CASE_STUDY

Katerina Dimitrova, Jose Romero, Sergi Mu±oz

March 18, 2018

Introduction

Load requiered packages

Select 5000 samples

```
#Load samples

### Use birthday of 1 member of the group
set.seed(03101994)
nrow(df)
```

```
## [1] 4866
```

```
#sam<-sample(1:nrow(df),5000)
#sam<-as.vector(sort(sam))

#df<-df[sam,]
#setwd("/Users/Sergi/Desktop/Sergi/CABS")
#df<-read.table("green_tripdata_2016-01.csv",header=T, sep=",")
#save.image("Taxi5000_raw.RData") # Dont execute again since it will create a new data and the followin
```

Load functions

```
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

```
load("Taxi5000_raw2.RData")
table(df$Ehail_fee) ##Delete unnecessary row
```

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

```
df$Ehail_fee<-NULL

# Now one by one describe vars
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"
```

Converting numeric variables corresponding to qualitative concepts to factors

VendorID

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

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

```
errors<-which(df$VendorID==0.0);length(errors) #No errors
```

```
## [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)))
```

Code_Doc_files/figure-latex/unnamed-chunk-7-1.pdf

RateCodeID

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

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

```
errors<-which(df$RateCodeID==0.0);length(errors) #No errors
```

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

```
df$RateCodeID<-factor(df$RateCodeID,labels=c("Standard rate","JFK","Newark","Nassau or Westchester","Neg
table(df$RateCodeID)
```

```
##
##         Standard rate                   JFK                Newark
##                  4874                    14                     6
## Nassau or Westchester       Negotiated fare
##                     4                   102
```

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

Code_Doc_files/figure-latex/unnamed-chunk-8-1.pdf

Store_and_fwd_flag

```
#//first the N and than Y
missingData<-which(is.na(df$Store_and_fwd_flag));length(missingData) #No missing Data
```

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

```
erors<-which(df$Store_and_fwd_flag==0.0);length(errors) #No errors
```

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

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

```
##
## not a store and forward trip       store and forward trip
##                         4982                           18
```

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

Code_Doc_files/figure-latex/unnamed-chunk-9-1.pdf

## Payment_type

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

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

```
errors<-which(df$Payment_type==0.0);length(errors) #No errors
```

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

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

```
##
## Credit card        Cash   No charge     Dispute
##        2469        2485          23          23
```

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

Code_Doc_files/figure-latex/unnamed-chunk-10-1.pdf

Trip_type

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

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

```
errors<-which(df$Trip_type==0.0);length(errors) #No errors
```

```
## [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)))
```

Code_Doc_files/figure-latex/unnamed-chunk-11-1.pdf

Univariant Descriptive Analysis

Passenger_count

```
## Number of missing values:

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

Code_Doc_files/figure-latex/unnamed-chunk-12-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-12-2.pdf

Trip_distance

```
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
```

```
dfaux<-df
ll<-which(is.na(df$Trip_distance));ll
```

```
## integer(0)
```

```
if(length(ll)>0){
  dfaux<-df[-ll,]
}
iqrvar<-IQR(dfaux$Trip_distance)
quantil3<-quantile(dfaux$Trip_distance, .75);quantil3 #get 3rd quartile
```

```
##    75%
## 3.4125
```

```
outliers<-which(df$Trip_distance>(iqrvar*3)+quantil3);length(outliers) #138 extreme outliers
```

```
## [1] 138
```

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

Code_Doc_files/figure-latex/unnamed-chunk-13-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-13-2.pdf

Pickup_longitude

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

```
## [1] 197
```

```
#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
```

```
df[errors,"Pickup_longitude"]<-NA #12 errors

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

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

```
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
```

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

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

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

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

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

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

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

Code_Doc_files/figure-latex/unnamed-chunk-14-1.pdf

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -74.04  -73.96  -73.95  -73.94  -73.92  -73.79      22
```

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"]<-NA #17 errors

ll<-which(is.na(df$Pickup_latitude));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] 0
```

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

Code_Doc_files/figure-latex/unnamed-chunk-15-1.pdf

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   40.58   40.69   40.75   40.75   40.80   40.91      12
```

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"]<-NA

ll<-which(is.na(df$Dropoff_longitude));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$Dropoff_longitude)
quantil3<-quantile(dfaux$Dropoff_longitude, .75);quantil3 #get 3rd quartile
```

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

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

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

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

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

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

```
## [1] 5
```

```
df[UpperOutlier,"Dropoff_longitude"]<-NA
df[LowerOutlier,"Dropoff_longitude"]<-NA

boxplot(df$Dropoff_longitude)
```

Code_Doc_files/figure-latex/unnamed-chunk-16-1.pdf

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"]<-NA

ll<-which(is.na(df$Dropoff_latitude));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] 0
```

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

boxplot(df$Dropoff_latitude)
```

Code_Doc_files/figure-latex/unnamed-chunk-17-1.pdf

Fare_amount

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

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

```
sel<-which(df$Fare_amount<=0.0);length(sel) #10 missings
```

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

```
outlier<-which(df$Fare_amount>100);length(outlier) #1 outlier
```

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

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

Code_Doc_files/figure-latex/unnamed-chunk-18-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-18-2.pdf

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     0.1     6.0     9.0    12.0    14.5    95.5      26
```

Extra

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

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

```
sel<-which(df$Extra<0.0);length(sel) #10 missings
```

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

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

Code_Doc_files/figure-latex/unnamed-chunk-19-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-19-2.pdf

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) #103 missings
```

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

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

Code_Doc_files/figure-latex/unnamed-chunk-20-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-20-2.pdf

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)
```

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

```
df[sel,"improvement_surcharge"]<-NA
boxplot(df$improvement_surcharge)
```

Code_Doc_files/figure-latex/unnamed-chunk-21-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-21-2.pdf

Tip_amount

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

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

```
sel<-which(df$Tip_amount<0.0);length(sel) #107 missings
```

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

```
outlier<-which(df$Tip_amount>60.0);length(outlier) #1 missings
```

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

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

Code_Doc_files/figure-latex/unnamed-chunk-22-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-22-2.pdf

Tolls_amount

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

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

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

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

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

Code_Doc_files/figure-latex/unnamed-chunk-23-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-23-2.pdf

Total_amount

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

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

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

```
## integer(0)
```

```
if(length(ll)>0){
  dfaux<-df[-ll,]
}
iqrvar<-IQR(dfaux$Total_amount)
quantil3<-quantile(dfaux$Total_amount, .75) #get 3rd quartile
sel<-which(df$Total_amount<=0.0);length(sel) #22 errors
```

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

```
df[sel,"Total_amount"]<-NA
outlier<-which(df$Total_amount>(iqrvar*3)+quantil3);length(outlier) #72 extreme outliers
```

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

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

Code_Doc_files/figure-latex/unnamed-chunk-24-1.pdf

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

Code_Doc_files/figure-latex/unnamed-chunk-24-2.pdf

Number of missing values:

```r
mis1<-countNA(df)
attributes(mis1)
```

```
## $names
## [1] "mis_col" "mis_ind"
```

```r
#sort(mis1$mis_col)
df$mis_ind <- mis1$mis_ind # new attribute missing values
mis1$mis_col
```

```
##                         mis_x
## VendorID                    0
## lpep_pickup_datetime        0
## Lpep_dropoff_datetime       0
## Store_and_fwd_flag          0
## RateCodeID                  0
## Pickup_longitude           22
## Pickup_latitude            12
## Dropoff_longitude          32
## Dropoff_latitude           20
## Passenger_count             2
## Trip_distance             197
## Fare_amount                26
## Extra                       4
## MTA_tax                    10
## Tip_amount                  4
## Tolls_amount                0
## improvement_surcharge      10
## Total_amount              134
## Payment_type                0
## Trip_type                   0
```

Declaring vectors of data

```r
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"
```

```r
vars_con<-names(df)[c(6:9,11:18)]
vars_dis<-names(df)[c(1,4,5,19,20:23)]
vars_res<-names(df)[c(18,23)]

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"
```

```r
vars_con<-names(df)[c(6,7,8,9,11,12,13,14,15,16,18)]
vars_dis<-names(df)[c(1,5,10,20,21)]
vars_res<-names(df)[c(19,22)]
vars_res
```

```
## [1] "Payment_type" NA
```

```r
vars_dis
```

```
## [1] "VendorID"        "RateCodeID"      "Passenger_count" "Trip_type"
## [5] "mis_ind"
```

```r
vars_con
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

Multivariant Outlier Detection

```
#install.packages("mvoutlier")
library(sgeostat)
library(mvoutlier)

vars_con # Problems c(5,8,9,10,11,12)
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

```
summary(df[,vars_con])
```

```
##  Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
##  Min.   :-74.04   Min.   :40.58   Min.   :-74.05    Min.   :40.57
##  1st Qu.:-73.96   1st Qu.:40.69   1st Qu.:-73.97    1st Qu.:40.70
##  Median :-73.95   Median :40.75   Median :-73.95    Median :40.75
##  Mean   :-73.94   Mean   :40.75   Mean   :-73.94    Mean   :40.74
##  3rd Qu.:-73.92   3rd Qu.:40.80   3rd Qu.:-73.91    3rd Qu.:40.79
##  Max.   :-73.79   Max.   :40.91   Max.   :-73.75    Max.   :41.02
##  NA's   :22       NA's   :12      NA's   :32        NA's   :20
##  Trip_distance     Fare_amount       Extra           MTA_tax
##  Min.   : 0.010   Min.   : 0.1   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 1.010   1st Qu.: 6.0   1st Qu.:0.0000   1st Qu.:0.5000
##  Median : 1.790   Median : 9.0   Median :0.5000   Median :0.5000
##  Mean   : 2.482   Mean   :12.0   Mean   :0.3461   Mean   :0.4889
##  3rd Qu.: 3.285   3rd Qu.:14.5   3rd Qu.:0.5000   3rd Qu.:0.5000
##  Max.   :10.610   Max.   :95.5   Max.   :2.0000   Max.   :0.5000
##  NA's   :197      NA's   :26     NA's   :4        NA's   :10
##    Tip_amount      Tolls_amount      Total_amount
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.10
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 7.80
##  Median : 0.000   Median : 0.0000   Median :11.00
##  Mean   : 1.277   Mean   : 0.1141   Mean   :13.49
##  3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.:16.62
##  Max.   :60.000   Max.   :12.5000   Max.   :45.42
##  NA's   :4                          NA's   :134
```

```
vars_con_out<-vars_con[c(1:4)]

dim(vars_con2)
```

```
## NULL
```

```
#aq.plot(df[,vars_con_out]) # Problems when few numeric values are present in one variable

# Use common sense, but technicalities might difficult the application of the procedure

vars_con_out<-vars_con[c(1:4)]
#mvout<-aq.plot(df[,vars_con_out])  # Problems when missing data are present

# Use common sense
vars_con
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

```
vars_con_out<-vars_con[c(6,9,12)]
#aq.plot(df[,vars_con_out])  # Problems when missing data are present
vars_con_out
```

```
## [1] "Fare_amount" "Tip_amount"  NA
```

Correlations error variable

```
#install.packages("polycor")
library(polycor)
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"
```

```
summary(df$mis_ind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.0946  0.0000 10.0000
```

```
corV <- cor(df[,vars_con], df$mis_ind,  use = "complete.obs")
corV
```

```
##                           [,1]
## Pickup_longitude   0.001458769
## Pickup_latitude    0.014443265
## Dropoff_longitude  0.026321051
## Dropoff_latitude   0.019551876
## Trip_distance      0.019762846
## Fare_amount        0.002546080
## Extra             -0.014100839
## MTA_tax           -0.127222906
## Tip_amount        -0.009256682
## Tolls_amount      -0.001667617
## Total_amount      -0.002372345
```

```
 # rank
rank(corV)
```

```
##  [1]  6  8 11  9 10  7  2  1  3  5  4
```

Imputation

Remove observations with NA at targets

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

```
##   [1]   57   74   82  145  176  247  323  333  351  404  454  460  468  472
##  [15]  526  553  609  637  690  734  745  825  831  883  907 1001 1022 1059
##  [29] 1062 1078 1082 1100 1105 1130 1159 1331 1361 1367 1368 1395 1421 1657
##  [43] 1689 1697 1723 1759 1761 1780 1854 1867 1905 2004 2069 2106 2140 2187
##  [57] 2249 2257 2334 2335 2411 2413 2428 2490 2506 2575 2634 2698 2722 2744
##  [71] 2840 2842 2845 2866 2874 2919 2971 2981 3005 3054 3067 3101 3181 3197
##  [85] 3293 3295 3346 3412 3484 3541 3705 3742 3759 3787 3788 3802 3803 3813
##  [99] 3874 3894 3933 3936 3947 3953 3988 4063 4075 4164 4206 4222 4252 4294
## [113] 4328 4348 4370 4391 4418 4431 4524 4574 4576 4597 4605 4659 4687 4700
## [127] 4714 4733 4778 4817 4890 4920 4923 4968
```

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

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

```
## integer(0)
```

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

Definition of binary outcome: AnyTip

```
# Binary Target: Any Tip?


df$AnyTip<-ifelse(df$Tip_amount<0.0001,0,1)
df$AnyTip<-factor(df$AnyTip,labels=paste("AnyTip",c("No","Yes")))
#IMPORTANT
#if you touch some "global" variables you  will modify this part
# Now one by one describe vars
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"
```

```r
vars_con<-names(df)[c(6,7,8,9,11,12,13,14,15,16,18)]
vars_dis<-names(df)[c(1,5,10,20,21)]
vars_res<-names(df)[c(19,22)]
vars_res
```

```
## [1] "Payment_type" "AnyTip"
```

```r
vars_dis
```

```
## [1] "VendorID"        "RateCodeID"      "Passenger_count" "Trip_type"
## [5] "mis_ind"
```

```r
vars_con
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

Imputation of numeric variables

```r
#install.packages("missMDA")
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] "mis_ind"               "AnyTip"
```

```r
vars_con_mis<-vars_con;length(vars_con_mis)
```

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

```r
vars_con_mis
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

```r
summary(df[,vars_con_mis])
```

```
##  Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
##  Min.   :-74.03   Min.   :40.58   Min.   :-74.03    Min.   :40.57
##  1st Qu.:-73.96   1st Qu.:40.69   1st Qu.:-73.97    1st Qu.:40.70
##  Median :-73.95   Median :40.75   Median :-73.95    Median :40.75
##  Mean   :-73.94   Mean   :40.75   Mean   :-73.94    Mean   :40.74
##  3rd Qu.:-73.92   3rd Qu.:40.80   3rd Qu.:-73.91    3rd Qu.:40.79
##  Max.   :-73.79   Max.   :40.91   Max.   :-73.75    Max.   :40.91
##  NA's   :19       NA's   :10      NA's   :23        NA's   :17
##  Trip_distance     Fare_amount        Extra           MTA_tax
##  Min.   : 0.010   Min.   : 0.10   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 1.010   1st Qu.: 6.00   1st Qu.:0.0000   1st Qu.:0.5000
##  Median : 1.790   Median : 9.00   Median :0.5000   Median :0.5000
##  Mean   : 2.467   Mean   :11.15   Mean   :0.3497   Mean   :0.4915
##  3rd Qu.: 3.260   3rd Qu.:14.00   3rd Qu.:0.5000   3rd Qu.:0.5000
##  Max.   :10.610   Max.   :42.50   Max.   :2.0000   Max.   :0.5000
##  NA's   :108
##    Tip_amount      Tolls_amount       Total_amount
##  Min.   : 0.000   Min.   : 0.00000   Min.   : 0.10
##  1st Qu.: 0.000   1st Qu.: 0.00000   1st Qu.: 7.80
##  Median : 0.000   Median : 0.00000   Median :11.00
##  Mean   : 1.124   Mean   : 0.07864   Mean   :13.49
##  3rd Qu.: 2.000   3rd Qu.: 0.00000   3rd Qu.:16.62
##  Max.   :22.000   Max.   :12.50000   Max.   :45.42
##
```

```
res.comp <- imputePCA(df[,vars_con], ncp=4)
attributes(res.comp$completeObs)
```

```
## $dim
## [1] 4866   11
##
## $dimnames
## $dimnames[[1]]
##     [1] "285"    "307"    "401"    "593"    "636"    "886"    "904"
##     [8] "978"    "1135"   "1282"   "1409"   "1475"   "1495"   "1905"
##    [15] "2126"   "2151"   "2201"   "2271"   "2747"   "3065"   "3089"
##    [22] "3130"   "3221"   "3420"   "3679"   "4310"   "4754"   "5241"
##    [29] "5277"   "5649"   "6353"   "6364"   "6755"   "6869"   "7079"
##    [36] "7211"   "7342"   "7802"   "8138"   "8443"   "8619"   "8891"
##    [43] "8960"   "9207"   "9503"   "9747"   "9765"   "9984"   "10034"
##    [50] "10199"  "10951"  "10955"  "10974"  "11189"  "11506"  "11713"
##    [57] "12492"  "12792"  "13043"  "13274"  "13332"  "13875"  "13927"
##    [64] "14874"  "14916"  "15407"  "15830"  "16080"  "16166"  "16345"
##    [71] "16391"  "17136"  "17355"  "18278"  "18596"  "18734"  "19101"
##    [78] "19344"  "19408"  "19991"  "20004"  "20009"  "20044"  "20077"
##    [85] "20271"  "20342"  "20361"  "20543"  "20621"  "20733"  "20917"
##    [92] "21425"  "21439"  "21539"  "21559"  "21735"  "22197"  "22332"
##    [99] "22825"  "22946"  "23091"  "23132"  "23811"  "24338"  "24863"
##   [106] "25262"  "25356"  "26062"  "26832"  "27216"  "27482"  "27495"
##   [113] "27594"  "27984"  "28083"  "28512"  "29375"  "29522"  "30659"
##   [120] "30856"  "31236"  "31456"  "31571"  "31583"  "31617"  "31726"
##   [127] "32873"  "32952"  "33882"  "34250"  "34280"  "34374"  "34390"
##   [134] "34922"  "35039"  "35207"  "35386"  "36076"  "36428"  "36540"
##   [141] "36696"  "36863"  "36933"  "37035"  "37273"  "37506"  "37517"
##   [148] "37561"  "37764"  "37821"  "37877"  "38445"  "38480"  "39213"
##   [155] "39623"  "39723"  "39943"  "40226"  "40245"  "40497"  "40560"
##   [162] "40802"  "40941"  "40943"  "40953"  "40969"  "42048"  "42779"
##   [169] "43577"  "43958"  "44992"  "46311"  "46572"  "46653"  "46790"
##   [176] "47428"  "47471"  "48166"  "48518"  "48796"  "48903"  "48915"
##   [183] "49242"  "49244"  "49383"  "49421"  "49783"  "49849"  "50027"
##   [190] "50328"  "50542"  "50979"  "50996"  "51868"  "51965"  "52728"
##   [197] "52825"  "52931"  "53452"  "53536"  "53680"  "54025"  "54342"
## [204] "54359"  "54794"  "54843"  "54958"  "54994"  "55030"  "55082" 
## [211] "55144"  "55353"  "55479"  "55718"  "56090"  "56696"  "56726" 
## [218] "56914"  "56920"  "57200"  "57278"  "57590"  "58422"  "58631" 
## [225] "59389"  "59449"  "59578"  "59938"  "60074"  "60146"  "60728" 
## [232] "61110"  "61236"  "61265"  "61370"  "61424"  "61547"  "61650" 
## [239] "61948"  "61959"  "62009"  "62273"  "62544"  "62605"  "62911" 
## [246] "62949"  "63123"  "63251"  "63256"  "63814"  "64004"  "64740" 
## [253] "64772"  "64773"  "65262"  "65285"  "65688"  "65815"  "65878" 
## [260] "66075"  "66344"  "66764"  "66777"  "66868"  "66995"  "67087" 
## [267] "67205"  "67465"  "67583"  "67752"  "67849"  "68112"  "69030" 
## [274] "69045"  "69361"  "69625"  "69718"  "70070"  "70657"  "71033" 
## [281] "71191"  "71590"  "71898"  "72223"  "72871"  "73135"  "73254" 
## [288] "73256"  "73529"  "73585"  "73593"  "73666"  "73710"  "74043" 
## [295] "74216"  "74784"  "75165"  "75448"  "75715"  "75730"  "76378" 
## [302] "76595"  "77764"  "77948"  "77969"  "78118"  "78368"  "78598" 
## [309] "78646"  "79148"  "79658"  "79861"  "80074"  "80093"  "80326" 
## [316] "80599"  "80754"  "81034"  "81302"  "81813"  "82015"  "82045" 
## [323] "82439"  "83356"  "83371"  "84398"  "84735"  "84843"  "85100" 
## [330] "85254"  "85340"  "85766"  "86980"  "87246"  "87458"  "87900" 
## [337] "89079"  "89243"  "89853"  "90225"  "90238"  "90794"  "91078" 
## [344] "91083"  "91243"  "91511"  "91550"  "92387"  "92587"  "92598" 
## [351] "93699"  "93809"  "94247"  "94305"  "94495"  "95014"  "95226" 
## [358] "95241"  "95530"  "96119"  "96194"  "96298"  "96544"  "96980" 
## [365] "97571"  "98401"  "98688"  "98906"  "98945"  "98956"  "98966" 
## [372] "98981"  "99572"  "99893"  "99988"  "100096" "100198" "100571"
## [379] "100795" "101048" "101193" "101216" "101421" "101664" "101790"
## [386] "102224" "102368" "102932" "103000" "103682" "103858" "104389"
## [393] "104792" "105704" "106216" "106842" "106937" "107062" "107417"
## [400] "107932" "108185" "108201" "108206" "108304" "108334" "108479"
## [407] "108515" "108606" "108839" "108929" "109040" "109260" "109333"
## [414] "109734" "110047" "110199" "110565" "110673" "110873" "110913"
## [421] "110937" "111064" "111150" "111223" "111543" "112353" "112893"
## [428] "112901" "113083" "113597" "113753" "113964" "114436" "115174"
## [435] "115978" "116095" "116206" "116369" "116640" "117031" "117517"
## [442] "118785" "119035" "119554" "120535" "120802" "121010" "121184"
## [449] "121442" "121485" "121530" "122037" "122299" "122910" "123241"
## [456] "124259" "124620" "124715" "126173" "126470" "126481" "126587"
## [463] "126592" "126715" "127134" "127652" "127683" "127966" "128224"
## [470] "128390" "128587" "128926" "128937" "129264" "129479" "129660"
## [477] "129793" "129938" "129958" "129974" "130639" "131180" "131369"
## [484] "131482" "131592" "131736" "132350" "132383" "132433" "132534"
## [491] "132670" "132761" "133262" "133422" "133475" "134419" "135413"
## [498] "135495" "135800" "135935" "136039" "136229" "136265" "136309"
## [505] "136416" "136544" "136888" "137150" "137172" "137527" "138047"
## [512] "139374" "139883" "140233" "140567" "141027" "141534" "141835"
## [519] "141905" "141983" "142214" "142290" "142398" "142615" "142824"
## [526] "143558" "144638" "144756" "145268" "145938" "147177" "147246"
## [533] "147283" "147486" "147510" "147625" "147678" "148258" "148591"
## [540] "149155" "149218" "149842" "149887" "149898" "150136" "151342"
## [547] "151412" "151661" "151767" "151963" "152346" "152470" "152513"
## [554] "152725" "152828" "153560" "153673" "153796" "153878" "154253"
## [561] "154323" "154578" "154581" "154751" "155031" "155155" "155371"
## [568] "155384" "155441" "155551" "155566" "155950" "156244" "156250"
## [575] "156393" "156707" "156954" "157064" "157195" "157324" "158033"
## [582] "158490" "158962" "159223" "159480" "159725" "159831" "160040"
## [589] "160092" "160337" "160606" "160748" "161313" "161421" "161481"
## [596] "161512" "162070" "162131" "162954" "163032" "163237" "163255"
## [603] "163666" "163693" "164248" "164387" "165000" "165535" "165920"
## [610] "166238" "167445" "167876" "169257" "169467" "169526" "169710"
## [617] "169822" "169826" "169833" "170195" "170331" "170391" "171098"
## [624] "171131" "171189" "171559" "171702" "172037" "172208" "172245"
## [631] "172772" "172876" "172903" "173364" "173415" "173788" "173824"
## [638] "173854" "174720" "174998" "175544" "175695" "175809" "175857"
## [645] "176099" "176276" "176434" "176666" "177236" "177344" "177354"
## [652] "177374" "177448" "177526" "177870" "177916" "178368" "178672"
## [659] "179677" "179782" "180180" "180797" "180853" "181044" "181235"
## [666] "182265" "182609" "182666" "183517" "183591" "183684" "184157"
## [673] "185409" "185603" "186082" "186192" "186363" "186546" "186713"
## [680] "186748" "187874" "188151" "188163" "188195" "188514" "188904"
## [687] "189301" "189474" "189845" "190000" "190068" "190073" "190658"
## [694] "190709" "190844" "191673" "191838" "192112" "192276" "192478"
## [701] "192654" "193397" "193682" "193690" "194697" "195428" "195550"
## [708] "195748" "196375" "196873" "197836" "198562" "198780" "198895"
## [715] "199161" "199326" "199973" "200506" "200522" "201276" "201411"
## [722] "201548" "201646" "201970" "202639" "203359" "204377" "205293"
## [729] "205522" "205842" "205885" "206032" "206062" "206324" "206917"
## [736] "207260" "207320" "207328" "207378" "207802" "207851" "208109"
## [743] "208269" "208378" "208517" "209135" "209271" "209395" "209436"
## [750] "209639" "209705" "210246" "210669" "210976" "211159" "211184"
## [757] "211461" "211659" "211709" "211761" "211788" "211829" "211844"
## [764] "211869" "211981" "212174" "212803" "212885" "213114" "213363"
## [771] "214466" "214549" "214603" "214791" "215091" "215122" "215378"
## [778] "215943" "216301" "216346" "216844" "217437" "217502" "217888"
## [785] "218240" "218849" "218901" "219337" "219716" "219899" "219915"
## [792] "220165" "220482" "220771" "221147" "221841" "222117" "222326"
## [799] "222512" "222532" "223057" "223380" "223816" "223895" "223913"
## [806] "224258" "226368" "226489" "226956" "227217" "227913" "227943"
## [813] "229968" "230460" "231340" "231370" "231515" "231737" "232196"
## [820] "232252" "232483" "233616" "233632" "233723" "233970" "234369"
## [827] "234503" "234526" "234673" "234872" "235081" "235882" "236757"
## [834] "236865" "237072" "237279" "237409" "237435" "237814" "237965"
## [841] "238052" "238235" "239055" "239215" "239225" "239528" "239602"
## [848] "239736" "239876" "240252" "240675" "241255" "241276" "241388"
## [855] "241584" "242154" "242161" "242239" "242437" "243189" "243433"
## [862] "243661" "244399" "244494" "244739" "244877" "245087" "245281"
## [869] "246099" "246217" "246340" "246419" "246460" "246634" "246786"
## [876] "246945" "247100" "247552" "247708" "247838" "247994" "248421"
## [883] "249032" "249075" "249129" "249205" "250064" "250526" "250637"
## [890] "251118" "251311" "251334" "251357" "251616" "251851" "251999"
## [897] "252342" "253448" "253799" "253862" "254014" "254346" "254467"
## [904] "254627" "254712" "254740" "254986" "255175" "255492" "256140"
## [911] "256536" "256597" "256631" "256760" "256818" "257408" "258499"
## [918] "259005" "260317" "260670" "260784" "261100" "261448" "261686"
## [925] "261907" "262144" "262320" "262413" "262434" "262746" "262772"
## [932] "263319" "263342" "263388" "263485" "264214" "264624" "264743"
## [939] "264770" "264845" "265453" "265881" "266621" "266722" "266875"
## [946] "267414" "267415" "267517" "267728" "267868" "268622" "269130"
## [953] "269201" "269391" "269793" "270412" "270431" "270465" "270792"
##  [960] "270897" "270951" "271545" "271635" "272072" "272448" "272451"
##  [967] "272766" "272929" "273093" "274245" "274842" "274998" "275072"
##  [974] "275525" "275582" "276163" "276264" "276876" "277459" "277578"
##  [981] "277871" "278060" "278068" "279353" "279553" "279563" "280282"
##  [988] "280326" "280521" "280727" "280735" "281116" "281558" "281956"
##  [995] "282382" "283662" "283914" "283974" "284313" "284717"
##  [ reached getOption("max.print") -- omitted 3866 entries ]
##
## $dimnames[[2]]
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

```r
summary(res.comp$completeObs)
```

```
##  Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
##  Min.   :-74.03   Min.   :40.58   Min.   :-74.03    Min.   :40.57
##  1st Qu.:-73.96   1st Qu.:40.69   1st Qu.:-73.97    1st Qu.:40.70
##  Median :-73.95   Median :40.75   Median :-73.95    Median :40.75
##  Mean   :-73.94   Mean   :40.75   Mean   :-73.94    Mean   :40.74
##  3rd Qu.:-73.92   3rd Qu.:40.80   3rd Qu.:-73.91    3rd Qu.:40.79
##  Max.   :-73.79   Max.   :40.91   Max.   :-73.75    Max.   :40.91
##  Trip_distance     Fare_amount        Extra           MTA_tax
##  Min.   : 0.010   Min.   : 0.10   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 1.010   1st Qu.: 6.00   1st Qu.:0.0000   1st Qu.:0.5000
##  Median : 1.800   Median : 9.00   Median :0.5000   Median :0.5000
##  Mean   : 2.524   Mean   :11.15   Mean   :0.3497   Mean   :0.4915
##  3rd Qu.: 3.320   3rd Qu.:14.00   3rd Qu.:0.5000   3rd Qu.:0.5000
##  Max.   :10.610   Max.   :42.50   Max.   :2.0000   Max.   :0.5000
##    Tip_amount      Tolls_amount       Total_amount
##  Min.   : 0.000   Min.   : 0.00000   Min.   : 0.10
##  1st Qu.: 0.000   1st Qu.: 0.00000   1st Qu.: 7.80
##  Median : 0.000   Median : 0.00000   Median :11.00
##  Mean   : 1.124   Mean   : 0.07864   Mean   :13.49
##  3rd Qu.: 2.000   3rd Qu.: 0.00000   3rd Qu.:16.62
##  Max.   :22.000   Max.   :12.50000   Max.   :45.42
```

```r
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$completeObs[,"Trip_distance"]
df[,"Fare_amount"]<-res.comp$completeObs[,"Fare_amount"]
df[,"Extra"]<-res.comp$completeObs[,"Extra"]
df[,"MTA_tax"]<-res.comp$completeObs[,"MTA_tax"]
df[,"Tip_amount"]<-res.comp$completeObs[,"Tip_amount"]
df[,"Tolls_amount"]<-res.comp$completeObs[,"Tolls_amount"]
#df[,"improvement_surcharge"]<-res.comp$completeObs[,"improvement_surcharge"]
```

Imputation of qualitative variables

```
library(missMDA)
#res.impute = imputeMCA(df[,vars_dis],ncp = 3)
#res.impute
#df[,"VendorID"]<-res.impute$completeObs[,"VendorID"]
#df[,"RateCodeID"]<-res.impute$completeObs[,"RateCodeID"]
#df[,"Passenger_count"]<-res.impute$completeObs[,"Passenger_count"]
#df[,"Payment_type"]<-res.impute$completeObs[,"Payment_type"]
#df[,"Trp_type"]<-res.impute$completeObs[,"Trp_type"]
```

Creating auxiliar variables and doing their analysis

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[
}

summary(df$trip_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.792   3.123   4.301   5.522  29.880
```

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

Code_Doc_files/figure-latex/unnamed-chunk-33-1.pdf

Trip distance in km

```
df$trip_distance_km<-df$Trip_distance*1.609344 # Miles to km

summary(df$trip_distance_km)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.01609  1.62544  2.89682  4.06140  5.34302 17.07514
```

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

Code_Doc_files/figure-latex/unnamed-chunk-34-1.pdf

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'))

summary(df$travel_time)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##    0.000    5.871    9.733   18.916   15.750 1438.317
```

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

Code_Doc_files/figure-latex/unnamed-chunk-35-1.pdf

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   15.24   20.34     Inf   27.19     Inf
```

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

Code_Doc_files/figure-latex/unnamed-chunk-36-1.pdf

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   13.48   19.00   23.00
```

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

Code_Doc_files/figure-latex/unnamed-chunk-37-1.pdf

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
##       787       977      1367      1735
```

Creating factors

Factorize function:

```
factorize<- function(x) {
  quantile(x,seq(0,1,0.1))
  pp<-quantile(x);pp
  breaks<-c(unique(pp))
  f.x<-factor(cut(x,breaks))
  return(f.x);
}
```

f.passenger

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

```
## (0,1] (1,6]  NA's
##  4122   743     1
```

f.distance

```
df$f.distance<-factorize(df$Trip_distance) # NO VA be?
summary(df$f.distance)
```

```
## (0.01,1.01]  (1.01,1.8]  (1.8,3.32] (3.32,10.6]        NA's
##        1223        1221        1206        1215           1
```

f.pickup_longitude

```
df$f.pickup_longitude<-factorize(df$Pickup_longitude)
summary(df$f.pickup_longitude)
```

```
## (-74.03,-73.96] (-73.96,-73.95] (-73.95,-73.92] (-73.92,-73.79]
##            1216            1216            1217            1216
##            NA's
##               1
```

f.pickup_latitude

```
df$f.pickup_latitude<-factorize(df$Pickup_latitude)
summary(df$f.pickup_latitude)
```

```
## (40.58,40.69] (40.69,40.75]  (40.75,40.8]  (40.8,40.91]          NA's
##          1217          1215          1216          1217             1
```

f.dropoff_longitude

```
df$f.dropoff_longitude<-factorize(df$Dropoff_longitude)
summary(df$f.dropoff_longitude)
```

```
## (-74.03,-73.97] (-73.97,-73.95] (-73.95,-73.91] (-73.91,-73.75]
##            1216            1216            1216            1217
##            NA's
##               1
```

f.dropoff_latitude

```
df$f.dropoff_latitude<-factorize(df$Pickup_latitude)
summary(df$f.dropoff_latitude)
```

```
## (40.58,40.69] (40.69,40.75]  (40.75,40.8]  (40.8,40.91]          NA's
##          1217          1215          1216          1217             1
```

f.fare_amount

```
df$f.fare_amount<-factorize(df$Fare_amount)
summary(df$f.fare_amount)
```

```
##   (0.1,6]     (6,9]    (9,14] (14,42.5]      NA's
##      1250      1254      1203      1158         1
```

f.extra

```
df$f.extra<-factorize(df$Extra)
summary(df$f.extra)
```

```
## (0,0.5] (0.5,2]    NA's 
##    1879     761    2226 
```

f.MTA_tax

```
df$f.MTA_tax<-factorize(df$MTA_tax)
summary(df$f.MTA_tax) #11 NA's -> values of -0.5 => Outliers?
```

```
## (0,0.5]    NA's 
##    4783      83 
```

f.Improvement_surcharge

```
df$f.Improvement_surcharge<-factorize(df$improvement_surcharge)
summary(df$f.Improvement_surcharge) #11 NA's -> values of -0.3 => Outliers?
```

```
##    (0,0.3] (0.3,0.77]       NA's 
##       4783          1         82 
```

f.tip_amount

```
df$f.tip_amount<-factor(df$Tip_amount)
summary(df$f.tip_amount) #2869 NA's
```

```
##       0       1       2    1.46    1.56       3    1.36    1.66    1.96 
##    2839     152     148      45      43      42      38      38      35 
##     1.7    1.76    2.16    2.26    2.06     1.5    2.36    1.86     2.7 
##      31      31      30      29      28      26      26      25      25 
##     1.2    1.26    2.46    2.66    1.16    1.45    1.95       4    1.06 
##      22      22      21      21      20      20      20      20      19 
##     2.2    2.45       5    2.86    1.55    2.96    1.85    2.76    2.32 
##      19      19      19      18      17      17      16      16      15 
##    2.56    2.95    3.06    3.36    3.46    1.82     3.2    2.08    2.58 
##      15      15      15      15      15      14      14      13      13 
##    3.16    1.25    2.05    3.32     0.7    3.56    3.86    4.06    1.32 
##      13      12      12      12      11      11      11      11      10 
##    3.26    3.58     3.7    3.96     0.5    0.96    1.35    1.58    1.65 
##      10      10      10      10       9       9       9       9       9 
##    1.75    2.19    1.15    1.89    2.15     2.5    3.76    4.26    1.74 
##       9       9       8       8       8       8       8       8       7 
##    3.05    3.15    3.95    4.16    2.25    2.34    2.55    3.45    3.66 
##       7       7       7       7       6       6       6       6       6 
##    4.08     5.2    5.66    5.76    1.59    2.04    2.35    2.75     3.8 
##       6       6       6       6       5       5       5       5       5 
##    4.32    4.36    4.45    4.55    4.56    4.76    4.86    6.46    0.86 
##       5       5       5       5       5       5       5       5       4 
## (Other) 
##     353 
```

f.tolls_amount

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

```
## (-1,1] (1,50] 
##   4799     67
```

f.total_amount

```
df$f.total<-factorize(df$Total_amount)
summary(df$f.total)
```

```
##   (0.1,7.8]    (7.8,11]   (11,16.6] (16.6,45.4]        NA's 
##        1252        1187        1216        1210           1
```

Profiling

```
#save(df,miss,vars_con,vars_dis,vars_res,file="MyTaxi5000_Clean.RData")
summary(df)
```

```
##                                VendorID            lpep_pickup_datetime
##  Creative Mobile Technologies, LLC:1049   2016-01-01 03:00:54:   2     
##  VeriFone Inc.                    :3817   2016-01-22 09:48:21:   2     
##                                           2016-01-22 19:49:32:   2     
##                                           2016-01-22 20:58:19:   2     
##                                           2016-01-27 20:37:18:   2     
##                                           2016-01-28 18:05:51:   2     
##                                           (Other)            :4854     
##         Lpep_dropoff_datetime                   Store_and_fwd_flag
##  2016-01-01 02:50:32:   2      not a store and forward trip:4848  
##  2016-01-01 05:49:43:   2      store and forward trip      :  18  
##  2016-01-19 19:02:43:   2                                         
##  2016-01-19 21:51:05:   2                                         
##  2016-01-21 14:52:58:   2                                         
##  2016-01-29 16:08:23:   2                                         
##  (Other)            :4854                                         
##                 RateCodeID   Pickup_longitude Pickup_latitude
##  Standard rate         :4783   Min.   :-74.03   Min.   :40.58  
##  JFK                   :   0   1st Qu.:-73.96   1st Qu.:40.69  
##  Newark                :   1   Median :-73.95   Median :40.75  
##  Nassau or Westchester:   1   Mean   :-73.94   Mean   :40.75  
##  Negotiated fare       :  81   3rd Qu.:-73.92   3rd Qu.:40.80  
##                                Max.   :-73.79   Max.   :40.91  
##                                                                
##  Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
##  Min.   :-74.03    Min.   :40.57    Min.   :1.000   Min.   : 0.010
##  1st Qu.:-73.97    1st Qu.:40.70    1st Qu.:1.000   1st Qu.: 1.010
##  Median :-73.95    Median :40.75    Median :1.000   Median : 1.800
##  Mean   :-73.94    Mean   :40.74    Mean   :1.349   Mean   : 2.524
##  3rd Qu.:-73.91    3rd Qu.:40.79    3rd Qu.:1.000   3rd Qu.: 3.320
##  Max.   :-73.75    Max.   :40.91    Max.   :6.000   Max.   :10.610  
##                                     NA's   :1                        
##   Fare_amount        Extra          MTA_tax         Tip_amount    
##  Min.   : 0.10   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000  
##  1st Qu.: 6.00   1st Qu.:0.0000   1st Qu.:0.5000   1st Qu.: 0.000  
##  Median : 9.00   Median :0.5000   Median :0.5000   Median : 0.000  
##  Mean   :11.15   Mean   :0.3497   Mean   :0.4915   Mean   : 1.124  
##  3rd Qu.:14.00   3rd Qu.:0.5000   3rd Qu.:0.5000   3rd Qu.: 2.000  
##  Max.   :42.50   Max.   :2.0000   Max.   :0.5000   Max.   :22.000  
##                                                                    
##   Tolls_amount      improvement_surcharge  Total_amount  
##  Min.   : 0.00000   Min.   :0.000        Min.   : 0.10  
##  1st Qu.: 0.00000   1st Qu.:0.300        1st Qu.: 7.80  
##  Median : 0.00000   Median :0.300        Median :11.00  
##  Mean   : 0.07864   Mean   :0.295        Mean   :13.49  
##  3rd Qu.: 0.00000   3rd Qu.:0.300        3rd Qu.:16.62  
##  Max.   :12.50000   Max.   :0.770        Max.   :45.42  
##                                                         
##      Payment_type        Trip_type       mis_ind             AnyTip    
##  Credit card:2384   Street-hail:4786   Min.   :0.00000   AnyTip No :2839  
##  Cash       :2448   Dispatch   :  80   1st Qu.:0.00000   AnyTip Yes:2027  
##  No charge  :  16                      Median :0.00000                    
##  Dispute    :  18                      Mean   :0.03658                    
##                                        3rd Qu.:0.00000                    
##                                        Max.   :5.00000                    
##                                                                           
##   trip_length     trip_distance_km    travel_time          espeed     
##  Min.   : 0.000   Min.   : 0.01609   Min.   :   0.000   Min.   : 0.00  
##  1st Qu.: 1.792   1st Qu.: 1.62544   1st Qu.:   5.871   1st Qu.:15.24  
##  Median : 3.123   Median : 2.89682   Median :   9.733   Median :20.34  
##  Mean   : 4.301   Mean   : 4.06140   Mean   :  18.916   Mean   :  Inf  
##  3rd Qu.: 5.522   3rd Qu.: 5.34302   3rd Qu.:  15.750   3rd Qu.:27.19  
##  Max.   :29.880   Max.   :17.07514   Max.   :1438.317   Max.   :  Inf  
##                                                                        
##   pick_up_hour     pick_up_period f.passenger       f.distance  
##  Min.   : 0.00   night    : 787   (0,1]:4122   (0.01,1.01]:1223  
##  1st Qu.: 9.00   morning  : 977   (1,6]: 743   (1.01,1.8] :1221  
##  Median :15.00   valley   :1367   NA's :   1   (1.8,3.32] :1206  
##  Mean   :13.48   afternoon:1735                (3.32,10.6]:1215  
##  3rd Qu.:19.00                                 NA's       :   1  
##  Max.   :23.00                                                   
##                                                                  
##       f.pickup_longitude     f.pickup_latitude      f.dropoff_longitude
##  (-74.03,-73.96]:1216     (40.58,40.69]:1217    (-74.03,-73.97]:1216  
##  (-73.96,-73.95]:1216     (40.69,40.75]:1215    (-73.97,-73.95]:1216  
##  (-73.95,-73.92]:1217     (40.75,40.8] :1216    (-73.95,-73.91]:1216  
##  (-73.92,-73.79]:1216     (40.8,40.91] :1217    (-73.91,-73.75]:1217  
##  NA's           :   1     NA's         :   1    NA's           :   1  
##                                                                       
##                                                                       
##      f.dropoff_latitude   f.fare_amount     f.extra       f.MTA_tax   
##  (40.58,40.69]:1217     (0.1,6]  :1250   (0,0.5]:1879   (0,0.5]:4783  
##  (40.69,40.75]:1215     (6,9]    :1254   (0.5,2]: 761   NA's   :  83  
##  (40.75,40.8] :1216     (9,14]   :1203   NA's   :2226                 
##  (40.8,40.91] :1217     (14,42.5]:1158
##  NA's         :    1     NA's      :    1
##
##
##  f.Improvement_surcharge  f.tip_amount     f.toll           f.total
##  (0,0.3]   :4783         0      :2839   (-1,1]:4799   (0.1,7.8]  :1252
##  (0.3,0.77]:    1         1      : 152   (1,50]:  67   (7.8,11]   :1187
##  NA's      :   82         2      : 148                 (11,16.6]  :1216
##                          1.46   :  45                 (16.6,45.4]:1210
##                          1.56   :  43                 NA's       :   1
##                          3      :  42
##                          (Other):1597
```

```
# Numeric Target Total_Amount
vars_con;vars_dis
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

```
## [1] "VendorID"        "RateCodeID"      "Passenger_count" "Trip_type"
## [5] "mis_ind"
```

```
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"             "espeed"
## [27] "pick_up_hour"            "pick_up_period"
## [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"
```

```
# condes(df[,c(vars_con,vars_dis)],1)
library(FactoMineR)
#condes(df,15)

# Binary Target AnyTip
vars_con;vars_dis
```

```
##  [1] "Pickup_longitude"  "Pickup_latitude"   "Dropoff_longitude"
##  [4] "Dropoff_latitude"  "Trip_distance"     "Fare_amount"
##  [7] "Extra"             "MTA_tax"           "Tip_amount"
## [10] "Tolls_amount"      "Total_amount"
```

```
## [1] "VendorID"        "RateCodeID"      "Passenger_count" "Trip_type"
## [5] "mis_ind"
```

```
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"             "espeed"
## [27] "pick_up_hour"            "pick_up_period"
## [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"
```

```
#catdes(df[,c(vars_dis,vars_con)],5)
#catdes(df,which(names(df)=="AnyTip"))
```