# 2. L'estimaci? de B1 ?s 0.123742 amb un error estandar 0.001059.

# 3. L'estimaci? de B2 ?s 0.110499 amb un error estandar 0.008970.

# 4. L'estimaci? de B3 ?s 0.065952 amb un error estandar 0.001967.

# 5. L'estimaci? de B4 ?s 0.015307 amb un error estandar 0.004860.

# La recta ajustada apareix, per tant, especificada a trav?s
# dels quatre coeficients:
# log(Total_amount) = 1.839703+0.123742*trip_distance_km+
# 0.110499*Extra+0.065952*Tip_amount+0.015307*Tolls_amount

# 6. L'error estandar de l'ajust te un valor de 0.2264.


Anova(m25)
```

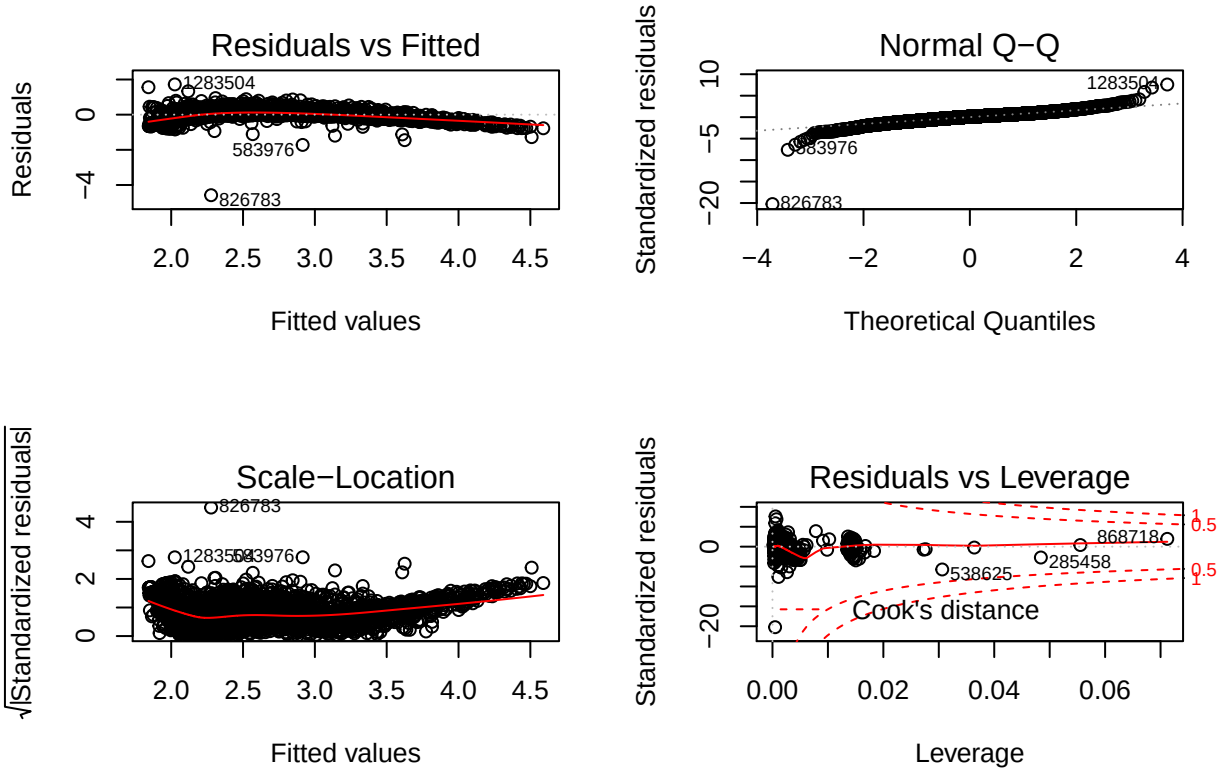
```
## Anova Table (Type II tests)
##
## Response: log(Total_amount)
```

```
##                   Sum Sq   Df    F value    Pr(>F)
## trip_distance_km 700.18    1 13665.5258 < 2.2e-16 ***
## Extra              7.78    1   151.7652 < 2.2e-16 ***
## Tip_amount        57.59    1  1123.9848 < 2.2e-16 ***
## Tolls_amount       0.51    1     9.9184  0.001646 **
## Residuals        249.06 4861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Observem que totes les variables s?n significant ja que tenen el p-valor inferir a 0.05

# apliquem l'exponencial per poder treure el logaritme aplicat anteriorment
# i aix? poder fer el predict de forma correcta.
vif(m25)
```

```
## trip_distance_km            Extra       Tip_amount     Tolls_amount
##         1.208775         1.003027         1.186837         1.049052
```

```
# No hi ha colinealitat entre les variables

par(mfrow=c(2,2))
plot(m25)
```




```
par(mfrow=c(1,1))
```

Creem el model m26.

```
# No cal transformar, pero cal termes de ordre superior
m26 <-lm(Total_amount~poly(trip_distance_km,2)
```

```
       +Extra+poly(Tip_amount,2)+Tolls_amount,data=df)
summary(m26)
```

```
##
## Call:
## lm(formula = Total_amount ~ poly(trip_distance_km, 2) + Extra +
##     poly(Tip_amount, 2) + Tolls_amount, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.158  -0.974  -0.316   0.478  35.580
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  13.07125    0.04879 267.924   <2e-16 ***
## poly(trip_distance_km, 2)1 449.41001    2.68540 167.353   <2e-16 ***
## poly(trip_distance_km, 2)2 -21.91576    2.45047  -8.943   <2e-16 ***
## Extra                        1.03055    0.09658  10.670   <2e-16 ***
## poly(Tip_amount, 2)1       137.04531    2.65556  51.607   <2e-16 ***
## poly(Tip_amount, 2)2        -5.94738    2.45251  -2.425   0.0153 *
## Tolls_amount                 0.79029    0.05243  15.073   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.436 on 4859 degrees of freedom
## Multiple R-squared:  0.9063, Adjusted R-squared:  0.9062
## F-statistic:  7833 on 6 and 4859 DF,  p-value: < 2.2e-16
```

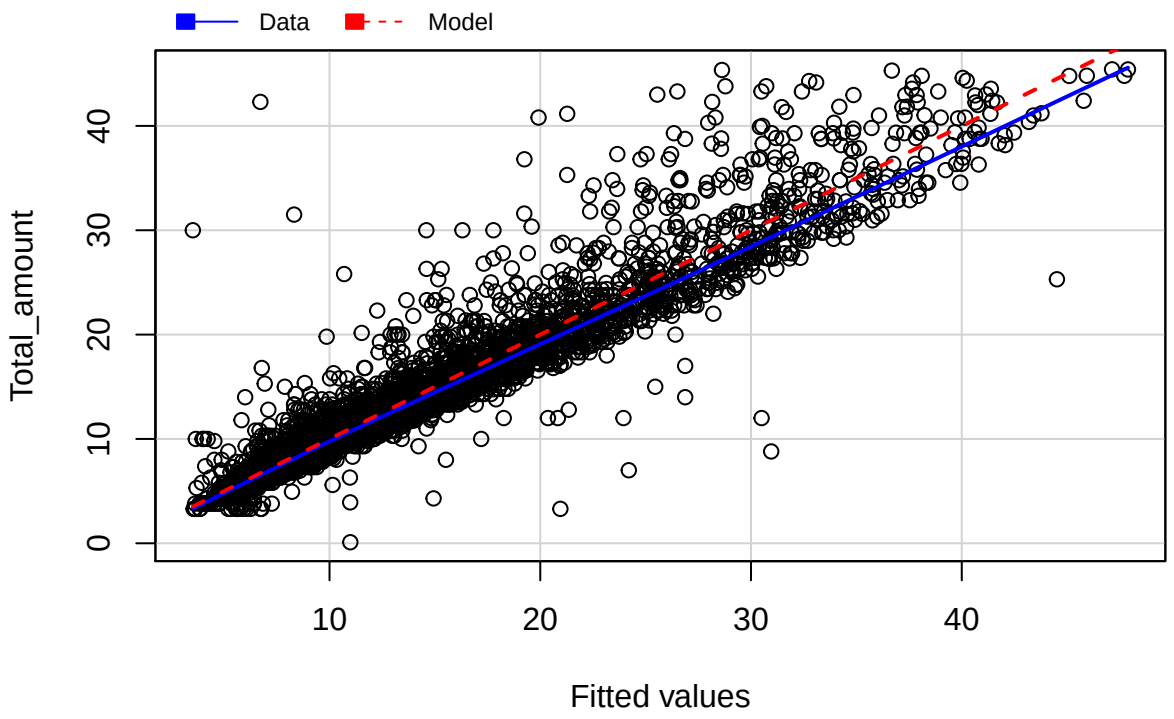
```
Anova(m26)
```

```
## Anova Table (Type II tests)
##
## Response: Total_amount
##                            Sum Sq   Df  F value    Pr(>F)
## poly(trip_distance_km, 2) 166762    2 14047.78 < 2.2e-16 ***
## Extra                        676    1   113.85 < 2.2e-16 ***
## poly(Tip_amount, 2)        15812    2  1331.98 < 2.2e-16 ***
## Tolls_amount                1349    1   227.20 < 2.2e-16 ***
## Residuals                  28841 4859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

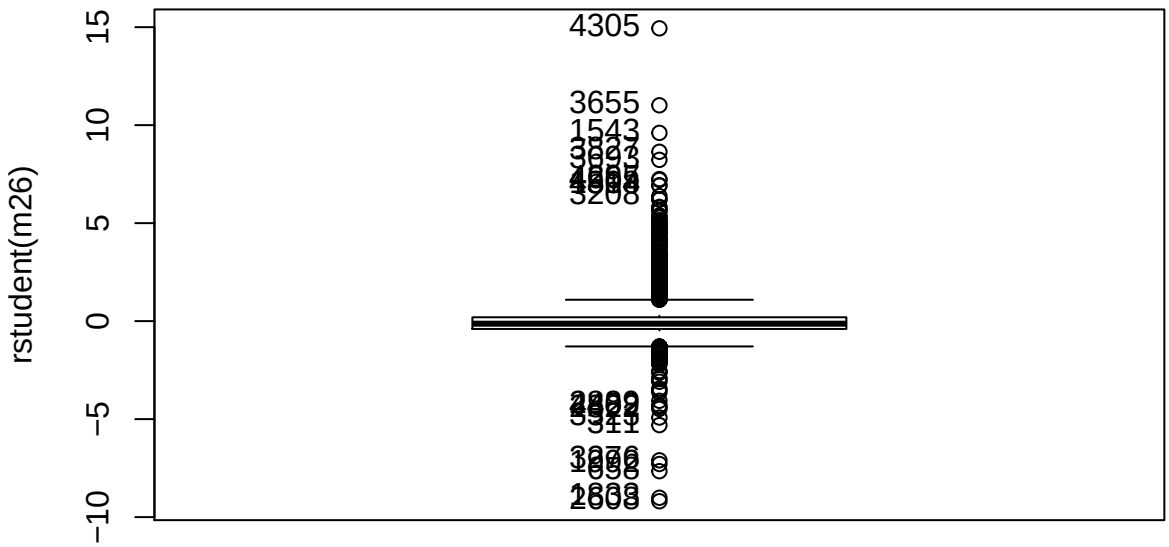
```
anova(m26,m23)
```

```
## Analysis of Variance Table
##
## Model 1: Total_amount ~ poly(trip_distance_km, 2) + Extra + poly(Tip_amount,
##     2) + Tolls_amount
## Model 2: Total_amount ~ trip_distance_km + Fare_amount + Extra + MTA_tax +
##     Tip_amount + Tolls_amount
##   Res.Df     RSS Df Sum of Sq F Pr(>F)
## 1   4859 28840.7
## 2   4859     0.9  0     28840
```

```
marginalModelPlot(m26) #EXTRA Y Tolls_amount com a variables factors
```




```
Boxplot(rstudent(m26))
```




```
##  [1] 2608 1833  658 1992 3276  311 3323 2822 4402 2889 4305 3655 1543 3827
## [15] 3693 4297 4665 4504 1818 3208
```

```
llout<-which(abs(rstudent(m26))>5); length(llout)
```

```
## [1] 30
```

Using factors as explanatory variables

Com hem vist un comportament extrany en les variables Extra i Tolls_amount probarem a utilitzarel com a factors.

Extra

```
df$f.extra<-0
df$f.extra[df$Extra>0]<-1
df$f.extra[df$Extra>0.5]<-2
m27<-lm(Total_amount~poly(trip_distance_km,2)
        +f.extra+poly(Tip_amount,2)+Tolls_amount,data=df)
summary(m27)
```

```
##
## Call:
## lm(formula = Total_amount ~ poly(trip_distance_km, 2) + f.extra +
##     poly(Tip_amount, 2) + Tolls_amount, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -22.160  -0.974  -0.316   0.477  35.580
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.07188    0.04882 267.754   <2e-16 ***
## poly(trip_distance_km, 2)1 449.44363    2.68578 167.342   <2e-16 ***
## poly(trip_distance_km, 2)2 -21.92484    2.45063  -8.947   <2e-16 ***
## f.extra                      0.51468    0.04839  10.637   <2e-16 ***
## poly(Tip_amount, 2)1       137.02420    2.65583  51.594   <2e-16 ***
## poly(Tip_amount, 2)2        -5.93981    2.45271  -2.422   0.0155 *
## Tolls_amount                 0.79019    0.05243  15.070   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.436 on 4859 degrees of freedom
## Multiple R-squared:  0.9063, Adjusted R-squared:  0.9062
## F-statistic:  7831 on 6 and 4859 DF,  p-value: < 2.2e-16
```

```
Anova(m27)
```

```
## Anova Table (Type II tests)
##
## Response: Total_amount
##                            Sum Sq   Df  F value    Pr(>F)
## poly(trip_distance_km, 2) 166764    2 14045.99 < 2.2e-16 ***
## f.extra                      672    1   113.15 < 2.2e-16 ***
## poly(Tip_amount, 2)        15806    2  1331.29 < 2.2e-16 ***
## Tolls_amount                1348    1   227.11 < 2.2e-16 ***
## Residuals                  28845 4859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

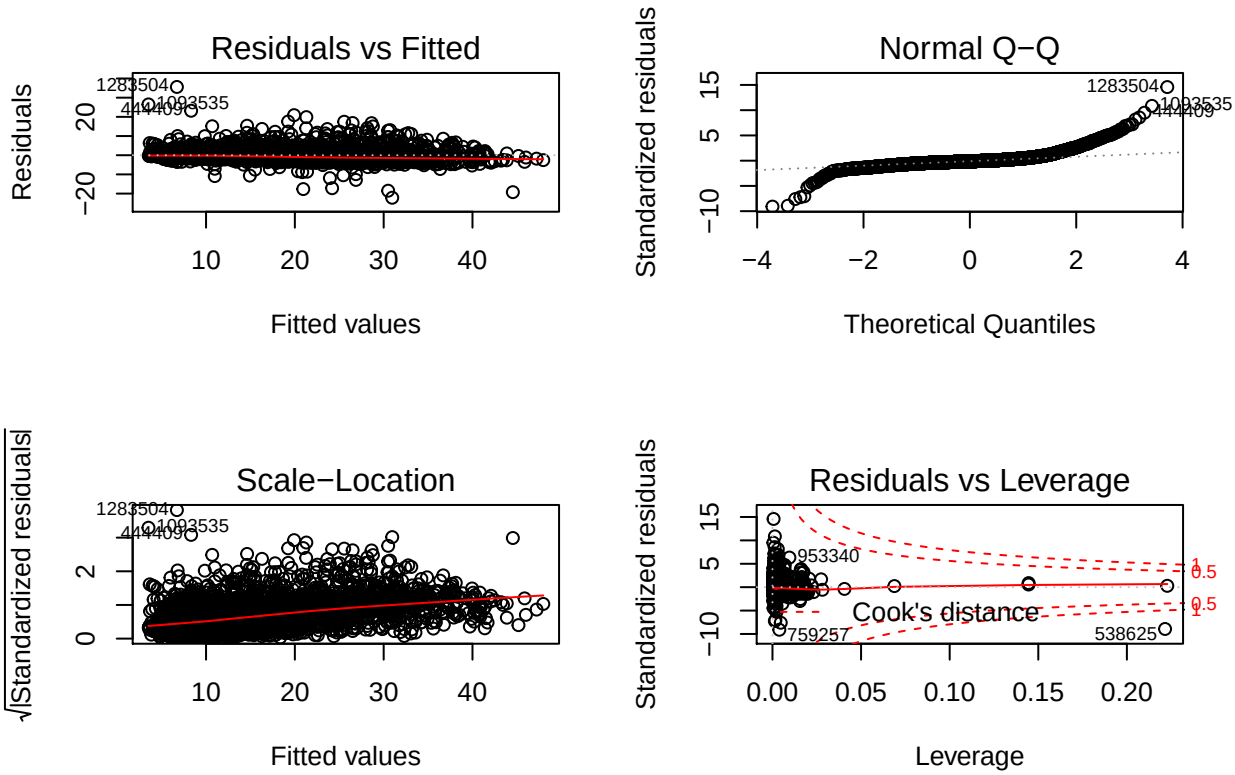
Tolls_amount com a factor no ens d?na una millora, ja que en el moment de fer el BIC amb el model m27 observem que el model m37 ?s m?s gran.

```
m37<-lm(Total_amount~poly(trip_distance_km,2)
        +f.extra+f.toll+poly(Tip_amount,2),data=df)
summary(m37)
```

```
##
## Call:
## lm(formula = Total_amount ~ poly(trip_distance_km, 2) + f.extra +
##     f.toll + poly(Tip_amount, 2), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.162  -0.975  -0.315   0.477  35.582
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.07192    0.04894 267.092   <2e-16 ***
## poly(trip_distance_km, 2)1 449.38409    2.69507 166.743   <2e-16 ***
## poly(trip_distance_km, 2)2 -21.75364    2.45545  -8.859   <2e-16 ***
## f.extra                      0.51626    0.04849  10.647   <2e-16 ***
## f.toll(1,50]                 4.42982    0.30861  14.354   <2e-16 ***
```

```
## poly(Tip_amount, 2)1       137.39996    2.66008  51.652   <2e-16 ***
## poly(Tip_amount, 2)2        -6.01048    2.45776  -2.446   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.442 on 4859 degrees of freedom
## Multiple R-squared:  0.9059, Adjusted R-squared:  0.9058
## F-statistic:  7796 on 6 and 4859 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m37)
```




```
par(mfrow=c(1,1))
BIC(m27,m37)
```

```
##     df      BIC
## m27  8 22536.84
## m37  8 22557.04
```

Creem el model m30

```
m30<-lm(log(Total_amount)~trip_distance_km+Tip_amount+(f.extra+f.passenger+Payment_type+Tolls_amount)*(
m30<-step(m30,k=log(nrow(df)))
```

```
## Start:  AIC=-18118.91
## log(Total_amount) ~ trip_distance_km + Tip_amount + (f.extra +
##     f.passenger + Payment_type + Tolls_amount) * (f.espeed) +
##     f.ttime
```

```
##
##                          Df Sum of Sq    RSS    AIC
## - f.extra:f.espeed        3     0.017 111.72 -18144
## - f.passenger:f.espeed    3     0.021 111.72 -18144
## - Payment_type:f.espeed   6     0.777 112.48 -18136
## - Tolls_amount:f.espeed   3     0.254 111.96 -18133
## <none>                                111.70 -18119
## - Tip_amount              1    12.577 124.28 -17608
## - trip_distance_km        1   101.970 213.68 -14971
## - f.ttime                 3   109.262 220.97 -14825
##
## Step:  AIC=-18143.62
## log(Total_amount) ~ trip_distance_km + Tip_amount + f.extra +
##     f.passenger + Payment_type + Tolls_amount + f.espeed + f.ttime +
##     f.passenger:f.espeed + Payment_type:f.espeed + Tolls_amount:f.espeed
##
##                          Df Sum of Sq    RSS    AIC
## - f.passenger:f.espeed    3     0.022 111.74 -18168
## - Payment_type:f.espeed   6     0.781 112.50 -18161
## - Tolls_amount:f.espeed   3     0.255 111.98 -18158
## <none>                                111.72 -18144
## - f.extra                 1     6.533 118.25 -17876
## - Tip_amount              1    12.596 124.32 -17632
## - trip_distance_km        1   101.964 213.69 -14996
## - f.ttime                 3   109.281 221.00 -14850
##
## Step:  AIC=-18168.15
## log(Total_amount) ~ trip_distance_km + Tip_amount + f.extra +
##     f.passenger + Payment_type + Tolls_amount + f.espeed + f.ttime +
##     Payment_type:f.espeed + Tolls_amount:f.espeed
##
##                          Df Sum of Sq    RSS    AIC
## - Payment_type:f.espeed   6     0.779 112.52 -18185
## - Tolls_amount:f.espeed   3     0.251 112.00 -18183
## - f.passenger             1     0.023 111.77 -18176
## <none>                                111.74 -18168
## - f.extra                 1     6.552 118.30 -17899
## - Tip_amount              1    12.596 124.34 -17657
## - trip_distance_km        1   102.058 213.80 -15019
## - f.ttime                 3   109.279 221.02 -14875
##
## Step:  AIC=-18185.27
## log(Total_amount) ~ trip_distance_km + Tip_amount + f.extra +
##     f.passenger + Payment_type + Tolls_amount + f.espeed + f.ttime +
##     Tolls_amount:f.espeed
##
##                          Df Sum of Sq    RSS    AIC
## - Tolls_amount:f.espeed   3     0.285 112.81 -18198
## - f.passenger             1     0.021 112.54 -18193
## <none>                                112.52 -18185
## - Payment_type            2     6.561 119.08 -17927
## - f.extra                 1     6.558 119.08 -17918
## - Tip_amount              1    12.673 125.19 -17675
## - trip_distance_km        1   102.077 214.60 -15052
```

```
## - f.ttime                3   109.479 222.00 -14904
##
## Step:  AIC=-18198.44
## log(Total_amount) ~ trip_distance_km + Tip_amount + f.extra +
##     f.passenger + Payment_type + Tolls_amount + f.espeed + f.ttime
##
##                    Df Sum of Sq    RSS    AIC
## - f.espeed          3     0.409 113.22 -18206
## - f.passenger       1     0.025 112.83 -18206
## <none>                          112.81 -18198
## - Tolls_amount      1     2.263 115.07 -18110
## - Payment_type      2     6.592 119.40 -17939
## - f.extra           1     6.494 119.30 -17935
## - Tip_amount        1    12.670 125.48 -17689
## - trip_distance_km  1   102.024 214.83 -15072
## - f.ttime           3   110.380 223.19 -14904
##
## Step:  AIC=-18206.28
## log(Total_amount) ~ trip_distance_km + Tip_amount + f.extra +
##     f.passenger + Payment_type + Tolls_amount + f.ttime
##
##                    Df Sum of Sq    RSS    AIC
## - f.passenger       1     0.025 113.24 -18214
## <none>                          113.22 -18206
## - Tolls_amount      1     2.272 115.49 -18118
## - Payment_type      2     6.520 119.74 -17951
## - f.extra           1     6.575 119.79 -17940
## - Tip_amount        1    12.944 126.16 -17688
## - trip_distance_km  1   121.612 234.83 -14665
## - f.ttime           3   129.212 242.43 -14527
##
## Step:  AIC=-18213.68
## log(Total_amount) ~ trip_distance_km + Tip_amount + f.extra +
##     Payment_type + Tolls_amount + f.ttime
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          113.24 -18214
## - Tolls_amount      1     2.266 115.51 -18126
## - Payment_type      2     6.515 119.76 -17959
## - f.extra           1     6.550 119.79 -17949
## - Tip_amount        1    12.947 126.19 -17695
## - trip_distance_km  1   121.661 234.90 -14672
## - f.ttime           3   129.330 242.57 -14532
```

```
vif(m30)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## trip_distance_km 2.596548  1        1.611381
## Tip_amount       2.012620  1        1.418668
## f.extra          1.003942  1        1.001969
## Payment_type     1.731343  2        1.147085
## Tolls_amount     1.055181  1        1.027220
## f.ttime          2.451407  3        1.161188
```

```
summary(m30)
```

```
##
## Call:
## lm(formula = log(Total_amount) ~ trip_distance_km + Tip_amount +
##     f.extra + Payment_type + Tolls_amount + f.ttime, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2492 -0.0641 -0.0008  0.0693  1.8568
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.774813   0.006395 277.529  < 2e-16 ***
## trip_distance_km    0.075598   0.001047  72.229  < 2e-16 ***
## Tip_amount          0.040721   0.001728  23.562  < 2e-16 ***
## f.extra             0.050821   0.003032  16.759  < 2e-16 ***
## Payment_typeCash   -0.095808   0.005767 -16.615  < 2e-16 ***
## Payment_typeOther  -0.118338   0.026687  -4.434 9.44e-06 ***
## Tolls_amount        0.032418   0.003289   9.858  < 2e-16 ***
## f.ttime(6,9.78]     0.265617   0.006314  42.068  < 2e-16 ***
## f.ttime(9.78,15.7]  0.459280   0.006943  66.149  < 2e-16 ***
## f.ttime(15.7,215]   0.635626   0.009297  68.369  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1527 on 4856 degrees of freedom
## Multiple R-squared:  0.9169, Adjusted R-squared:  0.9168
## F-statistic:  5954 on 9 and 4856 DF,  p-value: < 2.2e-16
```

```
Anova(m30)
```

```
## Anova Table (Type II tests)
##
## Response: log(Total_amount)
##                    Sum Sq   Df  F value    Pr(>F)
## trip_distance_km 121.661    1 5216.994 < 2.2e-16 ***
## Tip_amount        12.947    1  555.183 < 2.2e-16 ***
## f.extra            6.550    1  280.866 < 2.2e-16 ***
## Payment_type       6.515    2  139.693 < 2.2e-16 ***
## Tolls_amount       2.266    1   97.176 < 2.2e-16 ***
## f.ttime          129.330    3 1848.623 < 2.2e-16 ***
## Residuals        113.242 4856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

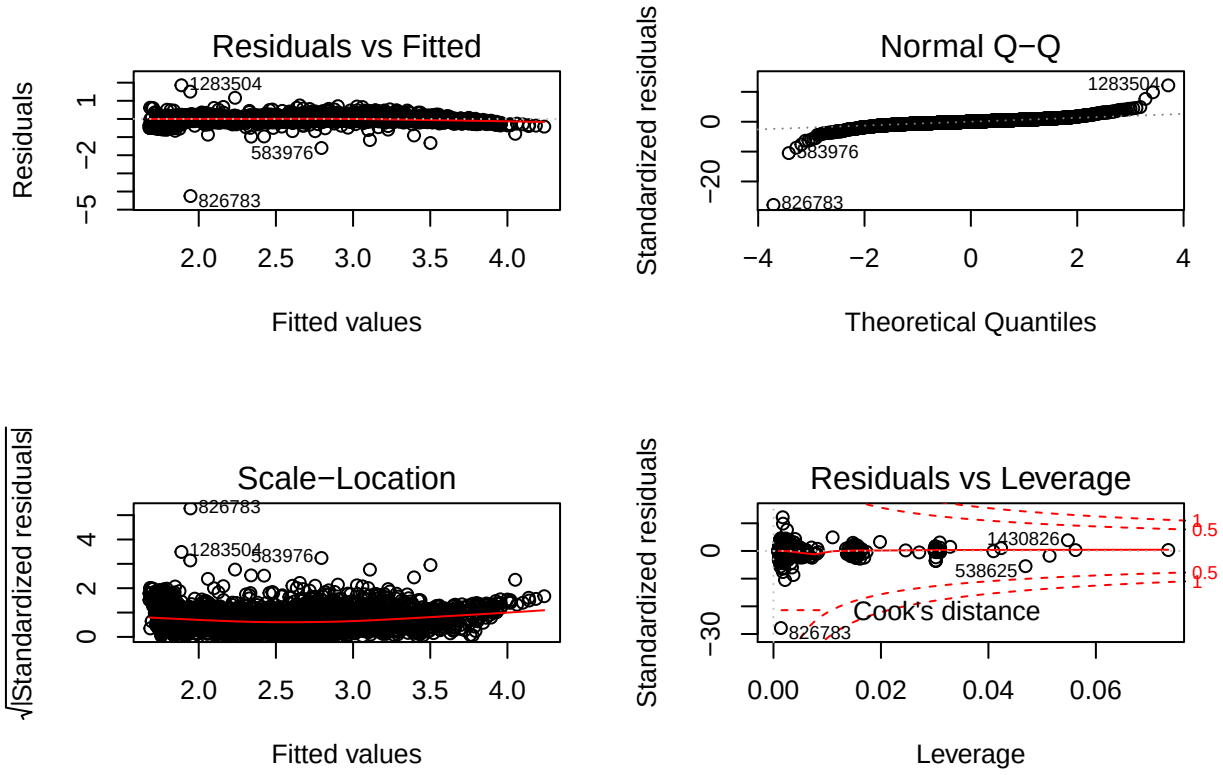
```
BIC(m30)
```

```
## [1] -4396.08
```

RESIDUAL ANALYSIS

Tenim el model m30 com a millor model.

```
# Residual Analysis
par(mfrow=c(2,2))
plot(m30)
```




```
par(mfrow=c(1,1))
```

RESIDUAL vs FITTED: En el gr?fic no podem identificar una tend?ncia de la vari?ncia dels residus a reduir-se ni augmentar a mesura que l'adher?ncia augmenta. Observem que la l?nia vermella ?s pr?cticament una l?nia recta. Tamb? observem uns punts que tenen algun valor estrany.

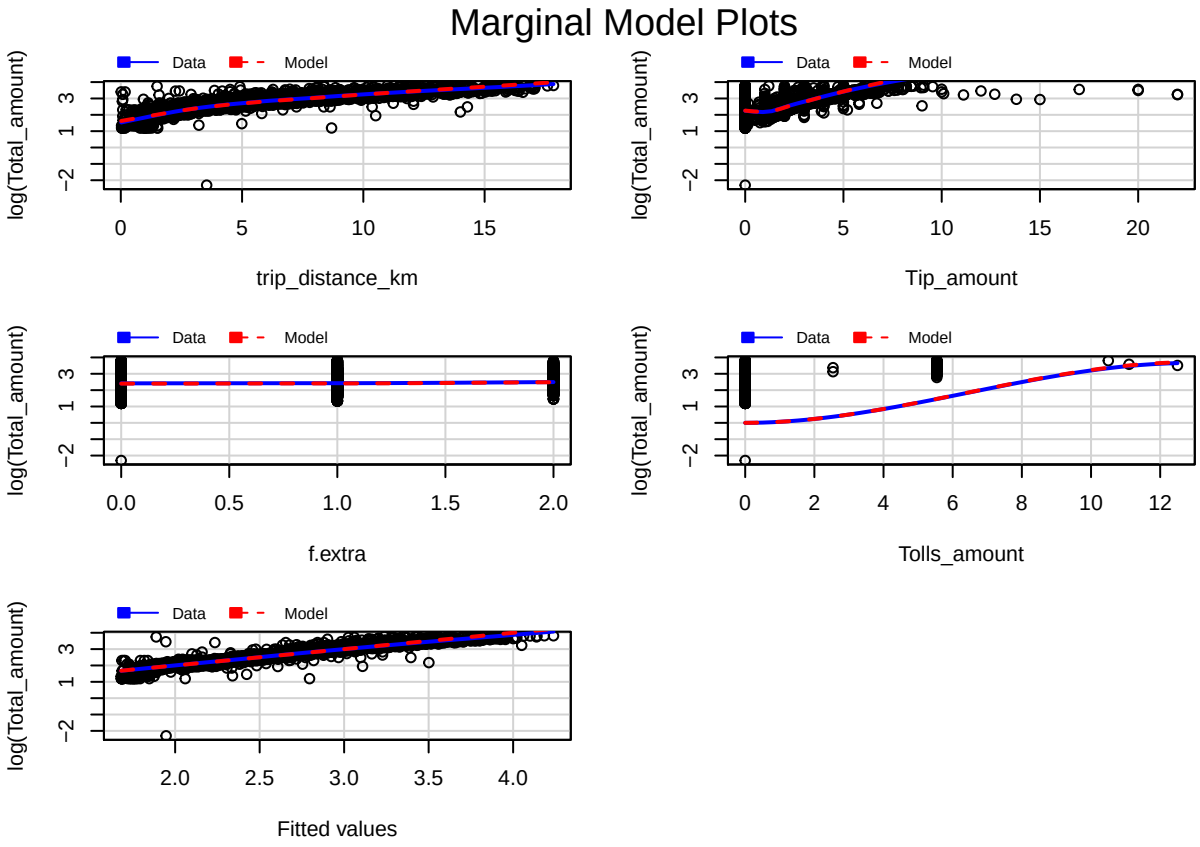
NORMAL Q-Q: Aquest grafic ens mostra les diferencies entre la distribuci? de probabilitat de la nostra poblaci? de la qual hem tret la mostra i la nostra distribuci? usada per la comparaci?. En el nostre grafic observem que en les puntes hi ha una separaci? marcada respecte a la distribuci? de la poblaci?, deguda a outliers.

SCALE-LOCATION: Aquest gr?fic ens mostra si els residus es distribueixen per igual al llarg del rang dels predictors. Aix? ?s com podem verificar la suposici? d'igual varian?a (homocedasticitat). Observem una l?nia horitzontal que ens diu que quan m?s augmenta Fitted values els residus augmenten.

RESIDUALS VS LEVERAGE: Aquest gr?fic ens ajuda a trobar casos influents si n'hi ha. No tots els outliers s?n influents en l'an?lisi de regressi? lineal. Encara que algunes dades tenen valors extrems, ?s possible que no siguin influents per determinar una l?nia de regressi?. Aix? significa que els resultats no serien molt diferents si els incloem o els exclouen de l'an?lisi. Segueixen la tend?ncia en la majoria dels casos i en realitat no els importa; no s?n influents. D'altra banda, alguns casos podrien ser molt influents, fins i tot si semblen estar dins d'un rang raonable dels valors. Poden ser casos extrems contra una l?nia de regressi? i poden alterar els resultats si els exclouen de l'an?lisi. Una altra forma de plantejar-la ?s que no s'acompanyen amb la tend?ncia en la majoria dels casos.

```
marginalModelPlots(m30)
```

```
## Warning in mmps(...): Interactions and/or factors skipped
```

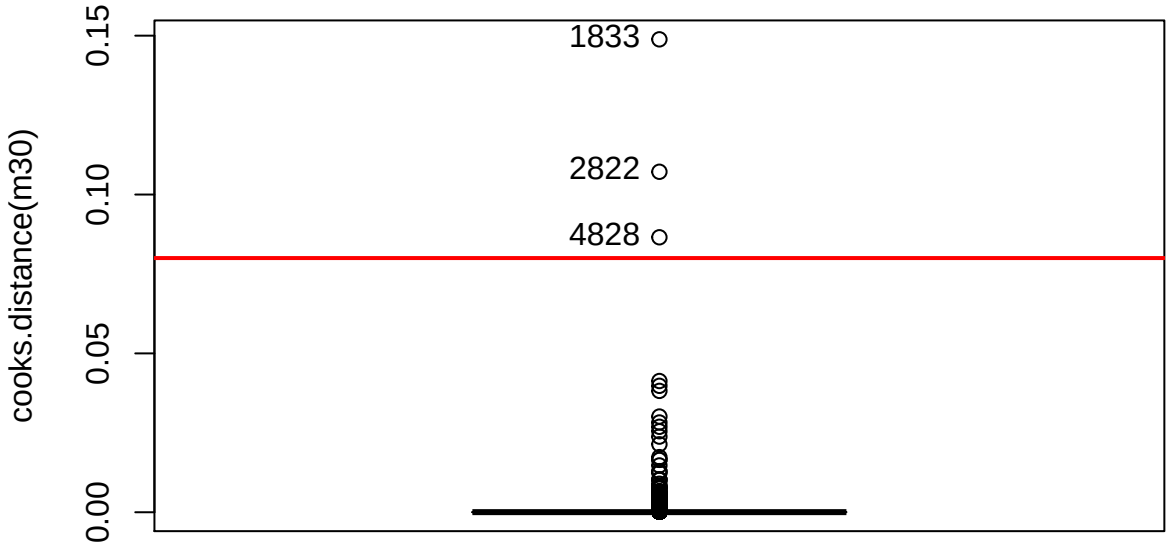


INFLUENT DATA

```
Boxplot(cooks.distance(m30),id.n=3)
```

```
## [1] 1833 2822 4828
```

```
abline(h=0.08,col="red",lwd=2)
```




```
llcoo<-which(cooks.distance(m30)>0.08);length(llcoo)
```

```
## [1] 3
```

```
df[llcoo,]
```

```
##              VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 538625  VeriFone Inc.  2016-01-12 11:12:03   2016-01-12 11:42:59
## 826783  VeriFone Inc.  2016-01-18 02:00:51   2016-01-19 01:06:50
## 1430826 VeriFone Inc.  2016-01-31 17:41:01   2016-01-31 17:42:43
##         Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 538625       Store_and_fwd Standard rate        -73.99138        40.69322
## 826783       Store_and_fwd  Special rate        -73.81512        40.75410
## 1430826      Store_and_fwd Standard rate        -73.85342        40.87118
##         Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 538625          -73.94975         40.68033               1      6.118366
## 826783          -73.81516         40.75413               1      2.199552
## 1430826         -73.85247         40.87116               1      0.050000
##         Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 538625          2.5     0     0.5         22            0
## 826783          0.1     0     0.0          0            0
## 1430826         3.0     0     0.5         22            0
##         improvement_surcharge Total_amount Payment_type   Trip_type
## 538625                    0.3         25.3  Credit card Street-hail
## 826783                    0.0          0.1         Cash    Dispatch
## 1430826                   0.3         25.8  Credit card Street-hail
##         mis_ind     AnyTip trip_length trip_distance_km travel_time
```

```
## 538625        3 AnyTip Yes   6.0631572        9.8465564    30.93333
## 826783        7  AnyTip No   2.8410192        3.5398359     5.89607
## 1430826       3 AnyTip Yes   0.1081646        0.0804672     1.70000
##         pick_up_hour pick_up_period    espeed f.passenger  f.distance
## 538625            11        morning 11.760434       (0,1] (3.31,11.5]
## 826783             2          night 28.910980       (0,1]  (1.8,3.31]
## 1430826           17         valley  3.817575       (0,1]    (0,1.01]
##         f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 538625      (-74.1,-73.96]       (40.5,40.7]     (-73.97,-73.94]
## 826783     (-73.92,-73.79]      (40.74,40.8]     (-73.91,-73.75]
## 1430826    (-73.92,-73.79]      (40.8,40.92]     (-73.91,-73.75]
##         f.dropoff_latitude f.fare_amount f.extra  f.MTA_tax
## 538625        (40.57,40.7]         (0,6]       0  (0.4,0.5]
## 826783       (40.75,40.79]         (0,6]       0 (-0.1,0.4]
## 1430826      (40.79,40.92]         (0,6]       0  (0.4,0.5]
##         f.Improvement_surcharge f.tip_amount f.toll   f.total    f.ttime
## 538625                (0.1,0.8]       (1,22] (-1,1] (16.6,46] (15.7,215]
## 826783               (-0.1,0.1]     (-0.1,1] (-1,1]  (-1,7.8]     (-1,6]
## 1430826               (0.1,0.8]       (1,22] (-1,1] (16.6,46]     (-1,6]
##          f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 538625   (0,15.3]      NoOutDim1      NoOutDim2      NoOutDim3
## 826783  (26.2,95]      NoOutDim1     YesOutDim2      NoOutDim3
## 1430826  (0,15.3]      NoOutDim1      NoOutDim2      NoOutDim3
##         f.outlierPCAd4 f.outlierPCA  claHP claKM
## 538625       NoOutDim4        NoOut kHP- 6 kKM-6
## 826783       NoOutDim4       YesOut kHP- 7 kKM-7
## 1430826      NoOutDim4        NoOut kHP- 1 kKM-3
```

```
df<-df[-llout]
```

Report: En les observacions que tenen una dist?ncia de Cook m?s gran a 0.8 observem que tenen una relaci? molt peculiar entre el total pagat i la dist?ncia del viatge. Observem que hi ha distancies molt petites amb un preu pagat total molt elevat. (fixar-se en Trip_distance i Total_amount)

Previous work

Load requiered packages

Statistical Modelling

Load your sample after data cleaning and validation

```
load("Taxi5000_raw_DataDefinitivev1.RData")
#load("C:/Users/Sergi/Desktop/Sergi/ADEI/Taxi5000_raw_DataDefinitivev1.RData")

names(df)
```

```
##  [1] "VendorID"               "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"  "Store_and_fwd_flag"
##  [5] "RateCodeID"             "Pickup_longitude"
##  [7] "Pickup_latitude"        "Dropoff_longitude"
##  [9] "Dropoff_latitude"       "Passenger_count"
```

```
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"          "f.outlierPCAd3"
## [47] "f.outlierPCAd4"          "f.outlierPCA"
## [49] "claHP"                   "claKM"
```

```
#summary(df)
#vars_con
#vars_dis
#vars_res
```

Split data 70-30

In order to build the model, and test it afterwords, we will generate work and test samples (consisting on a 70-30 split).

```
llwork<-sample(1:nrow(df),0.70*nrow(df),replace=FALSE)
llwork<-sort(llwork);length(llwork)
```

```
## [1] 3406
```

```
dfwork<-df[llwork,]
dftest<-df[-llwork,]
```

Selecting the best model

We start modelling and executing some basic tests of Bayesian information criterion (deviance) and colinearity to achieve the better suited model for our binary target. In first place, we take a look at catdes and cor outputs to grasp those numerical and discrete coefficients which explain better our targey Any Tip. In vars_cexp we keep the continuous explicative variables to look at their correlation table so we can decide wether to try to combine them or avoid it in order to don't carry colinearity in our model.

```
names(dfwork)
```

```
##  [1] "VendorID"               "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"  "Store_and_fwd_flag"
##  [5] "RateCodeID"             "Pickup_longitude"
##  [7] "Pickup_latitude"        "Dropoff_longitude"
##  [9] "Dropoff_latitude"       "Passenger_count"
```

```
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.ttime"
## [43] "f.espeed"                "f.outlierPCAd1"
## [45] "f.outlierPCAd2"          "f.outlierPCAd3"
## [47] "f.outlierPCAd4"          "f.outlierPCA"
## [49] "claHP"                   "claKM"
```

```
catdes(df,num.var=22)
```

```
## $test.chi2
##                               p.value df
## Payment_type             0.000000e+00  2
## f.tip_amount             0.000000e+00  1
## f.total                 4.794022e-105  3
## claKM                    1.981664e-77  6
## claHP                    5.819127e-72  6
## f.dropoff_longitude      1.497728e-67  3
## f.pickup_longitude       1.659845e-55  3
## f.outlierPCAd3           2.566411e-31  1
## f.outlierPCAd4           2.566411e-31  1
## f.dropoff_latitude       1.544341e-23  3
## f.distance               5.988343e-22  3
## f.pickup_latitude        2.235660e-19  3
## f.fare_amount            7.739366e-19  3
## f.ttime                  4.002046e-18  3
## f.MTA_tax                2.401669e-09  1
## f.outlierPCA             4.708698e-09  1
## f.outlierPCAd2           6.816601e-09  1
## Trip_type                6.988488e-09  1
## RateCodeID               9.277523e-09  1
## f.Improvement_surcharge  1.318599e-08  1
## f.toll                   6.355993e-06  1
## pick_up_period           5.569057e-04  3
##
## $category
## $category$`AnyTip No`
##                                      Cla/Mod     Mod/Cla     Global
## f.tip_amount=(-0.1,1]               93.20420 100.0000000 62.5976161
## Payment_type=Cash                  100.00000  86.2275449 50.3082614
## f.total=(-1,7.8]                    82.20271  36.2803804 25.7501028
## f.pickup_longitude=(-73.92,-73.79]  74.12531  32.0887636 25.2568845
```

```
## f.dropoff_longitude=(-73.91,-73.75]  73.99142  30.3628038 23.9416358
## claKM=kKM-2                          79.00147  18.9503346 13.9950678
## claHP=kHP- 2                         77.51323  20.6410708 15.5363748
## f.dropoff_latitude=(40.79,40.92]     70.75472  29.0595280 23.9621866
## f.ttime=(-1,6]                       67.86590  29.2356464 25.1335799
## f.distance=(0,1.01]                  67.45707  29.0595280 25.1335799
## f.dropoff_longitude=(-73.94,-73.91]  67.38411  28.6720676 24.8253185
## f.fare_amount=(0,6]                  66.82654  29.4469884 25.7090012
## claKM=kKM-7                          95.08197   2.0429729  1.2535964
## claHP=kHP- 7                         95.08197   2.0429729  1.2535964
## f.outlierPCA=YesOut                  95.08197   2.0429729  1.2535964
## f.MTA_tax=(-0.1,0.4]                 90.36145   2.6417753  1.7057131
## f.outlierPCAd2=YesOutDim2            95.00000   2.0077492  1.2330456
## claHP=kHP- 1                         65.04599  34.8714336 31.2782573
## claKM=kKM-3                          66.96252  23.9168721 20.8384710
## Trip_type=Dispatch                   90.00000   2.5361043  1.6440608
## RateCodeID=Special rate              89.15663   2.6065516  1.7057131
## f.Improvement_surcharge=(-0.1,0.1]   89.02439   2.5713279  1.6851624
## f.pickup_latitude=(40.74,40.8]       65.15495  30.3628038 27.1886560
## Payment_type=Other                  100.00000   1.1976048  0.6987259
## f.pickup_latitude=(40.8,40.92]       64.49704  26.8756604 24.3115495
## f.toll=(-1,1]                        58.72057  99.2603029 98.6230991
## f.pickup_longitude=(-73.95,-73.92]   63.49960  28.2493836 25.9556104
## pick_up_period=valley                62.76518  30.2219091 28.0928894
## claKM=kKM-1                          64.35786  15.7097570 14.2416769
## f.distance=(1.01,1.8]                62.35679  26.8404368 25.1130292
## f.fare_amount=(6,9]                  61.08453  26.9813315 25.7706535
## f.ttime=(9.78,15.7]                  55.90164  24.0225431 25.0719277
## f.fare_amount=(9,14]                 55.86035  23.6703064 24.7225647
## f.espeed=(20.1,26.2]                 55.76451  24.0225431 25.1335799
## pick_up_period=morning               55.06653  18.9503346 20.0780929
## f.distance=(1.8,3.31]                54.60581  23.1771751 24.7636663
## f.dropoff_latitude=(40.57,40.7]      54.69484  24.6213455 26.2638718
## f.toll=(1,50]                        31.34328   0.7396971  1.3769009
## claKM=kKM-5                          51.86413  22.0500176 24.8047678
## f.pickup_latitude=(40.7,40.74]       51.49123  20.6762945 23.4278668
## f.total=(11,16.6]                    51.66113  21.9091229 24.7431155
## f.pickup_latitude=(40.5,40.7]        51.39344  22.0852413 25.0719277
## f.Improvement_surcharge=(0.1,0.8]    57.81773  97.4286721 98.3148376
## RateCodeID=Standard rate             57.80891  97.3934484 98.2942869
## f.dropoff_latitude=(40.7,40.75]      51.15931  24.0929905 27.4763666
## Trip_type=Street-hail                57.81446  97.4638957 98.3559392
## f.outlierPCAd2=NoOutDim2             57.88598  97.9922508 98.7669544
## f.MTA_tax=(0.4,0.5]                  57.78800  97.3582247 98.2942869
## f.outlierPCA=NoOut                   57.87721  97.9570271 98.7464036
## f.ttime=(15.7,215]                   50.00000  21.3807679 24.9486231
## f.dropoff_longitude=(-73.97,-73.94]  50.07267  24.2691088 28.2778463
## f.fare_amount=(14,43]                48.79102  19.9013737 23.7977805
## f.pickup_longitude=(-73.96,-73.95]   48.61111  19.7252554 23.6744760
## f.distance=(3.31,11.5]               48.84868  20.9228602 24.9897246
## claHP=kHP- 3                         49.20962  25.2201479 29.9013564
## claKM=kKM-6                          38.35616   5.9175766  9.0012330
## f.pickup_longitude=(-74.1,-73.96]    46.31751  19.9365974 25.1130292
## claKM=kKM-4                          41.96891  11.4124692 15.8651870
```

```
## claHP=kHP- 6                            39.67862  11.3067982 16.6255651
## f.dropoff_longitude=(-74.1,-73.97]      42.43509  16.6960197 22.9551993
## f.total=(16.6,46]                       39.93453  17.1891511 25.1130292
## f.tip_amount=(1,22]                      0.00000   0.0000000 37.4023839
## Payment_type=Credit card                14.97483  12.5748503 48.9930127
##                                             p.value     v.test
## f.tip_amount=(-0.1,1]                  0.000000e+00        Inf
## Payment_type=Cash                      0.000000e+00        Inf
## f.total=(-1,7.8]                       8.443079e-95  20.657011
## f.pickup_longitude=(-73.92,-73.79]     5.353804e-40  13.237156
## f.dropoff_longitude=(-73.91,-73.75]    8.652981e-37  12.670167
## claKM=kKM-2                            2.600995e-34  12.214499
## claHP=kHP- 2                           3.744050e-33  11.995663
## f.dropoff_latitude=(40.79,40.92]       1.569737e-23   9.997069
## f.ttime=(-1,6]                         3.273821e-15   7.880000
## f.distance=(0,1.01]                    4.793014e-14   7.537436
## f.dropoff_longitude=(-73.94,-73.91]    1.211657e-13   7.415501
## f.fare_amount=(0,6]                    1.106603e-12   7.116552
## claKM=kKM-7                            6.469500e-11   6.532467
## claHP=kHP- 7                           6.469500e-11   6.532467
## f.outlierPCA=YesOut                    6.469500e-11   6.532467
## f.MTA_tax=(-0.1,0.4]                   9.937837e-11   6.467894
## f.outlierPCAd2=YesOutDim2              1.065158e-10   6.457402
## claHP=kHP- 1                           1.281134e-10   6.429394
## claKM=kKM-3                            2.752451e-10   6.312116
## Trip_type=Dispatch                     3.826056e-10   6.260964
## RateCodeID=Special rate                6.295785e-10   6.182842
## f.Improvement_surcharge=(-0.1,0.1]     9.751541e-10   6.113425
## f.pickup_latitude=(40.74,40.8]         3.174851e-09   5.922292
## Payment_type=Other                     1.018119e-08   5.727682
## f.pickup_latitude=(40.8,40.92]         7.051390e-07   4.959943
## f.toll=(-1,1]                          8.328194e-06   4.456569
## f.pickup_longitude=(-73.95,-73.92]     1.450835e-05   4.336023
## pick_up_period=valley                  8.786398e-05   3.921871
## claKM=kKM-1                            4.883997e-04   3.487039
## f.distance=(1.01,1.8]                  9.769241e-04   3.297089
## f.fare_amount=(6,9]                    2.213147e-02   2.288104
## f.ttime=(9.78,15.7]                    4.602998e-02  -1.995118
## f.fare_amount=(9,14]                   4.444180e-02  -2.009900
## f.espeed=(20.1,26.2]                   3.482537e-02  -2.110383
## pick_up_period=morning                 2.045298e-02  -2.317933
## f.distance=(1.8,3.31]                  2.478699e-03  -3.025930
## f.dropoff_latitude=(40.57,40.7]        2.117178e-03  -3.073283
## f.toll=(1,50]                          8.328194e-06  -4.456569
## claKM=kKM-5                            1.574083e-07  -5.243676
## f.pickup_latitude=(40.7,40.74]         9.352930e-08  -5.338867
## f.total=(11,16.6]                      6.719751e-08  -5.398508
## f.pickup_latitude=(40.5,40.7]          1.479187e-08  -5.663960
## f.Improvement_surcharge=(0.1,0.8]      9.751541e-10  -6.113425
## RateCodeID=Standard rate               6.295785e-10  -6.182842
## f.dropoff_latitude=(40.7,40.75]        4.620588e-10  -6.231477
## Trip_type=Street-hail                  3.826056e-10  -6.260964
## f.outlierPCAd2=NoOutDim2               1.065158e-10  -6.457402
## f.MTA_tax=(0.4,0.5]                    9.937837e-11  -6.467894
```

```
## f.outlierPCA=NoOut                   6.469500e-11  -6.532467
## f.ttime=(15.7,215]                   1.255269e-11  -6.773703
## f.dropoff_longitude=(-73.97,-73.94] 2.534785e-13  -7.317052
## f.fare_amount=(14,43]                5.794312e-14  -7.512648
## f.pickup_longitude=(-73.96,-73.95]  2.392093e-14  -7.627576
## f.distance=(3.31,11.5]               1.215939e-14  -7.714357
## claHP=kHP- 3                         4.239721e-17  -8.406072
## claKM=kKM-6                          1.208511e-18  -8.813914
## f.pickup_longitude=(-74.1,-73.96]   1.140791e-22  -9.798669
## claKM=kKM-4                          1.837298e-23  -9.981467
## claHP=kHP- 6                         1.211151e-31 -11.704328
## f.dropoff_longitude=(-74.1,-73.97]  2.751384e-34 -12.209926
## f.total=(16.6,46]                    6.840593e-51 -15.004694
## f.tip_amount=(1,22]                  0.000000e+00       -Inf
## Payment_type=Credit card             0.000000e+00       -Inf
##
## $category$`AnyTip Yes`
##                                        Cla/Mod     Mod/Cla     Global
## f.tip_amount=(1,22]                  100.000000  89.7878638 37.4023839
## Payment_type=Credit card              85.025168 100.0000000 48.9930127
## f.total=(16.6,46]                     60.065466  36.2111495 25.1130292
## f.dropoff_longitude=(-74.1,-73.97]    57.564906  31.7217563 22.9551993
## claHP=kHP- 6                          60.321384  24.0749877 16.6255651
## claKM=kKM-4                           58.031088  22.1016280 15.8651870
## f.pickup_longitude=(-74.1,-73.96]     53.682488  32.3630982 25.1130292
## claKM=kKM-6                           61.643836  13.3201776  9.0012330
## claHP=kHP- 3                          50.790378  36.4578194 29.9013564
## f.distance=(3.31,11.5]                51.151316  30.6857425 24.9897246
## f.pickup_longitude=(-73.96,-73.95]    51.388889  29.2057227 23.6744760
## f.fare_amount=(14,43]                 51.208981  29.2550567 23.7977805
## f.dropoff_longitude=(-73.97,-73.94]   49.927326  33.8924519 28.2778463
## f.ttime=(15.7,215]                    50.000000  29.9457326 24.9486231
## f.outlierPCA=NoOut                    42.122789  99.8519980 98.7464036
## f.MTA_tax=(0.4,0.5]                   42.212001  99.6053281 98.2942869
## f.outlierPCAd2=NoOutDim2              42.114024  99.8519980 98.7669544
## Trip_type=Street-hail                 42.185541  99.6053281 98.3559392
## f.dropoff_latitude=(40.7,40.75]       48.840688  32.2150962 27.4763666
## RateCodeID=Standard rate              42.191093  99.5559941 98.2942869
## f.Improvement_surcharge=(0.1,0.8]     42.182274  99.5559941 98.3148376
## f.pickup_latitude=(40.5,40.7]         48.606557  29.2550567 25.0719277
## f.total=(11,16.6]                     48.338870  28.7123828 24.7431155
## f.pickup_latitude=(40.7,40.74]        48.508772  27.2816971 23.4278668
## claKM=kKM-5                           48.135874  28.6630488 24.8047678
## f.toll=(1,50]                         68.656716   2.2693636  1.3769009
## f.dropoff_latitude=(40.57,40.7]       45.305164  28.5643809 26.2638718
## f.distance=(1.8,3.31]                 45.394191  26.9856931 24.7636663
## pick_up_period=morning                44.933470  21.6576221 20.0780929
## f.espeed=(20.1,26.2]                  44.235487  26.6896892 25.1335799
## f.fare_amount=(9,14]                  44.139651  26.1963493 24.7225647
## f.ttime=(9.78,15.7]                   44.098361  26.5416872 25.0719277
## f.fare_amount=(6,9]                   38.915470  24.0749877 25.7706535
## f.distance=(1.01,1.8]                 37.643208  22.6936359 25.1130292
## claKM=kKM-1                           35.642136  12.1854958 14.2416769
## pick_up_period=valley                 37.234821  25.1110015 28.0928894
```

```
## f.pickup_longitude=(-73.95,-73.92]   36.500396  22.7429699 25.9556104
## f.toll=(-1,1]                        41.279433  97.7306364 98.6230991
## f.pickup_latitude=(40.8,40.92]       35.502959  20.7202763 24.3115495
## Payment_type=Other                    0.000000   0.0000000  0.6987259
## f.pickup_latitude=(40.74,40.8]       34.845049  22.7429699 27.1886560
## f.Improvement_surcharge=(-0.1,0.1]   10.975610   0.4440059  1.6851624
## RateCodeID=Special rate              10.843373   0.4440059  1.7057131
## Trip_type=Dispatch                   10.000000   0.3946719  1.6440608
## claKM=kKM-3                          33.037475  16.5268870 20.8384710
## claHP=kHP- 1                         34.954008  26.2456833 31.2782573
## f.outlierPCAd2=YesOutDim2             5.000000   0.1480020  1.2330456
## f.MTA_tax=(-0.1,0.4]                  9.638554   0.3946719  1.7057131
## claKM=kKM-7                           4.918033   0.1480020  1.2535964
## claHP=kHP- 7                          4.918033   0.1480020  1.2535964
## f.outlierPCA=YesOut                   4.918033   0.1480020  1.2535964
## f.fare_amount=(0,6]                  33.173461  20.4736063 25.7090012
## f.dropoff_longitude=(-73.94,-73.91]  32.615894  19.4375925 24.8253185
## f.distance=(0,1.01]                  32.542927  19.6349285 25.1335799
## f.ttime=(-1,6]                       32.134096  19.3882585 25.1335799
## f.dropoff_latitude=(40.79,40.92]     29.245283  16.8228910 23.9621866
## claHP=kHP- 2                         22.486772   8.3867785 15.5363748
## claKM=kKM-2                          20.998532   7.0547607 13.9950678
## f.dropoff_longitude=(-73.91,-73.75]  26.008584  14.9481993 23.9416358
## f.pickup_longitude=(-73.92,-73.79]   25.874695  15.6882092 25.2568845
## f.total=(-1,7.8]                     17.797287  11.0014800 25.7501028
## f.tip_amount=(-0.1,1]                 6.795798  10.2121362 62.5976161
## Payment_type=Cash                     0.000000   0.0000000 50.3082614
##                                          p.value     v.test
## f.tip_amount=(1,22]                 0.000000e+00        Inf
## Payment_type=Credit card            0.000000e+00        Inf
## f.total=(16.6,46]                   6.840593e-51  15.004694
## f.dropoff_longitude=(-74.1,-73.97]  2.751384e-34  12.209926
## claHP=kHP- 6                        1.211151e-31  11.704328
## claKM=kKM-4                         1.837298e-23   9.981467
## f.pickup_longitude=(-74.1,-73.96]   1.140791e-22   9.798669
## claKM=kKM-6                         1.208511e-18   8.813914
## claHP=kHP- 3                        4.239721e-17   8.406072
## f.distance=(3.31,11.5]              1.215939e-14   7.714357
## f.pickup_longitude=(-73.96,-73.95]  2.392093e-14   7.627576
## f.fare_amount=(14,43]               5.794312e-14   7.512648
## f.dropoff_longitude=(-73.97,-73.94] 2.534785e-13   7.317052
## f.ttime=(15.7,215]                  1.255269e-11   6.773703
## f.outlierPCA=NoOut                  6.469500e-11   6.532467
## f.MTA_tax=(0.4,0.5]                 9.937837e-11   6.467894
## f.outlierPCAd2=NoOutDim2            1.065158e-10   6.457402
## Trip_type=Street-hail               3.826056e-10   6.260964
## f.dropoff_latitude=(40.7,40.75]     4.620588e-10   6.231477
## RateCodeID=Standard rate            6.295785e-10   6.182842
## f.Improvement_surcharge=(0.1,0.8]   9.751541e-10   6.113425
## f.pickup_latitude=(40.5,40.7]       1.479187e-08   5.663960
## f.total=(11,16.6]                   6.719751e-08   5.398508
## f.pickup_latitude=(40.7,40.74]      9.352930e-08   5.338867
## claKM=kKM-5                         1.574083e-07   5.243676
## f.toll=(1,50]                       8.328194e-06   4.456569
```

```
## f.dropoff_latitude=(40.57,40.7]     2.117178e-03   3.073283
## f.distance=(1.8,3.31]               2.478699e-03   3.025930
## pick_up_period=morning              2.045298e-02   2.317933
## f.espeed=(20.1,26.2]                3.482537e-02   2.110383
## f.fare_amount=(9,14]                4.444180e-02   2.009900
## f.ttime=(9.78,15.7]                 4.602998e-02   1.995118
## f.fare_amount=(6,9]                 2.213147e-02  -2.288104
## f.distance=(1.01,1.8]               9.769241e-04  -3.297089
## claKM=kKM-1                         4.883997e-04  -3.487039
## pick_up_period=valley               8.786398e-05  -3.921871
## f.pickup_longitude=(-73.95,-73.92]  1.450835e-05  -4.336023
## f.toll=(-1,1]                       8.328194e-06  -4.456569
## f.pickup_latitude=(40.8,40.92]      7.051390e-07  -4.959943
## Payment_type=Other                  1.018119e-08  -5.727682
## f.pickup_latitude=(40.74,40.8]      3.174851e-09  -5.922292
## f.Improvement_surcharge=(-0.1,0.1]  9.751541e-10  -6.113425
## RateCodeID=Special rate             6.295785e-10  -6.182842
## Trip_type=Dispatch                  3.826056e-10  -6.260964
## claKM=kKM-3                         2.752451e-10  -6.312116
## claHP=kHP- 1                        1.281134e-10  -6.429394
## f.outlierPCAd2=YesOutDim2           1.065158e-10  -6.457402
## f.MTA_tax=(-0.1,0.4]                9.937837e-11  -6.467894
## claKM=kKM-7                         6.469500e-11  -6.532467
## claHP=kHP- 7                        6.469500e-11  -6.532467
## f.outlierPCA=YesOut                 6.469500e-11  -6.532467
## f.fare_amount=(0,6]                 1.106603e-12  -7.116552
## f.dropoff_longitude=(-73.94,-73.91] 1.211657e-13  -7.415501
## f.distance=(0,1.01]                 4.793014e-14  -7.537436
## f.ttime=(-1,6]                      3.273821e-15  -7.880000
## f.dropoff_latitude=(40.79,40.92]    1.569737e-23  -9.997069
## claHP=kHP- 2                        3.744050e-33 -11.995663
## claKM=kKM-2                         2.600995e-34 -12.214499
## f.dropoff_longitude=(-73.91,-73.75] 8.652981e-37 -12.670167
## f.pickup_longitude=(-73.92,-73.79]  5.353804e-40 -13.237156
## f.total=(-1,7.8]                    8.443079e-95 -20.657011
## f.tip_amount=(-0.1,1]               0.000000e+00       -Inf
## Payment_type=Cash                   0.000000e+00       -Inf
##
##
## $quanti.var
##                              Eta2      P-value
## Tip_amount            0.548113548 0.000000e+00
## Total_amount          0.082638315 3.144687e-93
## Dropoff_longitude     0.049983053 3.547965e-56
## Pickup_longitude      0.041927982 3.202352e-47
## Trip_distance         0.018305708 2.562980e-21
## trip_distance_km      0.018305708 2.562980e-21
## trip_length           0.017051370 5.922417e-20
## Fare_amount           0.016699294 1.429460e-19
## Pickup_latitude       0.015267177 5.144548e-18
## Dropoff_latitude      0.014663201 2.330774e-17
## travel_time           0.013726241 2.428400e-16
## MTA_tax               0.007319590 2.266756e-09
## improvement_surcharge 0.006108764 4.787227e-08
```

```
## Tolls_amount          0.004248093 5.358685e-06
## mis_ind               0.003611919 2.728219e-05
##
## $quanti
## $quanti$`AnyTip No`
##                           v.test Mean in category Overall mean
## Dropoff_longitude      15.593831     -73.92694654 -73.93588159
## Pickup_longitude       14.282144     -73.92914435 -73.93626676
## Pickup_latitude         8.618284      40.75252018  40.74670866
## Dropoff_latitude        8.446092      40.75041804  40.74466297
## mis_ind                 4.191895       2.52307150   2.49280723
## Tolls_amount           -4.546094       0.04097922   0.07863954
## improvement_surcharge  -5.451526       0.29245157   0.29504110
## MTA_tax                -5.967395       0.48679112   0.49147143
## travel_time            -8.171791      11.17913278  12.14230352
## Fare_amount            -9.013438      10.39649172  11.15474312
## trip_length            -9.107959       3.69609873   4.00371795
## trip_distance_km       -9.437016       3.67902102   4.06433227
## Trip_distance          -9.437016       2.28603768   2.52545899
## Total_amount          -20.050820      11.56190560  13.49374846
## Tip_amount            -51.638865       0.00000000   1.12418208
##                       sd in category Overall sd      p.value
## Dropoff_longitude         0.04804361 0.04729786 8.017822e-55
## Pickup_longitude          0.04267174 0.04116523 2.827672e-46
## Pickup_latitude           0.05620848 0.05566292 6.796507e-18
## Dropoff_latitude          0.05871078 0.05624604 3.012162e-17
## mis_ind                   0.64257726 0.59595986 2.766335e-05
## Tolls_amount              0.47470577 0.68382164 5.465060e-06
## improvement_surcharge     0.04832210 0.03921036 4.993945e-08
## MTA_tax                   0.08018705 0.06474215 2.410710e-09
## travel_time               8.84916405 9.72933787 3.038452e-16
## Fare_amount               6.57686100 6.94416418 1.996956e-19
## trip_length               2.62022653 2.78797993 8.394639e-20
## trip_distance_km          3.11428790 3.37034414 3.835470e-21
## Trip_distance             1.93512878 2.09423476 3.835470e-21
## Total_amount              6.65419083 7.95310822 1.985436e-89
## Tip_amount                0.00000000 1.79703780 0.000000e+00
##
## $quanti$`AnyTip Yes`
##                           v.test Mean in category Overall mean
## Tip_amount             51.638865        2.6987025   1.12418208
## Total_amount           20.050820       16.1994721  13.49374846
## trip_distance_km        9.437016        4.6039961   4.06433227
## Trip_distance           9.437016        2.8607906   2.52545899
## trip_length             9.107959        4.4345670   4.00371795
## Fare_amount             9.013438       12.2167440  11.15474312
## travel_time             8.171791       13.4913128  12.14230352
## MTA_tax                 5.967395        0.4980266   0.49147143
## improvement_surcharge   5.451526        0.2986680   0.29504110
## Tolls_amount            4.546094        0.1313863   0.07863954
## mis_ind                -4.191895        2.4504193   2.49280723
## Dropoff_latitude       -8.446092       40.7366025  40.74466297
## Pickup_latitude        -8.618284       40.7385691  40.74670866
## Pickup_longitude      -14.282144      -73.9462423 -73.93626676
```

```
## Dropoff_longitude     -15.593831      -73.9483959 -73.93588159
##                       sd in category Overall sd      p.value
## Tip_amount                1.87167763 1.79703780 0.000000e+00
## Total_amount              8.79080963 7.95310822 1.985436e-89
## trip_distance_km          3.63120225 3.37034414 3.835470e-21
## Trip_distance             2.25632447 2.09423476 3.835470e-21
## trip_length               2.95386931 2.78797993 8.394639e-20
## Fare_amount               7.29685470 6.94416418 1.996956e-19
## travel_time              10.69783993 9.72933787 3.038452e-16
## MTA_tax                   0.03134941 0.06474215 2.410710e-09
## improvement_surcharge     0.01994570 0.03921036 4.993945e-08
## Tolls_amount              0.89563377 0.68382164 5.465060e-06
## mis_ind                   0.52078986 0.59595986 2.766335e-05
## Dropoff_latitude          0.05153073 0.05624604 3.012162e-17
## Pickup_latitude           0.05384535 0.05566292 6.796507e-18
## Pickup_longitude          0.03670311 0.04116523 2.827672e-46
## Dropoff_longitude         0.04323283 0.04729786 8.017822e-55
##
##
## attr(,"class")
## [1] "catdes" "list "
```

```
vars_cexp<-vars_con[c(1:4,5,6,7,8,11,13)]
dfX <- dfwork[,c("AnyTip",vars_cexp)]
cor(dfwork[,c(vars_cexp)])
```

```
##                   Pickup_longitude Pickup_latitude Dropoff_longitude
## Pickup_longitude      1.0000000000      0.25948936       0.791051889
## Pickup_latitude       0.2594893591      1.00000000       0.149194819
## Dropoff_longitude     0.7910518885      0.14919482       1.000000000
## Dropoff_latitude      0.2197872159      0.88296871       0.159171166
## Passenger_count       0.0054554547     -0.02024209       0.017463399
## Trip_distance        -0.0144891210     -0.05070494      -0.002327808
## Fare_amount          -0.0290667838     -0.06063673      -0.016225870
## Extra                -0.0088711740     -0.10939237       0.032909062
## Tolls_amount          0.0002816405      0.03377444      -0.014472099
## trip_length           0.0037536793     -0.04980288       0.016656400
##                   Dropoff_latitude Passenger_count Trip_distance
## Pickup_longitude        0.21978722     0.005455455  -0.014489121
## Pickup_latitude         0.88296871    -0.020242092  -0.050704940
## Dropoff_longitude       0.15917117     0.017463399  -0.002327808
## Dropoff_latitude        1.00000000    -0.017650710  -0.081884959
## Passenger_count        -0.01765071     1.000000000   0.021262647
## Trip_distance          -0.08188496     0.021262647   1.000000000
## Fare_amount            -0.09390701     0.016646290   0.938018859
## Extra                  -0.11479394     0.047964368  -0.052585066
## Tolls_amount            0.05130022     0.008874156   0.234701181
## trip_length            -0.09169987     0.017008804   0.903484449
##                   Fare_amount        Extra  Tolls_amount  trip_length
## Pickup_longitude  -0.02906678 -0.008871174  0.0002816405  0.003753679
## Pickup_latitude   -0.06063673 -0.109392367  0.0337744429 -0.049802878
## Dropoff_longitude -0.01622587  0.032909062 -0.0144720993  0.016656400
## Dropoff_latitude  -0.09390701 -0.114793945  0.0513002178 -0.091699868
## Passenger_count    0.01664629  0.047964368  0.0088741560  0.017008804
## Trip_distance      0.93801886 -0.052585066  0.2347011815  0.903484449
```

```
## Fare_amount        1.00000000 -0.048291164  0.1925195527  0.869134388
## Extra             -0.04829116  1.000000000 -0.0146097936 -0.049659889
## Tolls_amount       0.19251955 -0.014609794  1.0000000000  0.221172079
## trip_length        0.86913439 -0.049659889  0.2211720793  1.000000000
```

```
m50<-glm(AnyTip~.,family="binomial",data=dfX)
summary(m50)
```

```
##
## Call:
## glm(formula = AnyTip ~ ., family = "binomial", data = dfX)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8479  -1.0052  -0.7668   1.2007   2.0571
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.457e+02  8.615e+01  -7.495 6.64e-14 ***
## Pickup_longitude  -2.521e+00  1.512e+00  -1.667   0.0955 .
## Pickup_latitude   -2.958e+00  1.425e+00  -2.075   0.0379 *
## Dropoff_longitude -7.339e+00  1.285e+00  -5.711 1.12e-08 ***
## Dropoff_latitude   8.977e-01  1.408e+00   0.638   0.5236
## Passenger_count   -3.728e-02  3.585e-02  -1.040   0.2984
## Trip_distance      1.484e-01  5.910e-02   2.510   0.0121 *
## Fare_amount       -1.732e-02  1.554e-02  -1.115   0.2650
## Extra             -9.716e-03  9.992e-02  -0.097   0.9225
## Tolls_amount       1.450e-01  6.014e-02   2.412   0.0159 *
## trip_length        3.355e-02  3.045e-02   1.102   0.2706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4389.9  on 3395  degrees of freedom
## AIC: 4411.9
##
## Number of Fisher Scoring iterations: 4
```

```
Anova(m50,test="Wald")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: AnyTip
##                   Df   Chisq Pr(>Chisq)
## Pickup_longitude   1  2.7795    0.09548 .
## Pickup_latitude    1  4.3073    0.03795 *
## Dropoff_longitude  1 32.6135  1.124e-08 ***
## Dropoff_latitude   1  0.4067    0.52364
## Passenger_count    1  1.0812    0.29842
## Trip_distance      1  6.3004    0.01207 *
## Fare_amount        1  1.2422    0.26504
## Extra              1  0.0095    0.92254
## Tolls_amount       1  5.8174    0.01587 *
```

```
## trip_length        1  1.2137    0.27060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(m50)
```

```
##  Pickup_longitude   Pickup_latitude Dropoff_longitude  Dropoff_latitude
##          2.576036          4.901641          2.452523          4.813125
##   Passenger_count     Trip_distance       Fare_amount             Extra
##          1.003440         11.422166          8.750910          1.024343
##      Tolls_amount       trip_length
##          1.054317          5.406331
```

```
m51<-step(m50,k=log(nrow(dfwork)))
```

```
## Start:  AIC=4479.37
## AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Dropoff_latitude + Passenger_count + Trip_distance + Fare_amount +
##     Extra + Tolls_amount + trip_length
##
##                     Df Deviance    AIC
## - Extra              1   4389.9 4471.3
## - Dropoff_latitude   1   4390.3 4471.6
## - Passenger_count    1   4391.0 4472.3
## - trip_length        1   4391.1 4472.5
## - Fare_amount        1   4391.2 4472.5
## - Pickup_longitude   1   4392.7 4474.0
## - Pickup_latitude    1   4394.2 4475.6
## - Tolls_amount       1   4396.2 4477.6
## - Trip_distance      1   4396.3 4477.6
## <none>                   4389.9 4479.4
## - Dropoff_longitude  1   4423.2 4504.5
##
## Step:  AIC=4471.25
## AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Dropoff_latitude + Passenger_count + Trip_distance + Fare_amount +
##     Tolls_amount + trip_length
##
##                     Df Deviance    AIC
## - Dropoff_latitude   1   4390.3 4463.5
## - Passenger_count    1   4391.0 4464.2
## - trip_length        1   4391.1 4464.3
## - Fare_amount        1   4391.2 4464.4
## - Pickup_longitude   1   4392.7 4465.9
## - Pickup_latitude    1   4394.2 4467.4
## - Tolls_amount       1   4396.2 4469.4
## - Trip_distance      1   4396.3 4469.5
## <none>                   4389.9 4471.3
## - Dropoff_longitude  1   4423.4 4496.6
##
## Step:  AIC=4463.53
## AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Passenger_count + Trip_distance + Fare_amount + Tolls_amount +
##     trip_length
##
```

```
##                      Df Deviance    AIC
## - Passenger_count     1   4391.4 4456.5
## - trip_length         1   4391.5 4456.5
## - Fare_amount         1   4391.6 4456.7
## - Pickup_longitude    1   4393.4 4458.4
## - Tolls_amount        1   4396.8 4461.9
## - Trip_distance       1   4396.8 4461.9
## <none>                    4390.3 4463.5
## - Pickup_latitude     1   4400.4 4465.5
## - Dropoff_longitude   1   4423.5 4488.5
##
## Step:  AIC=4456.5
## AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Trip_distance + Fare_amount + Tolls_amount + trip_length
##
##                      Df Deviance    AIC
## - trip_length         1   4392.6 4449.5
## - Fare_amount         1   4392.7 4449.6
## - Pickup_longitude    1   4394.5 4451.4
## - Trip_distance       1   4397.8 4454.8
## - Tolls_amount        1   4397.9 4454.8
## <none>                    4391.4 4456.5
## - Pickup_latitude     1   4401.4 4458.3
## - Dropoff_longitude   1   4424.8 4481.7
##
## Step:  AIC=4449.49
## AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Trip_distance + Fare_amount + Tolls_amount
##
##                      Df Deviance    AIC
## - Fare_amount         1   4393.5 4442.3
## - Pickup_longitude    1   4395.5 4444.3
## - Tolls_amount        1   4399.2 4448.0
## <none>                    4392.6 4449.5
## - Pickup_latitude     1   4402.6 4451.4
## - Trip_distance       1   4405.0 4453.8
## - Dropoff_longitude   1   4425.7 4474.5
##
## Step:  AIC=4442.31
## AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Trip_distance + Tolls_amount
##
##                      Df Deviance    AIC
## - Pickup_longitude    1   4396.4 4437.1
## - Tolls_amount        1   4400.6 4441.2
## <none>                    4393.5 4442.3
## - Pickup_latitude     1   4403.4 4444.1
## - Dropoff_longitude   1   4426.5 4467.2
## - Trip_distance       1   4449.7 4490.4
##
## Step:  AIC=4437.08
## AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance +
##     Tolls_amount
##
```

```
##                       Df Deviance    AIC
## - Tolls_amount         1   4403.5 4436.1
## <none>                     4396.4 4437.1
## - Pickup_latitude      1   4410.2 4442.7
## - Trip_distance        1   4452.3 4484.8
## - Dropoff_longitude    1   4521.2 4553.8
##
## Step:  AIC=4436.08
## AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance
##
##                       Df Deviance    AIC
## <none>                     4403.5 4436.1
## - Pickup_latitude      1   4416.3 4440.7
## - Trip_distance        1   4472.8 4497.2
## - Dropoff_longitude    1   4529.1 4553.5
```

```
summary(m51)
```

```
##
## Call:
## glm(formula = AnyTip ~ Pickup_latitude + Dropoff_longitude +
##     Trip_distance, family = "binomial", data = dfX)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8621  -1.0020  -0.7852   1.2071   1.9845
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -561.46559   70.64578  -7.948  1.9e-15 ***
## Pickup_latitude     -2.32739    0.65317  -3.563 0.000366 ***
## Dropoff_longitude   -8.86703    0.82666 -10.726  < 2e-16 ***
## Trip_distance        0.14358    0.01749   8.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4403.5  on 3402  degrees of freedom
## AIC: 4411.5
##
## Number of Fisher Scoring iterations: 4
```

```
anova (m51,m50, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance
## Model 2: AnyTip ~ Pickup_longitude + Pickup_latitude + Dropoff_longitude +
##     Dropoff_latitude + Passenger_count + Trip_distance + Fare_amount +
##     Extra + Tolls_amount + trip_length
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3402     4403.5
## 2      3395     4389.9  7   13.636  0.05804 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m51,test="Wald")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: AnyTip
##                   Df   Chisq Pr(>Chisq)
## Pickup_latitude    1  12.697  0.0003663 ***
## Dropoff_longitude  1 115.054  < 2.2e-16 ***
## Trip_distance      1  67.406  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(m51)
```

```
##   Pickup_latitude Dropoff_longitude     Trip_distance
##          1.029823          1.030648          1.002859
```

So far we took general model m50 and applied step on it. The resulting one is the addition of Pickup_latitude + Dropoff_longitude + Trip_distance. We want to consider adding factors to our model. For that purpose we get those appearing in catdes as associated to anytip.

```
#We try claHP
m52<-glm(AnyTip~ (Pickup_latitude + Dropoff_longitude + Trip_distance)*claHP ,family="binomial",data=df
summary(m52)
```

```
##
## Call:
## glm(formula = AnyTip ~ (Pickup_latitude + Dropoff_longitude +
##     Trip_distance) * claHP, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6772  -1.0204  -0.7032   1.1495   2.7691
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -945.97274  253.86858  -3.726 0.000194 ***
## Pickup_latitude                 -7.58517    2.47657  -3.063 0.002193 **
## Dropoff_longitude              -16.96735    3.24379  -5.231 1.69e-07 ***
## Trip_distance                    0.19044    0.07564   2.518 0.011807 *
## claHPkHP- 2                   1193.25022  346.78523   3.441 0.000580 ***
## claHPkHP- 3                     38.98393  356.34849   0.109 0.912887
## claHPkHP- 4                    207.58449  444.62153   0.467 0.640586
## claHPkHP- 5                   -923.71294 1756.95623  -0.526 0.599065
## claHPkHP- 6                    555.49377  289.10408   1.921 0.054677 .
## claHPkHP- 7                   3219.56540 1459.24609   2.206 0.027362 *
## Pickup_latitude:claHPkHP- 2      1.44365    4.69911   0.307 0.758678
## Pickup_latitude:claHPkHP- 3     19.02985    3.93383   4.837 1.31e-06 ***
## Pickup_latitude:claHPkHP- 4      3.68189    4.24176   0.868 0.385390
## Pickup_latitude:claHPkHP- 5     21.47482   10.61196   2.024 0.043007 *
## Pickup_latitude:claHPkHP- 6      7.06958    2.90712   2.432 0.015023 *
## Pickup_latitude:claHPkHP- 7     20.12302   17.27343   1.165 0.244031
## Dropoff_longitude:claHPkHP- 2   16.94666    4.49593   3.769 0.000164 ***
## Dropoff_longitude:claHPkHP- 3   11.00336    4.21194   2.612 0.008990 **
```

```
## Dropoff_longitude:claHPkHP- 4    4.84658    5.27482   0.919 0.358193
## Dropoff_longitude:claHPkHP- 5   -0.66018   23.93442  -0.028 0.977995
## Dropoff_longitude:claHPkHP- 6   11.39360    3.63475   3.135 0.001721 **
## Dropoff_longitude:claHPkHP- 7   54.73184   25.66024   2.133 0.032929 *
## Trip_distance:claHPkHP- 2       -0.07228    0.11177  -0.647 0.517814
## Trip_distance:claHPkHP- 3       -0.12717    0.09638  -1.320 0.186985
## Trip_distance:claHPkHP- 4        0.13634    0.13658   0.998 0.318184
## Trip_distance:claHPkHP- 5       -0.46410    0.30301  -1.532 0.125608
## Trip_distance:claHPkHP- 6       -0.22601    0.08886  -2.543 0.010981 *
## Trip_distance:claHPkHP- 7        0.11744    0.29694   0.395 0.692474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4294.7  on 3378  degrees of freedom
## AIC: 4350.7
##
## Number of Fisher Scoring iterations: 6
```

```
#exaggerated colinearity with cluster factor
vif(m52)
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## Pickup_latitude         1.387303e+01  1        3.724651
## Dropoff_longitude       1.628560e+01  1        4.035543
## Trip_distance           1.921893e+01  1        4.383940
## claHP                   6.910577e+40  6     2531.004167
## Pickup_latitude:claHP   2.472835e+36  6     1078.366173
## Dropoff_longitude:claHP 4.481814e+40  6     2441.299997
## Trip_distance:claHP     3.253744e+04  6        2.377015
```

```
#As this last model has quite big complexity, we apply step on it and it gets rid of claHP and any poss
m62<-step(m52,k=log(nrow(dfwork)))
```

```
## Start:  AIC=4522.41
## AnyTip ~ (Pickup_latitude + Dropoff_longitude + Trip_distance) *
##     claHP
##
##                            Df Deviance    AIC
## - Trip_distance:claHP       6   4310.2 4489.2
## - Dropoff_longitude:claHP   6   4317.7 4496.6
## - Pickup_latitude:claHP     6   4324.7 4503.6
## <none>                          4294.7 4522.4
##
## Step:  AIC=4489.16
## AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance +
##     claHP + Pickup_latitude:claHP + Dropoff_longitude:claHP
##
##                            Df Deviance    AIC
## - Dropoff_longitude:claHP   6   4334.3 4464.4
## - Pickup_latitude:claHP     6   4340.4 4470.5
## - Trip_distance             1   4315.5 4486.3
## <none>                          4310.2 4489.2
##
```

```
## Step:  AIC=4464.44
## AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance +
##     claHP + Pickup_latitude:claHP
##
##                          Df Deviance    AIC
## - Pickup_latitude:claHP  6   4370.3 4451.6
## - Trip_distance          1   4341.3 4463.3
## <none>                       4334.3 4464.4
## - Dropoff_longitude      1   4372.5 4494.5
##
## Step:  AIC=4451.63
## AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance +
##     claHP
##
##                      Df Deviance    AIC
## - claHP               6   4403.5 4436.1
## - Pickup_latitude     1   4372.1 4445.3
## - Trip_distance       1   4376.9 4450.1
## <none>                    4370.3 4451.6
## - Dropoff_longitude   1   4404.0 4477.2
##
## Step:  AIC=4436.08
## AnyTip ~ Pickup_latitude + Dropoff_longitude + Trip_distance
##
##                      Df Deviance    AIC
## <none>                    4403.5 4436.1
## - Pickup_latitude     1   4416.3 4440.7
## - Trip_distance       1   4472.8 4497.2
## - Dropoff_longitude   1   4529.1 4553.5
```

```
summary(m62)
```

```
##
## Call:
## glm(formula = AnyTip ~ Pickup_latitude + Dropoff_longitude +
##     Trip_distance, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8621  -1.0020  -0.7852   1.2071   1.9845
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -561.46559   70.64578  -7.948  1.9e-15 ***
## Pickup_latitude     -2.32739    0.65317  -3.563 0.000366 ***
## Dropoff_longitude   -8.86703    0.82666 -10.726  < 2e-16 ***
## Trip_distance        0.14358    0.01749   8.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4403.5  on 3402  degrees of freedom
## AIC: 4411.5
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
#We try again without trip distance
m53<-glm(AnyTip~ (Pickup_latitude + Dropoff_longitude)*claHP ,family="binomial",data=dfwork)
summary(m53)
```

```
##
## Call:
## glm(formula = AnyTip ~ (Pickup_latitude + Dropoff_longitude) *
##     claHP, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6185  -1.0290  -0.7193   1.1510   2.4415
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1017.159    252.775  -4.024 5.72e-05 ***
## Pickup_latitude                   -7.171      2.463  -2.912 0.003594 **
## Dropoff_longitude                 -17.705      3.251  -5.447 5.14e-08 ***
## claHPkHP- 2                      1320.125    345.373   3.822 0.000132 ***
## claHPkHP- 3                       113.815    355.502   0.320 0.748852
## claHPkHP- 4                       430.487    439.750   0.979 0.327612
## claHPkHP- 5                        82.855   1257.663   0.066 0.947473
## claHPkHP- 6                       613.264    287.617   2.132 0.032989 *
## claHPkHP- 7                      3784.627   1492.302   2.536 0.011209 *
## Pickup_latitude:claHPkHP- 2         1.655      4.741   0.349 0.726949
## Pickup_latitude:claHPkHP- 3        18.525      3.924   4.722 2.34e-06 ***
## Pickup_latitude:claHPkHP- 4         2.778      4.212   0.660 0.509528
## Pickup_latitude:claHPkHP- 5        22.696     10.483   2.165 0.030384 *
## Pickup_latitude:claHPkHP- 6         6.619      2.895   2.286 0.022247 *
## Pickup_latitude:claHPkHP- 7        15.209     15.403   0.987 0.323454
## Dropoff_longitude:claHPkHP- 2      18.781      4.463   4.208 2.58e-05 ***
## Dropoff_longitude:claHPkHP- 3      11.739      4.217   2.784 0.005372 **
## Dropoff_longitude:claHPkHP- 4       7.357      5.238   1.405 0.160120
## Dropoff_longitude:claHPkHP- 5      13.640     17.402   0.784 0.433159
## Dropoff_longitude:claHPkHP- 6      11.933      3.632   3.286 0.001016 **
## Dropoff_longitude:claHPkHP- 7      59.656     25.980   2.296 0.021663 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4315.5  on 3385  degrees of freedom
## AIC: 4357.5
##
## Number of Fisher Scoring iterations: 6
```

```
vif(m53)
```

```
##                                 GVIF Df GVIF^(1/(2*Df))
## Pickup_latitude         1.372656e+01  1        3.704937
## Dropoff_longitude       1.643748e+01  1        4.054316
## claHP                   4.497044e+40  6     2441.990262
```

```
## Pickup_latitude:claHP   2.481430e+36  6     1078.678048
## Dropoff_longitude:claHP 2.908446e+40  6     2354.896103
```

```
#And we apply again step on it, resulting Dropoff_longitude + claHP
m63<-step(m53,k=log(nrow(dfwork)))
```

```
## Start:  AIC=4486.34
## AnyTip ~ (Pickup_latitude + Dropoff_longitude) * claHP
##
##                            Df Deviance    AIC
## - Dropoff_longitude:claHP  6   4341.3 4463.3
## - Pickup_latitude:claHP    6   4344.7 4466.7
## <none>                         4315.5 4486.3
##
## Step:  AIC=4463.32
## AnyTip ~ Pickup_latitude + Dropoff_longitude + claHP + Pickup_latitude:claHP
##
##                           Df Deviance    AIC
## - Pickup_latitude:claHP  6   4376.9 4450.1
## <none>                       4341.3 4463.3
## - Dropoff_longitude      1   4376.8 4490.7
##
## Step:  AIC=4450.08
## AnyTip ~ Pickup_latitude + Dropoff_longitude + claHP
##
##                      Df Deviance    AIC
## - Pickup_latitude    1   4378.5 4443.6
## <none>                   4376.9 4450.1
## - Dropoff_longitude  1   4408.2 4473.2
## - claHP              6   4472.8 4497.2
##
## Step:  AIC=4443.58
## AnyTip ~ Dropoff_longitude + claHP
##
##                      Df Deviance    AIC
## <none>                   4378.5 4443.6
## - Dropoff_longitude  1   4410.3 4467.2
## - claHP              6   4488.2 4504.5
```

```
summary(m63)
```

```
##
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude + claHP, family = "binomial",
##     data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6184  -0.9889  -0.7594   1.1985   2.5296
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -455.64456   81.69733  -5.577 2.44e-08 ***
## Dropoff_longitude   -6.15458    1.10490  -5.570 2.54e-08 ***
## claHPkHP- 2         -0.13334    0.14468  -0.922  0.35670
```

```
## claHPkHP- 3          0.38949    0.09323   4.178 2.95e-05 ***
## claHPkHP- 4          0.06426    0.17191   0.374  0.70853
## claHPkHP- 5         -0.34408    0.53271  -0.646  0.51835
## claHPkHP- 6          0.98681    0.10865   9.083  < 2e-16 ***
## claHPkHP- 7         -1.58659    0.60762  -2.611  0.00902 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4378.5  on 3398  degrees of freedom
## AIC: 4394.5
##
## Number of Fisher Scoring iterations: 4
```

```
vif(m63)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude 1.789683  1        1.337790
## claHP             1.789683  6        1.049699
```

```
#This two models are significally differents (p.value under threshold).
anova(m53,m63, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: AnyTip ~ (Pickup_latitude + Dropoff_longitude) * claHP
## Model 2: AnyTip ~ Dropoff_longitude + claHP
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1      3385     4315.5
## 2      3398     4378.5 -13  -62.975 1.534e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Now we try to add a new factor to the model, even though we know it adds too much complexity and we wi
m54<-glm(AnyTip~ (Pickup_latitude + Dropoff_longitude)+(claHP*f.distance) ,family="binomial",data=dfwor
summary(m54)
```

```
##
## Call:
## glm(formula = AnyTip ~ (Pickup_latitude + Dropoff_longitude) +
##     (claHP * f.distance), family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8193  -1.0040  -0.7135   1.1528   2.3016
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -390.66732   93.63198  -4.172 3.01e-05
## Pickup_latitude                  -1.68577    1.05406  -1.599   0.1098
## Dropoff_longitude                -6.20256    1.10898  -5.593 2.23e-08
## claHPkHP- 2                      -0.02988    0.25310  -0.118   0.9060
## claHPkHP- 3                       0.14547    0.20247   0.718   0.4725
## claHPkHP- 4                      -0.29908    0.38972  -0.767   0.4428
```

```
## claHPkHP- 5                          2.06720    1.16175   1.779   0.0752
## claHPkHP- 6                         15.40991  882.74339   0.017   0.9861
## claHPkHP- 7                         -0.37971    1.10700  -0.343   0.7316
## f.distance(1.01,1.8]                 0.18322    0.15624   1.173   0.2409
## f.distance(1.8,3.31]                 0.71040    0.16442   4.321 1.56e-05
## f.distance(3.31,11.5]                0.11005    0.33448   0.329   0.7421
## claHPkHP- 2:f.distance(1.01,1.8]    -0.44035    0.33086  -1.331   0.1832
## claHPkHP- 3:f.distance(1.01,1.8]     0.28158    0.22979   1.225   0.2204
## claHPkHP- 4:f.distance(1.01,1.8]    -0.52478    0.53615  -0.979   0.3277
## claHPkHP- 5:f.distance(1.01,1.8]   -16.10699  440.45200  -0.037   0.9708
## claHPkHP- 6:f.distance(1.01,1.8]    -0.23460 1248.38767   0.000   0.9999
## claHPkHP- 7:f.distance(1.01,1.8]   -13.06352  390.46257  -0.033   0.9733
## claHPkHP- 2:f.distance(1.8,3.31]    -0.41230    0.31842  -1.295   0.1954
## claHPkHP- 3:f.distance(1.8,3.31]    -0.27978    0.23459  -1.193   0.2330
## claHPkHP- 4:f.distance(1.8,3.31]     0.39734    0.49584   0.801   0.4229
## claHPkHP- 5:f.distance(1.8,3.31]    -2.67360    1.69887  -1.574   0.1155
## claHPkHP- 6:f.distance(1.8,3.31]   -14.01178  882.74364  -0.016   0.9873
## claHPkHP- 7:f.distance(1.8,3.31]   -13.76389  264.41961  -0.052   0.9585
## claHPkHP- 2:f.distance(3.31,11.5]    0.09102    0.46454   0.196   0.8447
## claHPkHP- 3:f.distance(3.31,11.5]    0.21158    0.40302   0.525   0.5996
## claHPkHP- 4:f.distance(3.31,11.5]    1.16416    0.61222   1.902   0.0572
## claHPkHP- 5:f.distance(3.31,11.5]   -3.27450    1.62156  -2.019   0.0435
## claHPkHP- 6:f.distance(3.31,11.5]  -14.39974  882.74345  -0.016   0.9870
## claHPkHP- 7:f.distance(3.31,11.5]   -0.56971    1.37698  -0.414   0.6791
##
## (Intercept)                       ***
## Pickup_latitude
## Dropoff_longitude                 ***
## claHPkHP- 2
## claHPkHP- 3
## claHPkHP- 4
## claHPkHP- 5                       .
## claHPkHP- 6
## claHPkHP- 7
## f.distance(1.01,1.8]
## f.distance(1.8,3.31]              ***
## f.distance(3.31,11.5]
## claHPkHP- 2:f.distance(1.01,1.8]
## claHPkHP- 3:f.distance(1.01,1.8]
## claHPkHP- 4:f.distance(1.01,1.8]
## claHPkHP- 5:f.distance(1.01,1.8]
## claHPkHP- 6:f.distance(1.01,1.8]
## claHPkHP- 7:f.distance(1.01,1.8]
## claHPkHP- 2:f.distance(1.8,3.31]
## claHPkHP- 3:f.distance(1.8,3.31]
## claHPkHP- 4:f.distance(1.8,3.31]
## claHPkHP- 5:f.distance(1.8,3.31]
## claHPkHP- 6:f.distance(1.8,3.31]
## claHPkHP- 7:f.distance(1.8,3.31]
## claHPkHP- 2:f.distance(3.31,11.5]
## claHPkHP- 3:f.distance(3.31,11.5]
## claHPkHP- 4:f.distance(3.31,11.5] .
## claHPkHP- 5:f.distance(3.31,11.5] *
## claHPkHP- 6:f.distance(3.31,11.5]
```

```
## claHPkHP- 7:f.distance(3.31,11.5]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4308.9  on 3376  degrees of freedom
## AIC: 4368.9
##
## Number of Fisher Scoring iterations: 13
```

```
#Too much colinearity claHP
vif(m54)
```

```
##                           GVIF Df GVIF^(1/(2*Df))
## Pickup_latitude   2.620450e+00  1        1.618780
## Dropoff_longitude 1.812008e+00  1        1.346108
## claHP             8.532466e+10  6        8.145596
## f.distance        8.971911e+01  3        2.115830
## claHP:f.distance  1.149310e+12 18        2.162779
```

```
m64<-step(m54,k=log(nrow(dfwork)))
```

```
## Start:  AIC=4552.94
## AnyTip ~ (Pickup_latitude + Dropoff_longitude) + (claHP * f.distance)
##
##                      Df Deviance    AIC
## - claHP:f.distance  18   4348.5 4446.1
## - Pickup_latitude    1   4311.5 4547.4
## <none>                   4308.9 4552.9
## - Dropoff_longitude  1   4341.0 4576.8
##
## Step:  AIC=4446.15
## AnyTip ~ Pickup_latitude + Dropoff_longitude + claHP + f.distance
##
##                      Df Deviance    AIC
## - Pickup_latitude    1   4350.3 4439.8
## <none>                   4348.5 4446.1
## - claHP              6   4401.2 4450.0
## - f.distance         3   4376.9 4450.1
## - Dropoff_longitude  1   4379.5 4469.0
##
## Step:  AIC=4439.79
## AnyTip ~ Dropoff_longitude + claHP + f.distance
##
##                      Df Deviance    AIC
## <none>                   4350.3 4439.8
## - f.distance         3   4378.5 4443.6
## - claHP              6   4412.2 4452.9
## - Dropoff_longitude  1   4381.8 4463.2
```

```
#Dropoff_longitude + claHP + f.distance
summary(m64)
```

```
##
```

```
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude + claHP + f.distance,
##     family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6145  -1.0089  -0.7401   1.1877   2.5356
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.530e+02  8.153e+01  -5.556 2.76e-08 ***
## Dropoff_longitude     -6.116e+00  1.103e+00  -5.547 2.91e-08 ***
## claHPkHP- 2           -1.897e-01  1.464e-01  -1.295 0.195167
## claHPkHP- 3            3.454e-01  9.437e-02   3.660 0.000252 ***
## claHPkHP- 4            8.193e-03  1.740e-01   0.047 0.962446
## claHPkHP- 5           -3.841e-01  5.364e-01  -0.716 0.473908
## claHPkHP- 6            8.900e-01  1.622e-01   5.486 4.11e-08 ***
## claHPkHP- 7           -1.700e+00  6.106e-01  -2.784 0.005363 **
## f.distance(1.01,1.8]   1.882e-01  1.031e-01   1.826 0.067882 .
## f.distance(1.8,3.31]   5.398e-01  1.041e-01   5.184 2.17e-07 ***
## f.distance(3.31,11.5]  3.113e-01  1.448e-01   2.151 0.031493 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4350.3  on 3395  degrees of freedom
## AIC: 4372.3
##
## Number of Fisher Scoring iterations: 4
```

```
#we would choose m64

#but colinearity test points that the model is not good enough (claHP over 4 points)
vif(m64)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude 1.786572  1        1.336627
## claHP             4.317986  6        1.129640
## f.distance        2.418040  3        1.158539
```

From this moment, we try to simplify our interactions with factors to achieve some balanced model.

```
m55<-glm(AnyTip~ Pickup_latitude + claHP + f.distance + Pickup_latitude:claHP,family="binomial",data=df
summary(m55)
```

```
##
## Call:
## glm(formula = AnyTip ~ Pickup_latitude + claHP + f.distance +
##     Pickup_latitude:claHP, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5130  -1.0141  -0.7189   1.1402   2.5962
##
```

```
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   291.7873    94.0576   3.102  0.00192 **
## Pickup_latitude                -7.1712     2.3054  -3.111  0.00187 **
## claHPkHP- 2                   -57.8002   182.9969  -0.316  0.75211
## claHPkHP- 3                  -726.5149   154.7703  -4.694 2.68e-06 ***
## claHPkHP- 4                   -26.4633   165.8821  -0.160  0.87325
## claHPkHP- 5                 -1018.3933   411.9696  -2.472  0.01344 *
## claHPkHP- 6                  -258.5917   112.2021  -2.305  0.02118 *
## claHPkHP- 7                   332.4018   393.7910   0.844  0.39861
## f.distance(1.01,1.8]            0.1929     0.1035   1.864  0.06237 .
## f.distance(1.8,3.31]            0.5683     0.1047   5.428 5.69e-08 ***
## f.distance(3.31,11.5]           0.3102     0.1457   2.128  0.03330 *
## Pickup_latitude:claHPkHP- 2     1.3923     4.4902   0.310  0.75650
## Pickup_latitude:claHPkHP- 3    17.8468     3.7997   4.697 2.64e-06 ***
## Pickup_latitude:claHPkHP- 4     0.6387     4.0697   0.157  0.87529
## Pickup_latitude:claHPkHP- 5    24.9697    10.1051   2.471  0.01347 *
## Pickup_latitude:claHPkHP- 6     6.3597     2.7511   2.312  0.02080 *
## Pickup_latitude:claHPkHP- 7    -8.2022     9.6605  -0.849  0.39585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4345.8  on 3389  degrees of freedom
## AIC: 4379.8
##
## Number of Fisher Scoring iterations: 5
```

```
vif(m55)
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## Pickup_latitude       1.234887e+01  1        3.514096
## claHP                 8.511492e+35  6      986.659137
## f.distance            2.490219e+00  3        1.164232
## Pickup_latitude:claHP 8.490014e+35  6      986.451418
```

```
#Here, with this step, it seems to get some decent model taking into consideration just and addition of
m65<-step(m55,k=log(nrow(dfwork)))
```

```
## Start:  AIC=4484.1
## AnyTip ~ Pickup_latitude + claHP + f.distance + Pickup_latitude:claHP
##
##                          Df Deviance    AIC
## - Pickup_latitude:claHP  6   4379.5 4469.0
## <none>                       4345.8 4484.1
## - f.distance             3   4376.8 4490.7
##
## Step:  AIC=4468.99
## AnyTip ~ Pickup_latitude + claHP + f.distance
##
##                    Df Deviance    AIC
## - Pickup_latitude  1   4381.8 4463.2
## <none>                 4379.5 4469.0
```

```
## - f.distance       3   4408.2 4473.2
## - claHP            6   4525.3 4566.0
##
## Step:  AIC=4463.16
## AnyTip ~ claHP + f.distance
##
##              Df Deviance    AIC
## <none>            4381.8 4463.2
## - f.distance  3   4410.3 4467.2
## - claHP       6   4552.1 4584.6
```

```
summary(m65)
```

```
##
## Call:
## glm(formula = AnyTip ~ claHP + f.distance, family = "binomial",
##     data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4777  -1.0537  -0.7477   1.2238   2.3938
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.79064    0.08650  -9.141  < 2e-16 ***
## claHPkHP- 2            -0.64198    0.12238  -5.246 1.56e-07 ***
## claHPkHP- 3             0.49250    0.09049   5.442 5.26e-08 ***
## claHPkHP- 4            -0.03327    0.17274  -0.193 0.847285
## claHPkHP- 5            -0.42095    0.53365  -0.789 0.430218
## claHPkHP- 6             0.93166    0.16082   5.793 6.91e-09 ***
## claHPkHP- 7            -2.01591    0.60680  -3.322 0.000893 ***
## f.distance(1.01,1.8]    0.18978    0.10290   1.844 0.065131 .
## f.distance(1.8,3.31]    0.54193    0.10387   5.217 1.81e-07 ***
## f.distance(3.31,11.5]   0.30084    0.14387   2.091 0.036527 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4381.8  on 3396  degrees of freedom
## AIC: 4401.8
##
## Number of Fisher Scoring iterations: 4
```

```
vif(m65)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## claHP      2.429199  6        1.076767
## f.distance 2.429199  3        1.159428
```

```
# We change to two numerical again and let's see which results we have from this interaction.
m56<-glm(AnyTip~ Pickup_latitude + Dropoff_longitude * f.distance, family="binomial",data=dfwork)
summary(m56)
```

```
##
```

```
## Call:
## glm(formula = AnyTip ~ Pickup_latitude + Dropoff_longitude *
##     f.distance, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5685  -1.0135  -0.7724   1.1661   2.1200
##
## Coefficients:
##                                      Estimate Std. Error z value
## (Intercept)                          -402.4197   156.3319  -2.574
## Pickup_latitude                        -2.0873     0.6569  -3.178
## Dropoff_longitude                      -6.5834     2.0387  -3.229
## f.distance(1.01,1.8]                 -366.0515   209.6026  -1.746
## f.distance(1.8,3.31]                 -364.3622   195.2787  -1.866
## f.distance(3.31,11.5]                 -26.0668   175.8894  -0.148
## Dropoff_longitude:f.distance(1.01,1.8]   -4.9532     2.8348  -1.747
## Dropoff_longitude:f.distance(1.8,3.31]   -4.9352     2.6411  -1.869
## Dropoff_longitude:f.distance(3.31,11.5]  -0.3631     2.3788  -0.153
##                                     Pr(>|z|)
## (Intercept)                          0.01005 *
## Pickup_latitude                      0.00148 **
## Dropoff_longitude                    0.00124 **
## f.distance(1.01,1.8]                 0.08074 .
## f.distance(1.8,3.31]                 0.06206 .
## f.distance(3.31,11.5]                0.88218
## Dropoff_longitude:f.distance(1.01,1.8]  0.08059 .
## Dropoff_longitude:f.distance(1.8,3.31]  0.06167 .
## Dropoff_longitude:f.distance(3.31,11.5] 0.87867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4393.2  on 3397  degrees of freedom
## AIC: 4411.2
##
## Number of Fisher Scoring iterations: 4
```

```
vif(m56)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## Pickup_latitude              1.043921e+00  1        1.021725
## Dropoff_longitude            6.245008e+00  1        2.499001
## f.distance                   3.150504e+19  3     1777.174198
## Dropoff_longitude:f.distance 3.150503e+19  3     1777.174126
```

```
#Since m56 it's not fitting well, we try a new step. m66 seems interesting, all their coefficients are
m66<-step(m56,k=log(nrow(dfwork)))
```

```
## Start:  AIC=4466.44
## AnyTip ~ Pickup_latitude + Dropoff_longitude * f.distance
##
##                                Df Deviance    AIC
## - Dropoff_longitude:f.distance  3   4401.2 4450.0
```

```
## <none>                               4393.2 4466.4
## - Pickup_latitude               1   4403.4 4468.4
##
## Step:  AIC=4450.03
## AnyTip ~ Pickup_latitude + Dropoff_longitude + f.distance
##
##                      Df Deviance    AIC
## <none>                    4401.2 4450.0
## - Pickup_latitude     1   4412.2 4452.9
## - f.distance          3   4472.8 4497.2
## - Dropoff_longitude   1   4525.3 4566.0
```

```
summary(m66)
```

```
##
## Call:
## glm(formula = AnyTip ~ Pickup_latitude + Dropoff_longitude +
##     f.distance, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6506  -1.0083  -0.7783   1.1852   1.9552
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -563.5696    70.5772  -7.985 1.40e-15 ***
## Pickup_latitude         -2.1579     0.6531  -3.304 0.000952 ***
## Dropoff_longitude       -8.8018     0.8264 -10.651  < 2e-16 ***
## f.distance(1.01,1.8]     0.1868     0.1028   1.818 0.069060 .
## f.distance(1.8,3.31]     0.5415     0.1033   5.241 1.59e-07 ***
## f.distance(3.31,11.5]    0.7836     0.1032   7.596 3.05e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4401.2  on 3400  degrees of freedom
## AIC: 4413.2
##
## Number of Fisher Scoring iterations: 4
```

```
vif(m66)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Pickup_latitude   1.032079  1        1.015913
## Dropoff_longitude 1.028528  1        1.014164
## f.distance        1.004396  3        1.000731
```

```
# With all models built so far we got the best estimator on m51 and m62 (which are actually the same).
BIC(m51,m52,m53,m54,m55,m56,m62,m63,m64,m65,m66)
```

```
##     df      BIC
## m51  4 4436.077
## m52 28 4522.410
## m53 21 4486.338
```

```
## m54 30 4552.942
## m55 17 4484.098
## m56  9 4466.435
## m62  4 4436.077
## m63  8 4443.580
## m64 11 4439.794
## m65 10 4463.164
## m66  6 4450.032
```

```
m58<-glm(AnyTip~Dropoff_longitude:claHP + Dropoff_longitude:f.distance, family="binomial",data=dfwork)
summary(m58)
```

```
##
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude:claHP + Dropoff_longitude:f.distance,
##     family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6147  -1.0089  -0.7401   1.1877   2.5348
##
## Coefficients:
##                                         Estimate Std. Error z value
## (Intercept)                            -4.522e+02  8.153e+01  -5.546
## Dropoff_longitude:claHPkHP- 1          -6.105e+00  1.103e+00  -5.537
## Dropoff_longitude:claHPkHP- 2          -6.102e+00  1.104e+00  -5.529
## Dropoff_longitude:claHPkHP- 3          -6.110e+00  1.102e+00  -5.543
## Dropoff_longitude:claHPkHP- 4          -6.105e+00  1.103e+00  -5.536
## Dropoff_longitude:claHPkHP- 5          -6.100e+00  1.103e+00  -5.531
## Dropoff_longitude:claHPkHP- 6          -6.117e+00  1.103e+00  -5.548
## Dropoff_longitude:claHPkHP- 7          -6.082e+00  1.103e+00  -5.512
## Dropoff_longitude:f.distance(1.01,1.8] -2.546e-03  1.394e-03  -1.827
## Dropoff_longitude:f.distance(1.8,3.31] -7.302e-03  1.408e-03  -5.186
## Dropoff_longitude:f.distance(3.31,11.5] -4.210e-03  1.958e-03  -2.150
##                                        Pr(>|z|)
## (Intercept)                            2.92e-08 ***
## Dropoff_longitude:claHPkHP- 1          3.08e-08 ***
## Dropoff_longitude:claHPkHP- 2          3.22e-08 ***
## Dropoff_longitude:claHPkHP- 3          2.98e-08 ***
## Dropoff_longitude:claHPkHP- 4          3.09e-08 ***
## Dropoff_longitude:claHPkHP- 5          3.18e-08 ***
## Dropoff_longitude:claHPkHP- 6          2.89e-08 ***
## Dropoff_longitude:claHPkHP- 7          3.55e-08 ***
## Dropoff_longitude:f.distance(1.01,1.8]   0.0677 .
## Dropoff_longitude:f.distance(1.8,3.31] 2.15e-07 ***
## Dropoff_longitude:f.distance(3.31,11.5]  0.0316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 4350.3  on 3395  degrees of freedom
## AIC: 4372.3
##
```

```
## Number of Fisher Scoring iterations: 4
```

```
vif(m58)
```

```
##                                   GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude:claHP      2.418369  7        1.065110
## Dropoff_longitude:f.distance 2.418369  3        1.158565
```

```
BIC(m51,m58)
```

```
##     df      BIC
## m51  4 4436.077
## m58 11 4439.773
```

We decide to incorporate f.total as a coefficient

```
m59<-glm(AnyTip~ Dropoff_longitude + Pickup_latitude + f.total + f.distance + claHP, family="binomial",
summary(m59)
```

```
##
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude + Pickup_latitude +
##     f.total + f.distance + claHP, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5438  -0.8904  -0.4544   0.9914   3.4159
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -355.35413   97.02002  -3.663 0.000250 ***
## Dropoff_longitude       -5.48792    1.14683  -4.785 1.71e-06 ***
## Pickup_latitude         -1.26853    1.09090  -1.163 0.244902
## f.total(7.8,11]          2.39621    0.17318  13.836  < 2e-16 ***
## f.total(11,16.6]         3.97373    0.23384  16.993  < 2e-16 ***
## f.total(16.6,46]         5.49009    0.29792  18.428  < 2e-16 ***
## f.distance(1.01,1.8]    -1.51894    0.17029  -8.920  < 2e-16 ***
## f.distance(1.8,3.31]    -2.78687    0.22657 -12.301  < 2e-16 ***
## f.distance(3.31,11.5]   -4.00206    0.28715 -13.937  < 2e-16 ***
## claHPkHP- 2             -0.39754    0.17196  -2.312 0.020787 *
## claHPkHP- 3              0.13937    0.15645   0.891 0.373019
## claHPkHP- 4             -0.05168    0.19885  -0.260 0.794945
## claHPkHP- 5             -2.10703    0.72086  -2.923 0.003467 **
## claHPkHP- 6              0.18241    0.19412   0.940 0.347386
## claHPkHP- 7             -2.14548    0.65109  -3.295 0.000983 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 3840.0  on 3391  degrees of freedom
## AIC: 3870
##
## Number of Fisher Scoring iterations: 5
```

```r
vif(m59)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude  1.746741  1        1.321643
## Pickup_latitude    2.427871  1        1.558163
## f.total           33.560047  3        1.795989
## f.distance        31.691401  3        1.778922
## claHP             12.525296  6        1.234475
```

```r
BIC(m51,m59)
```

```
##     df      BIC
## m51  4 4436.077
## m59 15 3961.972
```

```r
#We apply step in order to simplify
m69<-step(m59,k=log(nrow(dfwork)))
```

```
## Start:  AIC=3961.97
## AnyTip ~ Dropoff_longitude + Pickup_latitude + f.total + f.distance +
##     claHP
##
##                      Df Deviance    AIC
## - claHP               6   3877.4 3950.6
## - Pickup_latitude     1   3841.3 3955.2
## <none>                    3840.0 3962.0
## - Dropoff_longitude   1   3863.3 3977.1
## - f.distance          3   4094.4 4192.0
## - f.total             3   4348.5 4446.1
##
## Step:  AIC=3950.56
## AnyTip ~ Dropoff_longitude + Pickup_latitude + f.total + f.distance
##
##                      Df Deviance    AIC
## - Pickup_latitude     1   3884.0 3949.1
## <none>                    3877.4 3950.6
## - Dropoff_longitude   1   3971.5 4036.6
## - f.distance          3   4129.0 4177.8
## - f.total             3   4401.2 4450.0
##
## Step:  AIC=3949.08
## AnyTip ~ Dropoff_longitude + f.total + f.distance
##
##                      Df Deviance    AIC
## <none>                    3884.0 3949.1
## - Dropoff_longitude   1   3988.3 4045.2
## - f.distance          3   4135.3 4176.0
## - f.total             3   4412.2 4452.9
```

```r
summary(m69)
```

```
##
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude + f.total + f.distance,
##     family = "binomial", data = dfwork)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7649  -0.9205  -0.5026   1.0055   2.9028
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -633.5984    64.2131  -9.867   <2e-16 ***
## Dropoff_longitude       -8.5518     0.8684  -9.847   <2e-16 ***
## f.total(7.8,11]          2.3265     0.1693  13.741   <2e-16 ***
## f.total(11,16.6]         3.8754     0.2298  16.861   <2e-16 ***
## f.total(16.6,46]         5.4493     0.2853  19.102   <2e-16 ***
## f.distance(1.01,1.8]    -1.4523     0.1665  -8.721   <2e-16 ***
## f.distance(1.8,3.31]    -2.7048     0.2230 -12.131   <2e-16 ***
## f.distance(3.31,11.5]   -3.8591     0.2778 -13.890   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 3884.0  on 3398  degrees of freedom
## AIC: 3900
##
## Number of Fisher Scoring iterations: 4
```

```
#colinearity
vif(m69)
```

```
##                        GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude  1.002816  1        1.001407
## f.total           27.270596  3        1.734932
## f.distance        27.228655  3        1.734487
```

```
m1<-glm(AnyTip~ (Dropoff_longitude + Trip_distance)*f.total, family="binomial",data=dfwork)
summary(m1)
```

```
##
## Call:
## glm(formula = AnyTip ~ (Dropoff_longitude + Trip_distance) *
##     f.total, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2991  -0.8723  -0.4640   0.9330   6.2709
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -695.0903   196.3820  -3.539 0.000401
## Dropoff_longitude                      -9.3967     2.6559  -3.538 0.000403
## Trip_distance                          -1.6036     0.3094  -5.183 2.18e-07
## f.total(7.8,11]                       165.5282   245.3239   0.675 0.499845
## f.total(11,16.6]                     -162.4681   240.7726  -0.675 0.499817
## f.total(16.6,46]                      240.2689   217.8724   1.103 0.270116
## Dropoff_longitude:f.total(7.8,11]       2.1931     3.3179   0.661 0.508607
## Dropoff_longitude:f.total(11,16.6]     -2.2421     3.2562  -0.689 0.491097
## Dropoff_longitude:f.total(16.6,46]      3.2330     2.9465   1.097 0.272538
```

```
## Trip_distance:f.total(7.8,11]        -0.8752     0.3830  -2.285 0.022317
## Trip_distance:f.total(11,16.6]        0.3693     0.3340   1.106 0.268846
## Trip_distance:f.total(16.6,46]        1.5163     0.3114   4.869 1.12e-06
##
## (Intercept)                          ***
## Dropoff_longitude                    ***
## Trip_distance                        ***
## f.total(7.8,11]
## f.total(11,16.6]
## f.total(16.6,46]
## Dropoff_longitude:f.total(7.8,11]
## Dropoff_longitude:f.total(11,16.6]
## Dropoff_longitude:f.total(16.6,46]
## Trip_distance:f.total(7.8,11]        *
## Trip_distance:f.total(11,16.6]
## Trip_distance:f.total(16.6,46]       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 3802.6  on 3394  degrees of freedom
## AIC: 3826.6
##
## Number of Fisher Scoring iterations: 5
```

```
vif(m1)
```

```
##                                   GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude         9.137227e+00  1        3.022785
## Trip_distance             3.023492e+02  1       17.388191
## f.total                   3.370520e+19  3     1797.281711
## Dropoff_longitude:f.total 3.369538e+19  3     1797.194387
## Trip_distance:f.total     4.741568e+04  3        6.016173
```

```
m11<-step(m1,k=log(nrow(dfwork)))
```

```
## Start:  AIC=3900.17
## AnyTip ~ (Dropoff_longitude + Trip_distance) * f.total
##
##                             Df Deviance    AIC
## - Dropoff_longitude:f.total  3   3808.9 3882.1
## <none>                           3802.6 3900.2
## - Trip_distance:f.total      3   4044.9 4118.1
##
## Step:  AIC=3882.13
## AnyTip ~ Dropoff_longitude + Trip_distance + f.total + Trip_distance:f.total
##
##                           Df Deviance    AIC
## <none>                         3808.9 3882.1
## - Dropoff_longitude       1   3896.9 3962.0
## - Trip_distance:f.total   3   4055.4 4104.2
```

```r
summary(m11)
```

```
##
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude + Trip_distance + f.total +
##     Trip_distance:f.total, family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2960  -0.8711  -0.4725   0.9280   6.2921
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -591.1911    64.9355  -9.104  < 2e-16 ***
## Dropoff_longitude                 -7.9916     0.8782  -9.100  < 2e-16 ***
## Trip_distance                     -1.6044     0.3089  -5.194 2.06e-07 ***
## f.total(7.8,11]                    3.3681     0.3898   8.640  < 2e-16 ***
## f.total(11,16.6]                   3.3424     0.3952   8.458  < 2e-16 ***
## f.total(16.6,46]                   1.2074     0.3141   3.844 0.000121 ***
## Trip_distance:f.total(7.8,11]     -0.8754     0.3827  -2.287 0.022189 *
## Trip_distance:f.total(11,16.6]     0.3560     0.3332   1.069 0.285244
## Trip_distance:f.total(16.6,46]     1.5195     0.3109   4.887 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 3808.9  on 3397  degrees of freedom
## AIC: 3826.9
##
## Number of Fisher Scoring iterations: 5
```

```r
#still colinearity present
vif(m11)
```

```
##                               GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude         1.003779  1        1.001888
## Trip_distance           296.021568  1       17.205277
## f.total                1408.443378  3        3.348034
## Trip_distance:f.total 46745.871748  3        6.001925
```

```r
#This seems to be also a good model
m2<-glm(AnyTip~ Dropoff_longitude:f.espeed + Trip_distance:f.total, family="binomial",data=dfwork)
summary(m2)
```

```
##
## Call:
## glm(formula = AnyTip ~ Dropoff_longitude:f.espeed + Trip_distance:f.total,
##     family = "binomial", data = dfwork)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1038  -0.9489  -0.4181   1.0016   5.0535
##
```

```
## Coefficients:
##                                      Estimate Std. Error z value
## (Intercept)                          -645.25233   64.77368  -9.962
## Dropoff_longitude:f.espeed(0,15.3]      -8.74473    0.87601  -9.982
## Dropoff_longitude:f.espeed(15.3,20.1]   -8.74814    0.87600  -9.987
## Dropoff_longitude:f.espeed(20.1,26.2]   -8.75267    0.87606  -9.991
## Dropoff_longitude:f.espeed(26.2,95]     -8.75510    0.87621  -9.992
## Trip_distance:f.total(-1,7.8]           -4.24481    0.22873 -18.558
## Trip_distance:f.total(7.8,11]           -1.52553    0.10515 -14.509
## Trip_distance:f.total(11,16.6]          -0.76745    0.05926 -12.951
## Trip_distance:f.total(16.6,46]          -0.22531    0.02536  -8.886
##                                     Pr(>|z|)
## (Intercept)                           <2e-16 ***
## Dropoff_longitude:f.espeed(0,15.3]    <2e-16 ***
## Dropoff_longitude:f.espeed(15.3,20.1] <2e-16 ***
## Dropoff_longitude:f.espeed(20.1,26.2] <2e-16 ***
## Dropoff_longitude:f.espeed(26.2,95]   <2e-16 ***
## Trip_distance:f.total(-1,7.8]         <2e-16 ***
## Trip_distance:f.total(7.8,11]         <2e-16 ***
## Trip_distance:f.total(11,16.6]        <2e-16 ***
## Trip_distance:f.total(16.6,46]        <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4631.2  on 3405  degrees of freedom
## Residual deviance: 3879.0  on 3397  degrees of freedom
## AIC: 3897
##
## Number of Fisher Scoring iterations: 5
```

```
vif(m2)
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## Dropoff_longitude:f.espeed 1.196623  4        1.022692
## Trip_distance:f.total      1.196623  4        1.022692
```

```
BIC(m2)
```

```
## [1] 3952.2
```

```
BIC(m51,m2)
```

```
##     df      BIC
## m51  4 4436.077
## m2   9 3952.200
```

We finally choose m2.

Diagnostics

Influent Data

In the plots below two big elements can be clearly distinguished over the rest: 759257 and 199161 (in the boxplot 1844 and 512 but are the same elements). One of them (the first one) has almost a 7 times bigger

cook distance value compared to the second and it is a clearly outlier from rstudent boxplot.

```
# Potentially influent data (we don't see any from hatvalues analysis)
quantile(hatvalues(m2),seq(0,1,0.1))
```

```
##           0%          10%          20%          30%          40%
## 6.886773e-10 1.824522e-03 2.019740e-03 2.159474e-03 2.295755e-03
##          50%          60%          70%          80%          90%
## 2.451398e-03 2.631464e-03 2.846392e-03 3.123606e-03 3.665565e-03
##         100%
## 9.518260e-03
```

```
mean(hatvalues(m2))
```

```
## [1] 0.002642396
```

```
hh<-5*mean(hatvalues(m2))
llhat<-which(hatvalues(m2)>hh);length(llhat)
```

```
## [1] 0
```

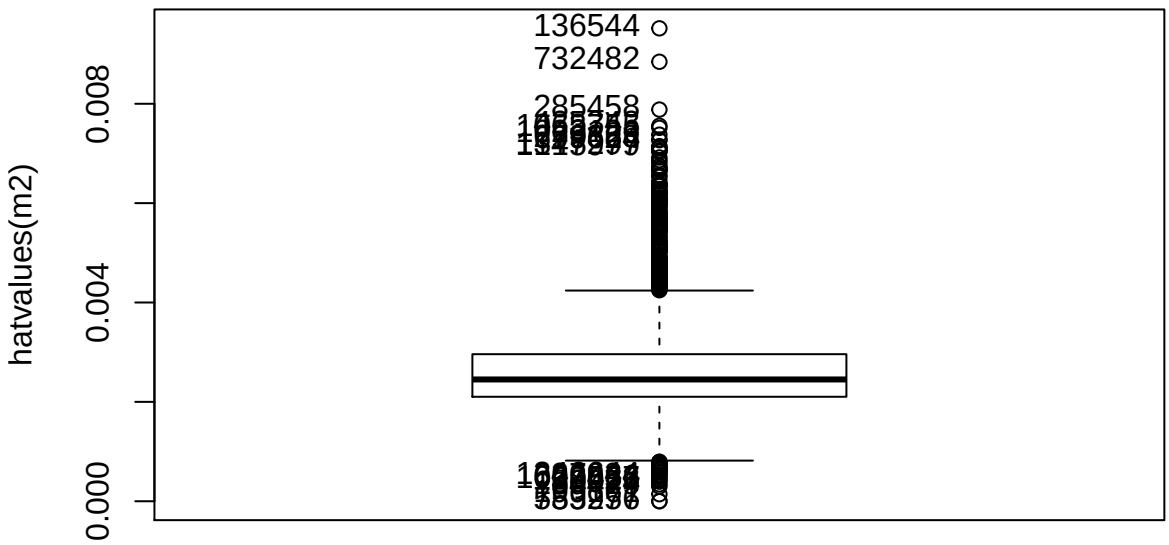
```
hh
```

```
## [1] 0.01321198
```

```
Boxplot(hatvalues(m2),labels=rownames(dfwork))
```

```
##  [1] "583976"  "759257"  "199161"  "134419"  "80074"   "39623"   "1008557"
##  [8] "622466"  "627934"  "1346644" "136544"  "732482"  "285458"  "425343"
## [15] "1063755" "1023126" "279353"  "699624"  "1347297" "1115979"
```

```
abline(h=hh,lwd=2,col="red",lty=2)
```

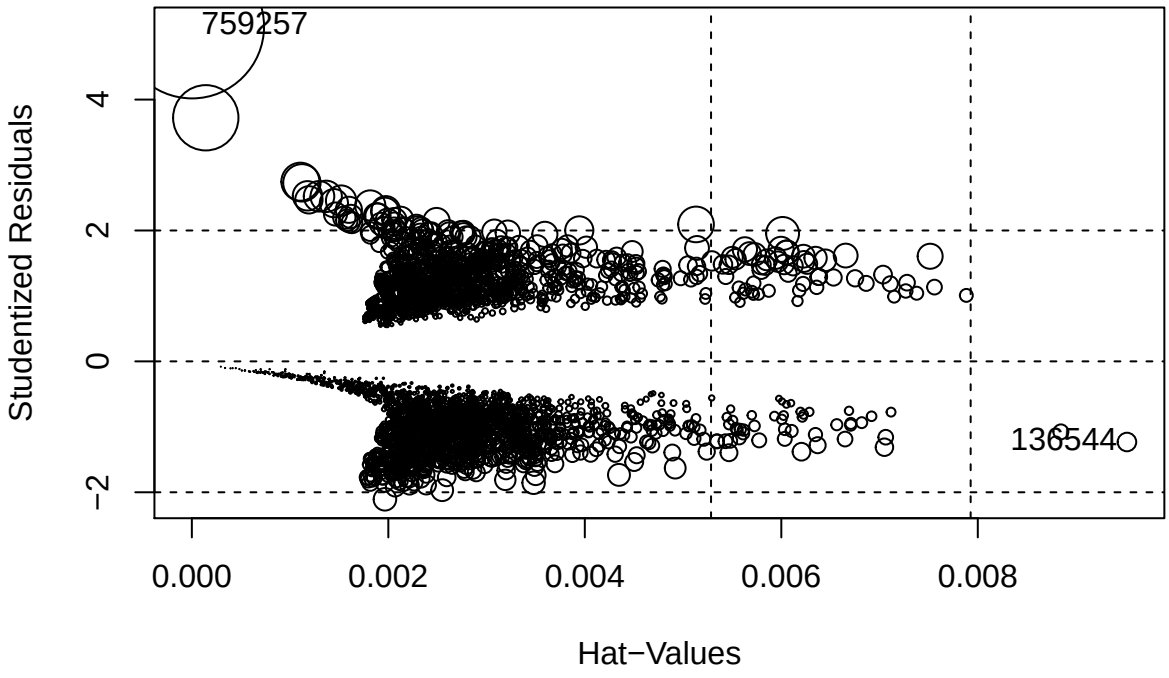



```
dfwork[llhat,]
```

```
##  [1] VendorID                lpep_pickup_datetime
##  [3] Lpep_dropoff_datetime   Store_and_fwd_flag
##  [5] RateCodeID              Pickup_longitude
##  [7] Pickup_latitude         Dropoff_longitude
##  [9] Dropoff_latitude        Passenger_count
## [11] Trip_distance           Fare_amount
## [13] Extra                   MTA_tax
## [15] Tip_amount              Tolls_amount
## [17] improvement_surcharge   Total_amount
## [19] Payment_type            Trip_type
## [21] mis_ind                 AnyTip
## [23] trip_length             trip_distance_km
## [25] travel_time             pick_up_hour
## [27] pick_up_period          espeed
## [29] f.passenger             f.distance
## [31] f.pickup_longitude      f.pickup_latitude
## [33] f.dropoff_longitude     f.dropoff_latitude
## [35] f.fare_amount           f.extra
## [37] f.MTA_tax               f.Improvement_surcharge
## [39] f.tip_amount            f.toll
## [41] f.total                 f.ttime
## [43] f.espeed                f.outlierPCAd1
## [45] f.outlierPCAd2          f.outlierPCAd3
## [47] f.outlierPCAd4          f.outlierPCA
```

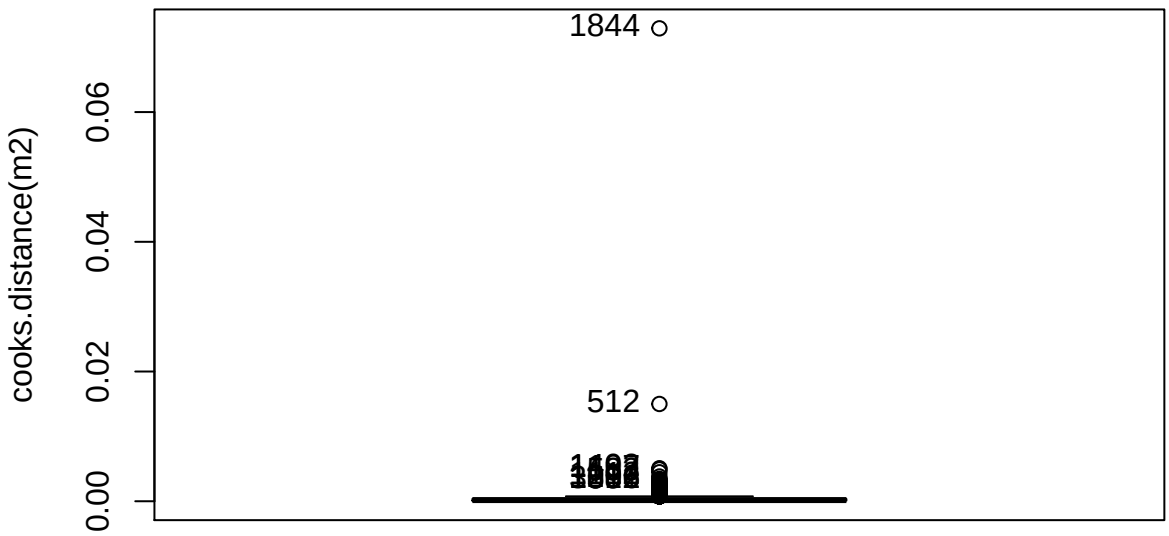
```
## [49] claHP                    claKM
## <0 rows> (or 0-length row.names)
```

```
ll1 <- influencePlot(m2); ll1
```



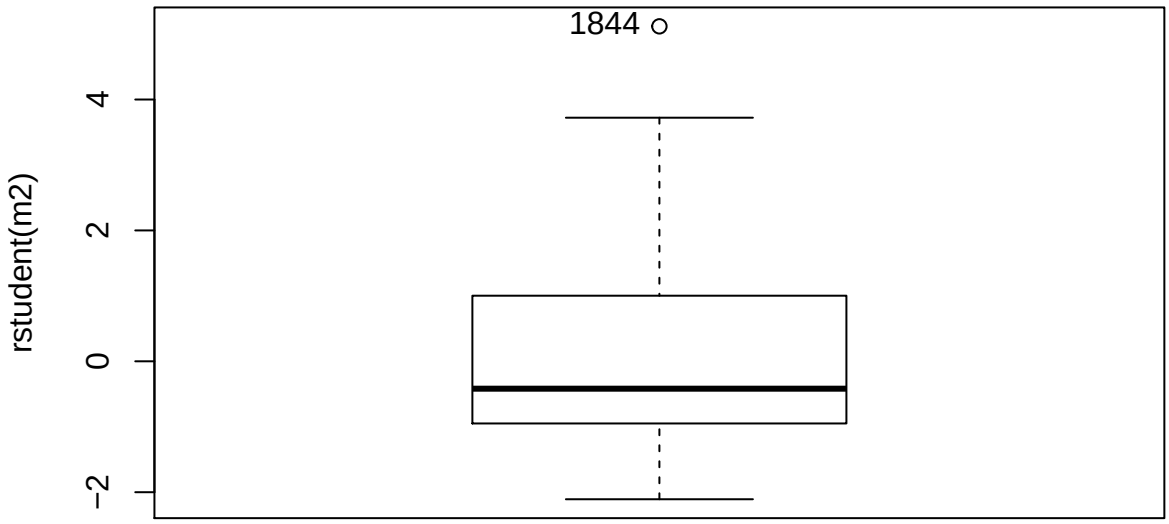

```
##          StudRes          Hat       CookD
## 136544 -1.231982 9.518260e-03 0.001212107
## 759257  5.118042 1.868453e-06 0.072909476
```

```
Boxplot(cooks.distance(m2))
```




```
##  [1] 1844  512 1402 1187  533 1112 1206 3012 1596 1881
```

```
Boxplot(rstudent(m2),id.n=15)
```




```
## [1] 1844
```

So we took a look at this very differentiate elements. We can observe that the most influential element (759257) it was outlier in the second dimension of PCA too.

```
llcoo<-which(cooks.distance(m2)>0.01);length(llcoo)
```

```
## [1] 2
```

```
dfwork[llcoo[1],]
```

```
##                                VendorID lpep_pickup_datetime
## 199161 Creative Mobile Technologies, LLC  2016-01-04 23:39:06
##        Lpep_dropoff_datetime Store_and_fwd_flag    RateCodeID
## 199161   2016-01-04 23:39:22  Not_Store_and_fwd Standard rate
##        Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
## 199161        -73.94524        40.75155         -73.94524         40.75155
##        Passenger_count Trip_distance Fare_amount Extra MTA_tax Tip_amount
## 199161               1             2         2.5   0.5     0.5       0.12
##        Tolls_amount improvement_surcharge Total_amount Payment_type
## 199161            0                   0.3         3.92  Credit card
##          Trip_type mis_ind     AnyTip trip_length trip_distance_km
## 199161 Street-hail       2 AnyTip Yes    2.119846         3.218688
##        travel_time pick_up_hour pick_up_period   espeed f.passenger
## 199161    6.805845           23      afternoon 18.68846       (0,1]
##        f.distance f.pickup_longitude f.pickup_latitude f.dropoff_longitude
## 199161 (1.8,3.31]    (-73.95,-73.92]      (40.74,40.8]     (-73.97,-73.94]
##        f.dropoff_latitude f.fare_amount    f.extra f.MTA_tax
```

```
## 199161      (40.75,40.79]         (0,6] (-0.1,0.5] (0.4,0.5]
##        f.Improvement_surcharge f.tip_amount f.toll  f.total  f.ttime
## 199161               (0.1,0.8]     (-0.1,1] (-1,1] (-1,7.8] (6,9.78]
##           f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 199161 (15.3,20.1]      NoOutDim1      NoOutDim2      NoOutDim3
##        f.outlierPCAd4 f.outlierPCA  claHP claKM
## 199161      NoOutDim4        NoOut kHP- 1 kKM-1
```

```
dfwork[llcoo[2],]
```

```
##                                VendorID lpep_pickup_datetime
## 759257 Creative Mobile Technologies, LLC  2016-01-16 20:50:09
##        Lpep_dropoff_datetime Store_and_fwd_flag   RateCodeID
## 759257   2016-01-16 20:50:51      Store_and_fwd Special rate
##        Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
## 759257        -73.85313        40.68809         -73.77808         40.68809
##        Passenger_count Trip_distance Fare_amount Extra MTA_tax Tip_amount
## 759257               1           8.7           8     0       0        0.8
##        Tolls_amount improvement_surcharge Total_amount Payment_type
## 759257            0                     0          8.8  Credit card
##        Trip_type mis_ind     AnyTip trip_length trip_distance_km
## 759257  Dispatch       5 AnyTip Yes    8.345335         14.00129
##        travel_time pick_up_hour pick_up_period   espeed f.passenger
## 759257    20.72006           20      afternoon 24.16596       (0,1]
##         f.distance f.pickup_longitude f.pickup_latitude
## 759257 (3.31,11.5]    (-73.92,-73.79]       (40.5,40.7]
##        f.dropoff_longitude f.dropoff_latitude f.fare_amount    f.extra
## 759257     (-73.91,-73.75]       (40.57,40.7]         (6,9] (-0.1,0.5]
##         f.MTA_tax f.Improvement_surcharge f.tip_amount f.toll  f.total
## 759257 (-0.1,0.4]              (-0.1,0.1]     (-0.1,1] (-1,1] (7.8,11]
##           f.ttime    f.espeed f.outlierPCAd1 f.outlierPCAd2 f.outlierPCAd3
## 759257 (15.7,215] (20.1,26.2]      NoOutDim1     YesOutDim2      NoOutDim3
##        f.outlierPCAd4 f.outlierPCA  claHP claKM
## 759257      NoOutDim4       YesOut kHP- 7 kKM-7
```

We decide to remove only the outlier.

```
df1<-dfwork[-llcoo[2],]
```

Predicting

Finally, we execute the predictions and generate a confusion matrix with the results of it. To conclude, the diagonal of this confusion matrix shows a performance of 70% hit rate in our prediction model against our testing data set, while the model 0 without any coefficient would give us only a 41% of hit rate.

Curve ROC of our choosen method is displayed at the end.

```
tfit2<-predict(m2,type="response",newdata=dftest)
fit.AnyTip<-factor(ifelse(predict(m2,type="response")<0.5,0,1),labels=c("fit.No","fit.Yes"))
tt<-table(fit.AnyTip,dfwork$AnyTip);tt
```

```
##
## fit.AnyTip AnyTip No AnyTip Yes
##    fit.No       1525        533
##    fit.Yes       455        893
```

```r
sum(tt)
```

```
## [1] 3406
```

```r
100*sum(diag(tt))/sum(tt)
```

```
## [1] 70.99237
```

```r
m0<-glm(AnyTip~1, family="binomial", data=dfwork)
fit0<-predict(m0,type="response")
fit.AnyTip0<-factor(ifelse(fit0<0.5,0,1),labels=c("fit.Yes"))
tt0<-table(fit.AnyTip0,dfwork$AnyTip);tt0;sum(tt0)
```

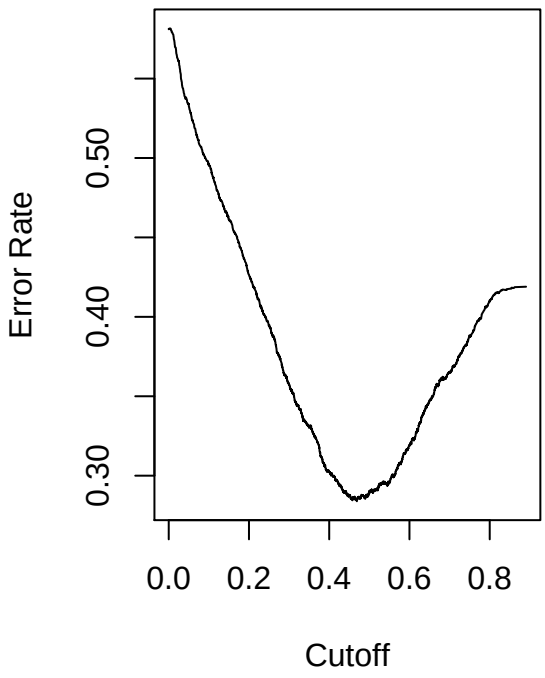
```
##
## fit.AnyTip0 AnyTip No AnyTip Yes
##     fit.Yes      1980       1426
```

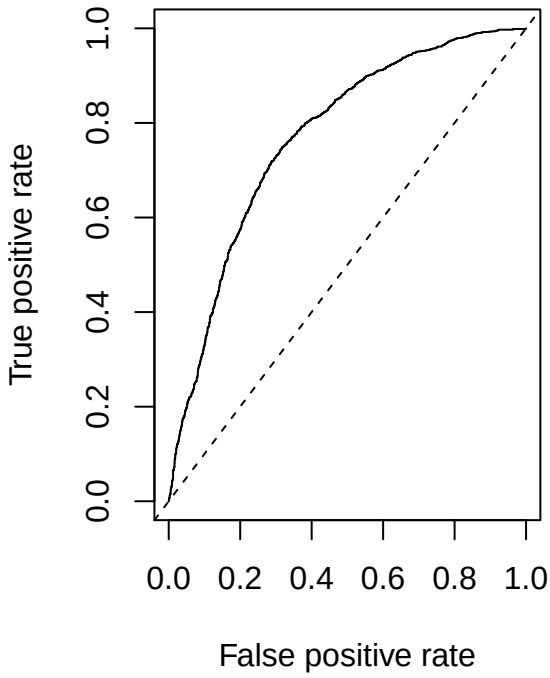
```
## [1] 3406
```

```r
100*sum(tt0[1,2])/sum(tt0)
```

```
## [1] 41.86729
```

```r
# Plot ROC curve
library("ROCR")
dadesroc<-prediction(predict(m2,type="response"),dfwork$AnyTip)
par(mfrow=c(1,2))
plot(performance(dadesroc,"err"))
plot(performance(dadesroc,"tpr","fpr"))
abline(0,1,lty=2)
```


Error Rate
0.30
0.40
0.50
0.0
0.2
0.4
0.6
0.8
Cutoff


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate