

# Deliverable I

Carles Capilla Cànovas  
Jesús Molina Roldán

<b>Descripció de les dades</b>	<b>4</b>
Paquets	5
Mostra	5
Anàlisis descriptiva univariant	6
VendorID	7
Trip_type	7
Lpep_pickup_time i Lpep_dropoff_time	7
RateCodeID	8
Passenger_count	8
Trip_distance	9
Pickup_latitude i Dropoff_latitude	9
Pickup_longitude i Dropoff_longitude	11
Store_and_fwd_flag	12
Payment_type	12
Fare_amount	13
Extra	14
MTA_tax	14
Improvement_surcharge	15
Tip_amount	15
Tolls_amount	16
Total_amount	17
Ehail_fee	18
Noves variables	18
tlenkm	18
traveltime	19
Effective speed(espeed)	19
lpep_pickup_period i lpep_dropoff_period	20
lpep_pickup_date	21
<b>Imputació</b>	<b>21</b>
Imputació variables categòriques	21
Imputació de variables numèriques	22
Multivariant outliers	23
<b>Data quality report</b>	<b>25</b>
Missings	25
Errors	25
Outliers	26
Discretització	27
espeed	27
tlenkm	27
traveltime	28
distHaversine	28
Fare_amount	28

<b>Univariant exploratory analysis (EDA)</b>	<b>29</b>
lpep_pickup_date	29
VendorID	30
lpep_pickup_time, lpep_pickup_period, travelttime, f.travelttime	30
tlenkm, distHaversine	31
espeed	31
lpep_pickup_latitude, lpep_pickup_longitude, lpep_dropoff_latitude,	
lpep_dropoff_longitude	32
Passenger_count	32
Payment_type	32
Extra, f.Extra, MTA_tax, Improvement_surcharge	33
Fare_amount, Tip_amount, Tolls_amount, Total_amount, AnyTip i AnyToll	33
<b>Profiling</b>	<b>35</b>
Profiling Total_amount	35
Associació global variables quantitatives	35
Associació global variables qualitatives	35
Profiling de les categories	35
Profiling AnyTip	36
Associació global variables quantitatives	36
Associació global variables categòriques	36
Profiling de les categories	36

# Descripció de les dades

1.VendorID	A code indicating the LPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc
2.Ipep_pickup_datetime	The date and time when the meter was engaged.
3.Ipep_dropoff_datetime	The date and time when the meter was disengaged.
4.Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
5.Trip_distance	The elapsed trip distance in miles reported by the taximeter.
6.Pickup_longitude	Longitude where the meter was engaged.
7.Pickup_latitude	Latitude where the meter was engaged.
8.RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
9.Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server: Y= store and forward trip N= not a store and forward trip
10.Dropoff_longitude	Longitude where the meter was timed off.
11.Dropoff_latitude	Latitude where the meter was timed off.
12.Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
13.Fare_amount	The time-and-distance fare calculated by the meter.
14.Extra	Miscellaneous extras and surcharges. Currently, this only includes the \$0.50 and \$1 rush hour and overnight charges.
15.MTA_tax	\$0.50 MTA tax that is automatically triggered based on the metered rate in use. Improvement_surcharge

16.Tip_amount	This field is automatically populated for credit card tips. Cash tips are not included.
17.Tolls_amount	Total amount of all tolls paid in trip.
18.Total_amount	The total amount charged to passengers. Does not include cash tips.
19.Trip_type	A code indicating whether the trip was a street-hail or a dispatch that is automatically assigned based on the metered rate in use but can be altered by the driver. 1= Street-hail 2= Dispatch
20. Ehail_fee	This fee is applied when a taxi is ordered via a virtual device.
21. improvement_surcharge	\$0.30 improvement surcharge assessed on hailed trips at the flag drop. The improvement surcharge began being levied in 2015.

## Paquets

Per tal de realitzar el projecte, s'utilitzarà el conjunt de llibreries que es veu en la imatge. Dins d'aquests paquets trobarem diferents funcions que ens permetrà facilitar-nos el treball

```
# Load Required Packages: to be increased over the course
options(contrasts=c("contr.treatment","contr.treatment"))

requiredPackages <- c("missMDA", "chemometrics", "mvoutlier", "effects", "FactoMineR", "car",
"factoextra", "RColorBrewer", "ggplot2", "dplyr", "ggmap", "ggthemes", "knitr")
missingPackages <- requiredPackages[!(requiredPackages %in%
installed.packages()[,"Package"])] 

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

## Mostra

De la mostra total que es carregarà, només s'utilitzarà un subconjunt amb 5000 elements. Com es pot veure en la imatge, la mostra serà agafada aleatoriament amb la llavor 02041997.

```
set.seed(02041997)
sam<-as.vector(sort(sample(1:nrow(df2),5000)))
df<-df2[sam,]
summary(df)
```

A continuació podem veure un resum de les dades que tractarem.

```
> summary(df_total)
  vendorID    lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag
Min.   :1.000  Length:5000          Length:5000          Length:5000
1st Qu.:2.000  Class :character   Class :character   Class :character
Median :2.000  Mode  :character   Mode  :character   Mode  :character
Mean   :1.779
3rd Qu.:2.000
Max.   :2.000

  RateCodeID   Pickup_longitude Pickup_latitude Dropoff_longitude
Min.   :1.000  Min.  :-74.16   Min.   : 0.00  Min.  :-74.18
1st Qu.:1.000  1st Qu.:-73.96   1st Qu.:40.69  1st Qu.:-73.97
Median :1.000  Median :-73.95   Median :40.75  Median :-73.95
Mean   :1.104  Mean   :-73.83   Mean   :40.69  Mean   :-73.88
3rd Qu.:1.000  3rd Qu.:-73.92   3rd Qu.:40.80  3rd Qu.:-73.91
Max.   :5.000  Max.   : 0.00   Max.   :40.89  Max.   : 0.00

  Dropoff_latitude Passenger_count Trip_distance      Fare_amount
Min.   : 0.00  Min.   :0.0000  Min.   : 0.000  Min.   :-50.00
1st Qu.:40.70  1st Qu.:1.000  1st Qu.: 1.020  1st Qu.: 6.50
Median :40.75  Median :1.000  Median : 1.850  Median : 9.00
Mean   :40.71  Mean   :1.375  Mean   : 2.807  Mean   :12.09
3rd Qu.:40.79  3rd Qu.:1.000  3rd Qu.: 3.583  3rd Qu.:14.50
Max.   :40.94  Max.   :6.000  Max.   :42.200  Max.   :400.00

  Extra        MTA_tax       Tip_amount      Tolls_amount
Min.   :-1.0000  Min.  :-0.5000  Min.   : 0.000  Min.   : 0.00000
1st Qu.: 0.0000  1st Qu.: 0.5000  1st Qu.: 0.000  1st Qu.: 0.00000
Median : 0.5000  Median : 0.5000  Median : 0.000  Median : 0.00000
Mean   : 0.3481  Mean   : 0.4858  Mean   : 1.319  Mean   : 0.09719
3rd Qu.: 0.5000  3rd Qu.: 0.5000  3rd Qu.: 2.000  3rd Qu.: 0.00000
Max.   : 1.0000  Max.   : 0.5000  Max.   :98.880  Max.   :11.08000

  Ehail_fee      improvement_surcharge Total_amount      Payment_type
Mode:logical   Min.   :-0.3000   Min.  :-50.000  Min.   :1.000
NA's:5000      1st Qu.: 0.3000   1st Qu.: 7.872  1st Qu.:1.000
               Median : 0.3000   Median :11.300  Median :2.000
               Mean   : 0.2912   Mean   :14.633  Mean   :1.515
               3rd Qu.: 0.3000   3rd Qu.:17.300  3rd Qu.:2.000
               Max.   : 0.3000   Max.   :498.880  Max.   :4.000

  Trip_type
Min.   :1.000
1st Qu.:1.000
Median :1.000
Mean   :1.025
3rd Qu.:1.000
Max.   :2.000
```

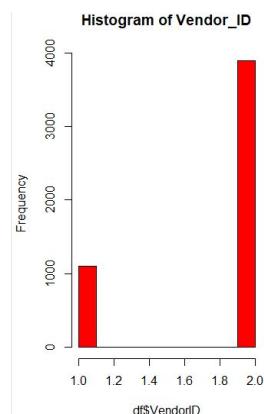
En aquesta primera vista de la mostra podem observar com, per exemple, la variable ehail\_fee conté 5000 valors faltants identificats amb NA.

També podem veure que hi ha valors irregulars pel que fa a les longituds i latituds, ja que pel que fa a les longituds, per exemple, veiem un valor màxim de 0 que, tenint en compte que es tracta d'unes mostres que afecten a la zona de Nova York, dista massa de la resta de valors.

## Anàlisis descriptiva univariant

Com podem veure per la tipologia de les variables, així com per les observacions d'aquestes, hi ha variables que comprenen o bé una sèrie de categories, o bé uns valors concrets. És per això que procedirem a tractar-les com variables categòriques transformant-les a variables factor.

## VendorID



Començant per la variable VendorID on hi ha els valors 1 i 2 depenent de si es tracta de Mobile o Verizon respectivament, el que fem és assignar cada valor a un nivell( valors possibles de la variable) diferent. A més, modificarem el nom de les etiquetes per indicar que es tracta d'un factor de tipus VendorID i poder facilitar la feina posteriorment.

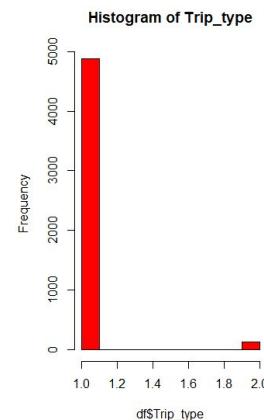
```
df$VendorID<-factor(df$VendorID,labels=c("Mobile","veriFone"))
levels(df$VendorID)<-paste0("f.vendor-",levels(df$VendorID))
```

f.vendor-Mobile f.vendor-veriFone

1103

3897

## Trip\_type



Igual que el VendorID, el Trip\_type consta de dos valors possibles 1 i 2 depenent de si es tracta d'un viatge del tipus Street-Hail o bé Dispatch. Ens disposem a convertir la variable a categòrica amb els valors comentats i modificar els seus noms per identificar-los posteriorment i indicar que es tracta d'un factor.

A sota, al summary, podem observar com al final disposem dels valors possibles que comprenia la variable i no 1 i 2 com teníem abans.

```
df$Trip_type<-factor(df$Trip_type,labels=c("Street-Hail","Dispatch"))
levels(df$Trip_type)<-paste0("f.TripType-",levels(df$Trip_type))

> summary(df$Trip_type)
f.TripType-Street-Hail      f.TripType-Dispatch
        4877                  123
```

## Lpep\_pickup\_time i Lpep\_dropoff\_time

Pel que fa a les variables que ens donen la informació sobre l'hora i data en què es recull i deixa als clients, hem considerat separar-les en dues variables cada una, data i hora, ja que ens facilitaran l'estudi i depuració d'aquestes i creiem que ens poden aportar més informació per separat.

```
df_datatime <- t(as.data.frame(strsplit(as.character(df$lpep_pickup_datetime), " ")))
df$lpep_pickup_date <- factor(df_datatime[,1])
df$lpep_pickup_time <- factor(df_datatime[,2])

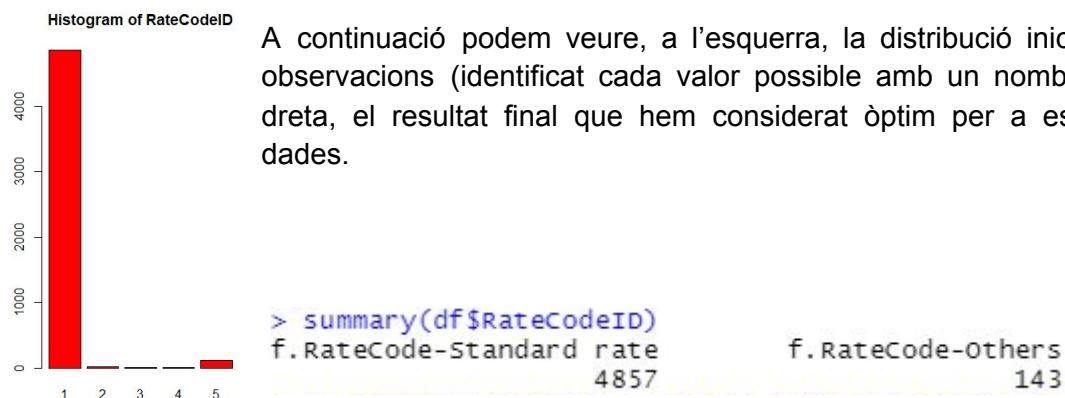
colnames(df)[which(names(df) == "lpep_dropoff_datetime")] <- "lpep_dropoff_datetime"
df_datatime <- t(as.data.frame(strsplit(as.character(df$lpep_dropoff_datetime), " ")))
df$lpep_dropoff_date <- factor(df_datatime[,1])
df$lpep_dropoff_time <- factor(df_datatime[,2])
```

## RateCodeID

Com aquesta variable, tot i estar codificada com numèrica, només disposa de 5 valors possibles, l'hem decidit factoritzar. Un cop factoritzada veiem que el RateCode "Standard rate" comprèn gairabé la totalitat d'observacions i, entre les altres quatre categories no arriben a juntar-ne 150 de les 5000 observacions. És per això que hem decidit agrupar les opcions de JFK, Newark, NassauOrWestChester i NegotiatedFare en una anomenada Others. A més, identifiquem els nivells amb la f de factor i el nom de la variable com hem fet amb les altres.

```
df$RateCodeID<-factor(df$RateCodeID)
barplot(table(df$RateCodeID))

# It is a categorical(factor) variable NO PROBLEM but not any interest
df$RateCodeID <- df$RateCodeID != 1
df$RateCodeID <- factor(df$RateCodeID, labels=c("standard rate","others"))
levels(df$RateCodeID)<-paste0("f.RateCode-",levels(df$RateCodeID))
```



## Passenger\_count

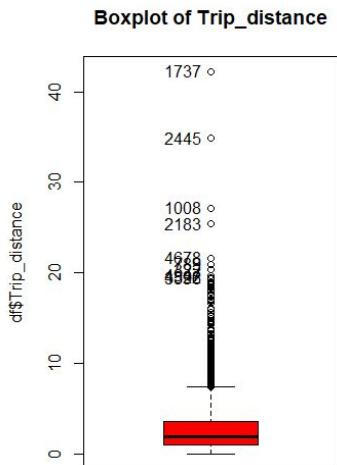


Pel que fa a la variable Passenger\_count mitjançant el hist i el boxplot observem com gairebé totes les observacions prenen el valor d'un passatger, tot i això no considerarem els altres valors com a outliers ni errors, ja que pot haver-se reservat un taxi, per exemple, i registrar-se el servei però no agafar cap passatger finalment.

```
#variable Passenger_count
hist(df$Passenger_count, main="Histogram of Passenger_count")
boxplot(df$Passenger_count, main="Boxplot of Passenger_count")
summary(df$Passenger_count)

# errors
l <- which(df$Passenger_count<0)
if (length(l)>0) {
  ierrs[]<-ierrs[1]+1
  jerrs["Passenger_count"]<-length(l)
}
df[1,"Passenger_count"]<-NA
```

## Trip\_distance

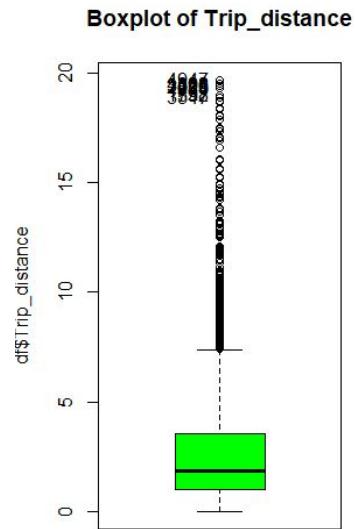


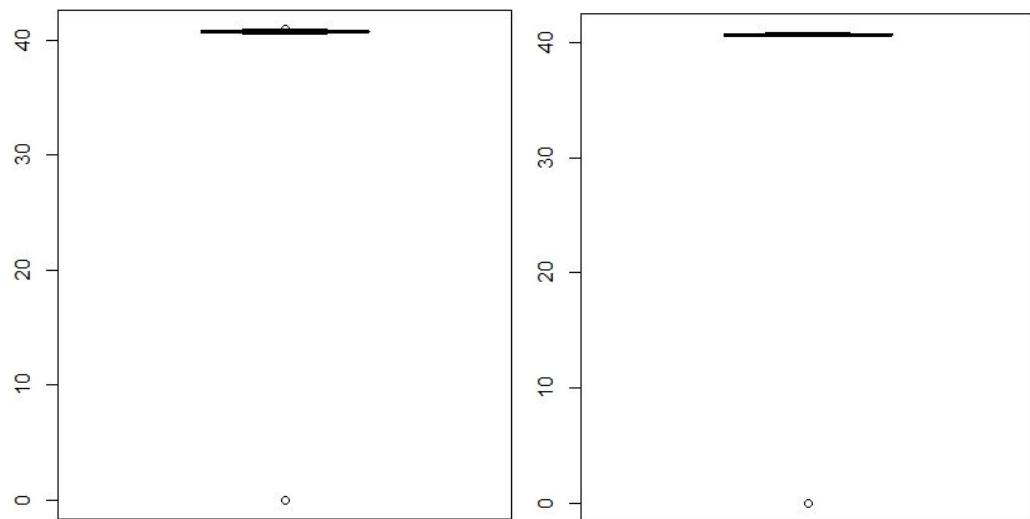
Al Trip\_distance tenim observacions amb distàncies de 0 que no tenen gaire sentit en relació a un viatge de taxi. Considerarem tot i això, vàlid tot valor immediatament superior a zero.

A més, en el boxplot inicial observem com a partir de les 20 milles, les observacions comencen a ésser disperses així que a partir de llavors les considerarem outliers i les deixarem com a NA.

```
#variable Trip_distance
summary(df$Trip_distance)
# errors
l <- which(df$Trip_distance<0.001); length(l)
if (length(l)>0) {
  ierrs[1]<-ierrs[1]+1
  jerrs["Trip_distance"]<-length(l)
}
df[1,"Trip_distance"]<-NA

#outliers
hist(df$Trip_distance, main="Histogram of Trip_distance")
boxplot(df$Trip_distance, main="Boxplot of Trip_distance")
var_out<-calcQ(df$Trip_distance)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")
l<-which(df$Trip_distance>20)
iouts[1]<-iouts[1]+1
jouts["Trip_distance"]<-length(l)
```



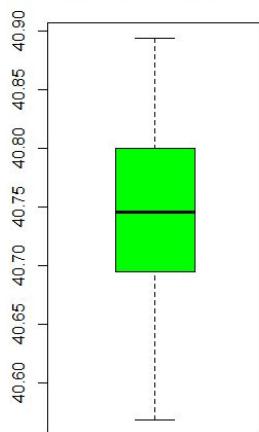


Vist l'error ens disposem a eliminar les observacions que tinguin per valor 0 i posar-les com a NA

```
summary(df$Pickup_latitude)
#0.00 looks to be an error
# Seeing the individuals with this "0" value:
df[which(df[,"Pickup_latitude"]==0),]

# It is a quantitative variable Non-possible values will be recoded to NA
sel<-which(df$Pickup_latitude ==0)
ierrs[sel]<-ierrs[sel]+1
jerrs["Pickup_latitude"]<-length(sel)
sel           ##### sel contains the rownames of the individuals with "0"
#                   as value for longitude
df[sel,"Pickup_latitude"]<-NA      # non-possible values are replaced by NA, missing value symbol in R
```

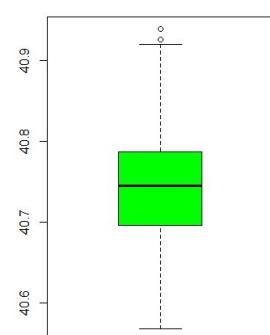
**Boxplot of Pickup\_latitude**



Aquí veiem com ha millorat la vista del boxplot un cop eliminats els valors de 0 a la vegada que podem observar que no tindríem outliers extra pel que fa a la variable de pickup\_latitude.

```
> summary(df$Pickup_latitude)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
40.57    40.70   40.75    40.75   40.80   40.89       7
```

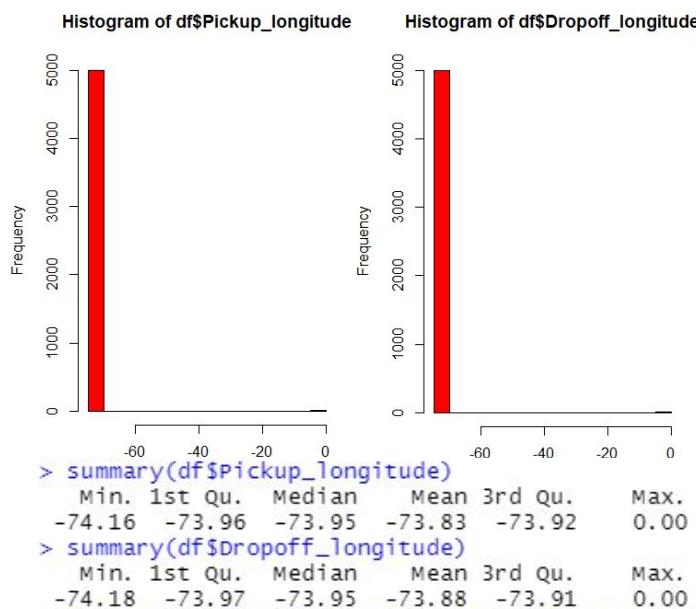
**Boxplot of Dropoff\_latitude**



Ara procediríem a fer exactament el mateix amb la variable Dropoff\_latitude, on acabaríem obtenint el següent boxplot un cop eliminades les observacions amb valor 0 i posades com a NA. Veiem que hi ha alguns valors que podríem considerar outliers però serien tan febles que els deixarem com a observacions vàlides.

```
> summary(df$Dropoff_latitude)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
40.57    40.70   40.75    40.74   40.79   40.94       4
```

## Pickup\_longitude i Dropoff\_longitude

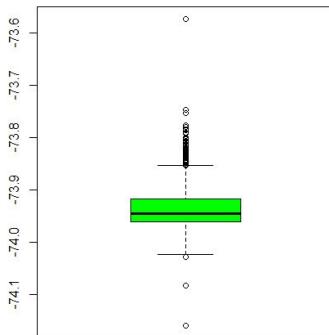


Es presentaria la mateixa situació que pel que feia a les latituds amb els valors de 0 tot i que les longituds rondarien el valor de -74. Aquest cop mostrem visualment les variables amb un hist en comptes d'un boxplot, ja que els boxplots serien semblants als dos anteriors i així mostrem d'un altre manera el nombre d'observacions d'aquestes dades.

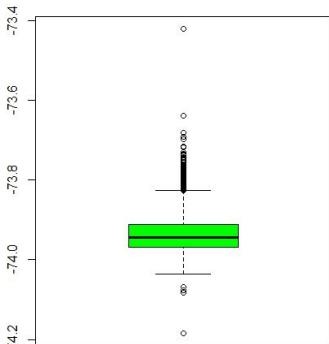
Procedim a eliminar les observacions amb valor 0 i marcar-les com NA de la variable Pickup\_longitude.

```
### variable Pickup_longitude
summary(df$Pickup_longitude)
#0.00 looks to be an error
# Seeing the individuals with this "0" value:
df[which(df[,"Pickup_longitude"]==0),]

# It is a quantitative variable Non-possible values will be recoded to NA
sel<-which(df$pickup_longitude ==0)
ierrs[sel]<-ierrs[sel]+1
jerrs["Pickup_longitude"]<-length(sel)
sel           ##### sel contains the rownames of the individuals with "0"
#                               as value for longitude
df[sel,"Pickup_longitude"]<-NA    # non-possible values are replaced by NA, missing value symbol in R
```



Aquesta seria la distribució de les observacions mostrada pel boxplot a falta d'eliminar els outliers visibles.



Un cop repetit el procés per a la variable Dropoff\_longitude aquest seria el boxplot resultant, on tot i això, podem seguir detectant outliers que tractarem més endavant.

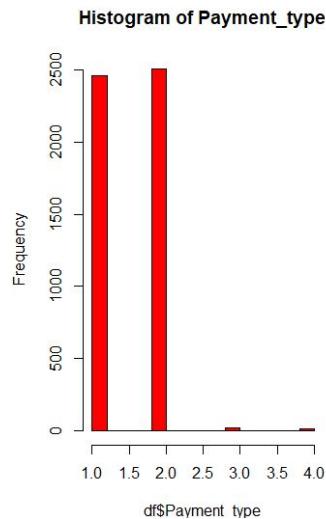
## Store\_and\_fwd\_flag

Pel que fa a aquesta variable, per tal de factoritzar-la hem decidit passar-la a booleà ja que només consta de dos valors possibles que equivalen a si o no i està codificada com a variable amb caràcters per representar aquests valors.

```
df$store_and_fwd_flag <- df$store_and_fwd_flag == "Y"
```

```
> summary(df$store_and_fwd_flag)
  Mode   FALSE    TRUE
logical  4983     17
```

## Payment\_type

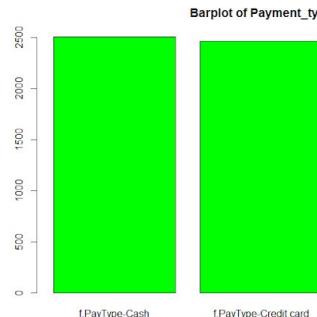


Com podem observar la variable Payment\_type és una variable categòrica i com diu la seva definició, té 6 valors possibles. Degut a que en la primera vista que tenim de les observacions de la variable en aquesta llavor només tenim 4 d'aquests 6 possibles valors, procedirem a factoritzar-la ambaquests 4 valors

```
df$Payment_type<-factor(df$Payment_type,labels=c("credit card","cash","No charge","Dispute"))
levels(df$Payment_type)<-paste0("f.PayType-",levels(df$Payment_type))
summary(df$Payment_type)
```

f.PayType-Credit card	2464	f.PayType-Cash	2507	f.PayType-No charge	19	f.PayType-Dispute	10
-----------------------	------	----------------	------	---------------------	----	-------------------	----

Després de veure l'ínfima quantitat d'observacions pel que faria als valors de No Charge i Dispute, hem decidit agrupar-les en una nova categoria anomenada Others. Al summary i barplot següents podem veure com queden distribuïdes les observacions de la variable.



## Fare\_amount

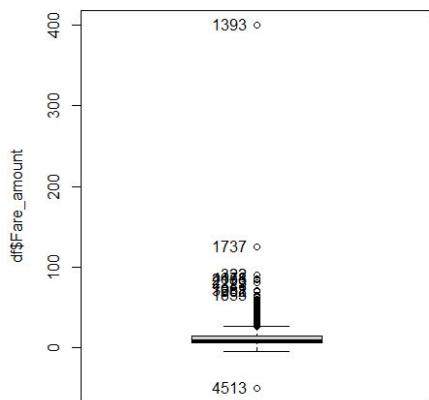
A l'hora de fer un summary de la variable Fare\_amount com a primer error que salta a la vista

tindríem que apareix alguna observació amb valors negatius. El que farem abans de tot és eliminar-les i posar-les com a NA.

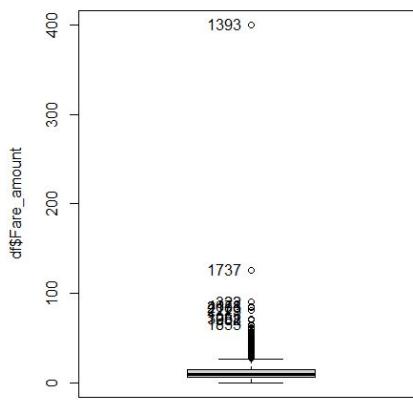
```
> summary(df$Fare_amount)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-50.00	6.50	9.00	12.09	14.50	400.00

Boxplot of Fare\_amount



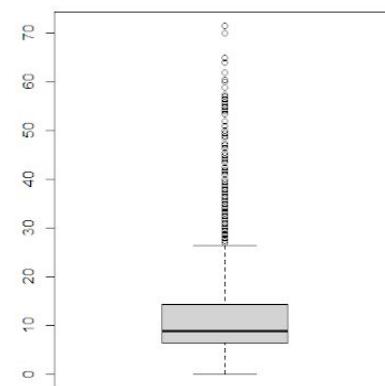
Boxplot of Fare\_amount



Un cop eliminades les observacions negatives procedirem a detectar i eliminar els valors que excedeixin de 80, els quals considerarem outliers.

```
# outlier detection
Boxplot(df$Fare_amount)
var_out<-calcQ(df$Fare_amount)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")
var_out$souts
l <-which(df$Fare_amount>100)
iouts[l ]<-iouts[l ]+1
jouts["Fare_amount"]<-length(l)
df[1,"Fare_amount"]<-NA
```

Boxplot of Fare\_amount

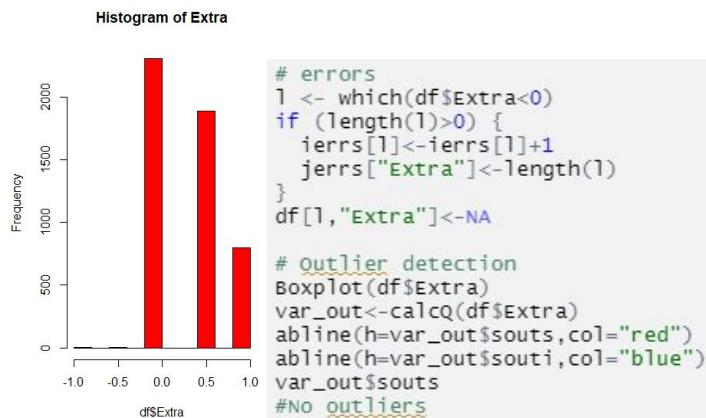


```
> summary(df$Fare_amount)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
	0.00	6.50	9.00	12.03	14.50	90.00	11

## Extra

Com podem veure al gràfic següent tenim imports Extra negatius que considerarem errors així que passarem a eliminar-los i deixar-los com a NA.

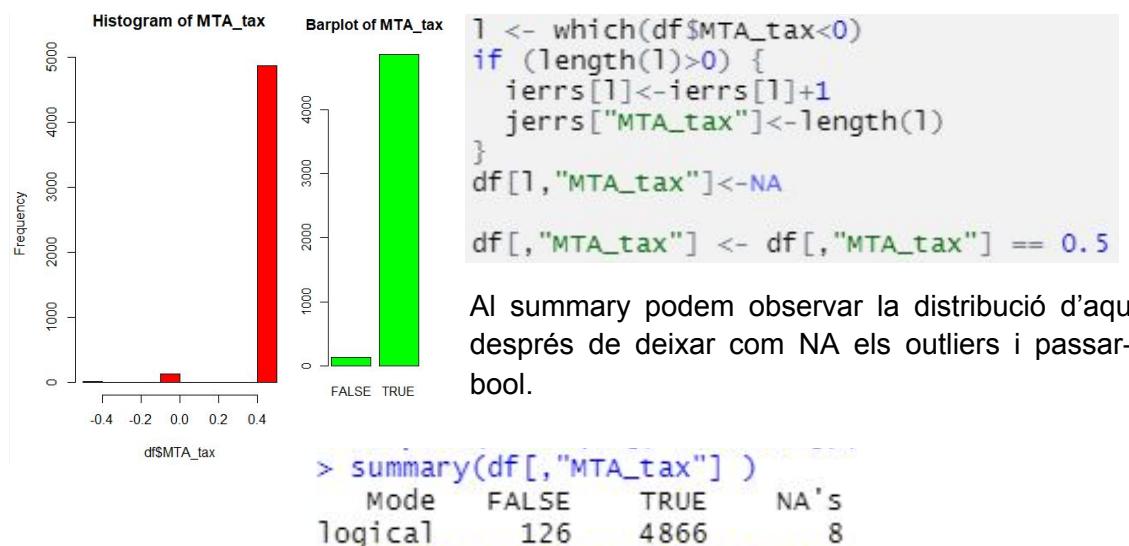


A part de les observacions negatives no trobem d'altres que podem considerar outliers pel que faria a aquesta variable.

```
> summary(df$Extra)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
0.000  0.000  0.500  0.349  0.500  1.000      5
```

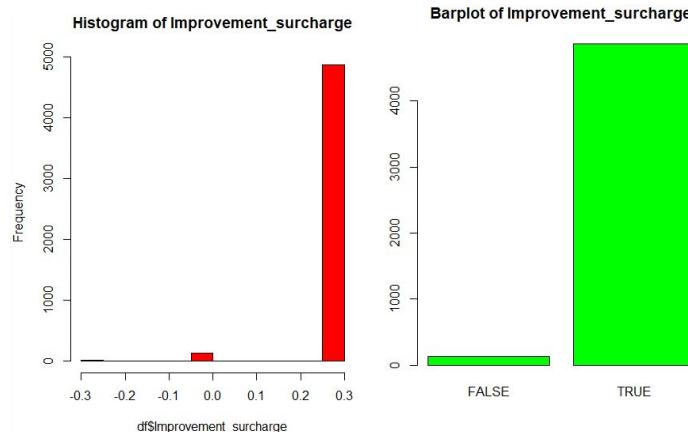
## MTA\_tax

Pel que fa a les observacions de la variable MTA\_tax, tenim, com podem veure al següent gràfic, valors negatius els quals considerarem errors i els eliminarem per deixar-los com a NA. A més, com es tracta d'una taxa d'import fixa de 0.5\$ l'hem convertida a booleà per determinar si ha estat o no cobrada.



## Improvement\_surcharge

Amb el Improvement\_surcharge tindriem valors negatius com ens passava amb el MTA\_tax i, al ser un import, no hauria d'ésser negatiu. És per això que procedim a eliminar les observacions negatives i deixar-les com a NA. A més com es tracta d'un import fixe de 0.3\$ que es cobra dependent del tipus de viatge o no, el convertirem a booleà per diferenciar de si es cobra o no.

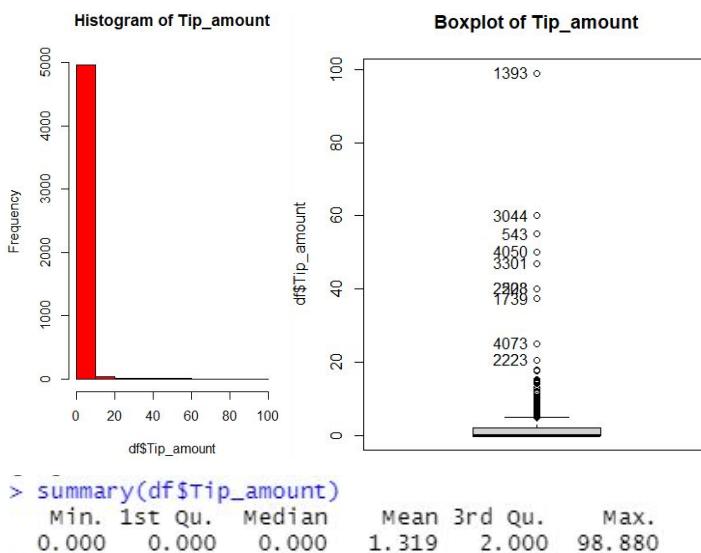


Al Barplot de color verd i al summary podem veure la distribució final que pren la variable.

```
> summary(df[, "Improvement_surcharge"] )  
  Mode   FALSE    TRUE    NA's  
logical  130     4862      8
```

## Tip\_amount

A les observacions de la variable Tip\_amount, com podem veure al hist de color vermell, no apareixen propines amb valor negatiu. Tot i això, al boxplot podem observar com hi ha valors molt dispersos que poden ser considerats com a outliers molt llunyans dels quartils.



```
> summary(df$Tip_amount)  
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
 0.000  0.000  0.000  1.319  2.000 98.880
```

És per això que considerem evaluar com a outliers totes les observacions que sobrepassin els 20\$ i marcar-les com a NA a la vegada que comprovem si hi hagués algun valor negatiu per fer el mateix. No considerarem avaluar com a outliers totes les observacions que sobrepassin les línies que indiquen a partir d'on comencen els outliers severs, ja que ens carregaríem una gran quantitat informativa d'aquestes.

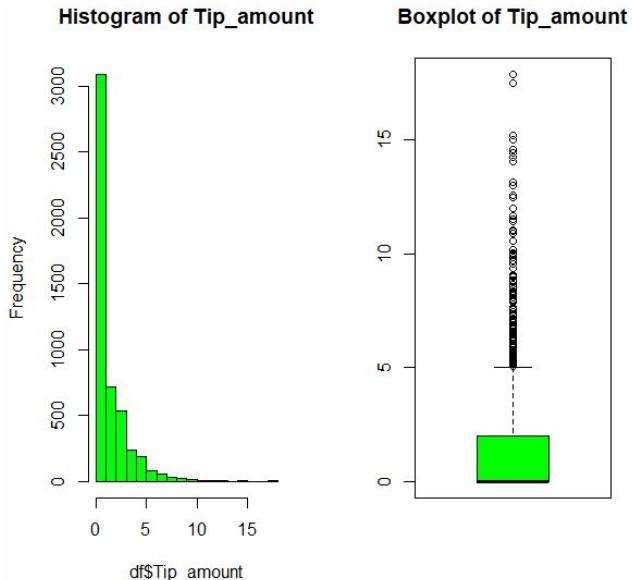
```

# errors
l <- which(df$Tip_amount<0)
if (length(l)>0) {
  ierrs[]<-ierrs[]+1
  jerrs["Tip_amount"]<-length(l)
}
df[l,"Tip_amount"]<-NA

# Outlier detection
boxplot(df$Tip_amount)
var_out<-calcQ(df$Tip_amount)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")
var_out$souts
l <-which(df$Tip_amount>20)
iouts[1 ]<-iouts[1]+1
jouts["Tip_amount"]<-length(l)
df[l,"Tip_amount"]<-NA

> summary(df$Tip_amount)
   Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
0.0000 0.0000 0.0000 1.227 2.000 17.880 10

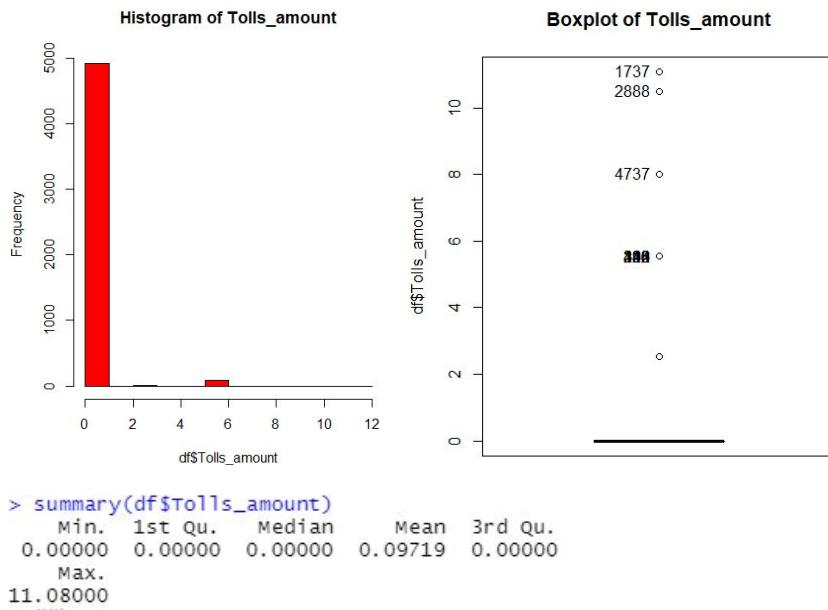
```



Finalment, a l'histograma i al boxplot podem veure com queden distribuïdes d'una manera menys dispersa les observacions un cop reduït el rang de valors a considerar. Podem observar al summary que de tots els valors que teníem, n'hem considerat 10 com a NA després d'aquesta avaluació d'errors i outliers.

### Tolls\_amount

Com podem observar, la variable referent al nombre de peatges creuats no comprèn valors negatius tot i que, per estar segurs, veiem que el valor mínim al summary és de 0. Per tant només haurem de centrar-nos en outliers a considerar segons la distribució de les observacions.



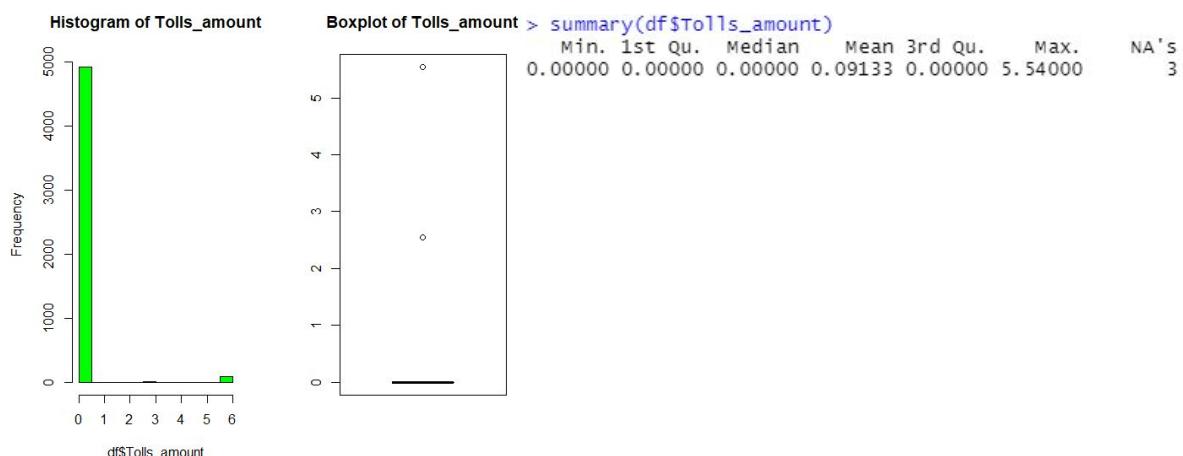
```

# errors
l <- which(df$Tolls_amount<0)
if (length(l)>0) {
  ierrs[1]<-ierrs[1]+1
  jerrs["Tolls_amount"]<-length(l)
}
df[l,"Tolls_amount"]<-NA

# outlier detection
boxplot(df$Tolls_amount)
var_out<-calcQ(df$Tolls_amount)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")
var_out$souts
l <-which(df$Tolls_amount>7)
iouts[1 ]<-iouts[1 ]+1
jouts["Tolls_amount"]<-length(l)
df[l,"Tolls_amount"]<-NA

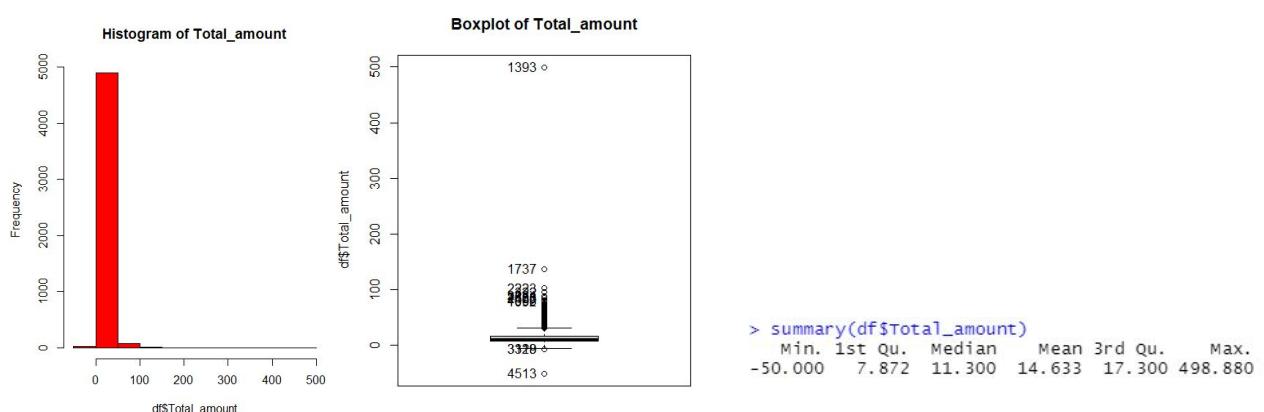
```

Vist el boxplot anterior i a fi de no considerar outlier tot i no disposar d'una variable amb totes les observacions amb valor 0, hem decidit delimitar com a outliers totes les observacions que superin el valor de 7 peatges. Un cop descrit els outliers ens quedarà una distribució com la que podem observar als gràfics i summary següents.



## Total\_amount

Per començar amb la variable referent a l'import total pagat pel servei de taxi eliminarem els valors negatius, ja que no té sentit cobrar per agafar un taxi. En canvi, deixarem els imports de 0\$ per si hi haguessin cops en que al final no s'ha donat el servei tot i haver-se demanat o reservat o bé es tracta d'un familiar al qual no se li pot haver cobrat.



A més, veiem com a partir dels 100\$ ja no hi ha observacions homogènies així que considerarem tota observació amb valor superior com a outlier.

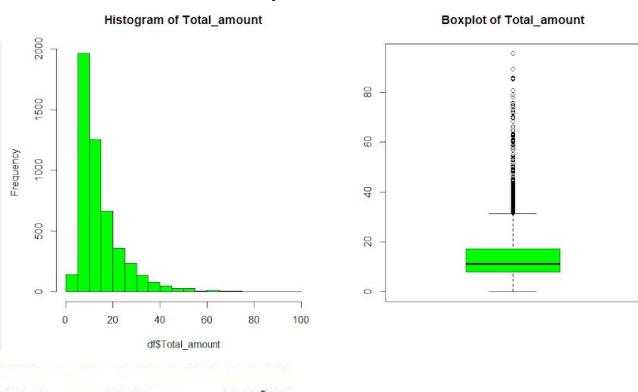
```

1<-which(df$Total_amount<0)
if (length(l)>0) {
  ierrs[1]<-ierrs[1]+1
  jerrs["Total_amount"]<-length(l)
}
df[1,"Total_amount"]<-NA

# outlier detection
boxplot(df$total_amount)
var_out<-calcq(df$total_amount)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")
l<-which((df$total_amount<0) | (df$total_amount>100))
iouts[1]<-iouts[1]+1
jouts["Total_amount"]<-length(l)
df[1,"Total_amount"]<-NA

> summary(df$Total_amount)
   Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
0.00    7.88   11.30 14.54 17.30 95.54 12

```



A sobre podem veure com s'han eliminat 12 valors, pel que fa al summary, i deixat com a NA i la distribució final que descriu la variable d'import total pagat.

## Ehail\_fee

Pel que fa a la variable Ehail\_fee observem que totes les observacions que conté són NA per tant, al ser una variable no informativa hem decidit eliminar-la, ja que no té sentit imputar una variable sencera.

```

> summary(df$Ehail_fee)
  Mode   NA's
logical 5000
          df$Ehail_fee<-NULL
> summary(df$Ehail_fee)
  Length Class Mode
0       NULL NULL

```

## Noves variables

### tlenkm

Pel fet que nosaltres treballem en sistema mètric, hem decidit crear les variables de Trip\_distance en quilòmetres la qual anomenarem "tlenkm".

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

```

> summary(df$tlenkm)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
0.01609 1.68981 3.03361 4.57586 5.79364 67.91432
NA's
64
> summary(df$Trip_distance)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
0.010 1.050 1.885 2.843 3.600 42.200 64

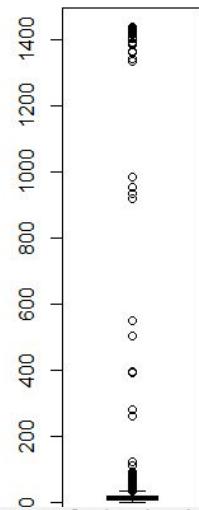
```

Aquí podem veure la diferència entre la variable en km i milles respectivament.

## traveltime

Pel fet que tenim l'hora de recollida del passatger així com l'hora de deixada, hem considerat útil tenir en una variable el temps de viatge en minuts de cada observació. A més, ens permetrà realitzar càlculs amb altres variables.

```
> summary(df$traveltime)
Boxplot of traveltime
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
0.000  5.963 10.033 23.546 16.554 1438.450
```



Al boxplot podem observar com apareixen valors molt distants a la gran majoria d'observacions que arribarien fins als 200 min. Tot i això serem una mica més permissius i considerarem outlier tota observació amb valor més gran a 800 minuts.

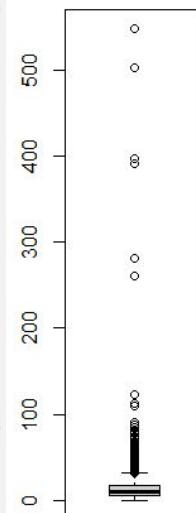
```
# Travel time in min
df$traveltime<-as.numeric(as.POSIXct(df$lpep_dropoff_datetime)) -
as.numeric(as.POSIXct(df$lpep_pickup_datetime))/60

#errors
summary(df$traveltime)
l<-which(df$traveltime<0);length(l)
if (length(l)>0) {
  ierrs[1]<-ierrs[1]+3
  jerrs["traveltime"]<-length(l)
  jerrs["lpep_dropoff_datetime"] <- length(l)
  jerrs["lpep_pickup_datetime"] <- length(l)
}
df[, "traveltime"]<-NA
df[, "lpep_dropoff_datetime"] <- NA
df[, "lpep_pickup_datetime"] <- NA

#outliers
boxplot(df$traveltime, main = "Boxplot of traveltime")
var_out<-calco(df$traveltime)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")
l<-which((df$traveltime<0)|(df$traveltime>800))
iouts[1]<-iouts[1]+1
jouts["traveltime"]<-length(l)
df[, "traveltime"]<-NA
df[, "lpep_dropoff_datetime"] <- NA
df[, "lpep_pickup_datetime"] <- NA

> summary(df$traveltime)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
0.000  5.933  9.967 13.014 16.383 548.117      39
```

Boxplot of traveltime



Al script podem veure que aquesta variable l'obtenim de la resta de l'hora en què es deixa a un passatger amb l'hora en què es recull i es divideix entre 60 per obtenir-ne els minuts.

Després es comprova si existeixen valors negatius per considerar-los errors i finalment se n'eliminen els outliers comentats i es deixen juntament amb possibles errors com NA com podem veure al summary.

## Effective speed(espeed)

Un cop definides les variables anteriors podem derivar la velocitat efectiva dels taxis en els seus serveis en km/h la qual obtindrem dividint la distància tlenkm entre el travelttime passat a hores.

```
> summary(df$espeed)
   Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
0.176 15.044 19.058 22.663 24.249 4152.108 101
```

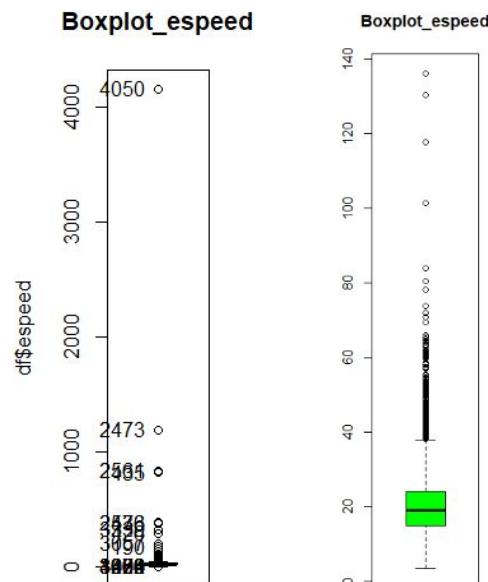
Com podem observar al summary i al primer boxplot, apareixen valors que no tenen cap sentit al voltant de 1000km/h i fins a 4000km/h. És per això que hem decidit comptar com a error i computar com a NA tota velocitat efectiva que superi els 140 km/h, ja que tot i ser una velocitat bastant elevada, podria assolir-se per carretera en algun trajecte en concret.

```
# Effective speed (km/h)
df$espeed<- (df$tlenkm/(df$travelttime))*60
summary(df$espeed)

# errors
summary(df$espeed)
l1<-which((df$espeed<=0) | (df$espeed=="Inf"))
ierrs[1]<-ierrs[1]+1
jerrs["espeed"]<-length(l1)
df[1,"espeed"]<-NA

# outliers
summary(df$espeed)
Boxplot(df$espeed)
var_out<-calcQ(df$espeed)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="blue")

l1<-which((df$espeed<=3) | (df$espeed>140))
iouts[1]<-iouts[1]+1
jouts["espeed"]<-length(l1)
df[1,"espeed"]<-NA
```



## lpep\_pickup\_period i lpep\_dropoff\_period

Pel que fa a les variables lpep\_pickup\_datetime i lpep\_dropoff\_datetime hem considerat, a més, crear una nova variable per cadascuna que ens indicarà la franja horaria en la que es realitza la recollida o es deixa als passatgers. Les franges que hem decidit crear són matí, vall del migdia, tarda i nit. Aquestes ens permetran estudiar millor els horaris en que més afluències de serveis es produueixen.

```

# lpep_pickup_time
df$lpep_pickup_time<-as.numeric(substr(strptime(df$lpep_pickup_datetime, "%Y-%m-%d %H:%M:%S"),12,13))
df$lpep_pickup_period<-1
df$lpep_pickup_period[df$lpep_pickup_time>7]<-2
df$lpep_pickup_period[df$lpep_pickup_time>10]<-3
df$lpep_pickup_period[df$lpep_pickup_time>16]<-4
df$lpep_pickup_period[df$lpep_pickup_time>19]<-1
df$lpep_pickup_period<-factor(df$lpep_pickup_period,labels=paste("Period",c("night","morning","valley","afternoon")))
df$lpep_dropoff_time<-as.numeric(substr(strptime(df$lpep_dropoff_datetime, "%Y-%m-%d %H:%M:%S"),12,13))
df$lpep_dropoff_period<-1
df$lpep_dropoff_period[df$lpep_dropoff_time>7]<-2
df$lpep_dropoff_period[df$lpep_dropoff_time>10]<-3
df$lpep_dropoff_period[df$lpep_dropoff_time>16]<-4
df$lpep_dropoff_period[df$lpep_dropoff_time>19]<-1
df$lpep_dropoff_period<-factor(df$lpep_dropoff_period,labels=paste("Period",c("night","morning","valley","afternoon")))

library(geosphere)
df$distHaversine <-distHaversine(df[,c("Pickup_longitude", "Pickup_latitude")],
                                    df[,c("dropoff_longitude", "dropoff_latitude")])/1000
# Errors
summary(df$distHaversine)
l<-which((df$distHaversine<0)|(df$distHaversine=="Inf"))
ierrs[1]<-ierrs[1]+1
jerrs["distHaversine"]<-length(l)
df[,"distHaversine"]<-NA

```

Al següent summary podem veure com queden distribuïdes les hores en les franges horàries que hem definit d'una manera més intuïtiva que observant hores a l'atzar amb certes observacions cadascuna

```
> summary(df$lpep_pickup_period)
   Period night    Period morning    Period valley Period afternoon
      2167          608          1297          928
> summary(df$lpep_dropoff_period)
   Period night    Period morning    Period valley Period afternoon
      2185          599          1284          932
```

Ipep\_pickup\_date

Creiem raonable crear una nova variable que comprengui les dates en les quals s'han realitzat els serveis de taxi extrets de la variable lpep\_pickup\_datetime la qual anomenarem lpep\_pickup\_date i convertirem en factor.

```
## Variable lpep_pickup_date,
df_datatime <- t(as.data.frame(strsplit(as.character(df$lpep_pickup_datetime), " ")))
df$lpep_pickup_date <- factor(df_datatime[,1])
```

## Imputació

## Imputació variables categòriques

Per tal de determinar uns valors que segueixin la distribució de les variables que ens interessen, ens disposem a imputar les que són categòriques i assignar-les-hi als valors determinats com a NA. Com podem observar al summary inferior només s'imputaran els valors que pertanyen a la variable lpep\_pickup\_date.

```

:summary(df[,vars_dis])
   VendorID      Payment_type Store_and_fwd_Flag     RateCodeID     f.Extra     f.MTA_tax     f.Improvement_surcharge
f.Vendor-Mobile :1103 f.PayType-Cash    :2507 FALSE:4983 Standard rate:4857 f.Extra-0 :1211 f.MTA_tax_NO : 134 f.Improvement_surcharge_NO : 138
f.Vendor-VeriFone:3897 f.PayType-Credit card:2464 TRUE : 17          Others : 143 f.Extra-0.5:1891 f.MTA_tax_YES:4866 f.Improvement_surcharge_YES:4862
f.PayType-Others   : 29

   lpep_pickup_period     Trip_type lpep_pickup_date
Period night :2180 f.TripType-Street-Hail:4877 2016-01-30: 220
Period morning : 607 f.TripType-Dispatch :123 2016-01-16: 217
Period valley :1292 2016-01-22: 197
Period afternoon: 921 2016-01-01: 195
                           2016-01-31: 188
                           (Other) :13919
                           NA's   : 64

```

```

## Imputation of qualitative variables
````{r}

vars_dis <- c("VendorID", "Payment_type", "Store_and_fwd_flag", "RateCodeID",
"f.Extra", "f.MTA_tax", "f.Improvement_surcharge", "lpep_pickup_period",
"Trip_type", "lpep_pickup_date")

summary(df[,vars_dis])
res.immc<-imputeMCA(df[,vars_dis],ncp=10)
summary(res.immc$completeObs)

# Check one by one
df[, vars_dis]<-res.immc$completeObs
summary(df[,vars_dis])
````
```

Un cop feta la imputació comprovem si pot haver-se originat algun valor erroni pel que fa a la variable lpep\_pickup\_date i l'eliminem si fos el cas. Al summary següent podem observar com han desaparegut els NA i s'han distribuït en diversos dels valors pertanyents a la variable.

```

> summary(res.immc$completeObs)
      VendorID      Payment_type   Store_and_fwd_flag    RateCodeID      f.Extra      f.MTA_tax
f.Vendor-Mobile :1103  f.PayType-Cash    :2507 FALSE:4983 Standard rate:4857 f.Extra-0 :2311  f.MTA_tax_NO : 134
f.Vendor-VeriFone:3897 f.PayType-Credit card:2464 TRUE : 17    Others       :143 f.Extra-0.5:1891 f.MTA_tax_YES:4866
f.PayType-Others     : 29         

      f.Improvement_surcharge   lpep_pickup_period      Trip_type      lpep_pickup_date
f.Improvement_surcharge_NO : 138    Period night     :2180    f.TripType-Street-Hail:4877 2016-01-30: 238
f.Improvement_surcharge_YES:4862  Period morning   : 607    f.TripType-Dispatch : 123 2016-01-01: 227
                                         Period valley    :1292
                                         Period afternoon: 921 2016-01-16: 222
                                         (Other)           :3738 2016-01-22: 201
                                         (Other)           :3738 2016-01-31: 168
                                         (Other)           :3738 2016-01-09: 186
                                         (Other)           :3738
```

## Imputació de variables numèriques

Un cop descartats tots els outliers i errors i marcats com a NA (missings), ens disposem a imputar les variables numèriques amb l'objectiu de crear uns valors que segueixin la distribució de les observacions vàlides:

- "Pickup\_longitude", "Pickup\_latitude", "Dropoff\_longitude", "Dropoff\_latitude", "Fare\_amount", "Tip\_amount", "Tolls\_amount", "espeed", "traveltime", "tlenkm", "distHaversine".

Al següent summary podríem observar com ha quedat cada variable de les que imputarem després d'eliminar-ne els errors i outliers i abans de realitzar-hi la imputació.

```

> summary(df[,vars_con])
   Pickup_longitude   Pickup_latitude   Dropoff_longitude   Dropoff_latitude   Fare_amount      Tip_amount      Tolls_amount
Min. :-74.16        Min. :40.57        Min. :-74.18        Min. :40.57        Min. : 0.00        Min. : 0.00000
1st Qu.:-73.96      1st Qu.:40.70      1st Qu.:-73.97      1st Qu.:40.70      1st Qu.: 6.50        1st Qu.: 0.00000  1st Qu.:0.00000
Median :-73.95      Median :40.75      Median :-73.95      Median :40.75      Median : 9.00        Median : 0.00000  Median :0.00000
Mean  :-73.94      Mean :40.75      Mean :-73.93      Mean :40.74      Mean :11.97        Mean : 1.227    Mean :0.09133
3rd Qu.:-73.92      3rd Qu.:40.80      3rd Qu.:-73.91      3rd Qu.:40.79      3rd Qu.:14.50      3rd Qu.: 2.000    3rd Qu.:0.00000
Max.  :-73.57      Max. :40.89      Max. :-73.42      Max. :40.94      Max. :71.50        Max. :17.880    Max. :5.54000
NA's   :7           NA's :7           NA's :4           NA's :15           NA's :10           NA's :3

   espeed      traveltime      tlenkm      distHaversine
Min. : 3.559    Min. : 0.000    Min. : 0.01609  Min. : 0.000
1st Qu.: 15.064   1st Qu.: 5.933    1st Qu.: 1.68981  1st Qu.: 1.242
Median : 19.059   Median : 9.967    Median : 3.03361  Median : 2.235
Mean  : 20.956   Mean :13.034    Mean : 4.57586  Mean : 3.212
3rd Qu.: 24.216   3rd Qu.:16.383    3rd Qu.: 5.79364  3rd Qu.: 4.127
Max.  :130.357   Max. :548.117    Max. :67.91432  Max. :26.106
NA's  :126       NA's :39         NA's :64         NA's :10
```

```

## Imputation of numeric variables
``{r}
names(df)
summary(df)
#vars_con
vars_con <- c("Pickup_longitude", "Pickup_latitude", "Dropoff_longitude",
"Dropoff_latitude", "Fare_amount", "Tip_amount", "Tolls_amount", "espeed",
"traveltime", "tlenkm", "disthaversine")

summary(df[,vars_con])
res_impc<-imputePCA(df[,vars_con],ncp=6)
summary(res_impc$completeObs)
df[,vars_con]<-res_impc$completeObs
summary(df[,vars_con])

```

Una cop realitzada la imputació d'aquestes variables ens dediquem a comprovar una per una que no presentin possibles errors en l'assignació de valors com per exemple que una velocitat efectiva sigui negativa.

Als dos summary que tenim a continuació podem veure la diferència entre les dades acabades d'imputar (summary 1) i les dades després d'eliminar-ne errors sorgits de la mateixa imputació.

```

summary(df[,vars_con])
Passenger_count      tlenkm      Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Fare_amount      espeed      Tip_amount
Min. :0.0000  Min. : 0.01609  Min. :-74.16   Min. :40.57  Min. :-74.18   Min. :40.57  Min. : 0.00  Min. : 3.559  Min. : 0.000
1st Qu.:1.0000  1st Qu.: 1.68981  1st Qu.:-73.96   1st Qu.:40.70  1st Qu.:-73.97   1st Qu.:40.70  1st Qu.: 6.50  1st Qu.: 15.064  1st Qu.: 0.000
Median :1.0000  Median : 3.03361  Median :-73.95   Median :40.75  Median :-73.95   Median :40.75  Median : 9.00  Median : 19.059  Median : 0.000
Mean   :1.375   Mean   : 4.57586  Mean  :-73.94   Mean  :40.75  Mean  :-73.93   Mean  :40.74  Mean  :11.97  Mean  : 20.956  Mean  : 1.227
3rd Qu.:1.0000  3rd Qu.: 5.79364  3rd Qu.:-73.92   3rd Qu.:40.80  3rd Qu.:-73.91   3rd Qu.:40.79  3rd Qu.:14.50  3rd Qu.: 24.216  3rd Qu.: 2.000
Max.  :6.0000  Max.  :67.91432  Max. :-73.57   Max. :40.89  Max. :-73.42   Max. :40.94  Max. :71.50  Max. :130.357  Max. :17.880
NA's   :64      NA's   : 7      NA's  : 7      NA's  : 4      NA's  : 15     NA's  :126    NA's  : 10      NA's  : 10

Tolls_amount      tpep_pickup_time traveltime      disthaversine      Total_amount
Min. :0.00000  Min. : 0.00  Min. : 0.000  Min. : 0.000  Min. : 0.000
1st Qu.:0.00000 1st Qu.: 9.00  1st Qu.: 5.933  1st Qu.: 1.242  1st Qu.: 7.88
Median :0.00000  Median :15.00  Median : 9.967  Median : 2.238  Median :11.30
Mean   :0.09133  Mean   :13.49  Mean   :13.014  Mean   : 3.212  Mean   :14.54
3rd Qu.:0.00000  3rd Qu.:19.00  3rd Qu.:16.383  3rd Qu.: 4.127  3rd Qu.:17.30
Max.  :5.54000  Max.  :23.00  Max.  :548.117  Max.  :26.106  Max.  :95.54
NA's   :3         NA's   :39    NA's  :10      NA's  :12      NA's  : 12

> summary(df[,vars_con])
Passenger_count      tlenkm      Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Fare_amount      espeed      Tip_amount
Min. :0.0000  Min. : 0.001  Min. :-74.16   Min. :40.57  Min. :-74.18   Min. :40.57  Min. : 0.0  Min. : 0.001  Min. : 0.00
1st Qu.:1.0000  1st Qu.: 1.690  1st Qu.:-73.96   1st Qu.:40.70  1st Qu.:-73.97   1st Qu.:40.70  1st Qu.: 6.5  1st Qu.: 15.167  1st Qu.: 0.00
Median :1.0000  Median : 3.034  Median :-73.95   Median :40.75  Median :-73.95   Median :40.75  Median : 9.0  Median : 19.088  Median : 0.00
Mean   :1.375   Mean   : 4.582  Mean  :-73.94   Mean  :40.75  Mean  :-73.93   Mean  :40.74  Mean  :12.0  Mean  : 20.980  Mean  : 1.23
3rd Qu.:1.0000  3rd Qu.: 5.810  3rd Qu.:-73.92   3rd Qu.:40.80  3rd Qu.:-73.91   3rd Qu.:40.79  3rd Qu.:14.5  3rd Qu.: 24.241  3rd Qu.: 2.00
Max.  :6.0000  Max.  :67.914  Max. :-73.57   Max. :40.89  Max. :-73.42   Max. :40.94  Max. :71.5  Max. :130.357  Max. :17.88
NA's   :3         NA's   :39    NA's  :10      NA's  :12      NA's  : 12

```

## Multivariant outliers

Per la realització de Moutlier, volem agafar aquelles variables no senceres. Les úniques variables que ens van permetre realitzar Moutlier van ser tlenkm, Fare\_amount i Total\_amount.

```

> summary(df[,vars_con])
      tlenkm      Fare_amount      Total_amount
Min. : 0.001  Min. : 0.0  Min. : 0.000
1st Qu.: 1.690  1st Qu.: 6.5  1st Qu.: 7.872
Median : 3.034  Median : 9.0  Median :11.300
Mean   : 4.582  Mean   :12.0  Mean   :14.543
3rd Qu.: 5.810  3rd Qu.:14.5  3rd Qu.:17.300
Max.  :67.914  Max.  :71.5  Max.  :95.540

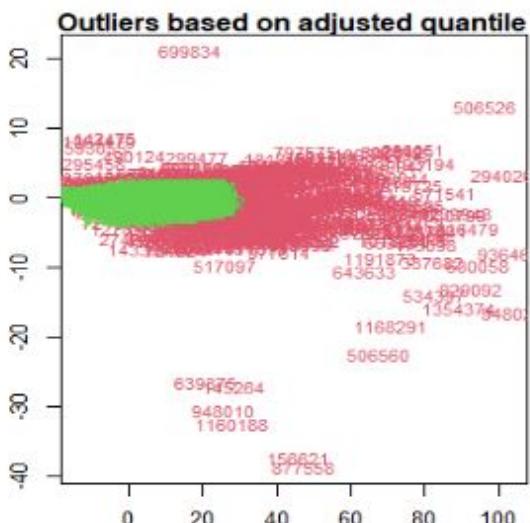
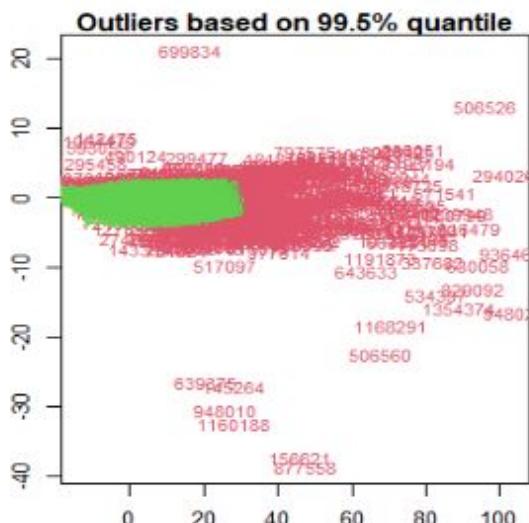
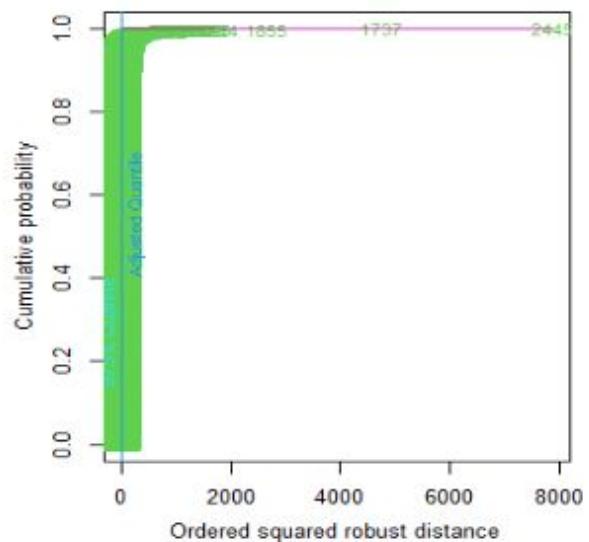
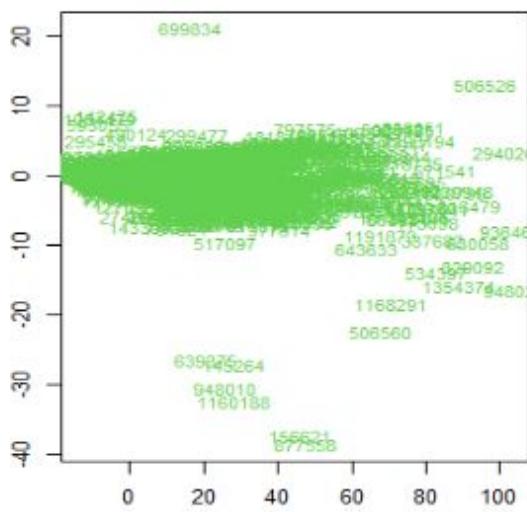
```

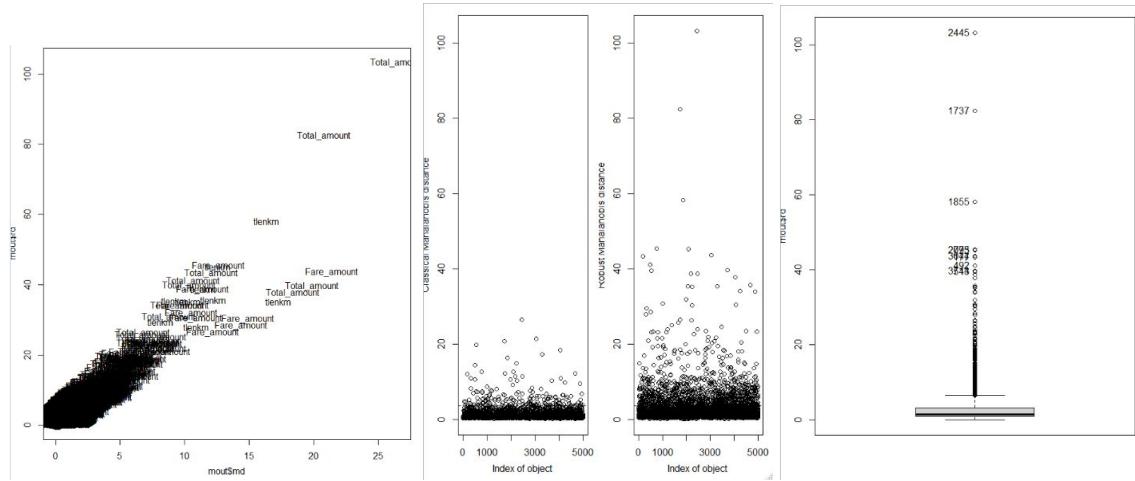
```

## Multivariate Outliers

``{r}
vars_con <- c( "tlenkm", "Fare_amount", "Total_amount")
summary(df[,vars_con])
aq.plot(df[,vars_con],delta=qchisq(0.995,length(vars_con)),quantile=0.75)
mout<-Moutlier(df[,vars_con],quantile = 0.995, plot = TRUE)
par(mfrow=c(1,1))
plot(mout$md,mout$rd, type="n")
text(mout$md,mout$rd,labels=vars_con)

Boxplot(mout$rd)
summary(mout$rd)
l<-which(mout$rd>50);length(l)
df[,"multiouts"] <- FALSE
df[l,"multiouts"] <- TRUE
df[,"multiouts"] <- as.factor(df[,"multiouts"])
``
```

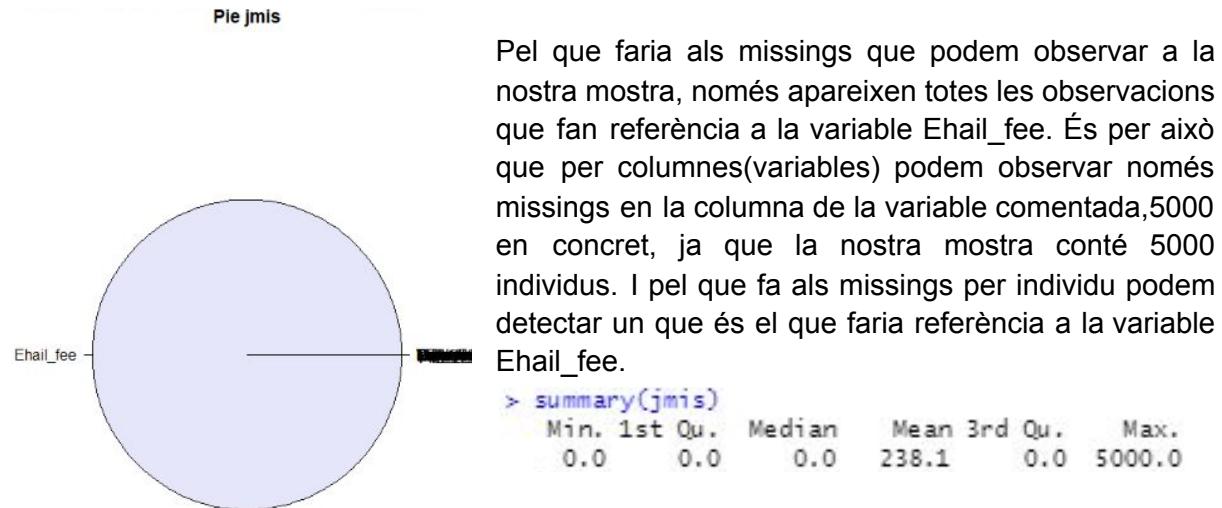




Un cop fet Moutlier, es va crear la variable multiouts. Aquells individus amb distància de Mahanalobis robusta superior a 50 se li va assignar el valor TRUE a la variable multiouts. Altrament, multiouts se li va assignar el valor FALSE.

## Data quality report

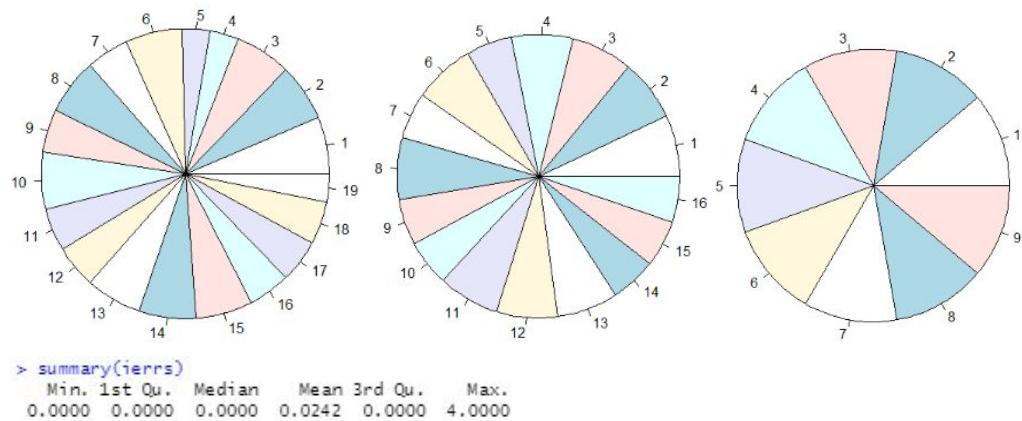
### Missings



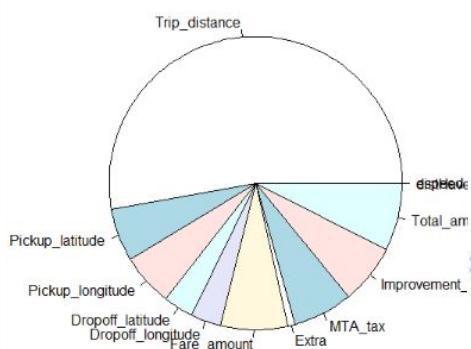
### Errors

El primer gràfic ens mostra el nombre d'individus amb més d'un error en les observacions de totes les variables, que puja a 19 individus. El segon el nombre d'individus que tenen més de dos errors, on tenim una baixada fins a 16 i, finalment, tenim el tercer gràfic que ens

mostra els individus que tenen més de tres errors en observacions de variables diferents. Aquest tercer gràfic ens mostra que el nombre d'individus amb més de tres errors baixa fins als nou d'un total de 5000, on aquests més de tres són estrictament quatre errors, cosa que podem confirmar gràcies al summary.



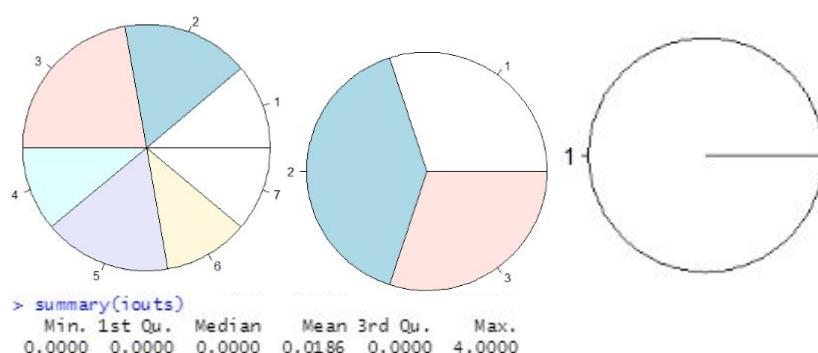
Pie jerrs

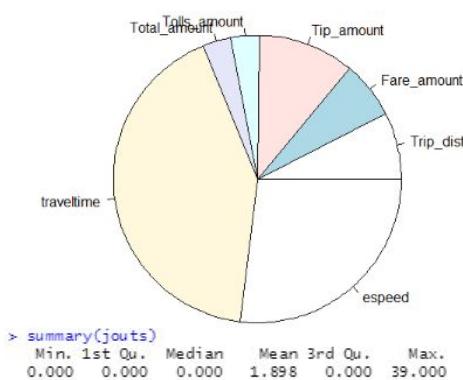


En aquest gràfic en forma de pastís podem veure una comparació entre els errors que presenta la mostra per variable. Podem observar que la que més errors presenta seria 64 errors la variable Trip\_distance i, que la mitjana d'errors per variable és de 2.24, un valor molt baix tenint en compte que disposem de 5000 individus en aquesta mostra.

## Outliers

Tindríem la mateixa situació que teníem abans pels errors on el primer gràfic indicaria el nombre d'individus amb més d'un outlier, el segon amb més de dos outlier i el tercer amb 4 outlier per la limitació del summary.





Pel que faria a les variables que disposen d'outliers, les veuríem representades en aquest altre gràfic en forma de pastís. A partir d'aquest i juntament amb el summary podem detectar que la variable amb més outliers seria la de travelltime amb 39 valors considerats com a tal de 39 individus diferents i la seguiria la espeed. La mitjana d'outliers queda definida en 1.9 aproximadament que és un valor molt baix considerant que per cada variable hi hauria dos outliers si aquests fossin distribuïts equitativament i disposem d'una mostra amb 5000 individus.

## Discretització

A continuació ens dediquem a fer la discretització de cadascuna de les variables numèriques per tal de dividir-les en els rangs que considerem.

### espeed

Al primer summary veiem indicats els valors distribuïts per rangs constraint un 25% de les observacions cadascun d'ells. Veiem que el primer rang comprèn valors de gairebé 0km/h fins a 15.1km/h mentre que al segon, només hi ha una diferència entre límits de 4km/h. Al tercer rang passa una cosa semblant ja que només hi ha una diferència de 5.1km/h, no com al tercer que hi caben totes les velocitats superiors a 24.2km/h. D'això en podem extreure que la meitat de valors estan situats entre els 15.1 i els 24.2km/h, on tindrem compresa la mitjana.

```
> ## variable espeed
> varaux<-factor(cut(df$espeed,breaks=quantile(df$espeed,seq(0,1,0.25),na.rm=TRUE),include.lowest = T ))
> summary(varaux)
[0.001,15.1] (15.1,19.1] (19.1,24.2] (24.2,130]
    1250      1251      1249      1250
> tapply(df$espeed,varaux,median) #tapply(X, INDEX, FUN = NULL) map function
[0.001,15.1] (15.1,19.1] (19.1,24.2] (24.2,130]
    12.81027   17.14094   21.27133   29.88982
> df$espeed<-factor(cut(df$espeed,breaks=c(0,25,max(df$espeed),na.rm=TRUE),include.lowest = T ))
> levels(df$espeed)<-paste("f.espeed-",levels(df$espeed),sep="")
```

Al segon summary podem observar la mediana de cadascun dels rangs, on podem destacar, sobretot en els primer i quart rang que aquesta està molt a prop dels límits superior i inferior respectivament.

### tlenkm

En la variable que descriu la distància en km podem veure una distribució més homogènia entre els tres primers quarts, ja que disten entre 1.7 i 2.75 km els seus límits. La mitjana la tenim situada en el rang pertanyent als valors del 50-75% per tant podem deduir que els valors situats en el quart rang seran elevats per tal de fer-la augmentar

```

> ## Variable tlenkm
> summary(df$tlenkm)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.001  1.690  3.034  4.581  5.810 67.914
> varaux<-factor(cut(df$tlenkm,breaks=quantile(df$tlenkm,seq(0,1,0.25),na.rm=TRUE),include.lowest = T ))
> summary(varaux)
[0.001,1.69] (1.69,3.03] (3.03,5.81] (5.81,67.9]
1258      1242     1254     1246
> df$df.tlenkm<-factor(cut(df$tlenkm,breaks=c(0,5,max(df$tlenkm),na.rm=TRUE),include.lowest = T ))
> levels(df$df.tlenkm)<-paste("f.tlenkm-",levels(df$df.tlenkm),sep="")

```

## traveltime

Pel que fa al temps en minuts de viatge, principalment veiem que gairebé un 75% dels valors estan entre 0 i 15 minuts fet que ens mostra, juntament amb la mitjana de 13 min, que els viatges no solen trigar gaire més normalment.

```

## Variable traveltime
summary(df$traveltime)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.000  5.950  9.967 13.001 16.350 548.117
varaux<-factor(cut(df$traveltime,breaks=c(0,15,40,max(df$traveltime)),include.lowest = T ))
summary(varaux)
[0,15] (15,40] (40,548]
3567    1323     110
df$df.traveltime <- varaux
levels(df$df.traveltime)<-paste("f.traveltime-",levels(df$df.traveltime),sep="")

```

## distHaversine

Pel que fa a la distància entre punts de recollida i arribada veiem que els quartils són prou parells respecte a rangs de valors menys el quart i que les seves medianes ens mostren una distribució equilibrada dins d'aquests.

```

> varaux<-factor(cut(df$distHaversine,breaks=quantile(df$distHaversine,seq(0,1,0.25),na.rm=TRUE),include.lowest = T ))
> summary(varaux)
[0,1.24] (1.24,2.23] (2.23,4.13] (4.13,26.1]
1250     1250     1250     1250
> tapply(df$distHaversine,varaux,median) #tapply(X, INDEX, FUN = NULL) map function
[0,1.24] (1.24,2.23] (2.23,4.13] (4.13,26.1]
0.8461959  1.6775889  3.0162302  6.2127188
> df$df.distHaversine<-factor(cut(df$distHaversine,breaks=c(0,5,10,max(df$distHaversine)),include.lowest = T ))
> levels(df$df.distHaversine)<-paste("f.distHaversine-",levels(df$df.distHaversine),sep="")

```

## Fare\_amount

Pel que faria a l'import del trajecte, tenint en compte que la mitjana se situaria al tercer quartil, que compren entre 9 i 14.5\$, podem veure que les observacions del quart quartil són prou significativament altes per condicionar més aquesta que el 50% que comprèn dels 0\$ als 9\$

```

varaux<-factor(cut(df$Fare_amount,breaks=quantile(df$Fare_amount,seq(0,1,0.25),na.rm=TRUE),include.lowest = T ))
summary(df$Fare_amount)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0     6.5     9.0    12.0   14.5    71.5
df$df.Fare_amount <- varaux
levels(df$df.Fare_amount)<-paste("f.Fare_amount-",levels(df$df.Fare_amount),sep="")
summary(df$df.Fare_amount)
f.Fare_amount-[0,6.5] f.Fare_amount-(6.5,9] f.Fare_amount-(9,14.5] f.Fare_amount-(14.5,71.5]
1470          1037          1262          1231

```

# Univariant exploratory analysis (EDA)

A continuació farem una breu descripció de cada una de les variables un cop fet el preprocessament. Només hem considerat fer l'estudi d'aquelles variables importants y que formaran part del nostre dataset final.

```
> summary(df)

      VendorID          Payment_type  Store_and_fwd_flag    RateCodeID      f.Extra      f.MTA_tax
f.Vendor-Mobile :1103  f.PayType-Cash   :2507  FALSE:4983  Standard rate:4857  f.Extra-0 :2311  f.MTA_tax_NO :134
f.Vendor-VeriFone:3897 f.PayType-Credit card:2464  TRUE : 17    Others : 143   f.Extra-0.5:1891  f.MTA_tax_YES:4866
f.PayType-Others     : 29                   f.Extra-1 : 798

      f.Improvement_surcharge    lpep_pickup_period       Trip_type    lpep_pickup_date multiouts      f.espeed
f.Improvement_surcharge_NO : 138 Period night :2180  f.TripType-Street-Hail:4877 2016-01-30: 238  FALSE:4997  f.espeed-[0,1] : 6
f.Improvement_surcharge_YES:4862 Period morning : 607  f.TripType-Dispatch : 123 2016-01-01: 227  TRUE : 3   f.espeed-(1,25] :3875
f.tlenkm             f.traveltime      f.distHaversine AnyToll      f.Fare_amount
f.tlenkm-[0,1] : 416 f.traveltime-[0,15] :3567 f.distHaversine-[0,5] :4079 AnyToll No : 87  f.Fare_amount-[0,6,5] :1470
f.tlenkm-(1,5] : 3100 f.traveltime-(15,40] :1323 f.distHaversine-(5,10] : 727 AnyToll Yes:4913  f.Fare_amount-(6,5,9] :1037
f.tlenkm-(5,67.9]:1484 f.traveltime-(40,548] : 110 f.distHaversine-(10,26.1] : 194 f.Fare_amount-(9,14,5] :1262
                                         f.Fare_amount-(14,5,71,5]:1231

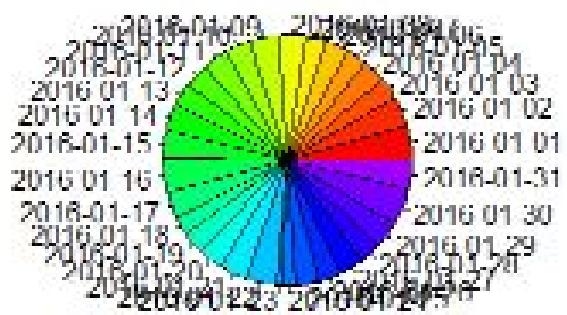
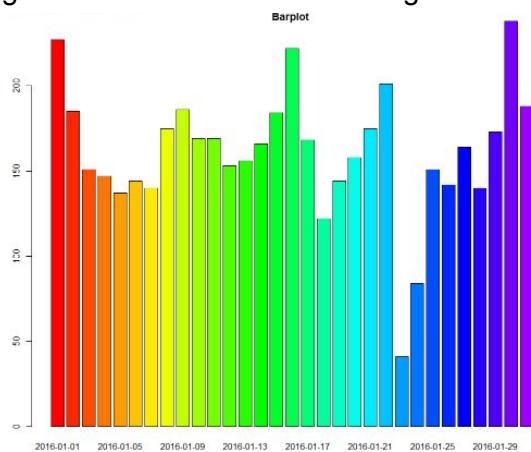
      f.Passenger_count Passenger_count      tlenkm      Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Fare_amount      espeed
f.Passenger_count-1 :4207 Min. :0.0000  Min. : 0.001  Min. :-74.16  Min. :40.57  Min. :74.18  Min. :40.57  Min. : 0.0  Min. : 0.001
f.Passenger_count-2 : 367 1st Qu.:1.0000  1st Qu.: 1.690  1st Qu.:-73.96  1st Qu.:40.70  1st Qu.:-73.97  1st Qu.:40.70  1st Qu.: 6.5  1st Qu.: 15.143
f.Passenger_count-Others: 426 Median :1.0000  Median : 3.034  Median :-73.95  Median :40.75  Median :-73.95  Median :40.75  Median : 9.0  Median : 19.081
Mean : 1.375  Mean : 4.581  Mean :-73.94  Mean :40.75  Mean :-73.93  Mean :40.74  Mean :12.0  Mean : 20.971
3rd Qu.:1.0000 3rd Qu.: 5.810  3rd Qu.:-73.92  3rd Qu.:40.80  3rd Qu.:-73.91  3rd Qu.:40.79  3rd Qu.:14.5  3rd Qu.: 24.249
Max. :6.0000  Max. :67.914  Max. :-73.57  Max. :40.89  Max. :-73.42  Max. :40.94  Max. :71.5  Max. :130.357

      Tip_amount      Tolls_amount    lpep_pickup_time traveltime      distHaversine AnyTip      Total_amount
Min. :-0.04124 Min. :0.000000  Min. : 0.000  Min. : 0.000  Min. : 0.0000  AnyTip No :2869  Min. : 0.000
1st Qu.: 0.00000 1st Qu.:0.000000 1st Qu.: 9.00  1st Qu.: 5.950  1st Qu.: 1.242  AnyTip Yes:2131 1st Qu.: 7.872
Median : 0.00000 Median :0.000000  Median :15.00  Median : 9.967  Median : 2.235  Median :11.300
Mean : 1.23015 Mean :0.09183  Mean :13.63  Mean :13.001  Mean : 3.211  Mean :14.541
3rd Qu.: 2.00000 3rd Qu.:0.000000 3rd Qu.:19.00  3rd Qu.:16.350  3rd Qu.: 4.126  3rd Qu.:17.300
Max. :17.88000 Max. :5.540000  Max. :168.53  Max. :548.117  Max. :26.106  Max. :95.540
```

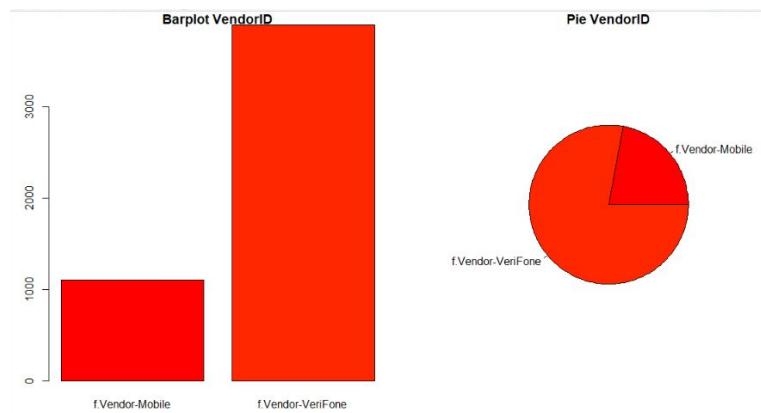
Aquest summary ens permetrà treure algunes conclusions del nostre dataset juntament amb les descripcions gràfiques.

## Lpep\_pickup\_date

Com podem veure en barplot, el nostre dataset conté alguns registres de taxis del mes de gener del 2016. Com es pot veure, el mes es manté constant excepte a finals de mes, on el nombre de registres dels dies 23 i 24 es molt baix. En contraposició els dies 1, 16 i 30 de gener son els dies amb més registres de taxis.

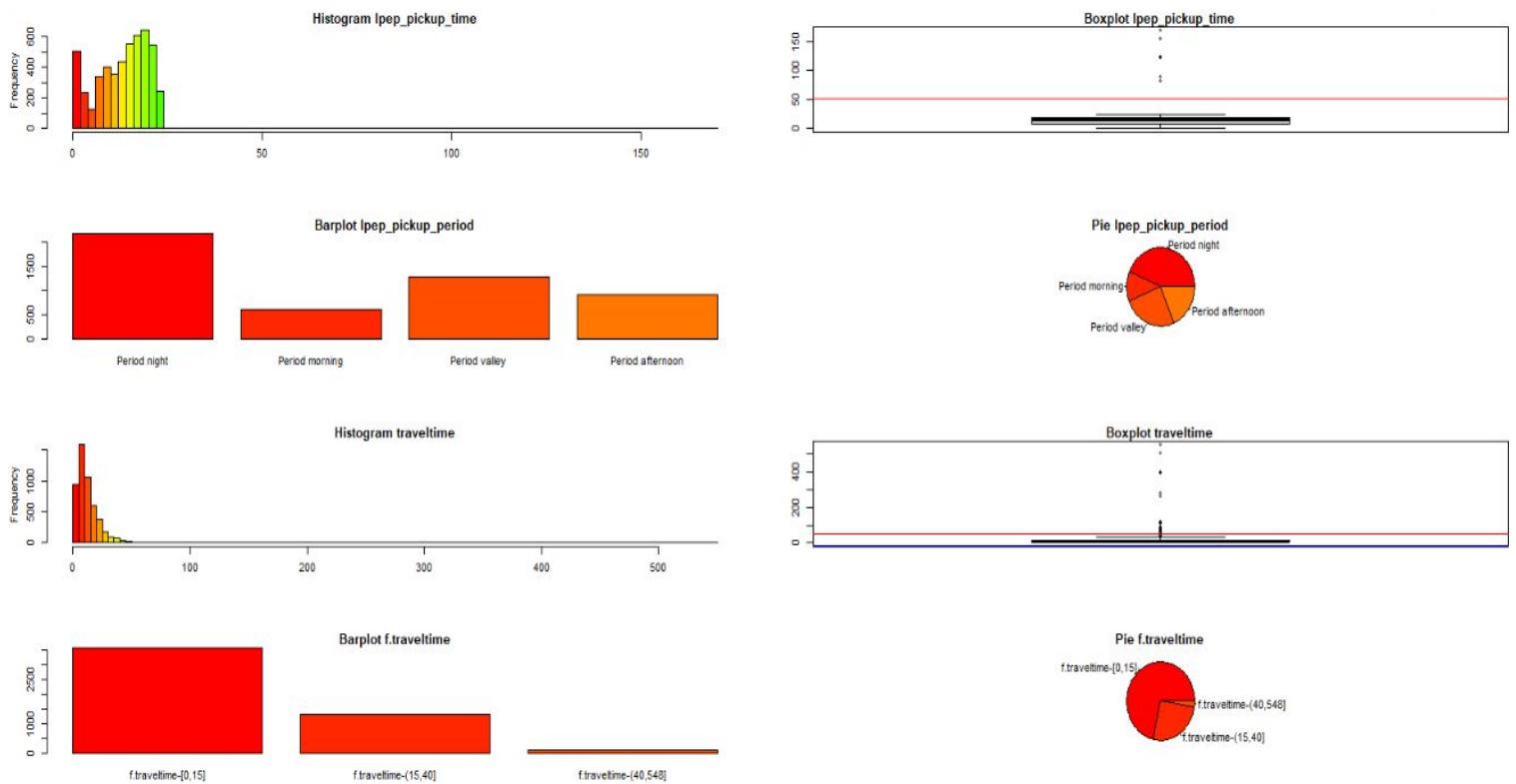


## VendorID



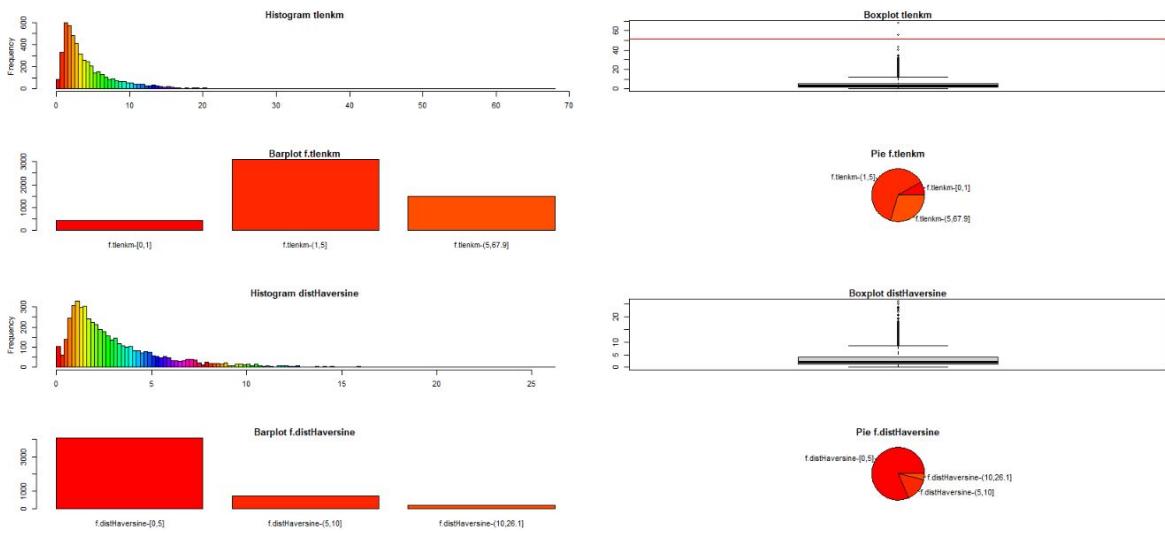
Com podem veure, la gran majoria de tuples del nostre dataset està proporcionat per VeriFone.

## lpep\_pickup\_time, lpep\_pickup\_period, traveltimes, f.traveltimes



Com podem veure tant en els gràfics com en el summary, la durada dels nostres viatges es de mitjana de 13 minuts, és a dir, com es veu en el barpot, la majoria de viatges té una duració de entre 0 i 15 minuts. Respecte a quina hora es solia agafar el taxi en el mes de gener del 2016, podem dir que era a partir de les 19 hores, predominant el horari de nit. Hem considerat eliminar les variables lpep\_pickup\_datatime, lpep\_dropoff\_datatime i altres relacionades amb aquestes, ja que les variables explicades anteriorment ja ens aporten tota la informació necessària.

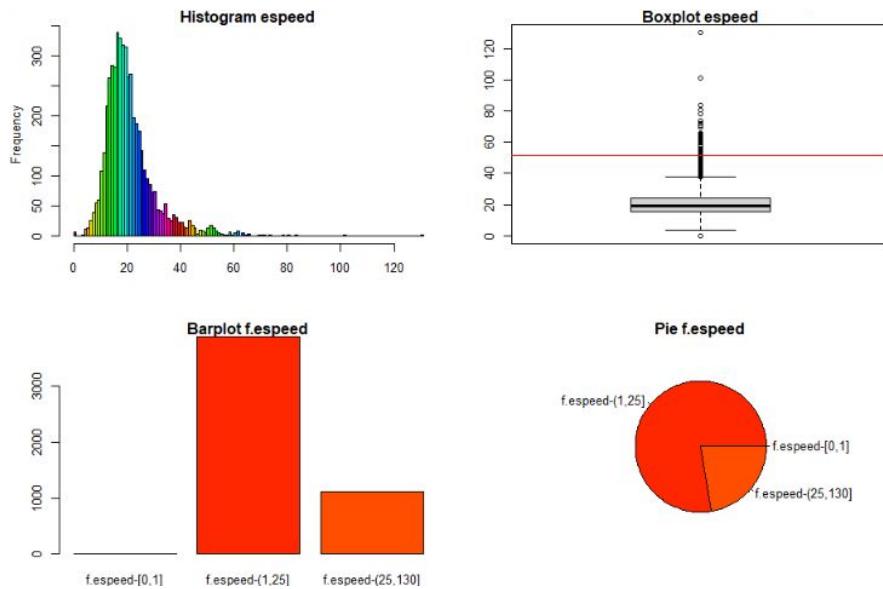
## tlenkm, distHaversine



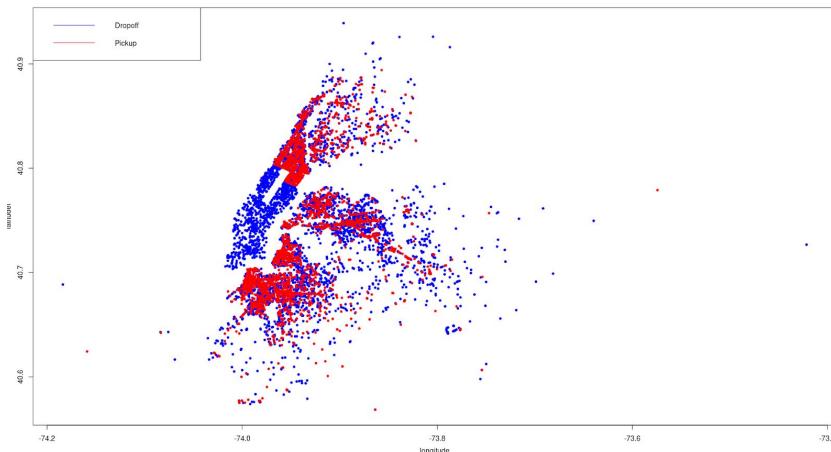
Com hem vist en els gràfics anteriors, la duració dels viatges era d'entre 0 i 15 minuts és per això, que la distància recorreguda pels taxis sigui inferior a 15 km i que la distància entre punts sigui inferior a 5 km. Aquesta informació ens permet dir que els viatges feien trajectes curtes.

## espeed

Com podem veure la velocitat mitjana dels taxis es de 21 km/h i predomina una velocitat de 19 km/h aproximadament. Com hem explicat abans, aquestes dades son dels taxis de New York, és per això que té molt de sentit que la velocitat dels taxis sigui baixa, ja que en les grans ciutats trobem molt tràfic.



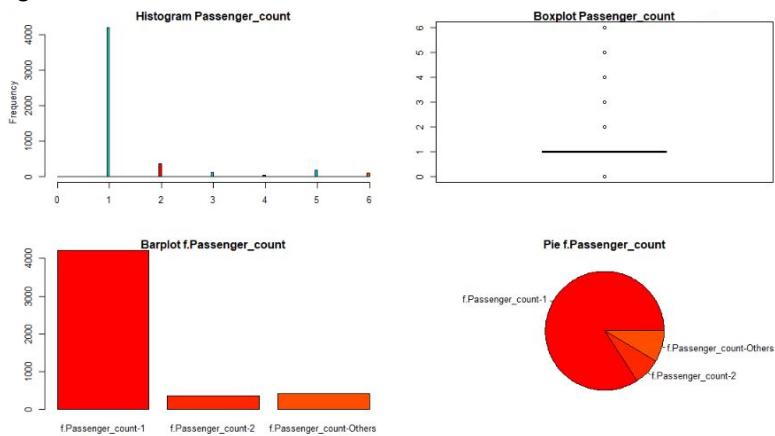
`lpep_pickup_latitude`, `lpep_pickup_longitude`, `lpep_dropoff_latitude`,  
`lpep_dropoff_longitude`



En el plot podem veure els punts d'agafada, punts vermells, i deixada dels passatgers, punts blaus. Com es pot admirar, molts dels passatgers es deixen en una zona on no s'agafen passatgers.

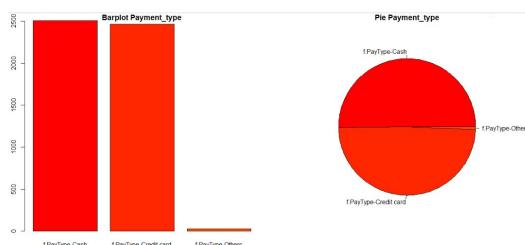
## Passenger\_count

Respecte al nombre de passatgers, podem veure que quasi sempre en el viatge es troba només un passatger.

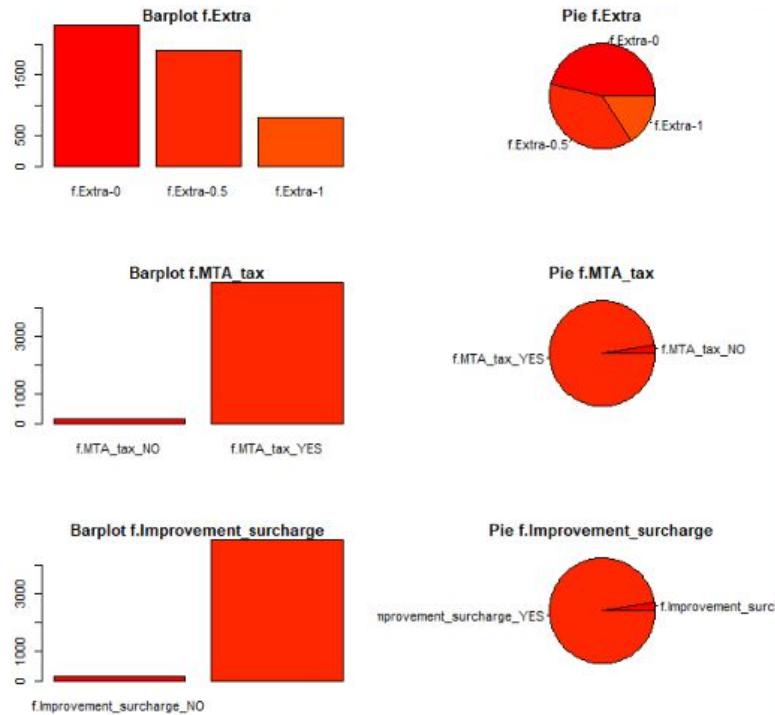


## Payment\_type

Amb relació a com efectuen els passatgers el pagament, podem veure que en la meitat de les vegades és en efectiu i altres amb tarja.



## Extra, f.Extra, MTA\_tax, Improvement\_surcharge

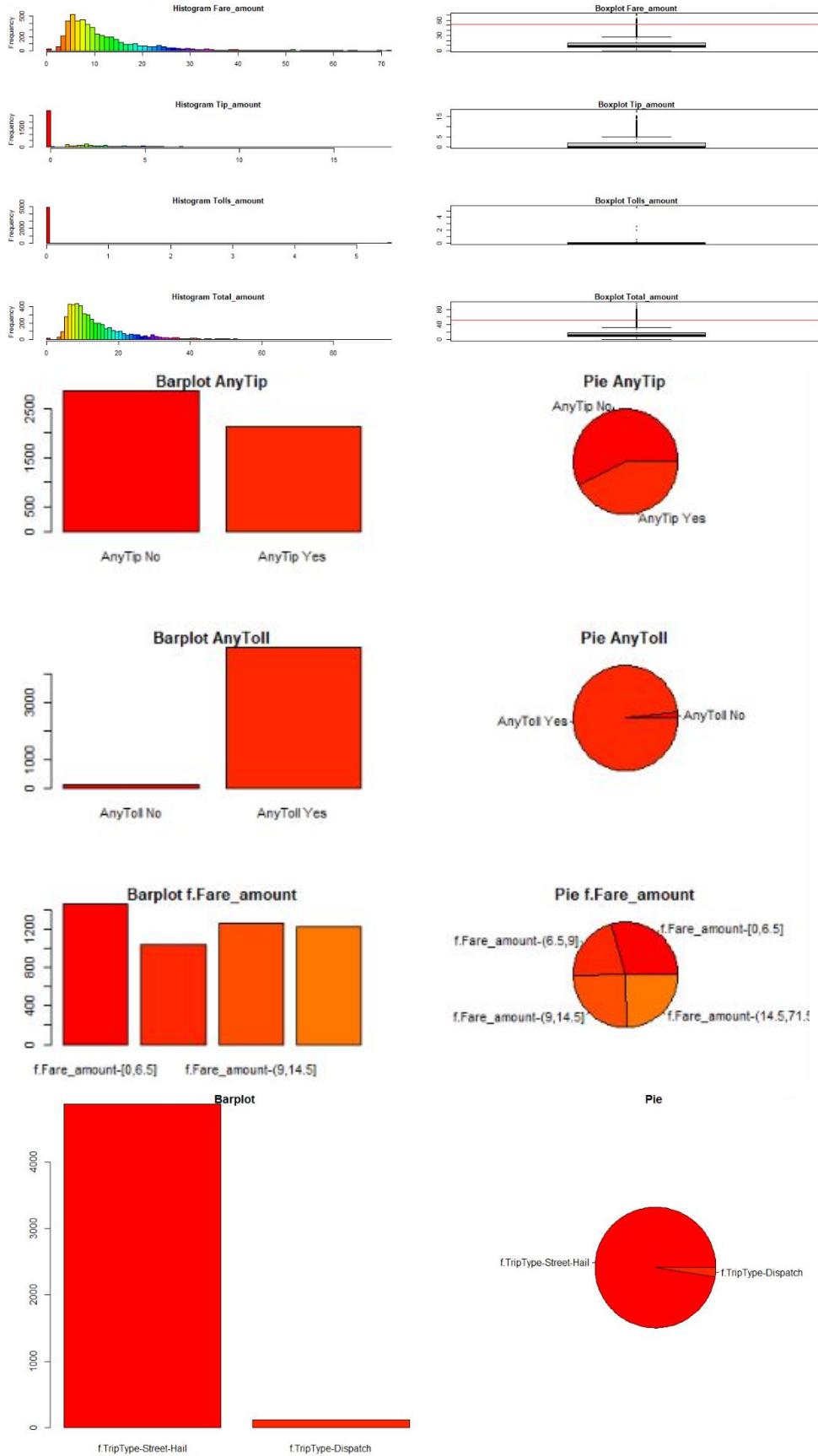


Respecte als extres, podem veure que en la meitat de les ocasions no s'aplica extra per nocturnitat i hores puntes. Aquesta informació no té cap sentit si observem la variable lpep\_pickup\_period, ja que ens deia que molts dels viatges eran nocturns. Respecte a les variables MTA\_tax i Improvement\_surcharge, podem dir que no ens dóna gaire informació, ja que quasi tots els passatgers han pagat les taxes.

Fare\_amount, Tip\_amount, Tolls\_amount, Total\_amount, AnyTip i AnyToll

Respecte als imports no relacionats amb taxes, trobem:

- Peatges: casi sempre el passatger o conductor ha hagut de pagar peatges. Tot i això, els costos dels peatges no són molt alts, ja que el valor mitjà es de 0.09 dòlars.
- Propina: aproximadament el 60% de les ocasions els usuaris no donaven propina. El valor mitjà de la propina és de 1.3 dòlars aproximadament.
- Tarifa calculada pel taxímetre: La similitud en les distribucions entre Fare\_amount i Total\_amount, ens informa que les variables poden estar correlacionades. Això té sentit, ja que el total que ha de pagar el passatger dependrà de la tarifa calculada pel taxímetre més petits pagaments (peatges, propines i taxes).
- Total: en total els passatgers paguen de mitja uns 15 dòlars.



# Profiling

## Profiling Total\_amount

```
> res.condes$quanti
      correlation      p.value
Fare_amount      0.96429378  0.000000e+00
tlenkm          0.88606012  0.000000e+00
distHaversine   0.81148547  0.000000e+00
Tip_amount       0.61867812  0.000000e+00
traveltime      0.45675475  2.818852e-256
espeed          0.43384823  1.245981e-228
Tolls_amount     0.31384500  1.037429e-114
Dropoff_longitude 0.03118898  2.742751e-02
lpep_pickup_time -0.04016539  4.503467e-03
Pickup_latitude   -0.10085123  8.824048e-13
Dropoff_latitude   -0.14384593  1.578804e-24
> res.condes$quali
      R2      p.value
f.tlenkm        0.508323923  0.000000e+00
f.traveltime    0.499311572  0.000000e+00
f.distHaversine 0.534138997  0.000000e+00
f.Fare_amount    0.631901913  0.000000e+00
f.espeed         0.144188846  1.112472e-169
AnyToll          0.104974746  1.505319e-122
AnyTip           0.067723891  3.371357e-78
Payment_type     0.068478740  1.066937e-77
RateCodeID       0.038918754  4.707810e-45
multiouts        0.008310979  1.059920e-10
Trip_type        0.008016047  2.267247e-10
f.MTA_tax        0.007307771  1.410197e-09
f.Improvement_surcharge 0.004970336  6.034515e-07
```

amb la discretització, són les més correlacionades. Tot i això aquest profiling no ens dóna molta informació per això utilitzarem el profiling de les categories per treure conclusions.

## Profiling de les categories

```
> res.condes$category
      Estimate      p.value
f.Fare_amount=f.Fare_amount-(14.5,71.5) 13.3466387  0.000000e+00
f.distHaversine=f.distHaversine-(10,26.1) 16.7349168  0.000000e+00
f.tlenkm=f.tlenkm-(5,67.9) 11.5397339  0.000000e+00
f.espeed=f.espeed-(25,138) 16.8990525  5.423368e-197
AnyToll=AnyToll_No 12.7638515  1.505319e-122
AnyTip=AnyTip_Yes 2.7107302  3.371357e-78
Payment_type=f.PayType-Credit card 1.8691105  2.183961e-77
RateCodeID=Others 6.0967688  4.707810e-45
multiouts=TRUE 19.1771081  1.059920e-10
Trip_type=f.TripType-Dispatch 2.9773064  2.267247e-10
f.MTA_tax=f.MTA_tax_NO 2.7266309  1.410197e-09
f.Improvement_surcharge=f.Improvement_surcharge_No 2.2167575  6.034515e-07
f.Extra=f.Extra_0 0.4013839  2.178649e-02
f.Extra=f.Extra_0.5 -0.3687408  2.834464e-02
lpep_pickup_date=2016-01-15 -1.8470697  1.300711e-02
f.Improvement_surcharge=f.Improvement_surcharge_YES -2.2167575  6.034515e-07
f.MTA_tax=f.MTA_tax_YES -2.7266309  1.410197e-09
Trip_type=f.TripType-Street-Hail -2.9773064  2.267247e-10
multiouts=FALSE -19.1771081  1.059920e-10
RateCodeID=Standard rate 6.0967688  4.707810e-45
f.tlenkm=f.tlenkm-[0,1] 7.5696279  1.311895e-61
f.Fare_amount=f.Fare_amount-(6.5,9] -4.8566275  7.049877e-62
AnyTip=AnyTip_No -2.7107302  3.371357e-78
Payment_type=f.PayType-Cash -3.5251722  4.458285e-79
AnyToll=AnyToll_Yes -12.7638515  1.505319e-122
f.espeed=f.espeed-(1,25] -3.3724891  3.584547e-170
f.distHaversine=f.distHaversine-(5,10] -1.4895434  5.052124e-214
f.Fare_amount=f.Fare_amount-[0,6.5] -7.8815863  7.882834e-289
f.distHaversine=f.distHaversine-[0,5] -15.2453734  0.000000e+00
f.traveltime=f.traveltime-(15,40] -1.6624443  0.000000e+00
f.traveltime=f.traveltime-[0,15] -15.2366082  0.000000e+00
f.tlenkm=f.tlenkm-(1,5] -3.9701060  0.000000e+00
```

### Associació global variables quantitatives

La distribució del Fare\_amount, juntament amb el profiling, ens ha permet conoure que la variable està correlacionada amb Total\_amount. També té sentit que la variable tlenkm estigui correlacionada, ja que el valor Fare\_amount es calcula a partir de la distància recorreguda pel taxi. Com podem veure el p-value és inferior a 0.05, el que provoca que hauríem de rebutjar la hipòtesis nul·la (no hi ha correlació).

### Associació global variables qualitatives

Com es pot veure, no hi trobem gran correlació entre les variables qualitatives amb la variable target Total\_amount. Només ens informa de que aquelles variables, creades

Com ja hem pogut veure la variable Fare\_amount estava correlacionada amb Total\_amount. Amb aquest profiling, podem extreure també que la categoria més correlacionada es Fare\_amount-(14.5,71.5). Això té sentit, ja que si trobem aquest valor en la variable f.Fare\_amount segurament trobarem un Total\_amount molt alt.

```

> res.cat$quanti.var
      Eta2      P-value
Tip_amount   0.52550476 0.000000e+00
Total_amount 0.067723891 3.371357e-78
Dropoff_longitude 0.041259489 1.032186e-47
Pickup_longitude 0.03939048 1.362968e-45
Dropoff_latitude 0.017667628 3.78345e-21
distHaversine 0.017181401 1.321598e-20
Fare_amount   0.016716278 4.369562e-20
Pickup_latitude 0.016346559 1.130280e-19
tlenkm       0.015338101 1.509292e-18
traveltime    0.004823040 8.857851e-07
espeed        0.004721182 1.155472e-06
Tolls_amount   0.003519621 2.696894e-05
> res.cat$quanti
$'AnyTip No'
      v.test Mean in category Overall mean sd in category Overall sd p.value
Dropoff_longitude 14.361622 -7.392603e+01 -73.93489422 4.946718e-02 0.000000e+00
Pickup_longitude 14.033027 -7.392857e+01 -73.93586026 4.305733e-02 0.04261069 9.788296e-47
Dropoff_latitude 9.397950 4.075110e+01 40.74449276 6.032266e-02 0.05767782 5.536322e-21
Pickup_latitude 9.397912 4.075271e+01 40.74645109 5.72185e-02 0.05689713 1.570844e-19
Tolls_amount   -4.194590 5.591573e-02 0.09183507 5.485642e-02 0.70251226 2.733656e-05
espeed        -4.858105 2.041876e+01 20.97100762 9.358228e-00 9.32574758 1.185144e-06
traveltime    -4.910232 1.198297e+01 13.00095521 1.760214e+01 17.00805101 9.096866e-07
tlenkm       -8.756436 4.087999e+00 4.58146199 4.398866e+00 4.63162831 2.015215e-18
Fare_amount   -9.141369 1.102509e+01 11.9993894 8.189565e+00 8.74051844 6.166654e-20
distHaversine -9.267676 2.869467e+00 3.21086536 2.748933e+00 3.62208213 1.992451e-20
Total_amount   -18.399775 1.223053e+01 14.54158184 8.562468e+00 10.30225574 1.319107e-75
Tip_amount     -51.256481 -1.437253e-05 1.23014878 7.697012e-04 1.96892443 0.000000e+00

$'AnyTip Yes'
      v.test Mean in category Overall mean sd in category Overall sd p.value
Tip_amount     51.256481 2.8883375 1.23014878 2.07738644 1.06892443 0.000000e+00
Total_amount   18.399775 17.651994 14.54158184 11.55210794 10.30225574 1.319107e-75
distHaversine 9.267676 3.6704961 3.20886536 2.29955331 2.02890213 1.992451e-20
Fare_amount   -9.141369 13.3111736 11.9993894 9.27136843 8.74051844 6.166654e-20
tlenkm       -8.756436 5.24703089 4.58146199 4.04077107 4.63162831 2.015215e-18
traveltime    4.910232 21.7145103 20.97100762 16.07227272 17.00805101 9.096866e-07
espeed        4.858105 13.00095521 9.358228e-00 9.32574758 1.185144e-06
Tolls_amount   0.1401939 0.09183507 0.86551269 0.70251226 2.733656e-05
Pickup_latitude 9.029712 48.7380328 40.74645109 0.05261794 0.05767782 5.536322e-21
Dropoff_latitude -9.397950 40.74645109 40.74449276 0.05261794 0.05767782 5.536322e-21
Pickup_longitude -14.033027 -73.93489422 40.74449276 0.03995419 0.04261069 9.788296e-45
Dropoff_longitude -14.361622 -73.93489422 0.04078992 0.05066112 9.099363e-05

```

## Profiling AnyTip

### Associació global variables quantitatives

Com podem veure, la mitjana de la variable Tip\_amount és diferent entre els AnyTip amb valor True i AnyTip amb valor False. Lògicament és correcte, ja que els passatgers que no han donat propina sempre tenen el total de propina amb valor zero. En definitiva, la variable Tip\_amount influeix sobre la variable AnyTip.

### Associació global variables categòriques

Per la gent que ha donat propina, AnyTip Yes, la mitjana de Tip\_amount està per damunt de la mitjana global. Això té sentit ja que el valor Mean in category, contabilitza tots Tip\_amount amb valor diferent de zero.

Per la gent que ha donat propina, la mitjana de Total\_amount es superior també respecte a la mitjana global. Respecte a les persones que no han donat propina, podem veure que la mitjana del Tolls\_amount es superior a la global. Tots els arguments mencionats funcionen de manera podem explicar-se de manera complementària, ja que AnyTip és una variable binària. Per exemple, les persones que no han donat propina estan per sota de la mitjana global per un 51.25%.

### Profiling de les categories

```

> res.cat$test.chi2
      p.value df
Payment_type 0.0000000e+00 2
f.Fare_amount 6.129444e-24 3
f.tlenkm 1.087294e-19 2
f.traveltime 1.429469e-17 2
f.distHaversine 7.945740e-11 2
f.Improvement_surcharge 3.570885e-09 1
f.MTA_tax 4.542332e-09 1
Trip_type 5.576360e-08 1
RateCodeID 3.777248e-06 1
lpep_pickup_period 5.610059e-06 3
AnyToll 3.502094e-05 1
f.espeed 1.425728e-04 2
VendorID 5.925000e-03 1
lpep_pickup_date 1.450254e-02 30
f.Extra 3.273592e-02 2
f.Payment_type=f.PayType-Credit card
f.tlenkm=f.tlenkm[5,67,9]
f.traveltime=f.traveltime[15,40]
f.Fare_amount=f.Fare_amount[14.5,71.5]
f.Improvement_surcharge=f.Improvement_surcharge_Yes
f.MTA_tax=f.MTA_tax_YEE
Trip_type=f.TripType-Street-Hail
f.distHaversine=f.distHaversine[5,10]
RateCodeID=f.RateCodeID
f.espeed=f.espeed[25,130]
AnyToll=AnyToll_No
f.distHaversine=f.distHaversine[10,26.1]
f.Fare_amount=f.Fare_amount[9,14.5]
f.Extra=f.Extra[0]
lpep_pickup_date=2016-01-30
lpep_pickup_period=Period morning
f.Extra=f.Extra[0.5]
f.traveltime=f.traveltime[40,548]
lpep_pickup_date=2016-01-21
lpep_pickup_date=2016-01-29
f.Passenger_count=f.Passenger_count[2
lpep_pickup_date=2016-01-22
f.Extra=f.Extra[0]
f.Extra=f.Extra[0]
f.Fare_amount=f.Fare_amount[(-6.5,9]
VendorID=VendorID_VerForamt
lpep_pickup_date=2016-01-01
AnyToll=AnyToll_Y
f.espeed=f.espeed[1,25]
f.tlenkm=f.tlenkm[1,5]
RateCodeID=0Others
23.77622 1.595495 2.86 2.021134e-02 4.751300
36.68731 22.243078 25.84 4.946405e-07 5.023900
39.05497 19.005161 20.74 8.983861e-03 2.612698
41.05806 17.605161 22.04 1.019565e-03 2.744956
41.27718 9.331769 10.26 1.021134e-02 4.751300
42.23489 97.377126 98.26 4.264542e-05 4.092659
41.05806 74.659784 77.59 3.615147e-05 4.130974
40.83226 58.235574 62.09 3.269828e-06 4.719898
23.77622 1.595495 2.86 2.021134e-02 4.751300
36.68731 22.243078 25.84 4.946405e-07 5.023900
0.00000 0.000000 0.58 9.00930e-08 5.336036
Trip_type=f.TripType-Dispatch
f.tlenkm=f.tlenkm[0,1]
f.MTA_tax=f.MTA_tax_NO
f.Improvement_surcharge=f.Improvement_surcharge_No
f.distHaversine=f.distHaversine[0,5]
f.Fare_amount=f.Fare_amount[0,6,5]
f.traveltime=f.traveltime[0,15]
Payment_type=f.PayType-Cash
0.00000 0.000000 50.14 0.000000e+00 -Inf

```

```

> res.cat$category
$ AnyTip_No

Payment_type=f_PayType-Cash          Cle/Mod    Mod/Cle Global      p-value      v-test
100.00000 87.385363 86.14 0.000000e+00      Inf
f.traveltime=f.traveltime-[0,15]      61.28399 76.193796 71.34 1.723608e-18 8.774045
f.Fare_amount=[0,6,5]                66.32652 33.083067 29.49 9.327654e-17 8.313946
f.distHaversine=f.distHaversine-[0,5] 59.62246 84.768212 81.58 1.928576e-11 6.711247
f.Improvement_surcharge=f.Improvement_surcharge_NO 81.88496 3.938655 2.76 6.745460e-10 6.171946
f.MTA_tax=f.MTA_tax_NO              82.08955 3.834089 2.68 8.573292e-10 6.133932
f.tlenkm=f.tlenkm-[0,1]             70.43269 10.212618 8.32 1.103123e-08 5.714060
Trip_type=f.TripType-Dispatch       81.30081 3.485355 2.46 1.472744e-08 5.664769
Payment_type=f_PayType-Others      100.00000 1.010805 0.58 9.500930e-08 5.336036
lpep_pickup_period=Period valley   63.31269 28.511677 25.84 4.946405e-07 5.028388
RateCodeID=Others                  76.22378 3.799233 2.86 2.021134e-06 4.751308
f.tlenkm=f.tlenkm-(1,5]            66.96774 64.710996 62.00 3.368100e-06 4.711908
f.lespeed=f.lespeed-[0,1,25]        88.94700 2.899999 77.14 1.928576e-10 4.070913
AnyToll=AnyToll_No                  57.72551 98.919484 98.26 4.264542e-05 4.092659
lpep_pickup_date=2016-01-01         68.72447 5.437435 4.54 3.443356e-04 3.579422
VendorID=f.Vendor-Verifone          58.40936 79.330777 77.94 6.051986e-03 2.744956
f.Fare_amount=f.Fare_amount-[6,5,9] 60.94503 22.028581 28.74 8.983061e-03 2.612698
f.Extra=f.Extra-0                  59.28178 47.751839 46.22 1.171496e-02 2.520616
lpep_pickup_date=2016-01-22         64.17918 4.496340 4.02 4.587826e-02 1.996512
f.Passenger_count=f.Passenger_count-2 52.31608 6.692227 7.34 4.246928e-02 -2.028981
lpep_pickup_date=2016-01-29         49.71098 2.997560 3.46 3.928834e-02 -2.061154
lpep_pickup_date=2016-01-21         49.71429 3.032415 3.50 3.820802e-02 -2.072616
f.traveltime=f.traveltime-[40,548]   47.27273 1.812478 2.20 3.182030e-02 -2.146661
f.Extra=f.Extra-0,5                 55.31465 36.458696 37.82 2.139829e-02 -2.300882
lpep_pickup_time=Period morning    52.721829 11.178112 18.14 4.264542e-02 -2.476539
lpep_pickup_date=2016-01-30         49.57473 4.112000 4.10 1.326216e-02 -2.476531
VendorID=f.Vendor-Mobile           53.76247 20.669232 22.06 6.051986e-03 -2.474916
f.Fare_amount=f.Fare_amount-[9,14,5] 53.88273 23.701638 25.24 3.743941e-03 -2.898967
f.distHaversine=f.distHaversine-[10,26,1] 44.32996 2.997560 3.88 2.031444e-04 3.715924
AnyToll=AnyToll_No                  35.63218 1.088516 1.74 4.264542e-05 -4.092659
f.lespeed=f.lespeed-[25,130]        51.92136 20.250959 22.38 2.059530e-05 -4.176558
RateCodeID=Standard rate           56.82528 96.200767 97.14 2.021134e-06 -4.751308
f.distHaversine=f.distHaversine-[5,10] 48.28661 12.234228 14.54 9.498968e-08 -5.336058
Trip_type=f.TripType-Street-Hail  56.77671 96.514465 97.54 1.472744e-08 -5.664769
f.MTA_tax=f.MTA_tax_YES            56.69956 96.165911 97.32 8.573292e-10 -6.133932
f.Improvement_surcharge=f.Improvement_surcharge_YES 56.68449 96.061345 97.24 6.745460e-10 -6.171946
f.Fare_amount=f.Fare_amount-[14,5,71,5] 47.27864 20.285914 24.62 2.082215e-16 -8.272582
f.traveltime=f.traveltime-[15,40]    47.69463 21.993726 26.46 1.310886e-16 -8.272582
f.tlenkm=f.tlenkm-[5,67,9]          48.31536 24.991286 29.68 4.817933e-17 -8.391056
Payment_type=f.Paytype-Credit card 13.51461 11.666832 49.28 0.000000e+00 -Inf

```

Un cop realitzat el profiling de les categories podem treure les següents conclusions:

- La variable categòrica que més caracteritza al factor AnyTip es la variable Payment\_type. Això és degut a que aquesta informació, tal com s'explica en la descripció de les variables, s'omplia automàticament per la targeta de crèdit. Respecte als pagaments de propines amb efectiu, no s'inclou.
- El 61.28% de les persones que viatgen entre 0 i 15 minuts no han donat propina i el 76.19% de les persones que no han donat propina han realitzat un viatge d'entre 0 i 15 minuts.
- El 51.68% de les persones que han realitzat un viatge d'entre 5 i 67.9 km han donat propina. El 35.99% de les persones que han donat propina han realitzat un viatge d'entre 5 i 67.9 km. El 29.69% de les persones ha realitzat un viatge d'entre 5 i 69.9 km.

# Deliverable II

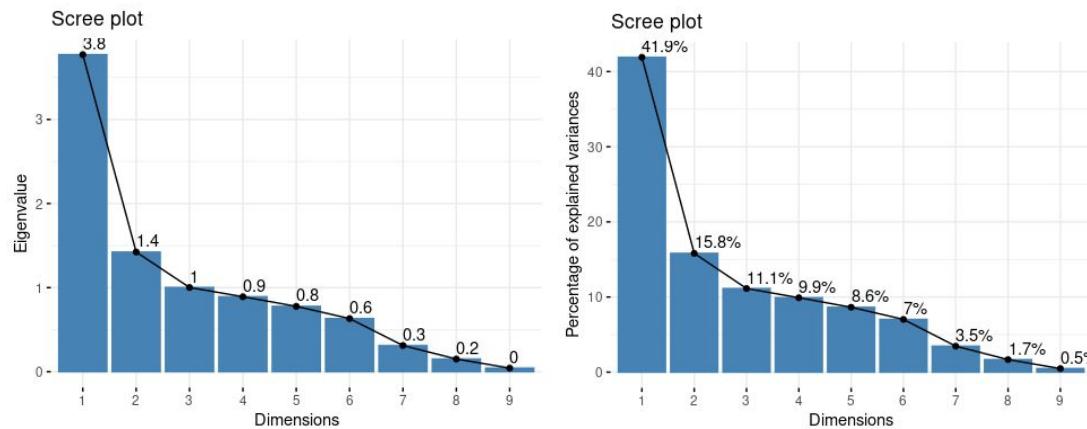
Carles Capilla Cànovas  
Jesús Molina Roldán

|   |           |
|---|-----------|
| <b>Principal Component Analysis (PCA)</b>                 | <b>4</b>  |
| Eigenvalues and dominant axes analysis.                   | 4         |
| Quality of representation                                 | 4         |
| Contribution  | 5         |
| Interpreting the axes                                     | 6         |
| Active numerical variables                                | 6         |
| Supplementary numerical variables                         | 7         |
| Supplementary categorical variables                       | 7         |
| Individuals   | 8         |
| Supplementary individuals                                 | 9         |
| <b>HCPC</b>   | <b>10</b> |
| Description of the clusters by the variables              | 10        |
| Categorical variables which characterizes the clusters    | 10        |
| Description of each cluster by the categories             | 10        |
| Quantitative variables which characterizes the clusters   | 14        |
| Description of each cluster by the quantitative variables | 15        |
| Description of the clusters by the individuals            | 17        |
| Characteristic individuals                                | 17        |
| Hierarchical tree result                                  | 19        |
| Ratio between within inertias                             | 19        |
| Inertia gain  | 19        |
| Partition quality   | 19        |
| <b>K-Means: Partitioning in k=6</b>                       | <b>20</b> |
| Profiling KM  | 20        |
| Global association variables numeric                      | 20        |
| Global association category categoricas                   | 23        |
| Confusion Table   | 25        |

|  |           |
|--|-----------|
| <b>Correspondence Analysis (CA)</b>  | <b>26</b> |
| CA in Total amount and Pick up period  | 26        |
| CA in Total amount and Travel time   | 27        |
| <b>MCA analysis</b>  | <b>28</b> |
| Quality of representation  | 28        |
| Contribution   | 29        |
| Eigenvalues and dominant axes analysis.  | 31        |
| Individuals  | 32        |
| Categorical variables, supplementary numerical variables and supplementary categorical variables | 32        |
| <b>Hierarchical Clustering (from MCA)</b>  | <b>33</b> |
| Categorical variables which characterizes the clusters   | 34        |
| Numerical variables which characterizes the clusters   | 34        |
| Description of each cluster by the categories  | 35        |
| Description of the clusters by the individuals   | 36        |
| Hierarchical tree result   | 37        |
| Ratio between within inertias  | 37        |
| Inertia gain   | 37        |
| Partition quality  | 37        |

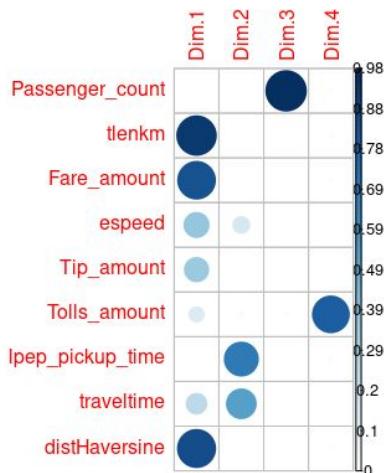
# Principal Component Analysis (PCA)

## Eigenvalues and dominant axes analysis.



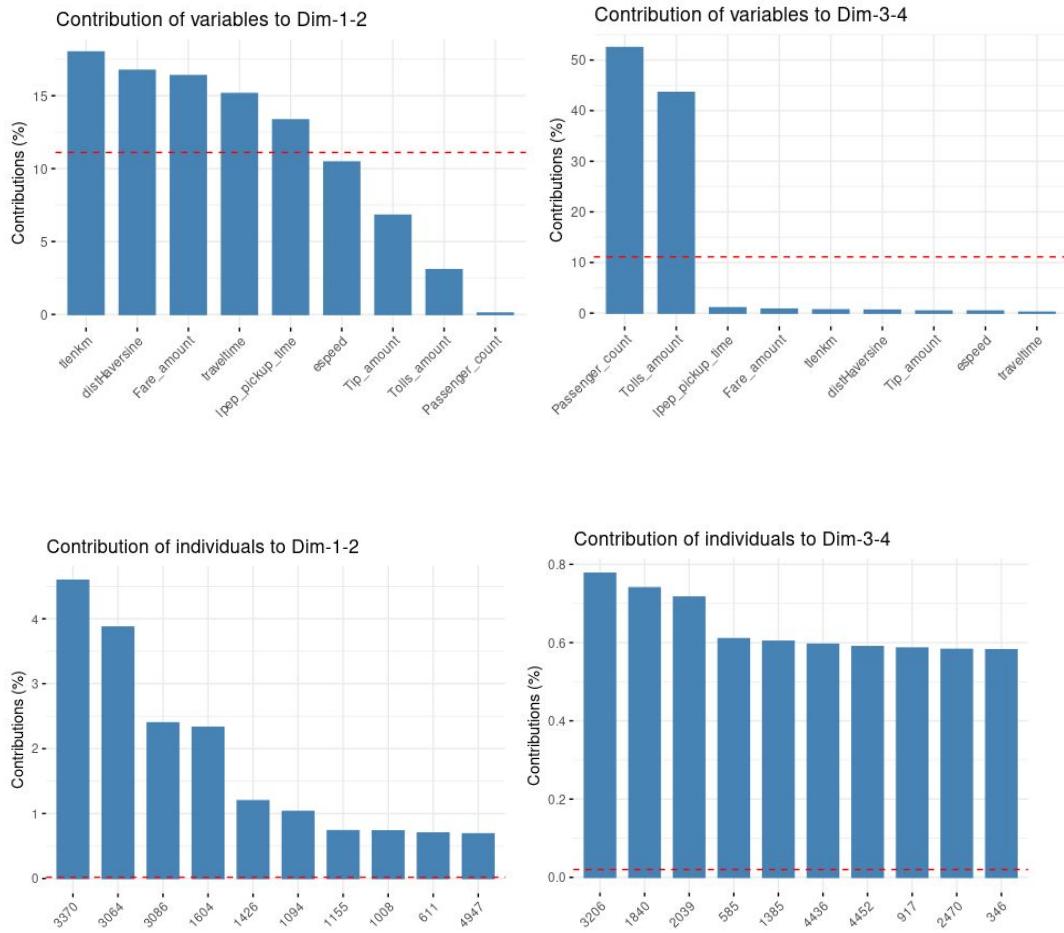
En esta imagen podemos ver los valores propios. Como podemos ver hasta la tercera dimensión tenemos un valor propio igual a 1. Según el criterio de Kaiser deberíamos eliminar todas las componentes con valor propio por debajo de 1, lo que significa que deberíamos coger hasta la tercera dimensión. Según la regla de Elbow, debemos coger hasta que no haya un descenso significativo, lo que significa que también se debería coger hasta la tercera dimensión. A pesar de todo, hemos decidido incluir hasta la cuarta dimensión, ya que nos facilita el estudio. Como podemos ver hasta la cuarta dimensión encontramos una varianza acumulada del 78.75%. También podemos admirar como la primera dimensión contribuye mucho en el PCA, explicando un 41.9% de la varianza.

## Quality of representation



A partir de la suma del coseno al cuadrado de la primera dimensión más el coseno al cuadrado de la segunda dimensión podemos obtener calidad de las variables en el primer plano factorial. Como podemos ver tlenkm y distHaversine son las dos variables que mejor se representan en la primera dimensión.

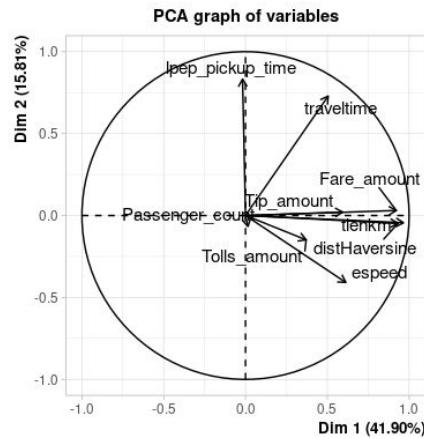
# Contribution



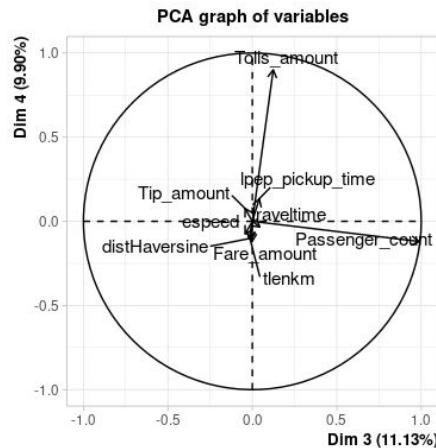
En este plot encontramos las 10 variables que contribuyen más en el primer plano factorial. Como podemos ver la variable tlenkm es la variable que contribuye más junto con la variable distHaversine. Si vemos el segundo plano factorial, vemos que el número total de pasajeros y el número total de pagos por peajes influyen bastante en el segundo plano factorial. En este plot encontramos los 10 individuos que contribuyen más en el primer plano factorial. Como podemos ver el individuo 3370 es el individuo que contribuye más, junto con el individuo 3064 al primer plano factorial. Si vemos el segundo plano factorial, vemos que el individuo 3206 y el individuo 1840, son los individuos que contribuyen más el segundo plano factorial.

## Interpreting the axes

### Active numerical variables

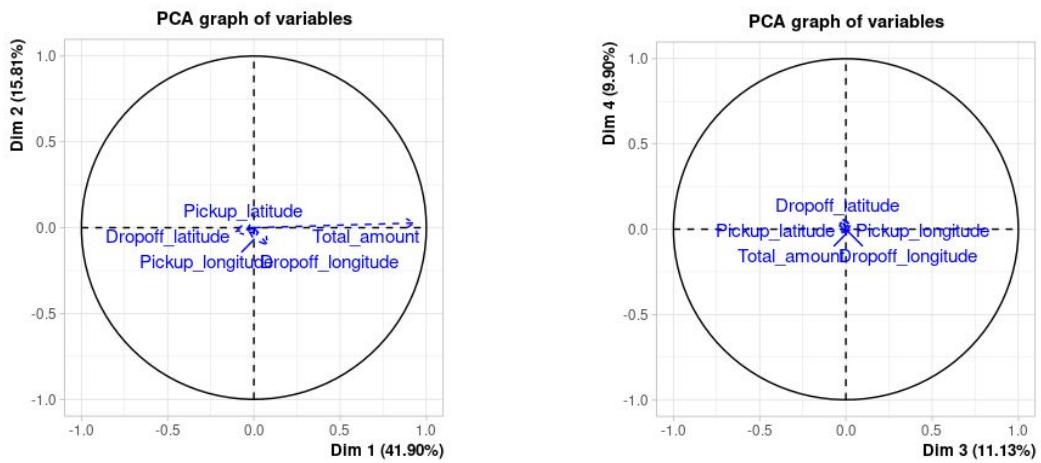


Si vemos el primer plano factorial podemos ver que la variable mejor representadas en la primera dimensión es la cantidad de km recorridos en el taxi y la tarifa pagada. Como podemos ver las variables tlenkm, Tip\_amount y Fare\_amount están agrupadas, lo que significa que están positivamente correlacionadas.



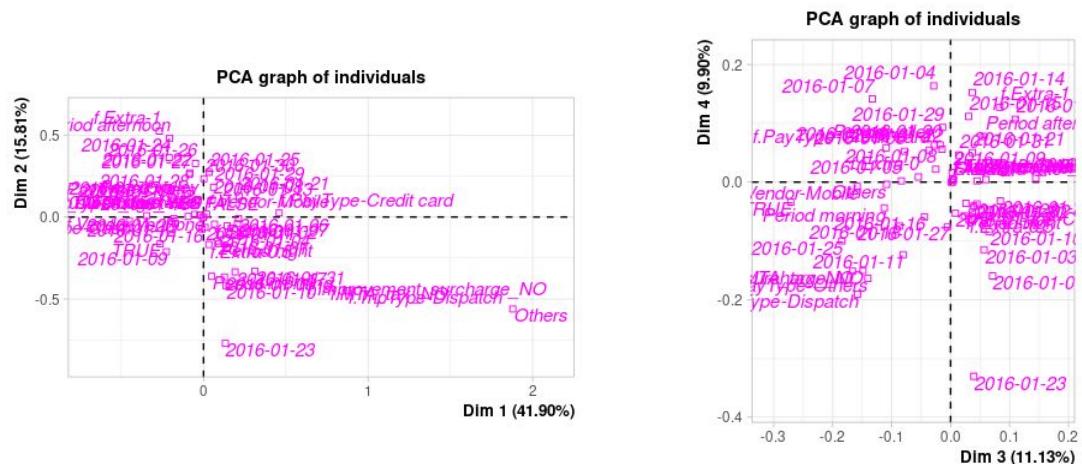
Si observamos el segundo plano factorial, podemos de nuevo que la variable Passenger contribuye mucho en la dimensión 3 y que la variable Tolls\_amount contribuye mucho en la dimensión 4.

## Supplementary numerical variables



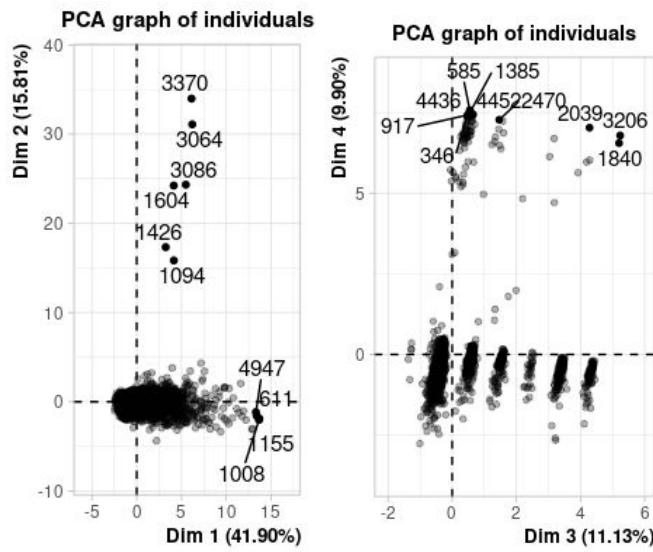
Como podemos ver, la única variable numérica complementaria que contribuye en la construcción del primer plano factorial es Total\_amount.

## Supplementary categorical variables



En estas dos imágenes nos encontramos las variables categóricas suplementarias situadas en las dos dimensiones más importantes. Si nos fijamos, los viajes realizados el día 23 de enero del 2016 se encuentran siempre alejados del conjunto. Las categorías son difíciles de distinguir pero si nos fijamos, en ambos planos factoriales los individuos con f.Extra-1 están cerca de los individuos con Period afternoon.

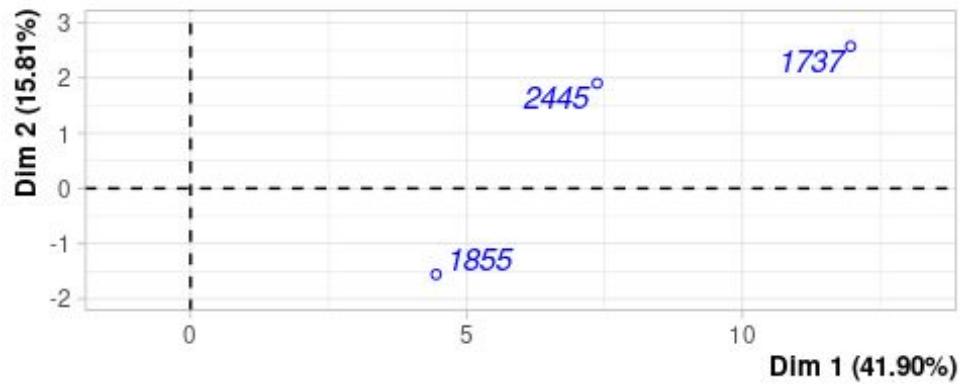
## Individuals



Si analizamos el primer plano factorial vemos que los individuos que contribuyen más en el plano forma dos grupos. El primer grupo está formado por los individuos 4947, 611, 1155 y 1008. El segundo grupo está formado por los individuos 3064, 3086, 1604, 1426 y 1094. Si vemos en el dataset las características de dichos individuos encontramos que son viajes de taxis con un largo recorrido en km, más de 30 km, tal como se puede analizar con el plano factorial de las variables activas numéricas. Si vemos los individuos que han contribuido al segundo plano factorial, vemos que son individuos con que han pagado 5.54 dólares en peajes.

## Supplementary individuals

PCA graph of individuals

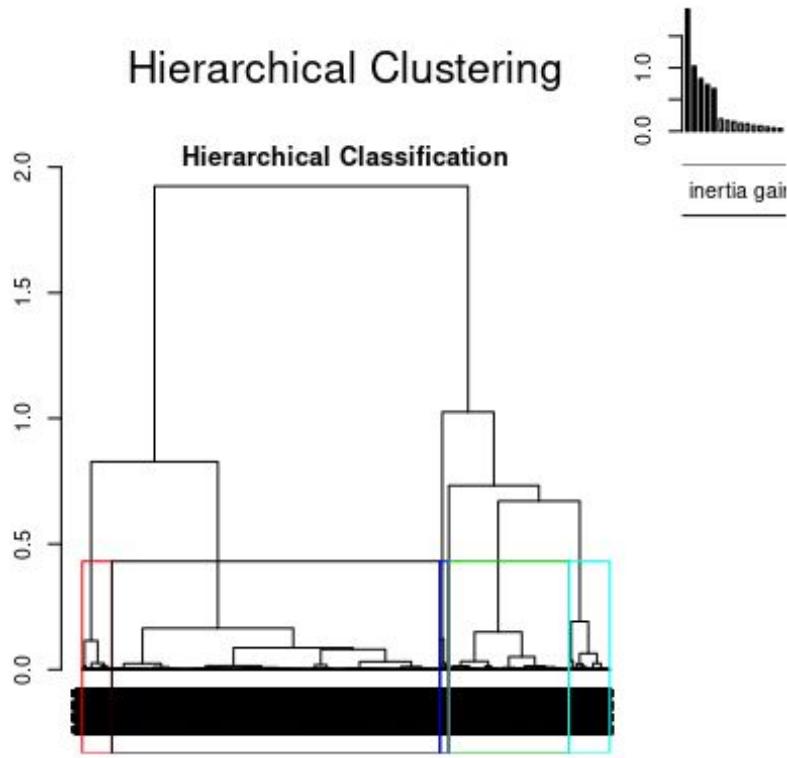


```
> df[c("1855", "2445", "1737"), ]
  VendorID Payment_type Store_and_fwd_flag RateCodeID f.Extra f.MTA_tax f.Improvement_surcharge
1855 f.Vendor-VeriFone f.PayType-Credit card FALSE Others f.Extra-0 f.MTA_tax_NO f.Improvement_surcharge_NO
2445 f.Vendor-VeriFone f.PayType-Cash FALSE Others f.Extra-0 f.MTA_tax_NO f.Improvement_surcharge_NO
1737 f.Vendor-Mobile f.PayType-Others FALSE Standard rate f.Extra-0 f.MTA_tax_YES f.Improvement_surcharge_YES
lpep_pickup_period Trip_type lpep_pickup_date multiopts f.espeed f.tlenkm f.traveltime
1855 Period morning f.TripType-Dispatch 2016-01-12 TRUE f.espeed-(25,130] f.tlenkm-[0,1] f.traveltime-[0,10]
2445 Period night f.TripType-Dispatch 2016-01-15 TRUE f.espeed-(25,130] f.tlenkm-(5,67.9] f.traveltime-(40,548]
1737 Period valley f.TripType-Street-Hail 2016-01-11 TRUE f.espeed-(25,130] f.tlenkm-(5,67.9] f.traveltime-(40,548]
f.distHaversine AnyToll f.Fare_amount f.Passenger_count f.Total_amount Passenger_count tlenkm
1855 f.distHaversine-[0,5] AnyToll Yes f.Fare_amount-(14.5,71.5] f.Passenger_count-2 f.Total_amount-(40,95.5] 2 0.0804672
2445 f.distHaversine-[0,5] AnyToll Yes f.Fare_amount-(9,14.5] f.Passenger_count-Others f.Total_amount-(8,11] 3 56.0856400
1737 f.distHaversine-(5,10] AnyToll No f.Fare_amount-(14.5,71.5] f.Passenger_count-1 f.Total_amount-(40,95.5] 1 67.9143168
  Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Fare_amount espeed Tip_amount Tolls_amount lpep_pickup_time
1855 -74.08338 40.64220 -74.08366 40.64284 65.00000 48.28032 13 0.00000 9
2445 -74.00069 40.59997 -73.96132 40.62992 10.00000 40.04528 0 0.00000 20
1737 -73.86628 40.87276 -73.86414 40.82361 61.50166 37.35547 0 0.22224 15
  traveltime distHaversine AnyTip Total_amount hcpck clkKM
1855 0.10000 0.07562334 AnyTip Yes 78.00000 KHP- 4 kKM-1
2445 84.03333 4.70946300 AnyTip No 10.00000 KHP- 4 kKM-3
1737 109.08333 5.47499787 AnyTip No 70.61709 KHP- 4 kKM-3

> res.pca$ind$contrib[c("1855", "2445", "1737"), ]
    Dim.1   Dim.2   Dim.3   Dim.4
1855 0.1058658 0.03712658 0.0003443667 0.006059698
2445 0.2744526 0.05496966 0.0474553876 0.083996708
1737 0.7316331 0.09785061 0.0016757330 0.152349937
```

Si vemos los individuos suplementarios observamos que el individuo 1737 es un individuo que contribuye más en el primer plano factorial. En contraposición el individuo que contribuye más al segundo plano factorial es el 2445. Si observamos las características de dichos individuos vemos que la distancia recorrida de ambos viajes es superior a 50 km.

## HCPC



Viendo la inertia gain (pérdida importante de ir entre n clusters a n+1 clusters) y aplicando Kaiser Rule podemos ver que el número de clusters óptimo es 6.

## Description of the clusters by the variables

### Categorical variables which characterizes the clusters

```
##          p.value df
## Payment_type    8.703621e-41 10
## RateCodeID      7.823482e-15  5
## VendorID       9.081168e-11  5
## f.Extra         7.914188e-08 10
## lpep_pickup_period 5.249054e-06 15
## Trip_type       9.975894e-06  5
## f.MTA_tax        2.966099e-05  5
## f.Improvement_surcharge 2.904864e-04  5
```

En esta imagen podemos encontrar las variables categóricas que contribuyen más en los clusters ordenadas por importancia. Podemos ver que las variables categóricas Payment\_type, RateCodeID y VendorID son muy importantes en la construcción de los clusters.

## Description of each cluster by the categories

|   | Cla/Mod  | Mod/Cla    | Global | p.value      | v.test     |
|---|----------|------------|--------|--------------|------------|
| Payment_type=f.PayType-Cash                         | 70.64220 | 56.817453  | 50.14  | 3.893540e-34 | 12.181644  |
| RateCodeID=Standard rate                            | 62.94009 | 98.075072  | 97.14  | 6.026448e-07 | 4.990368   |
| f.Extra=f.Extra-1                                   | 68.17043 | 17.452679  | 15.96  | 1.855149e-04 | 3.737966   |
| lpep_pickup_period=Period afternoon                 | 67.31813 | 19.890921  | 18.42  | 5.129457e-04 | 3.473903   |
| Trip_type=f.TripType-Street-Hail                    | 62.64097 | 98.010908  | 97.54  | 6.556979e-03 | 2.718545   |
| lpep_pickup_date=2016-01-05                         | 71.53285 | 3.144049   | 2.74   | 2.274727e-02 | 2.277653   |
| f.MTA_tax=f.MTA_tax_YES                             | 62.57707 | 97.690087  | 97.32  | 3.959358e-02 | 2.057964   |
| lpep_pickup_date=2016-01-19                         | 70.13889 | 3.240295   | 2.88   | 4.822322e-02 | 1.975396   |
| lpep_pickup_date=2016-01-31                         | 55.31915 | 3.336542   | 3.76   | 4.502308e-02 | -2.004439  |
| f.MTA_tax=f.MTA_tax_NO                              | 53.73134 | 2.309913   | 2.68   | 3.959358e-02 | -2.057964  |
| lpep_pickup_date=2016-01-17                         | 54.16667 | 2.919474   | 3.36   | 2.799497e-02 | -2.197357  |
| Trip_type=f.TripType-Dispatch                       | 50.40650 | 1.989092   | 2.46   | 6.556979e-03 | -2.718545  |
| f.Extra=f.Extra-0.5                                 | 59.43945 | 36.060314  | 37.82  | 9.855191e-04 | -3.294628  |
| lpep_pickup_period=Period night                     | 58.53211 | 40.936798  | 43.60  | 1.065019e-06 | -4.879228  |
| RateCodeID=Others                                   | 41.95804 | 1.924928   | 2.86   | 6.026448e-07 | -4.990368  |
| Payment_type=f.PayType-Credit card                  | 53.89610 | 42.605069  | 49.28  | 4.274103e-34 | -12.174037 |
| <b>\$`2`</b>  |          |            |        |              |            |
|   | Cla/Mod  | Mod/Cla    | Global | p.value      | v.test     |
| VendorID=f.Vendor-VeriFone                          | 6.954067 | 94.7552448 | 77.94  | 8.701382e-16 | 8.043914   |
| f.Extra=f.Extra-0.5                                 | 6.980434 | 46.1538462 | 37.82  | 3.089559e-03 | 2.958684   |
| lpep_pickup_period=Period night                     | 6.697248 | 51.0489510 | 43.60  | 9.268990e-03 | 2.601970   |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 5.841218 | 99.3006993 | 97.24  | 1.470478e-02 | 2.439569   |
| Payment_type=f.PayType-Cash                         | 6.501795 | 56.9930070 | 50.14  | 1.703725e-02 | 2.385903   |
| f.MTA_tax=f.MTA_tax_YES                             | 5.836416 | 99.3006993 | 97.32  | 1.784021e-02 | 2.368918   |
| Trip_type=f.TripType-Street-Hail                    | 5.823252 | 99.3006993 | 97.54  | 3.013982e-02 | 2.168248   |
| f.Extra=f.Extra-1                                   | 7.393484 | 20.6293706 | 15.96  | 3.102468e-02 | 2.156756   |
| RateCodeID=Standard rate                            | 5.826642 | 98.9510490 | 97.14  | 4.158928e-02 | 2.037607   |
| lpep_pickup_date=2016-01-07                         | 2.142857 | 1.0489510  | 2.80   | 4.729447e-02 | -1.983654  |
| RateCodeID=Others                                   | 2.097902 | 1.0489510  | 2.86   | 4.158928e-02 | -2.037607  |
| lpep_pickup_period=Period morning                   | 3.953871 | 8.3916084  | 12.14  | 3.913031e-02 | -2.062814  |
| Trip_type=f.TripType-Dispatch                       | 1.626016 | 0.6993007  | 2.46   | 3.013982e-02 | -2.168248  |
| lpep_pickup_date=2016-01-25                         | 1.986755 | 1.0489510  | 3.02   | 2.938094e-02 | -2.178337  |
| Payment_type=f.PayType-Credit card                  | 4.951299 | 42.6573427 | 49.28  | 2.107704e-02 | -2.306602  |
| f.MTA_tax=f.MTA_tax_NO                              | 1.492537 | 0.6993007  | 2.68   | 1.784021e-02 | -2.368918  |
| lpep_pickup_period=Period valley                    | 4.411765 | 19.9300699 | 25.84  | 1.654204e-02 | -2.396732  |
| f.Improvement_surcharge=f.Improvement_surcharge_NO  | 1.449275 | 0.6993007  | 2.76   | 1.470478e-02 | -2.439569  |
| f.Extra=f.Extra-0                                   | 4.110775 | 33.2167832 | 46.22  | 4.527082e-06 | -4.585589  |
| VendorID=f.Vendor-Mobile                            | 1.359927 | 5.2447552  | 22.06  | 8.701382e-16 | -8.043914  |
| <b>\$`3`</b>  |          |            |        |              |            |
|   | Cla/Mod  | Mod/Cla    | Global | p.value      | v.test     |
| Payment_type=f.PayType-Credit card                  | 29.82955 | 60.693642  | 49.28  | 5.929786e-20 | 9.145603   |
| lpep_pickup_period=Period night                     | 26.60550 | 47.894302  | 43.60  | 5.533695e-04 | 3.453497   |
| f.Extra=f.Extra-0.5                                 | 26.22951 | 40.957886  | 37.82  | 9.942040e-03 | 2.577839   |
| lpep_pickup_date=2016-01-01                         | 29.95595 | 5.615194   | 4.54   | 4.304996e-02 | 2.023225   |
| RateCodeID=Others                                   | 31.46853 | 3.715937   | 2.86   | 4.549479e-02 | 2.000051   |
| lpep_pickup_date=2016-01-19                         | 17.36111 | 2.064410   | 2.88   | 4.637866e-02 | -1.991931  |
| RateCodeID=Standard rate                            | 24.00659 | 96.284063  | 97.14  | 4.549479e-02 | -2.000051  |
| lpep_pickup_date=2016-01-12                         | 16.99346 | 2.146986   | 3.06   | 2.980772e-02 | -2.172636  |
| lpep_pickup_date=2016-01-14                         | 16.86747 | 2.312139   | 3.32   | 2.089799e-02 | -2.309822  |
| lpep_pickup_period=Period afternoon                 | 20.52117 | 15.606936  | 18.42  | 3.327487e-03 | -2.935744  |
| f.Extra=f.Extra-1                                   | 18.67168 | 12.303881  | 15.96  | 4.570701e-05 | -4.076559  |
| Payment_type=f.PayType-Cash                         | 18.74751 | 38.810900  | 50.14  | 1.065964e-19 | -9.082001  |

```

$`4`  

Payment_type=f.PayType-Credit card           Cla/Mod   Mod/Cla Global      p.value    v.test  

RateCodeID=Others                           8.847403  72.909699 49.28 1.216693e-17 8.551338  

Trip_type=f.TripType-Dispatch              17.482517  8.361204 2.86 1.043849e-06 4.883187  

f.MTA_tax=f.MTA_tax_NO                    16.260163  6.688963 2.46 4.211775e-05 4.095544  

f.Improvement_surcharge=f.Improvement_surcharge_NO 15.671642  7.023411 2.68 4.779382e-05 4.066161  

f.Improvement_surcharge=f.Improvement_surcharge_NO 14.492754  6.688963 2.76 2.297538e-04 3.683832  

f.Extra=f.Extra-1                         4.260652  11.371237 15.96 2.134395e-02 -2.301844  

f.Improvement_surcharge=f.Improvement_surcharge_YES 5.738379  93.311037 97.24 2.297538e-04 -3.683832  

f.MTA_tax=f.MTA_tax_YES                  5.713111  92.976589 97.32 4.779382e-05 -4.066161  

Trip_type=f.TripType-Street-Hail          5.720730  93.311037 97.54 4.211775e-05 -4.095544  

RateCodeID=Standard rate                 5.641342  91.638796 97.14 1.043849e-06 -4.883187  

Payment_type=f.PayType-Cash               3.071400  25.752508 50.14 8.539854e-19 -8.852737  

$`5`  

lpep_pickup_date=2016-01-01           Cla/Mod   Mod/Cla Global      p.value    v.test  

lpep_pickup_period=Period night       1.3215859  50.00000 4.54 0.001725940 3.133739  

lpep_pickup_date=2016-01-30           0.8403361  33.33333 4.76 0.031721963 2.147897  

$`6`  

Payment_type=f.PayType-Credit card           Cla/Mod   Mod/Cla Global      p.value    v.test  

RateCodeID=Others                       2.3944805  72.839506 49.28 1.580726e-05 4.317128  

f.Extra=f.Extra-0                      6.9930070  12.345679 2.86 1.039172e-04 3.881259  

f.Extra=f.Extra-0                      2.1202942  60.493827 46.22 9.859179e-03 2.580730  

lpep_pickup_date=2016-01-04             4.0816327  7.407407 2.94 4.036183e-02 2.050027  

lpep_pickup_date=2016-01-31             3.7234043  8.641975 3.76 4.227456e-02 2.030807  

lpep_pickup_period=Period valley       2.2445820  35.802469 25.84 4.573573e-02 1.997824  

f.Extra=f.Extra-0.5                   1.1105235  25.925926 37.82 2.397750e-02 -2.257489  

RateCodeID=Standard rate              1.4618077  87.654321 97.14 1.039172e-04 -3.881259  

Payment_type=f.PayType-Cash            0.8775429  27.160494 50.14 2.509186e-05 -4.213972

```

## Cluster 1

Por lo que hace a las categorías pertenecientes a las variables de tipo factor pertenecientes al cluster 1, y dado el tratamiento que hemos aplicado a nuestros datos, podemos observar como las dos categorías más características de este serían, por un lado el pago mediante efectivo. Este pago en efectivo tiene una representación global de un 50.15% cuando en el cluster 1 esta es de 56.78% por lo que podemos decir que esta categoría aparece sobrerepresentada en este cluster en concreto. Además, vemos como un 70.63% de todas las observaciones donde el pago es en efectivo, aparecen en el cluster 1, esto también podría justificarse por el gran número de observaciones pertenecientes a este cluster en comparación a los demás.

Debido a esta sobrerepresentación, podemos ver que, por el contrario, el pago con tarjeta está infrarepresentado pasando de una representación global del 49.29% al 42.64% dentro de este cluster. Y además, tenemos que del total de observaciones donde el pago se realiza mediante tarjeta de crédito, cerca de un 54% están en este cluster.

No solo son estas las categorías caracterizadas por este cluster ya que todas las que aparecen en el output de res.hcpc *desc.var* category, aun así, son las más destacables.

## Cluster 2

Para el cluster 2, la categoría que podemos destacar debido a su sobrerepresentación es la del VendorID = Verizone. Esta categoría representa un 77.95% de las observaciones globales cuando en este cluster número 2 estas observaciones ascienden a un 94.76%. Por el contrario, seguramente debido a las pocas observaciones pertenecientes a este cluster, las

pertenecientes al cluster 2 con VendorID = Verizone solo representan cerca del 7% de las observaciones totales con este VendorID.

Por otro lado, como categoría infrarepresentada, como es obvio ya que se trata de un factor binario, tendríamos la perteneciente a las observaciones donde VendorID = Mobile que representan el complementario en cuanto a observaciones globales, un 22.05% así como a las pertenecientes al propio cluster, un 5.24%. Además podemos ver como las observaciones de esta categoría pertenecientes a este cluster, solo suponen un 1.36% de las observaciones totales.

### **Cluster 3**

A diferencia del cluster 1, esta vez tenemos sobrerepresentada la categoría de pago mediante tarjeta de crédito. Esta presenta como ya hemos dicho un 49.29% de las observaciones globales y, en cambio, en este cluster ascienden hasta un 60.66% suponiendo un 29.8% las observaciones pertenecientes a este cluster del total de observaciones donde la tarjeta de crédito aparece como método de pago.

Por otro lado y como podemos suponer, tendremos como categoría infrarrepresentada la que denomina los pagos en efectivo de un 38.84% de observaciones dentro del cluster frente a un 50.15% de las observaciones globales. Finalmente mencionar que estas observaciones suponen un 18.75% de las observaciones totales de nuestra muestra donde los pagos son en efectivo.

### **Cluster 4**

Debido al pequeño número de observaciones que tiene el cluster 4, cualquier categoría podría llegar a considerarse caracterizada. En este caso tendríamos la fecha de recogida 1/1/2016 que sufre una sobrerepresentación del 50% en este cluster frente al 4.54% de representación global que tiene. Las observaciones de esta fecha en el cluster 4 suponen un 1,32% de sus observaciones totales.

Después tendríamos el periodo de recogida de noche que también aparece sobrerepresentado en este cluster de un 43.61% de observaciones en las que aparece a un 100% dentro de este cluster número 4, por lo que todas las observaciones dentro de este cluster tendrán como período de recogida, que este ha sido por la noche. Debido a que este periodo supone unas 2180 observaciones de nuestra muestra, por mucho porcentaje de estas que aparezca en el cluster, el número de observaciones que lo componen es mínimo, por lo que solo acaba representando un 0.84% de las observaciones donde este período es la noche.

Y para terminar la fecha 30/1/2016 aparece también sobrerepresentada aumentando de un 4.76% de representación global hasta un 33.3% dentro de este cluster. Vemos también como la participación de esta fecha en el cluster número 4 sólo supone un 0.84% del total de observaciones donde la fecha es la indicada.

### **Cluster 5**

Para el cluster 5, las categorías más caracterizadas serían la de pago mediante tarjeta y mediante efectivo de nuevo. El pago con tarjeta está sobrerepresentado en este cluster(73.06%) respecto al porcentaje de observaciones totales de la misma categoría(49.29%). Además, las observaciones con la categoría Tarjeta de crédito dentro del cluster 5 suponen un 8.81% de las observaciones totales del tipo de pago Credit card. Y, por otro lado, el pago en efectivo se ve infrarrepresentado pasando de un 50.15% de

observaciones globales en toda nuestra muestra a un 25.93% dentro del cluster, lo que supone un 3.07% de las observaciones globales donde Tipo de pago es efectivo.

### Cluster 6

Y para finalizar con el cluster 6 no tenemos dos/tres categorías que destaqueen sobre el resto de las que nos presenta el desc.var\$category. Vemos como el pago por tarjeta, el RateCodeID = Otros, el Extra 0, las fechas 4 y 31 de enero de 2016 así como el periodo de recogida del mediodía se ven sobrerepresentadas dentro de este último cluster. Y, por otro lado, el Extra = 0.5, la tarifa estándar y el tipo de pago en efectivo están infrarrepresentadas dentro de este cluster número 6.

### Quantitative variables which characterizes the clusters

```
##           Eta2   P-value
## Passenger_count 0.798826768 0.000000e+00
## tlenkm          0.701244237 0.000000e+00
## Fare_amount      0.669365956 0.000000e+00
## espeed          0.316696185 0.000000e+00
## Tip_amount       0.272609209 0.000000e+00
## Tolls_amount     0.990410306 0.000000e+00
## traveltimes      0.773024273 0.000000e+00
## distHaversine    0.677446046 0.000000e+00
## Total_amount     0.677248346 0.000000e+00
## lpep_pickup_time 0.241495539 2.011289e-296
## Dropoff_longitude 0.015093868 5.818893e-15
## Dropoff_latitude  0.012869428 1.289390e-12
## Pickup_latitude   0.010550394 3.388122e-10
## Pickup_longitude  0.003714435 2.291007e-03
```

En el output de quanti.var podemos observar que dentro de las variables numéricas más asociadas a la muestra globalmente, las que predominan serían:

|                     |                  |
|---------------------|------------------|
| Passenger_count     | espeed           |
| Tip_amount          | distHaversine    |
| tlenkm Tolls_amount | Total_amount     |
| Fare_amount         | lpep_pickup_time |
| Traveltimes         |                  |

## Description of each cluster by the quantitative variables

```
$`1`  

      v.test Mean in category Overall mean sd in category Overall sd p.value  

Dropoff_latitude 6.594838 40.7486743 40.74449280 0.0566303 0.05767784 4.257204e-11  

Pickup_latitude 5.278030 40.7497472 40.74645114 0.0559643 0.05680712 1.305799e-07  

Tolls_amount -11.891577 0.0000000 0.09183507 0.0000000 0.70251226 1.309107e-32  

Passenger_count -22.260872 1.1154957 1.37460000 0.3826614 1.05880822 8.850835e-110  

travelttime -26.744977 8.0004720 13.00095517 4.4246451 17.00805091 1.412584e-157  

Tip_amount -27.556518 0.6337164 1.23015723 0.9610917 1.96891907 3.697028e-167  

espeed -30.846707 17.8086743 20.97100764 5.4707493 9.32574766 6.201323e-209  

Total_amount -45.656047 9.3705000 14.54115814 3.4276610 10.30225573 0.000000e+00  

distHaversine -45.769105 1.6903393 3.21086535 0.8923275 3.02208211 0.000000e+00  

tlenkm -45.784707 2.2503168 4.58146198 1.1219635 4.63162829 0.000000e+00  

Fare_amount -46.386473 7.5429276 11.99993894 2.8486735 8.74051843 0.000000e+00  

$`2`  

      v.test Mean in category Overall mean sd in category Overall sd p.value  

Passenger_count 63.023610 5.2062937 1.37460000 0.59961722 1.05880822 0.000000000  

Pickup_latitude -2.269941 40.7390468 40.74645114 0.04718364 0.05680712 0.023211182  

Tolls_amount -2.276588 0.0000000 0.09183507 0.0000000 0.70251226 0.022810838  

Total_amount -2.311406 13.1738112 14.54115814 7.61900952 10.30225573 0.020810452  

tlenkm -2.387720 3.9464421 4.58146198 3.11639710 4.63162829 0.016953252  

distHaversine -2.470950 2.7820795 3.21086535 2.00794747 3.02208211 0.013475468  

Tip_amount -2.839549 0.9091259 1.23015723 1.42164079 1.96891907 0.004517734  

$`3`  

      v.test Mean in category Overall mean sd in category Overall sd p.value  

Fare_amount 24.305876 17.314863749 11.99993894 5.64395224 8.74051843 1.699059e-130  

distHaversine 23.075943 4.955539729 3.21086535 1.75914169 3.02208211 8.076955e-118  

Total_amount 22.699730 20.391775392 14.54115814 6.23967100 10.30225573 4.507511e-114  

tlenkm 22.069923 7.138770810 4.58146198 2.28777504 4.63162829 6.149616e-108  

espeed 15.722019 24.639103159 20.97100764 9.41146415 9.32574766 1.068635e-55  

travelttime 14.118847 19.008574895 13.00095517 8.36123658 17.00805091 2.907098e-45  

Tip_amount 13.046692 1.872810834 1.23015723 2.01981565 1.96891907 6.636683e-39  

Pickup_longitude -3.756357 -73.939864649 -73.93586028 0.04267852 0.04261074 1.724048e-04  

Pickup_latitude -3.938254 40.740854154 40.74645114 0.05809040 0.05680712 8.207682e-05  

Dropoff_latitude -4.694268 40.737719103 40.74449280 0.05783748 0.05767784 2.675637e-06  

Tolls_amount -4.958860 0.004681657 0.09183507 0.10449893 0.70251226 7.090812e-07  

Dropoff_longitude -6.669336 -73.943347103 -73.93489420 0.05234703 0.05066108 2.569625e-11  

Passenger_count -8.592768 1.146985962 1.37460000 0.43564971 1.05880822 8.489897e-18  

$`4`  

      v.test Mean in category Overall mean sd in category Overall sd p.value  

tlenkm 46.112918 16.559185 4.581462 6.79891777 4.63162829 0.000000e+00  

distHaversine 44.253180 10.710994 3.210865 3.78962368 3.02208211 0.000000e+00  

Fare_amount 43.840623 33.489714 11.999939 10.72037383 8.74051843 0.000000e+00  

Total_amount 43.009015 39.390208 14.541158 13.15926915 10.30225573 0.000000e+00  

espeed 29.466785 36.3.382134 20.971008 14.56778400 9.32574766 7.673989e-191  

Tip_amount 28.593030 4.387386 1.230157 3.88197499 1.96891907 8.202595e-180  

travelttime 19.402917 31.508105 13.000955 17.57953510 17.00805091 7.291369e-84  

Dropoff_longitude 5.704871 -73.918686 -73.934894 0.05120726 0.05066108 1.164315e-08  

lpep_pickup_time -3.121624 12.265912 13.633786 6.97177044 7.81353684 1.798568e-03  

Pickup_latitude -3.472807 40.735387 40.746451 0.06566803 0.05680712 5.150458e-04  

Dropoff_latitude -4.479670 40.730003 40.744493 0.06888997 0.05767784 7.475845e-06  

$`5`  

      v.test Mean in category Overall mean sd in category Overall sd p.value  

travelttime 55.286972 396.69444 13.00096 104.81932037 17.00805091 0.000000e+00  

lpep_pickup_time 34.472241 123.54051 13.63379 31.04230913 7.81353684 2.091014e-260  

Dropoff_latitude 2.302676 40.79869 40.74449 0.04491073 0.05767784 2.129710e-02  

espeed -5.510714 0.00100 20.97101 0.0000000 9.32574766 3.573815e-08  

$`6`  

      v.test Mean in category Overall mean sd in category Overall sd p.value  

Tolls_amount 70.362620 5.540000 0.09183507 0.0000000 0.70251226 0.000000e+00  

Total_amount 21.758712 39.248148 14.54115814 15.31579586 10.30225573 5.713080e-105  

distHaversine 18.784138 9.467659 3.21086535 6.11971400 3.02208211 1.018264e-78  

tlenkm 18.653711 14.104012 4.58146198 7.96807766 4.63162829 1.178083e-77  

Fare_amount 16.796420 28.181047 11.99993894 12.83914942 8.74051843 2.592256e-63  

espeed 13.854973 35.212112 20.97100764 14.02499384 9.32574766 1.187017e-43  

Tip_amount 13.187988 4.092099 1.23015723 3.83249289 1.96891907 1.028943e-39  

travelttime 5.944239 24.144033 13.00095517 9.11695699 17.00805091 2.777437e-09  

Pickup_latitude 3.273619 40.766948 40.74645114 0.04679207 0.05680712 1.061797e-03  

Dropoff_longitude 2.600295 -73.920375 -73.93489420 0.06862972 0.05066108 9.314357e-03  

lpep_pickup_time -2.212494 11.728395 13.63378632 5.85422153 7.81353684 2.693253e-02
```

**Cluster 1** Las variables más asociadas significativamente a este cluster son tlenkm, Fare\_amount, distHaversine y Total\_amount con medias inferiores en el propio cluster que las suyas en el global de las observaciones además de unas desviaciones inferiores también dentro del cluster que en general. Como ejemplo podemos considerar la variable de tlenkm cuya media global es de 4.56 vs la media dentro del cluster 1 que sería de 2.25, no solo significativa porque aparece en la lista sino que, observando la desviación estándar global que es 4.49 y la del propio cluster que sería de 1.12, esta diferencia entre las medias de  $4.56 - 2.25 = 2.3$  representa más del 50% de la desviación total que sufre la variable en todas sus observaciones.

**Cluster 2** La variable más asociada a este cluster es el número de pasajeros, también podríamos considerar el importe de propinas o incluso el distHaversine pero solo comentaremos la más significativa. Su media global se sitúa en 1.37 por lo que vemos que predominan los viajes con un solo pasajero, en cambio, en este cluster vemos que la media asciende hasta los 5.21 por lo que podemos considerar que si no son todas, la mayor parte de observaciones donde los pasajeros sean 6 y 5 estarán en este cluster.

**Cluster 3** Las variables más asociadas al cluster 3 serían el Fare\_amount seguida del distHaversine, el importe total y la distancia en km del trayecto. Observando la que nos indica el p valor que estaría más asociada(Fare\_amount) podemos ver como existe una diferencia considerable entre la media global de sus observaciones(11.98) y la media dentro del cluster(17.32), de esto podemos deducir que seguramente el total\_amount también tendrá una media mayor en este cluster y, por lo que podemos observar en el output del quanti para este cluster, esto es así y por tanto, podemos afirmar que los trayectos pertenecientes a este cluster son más caros que la media.

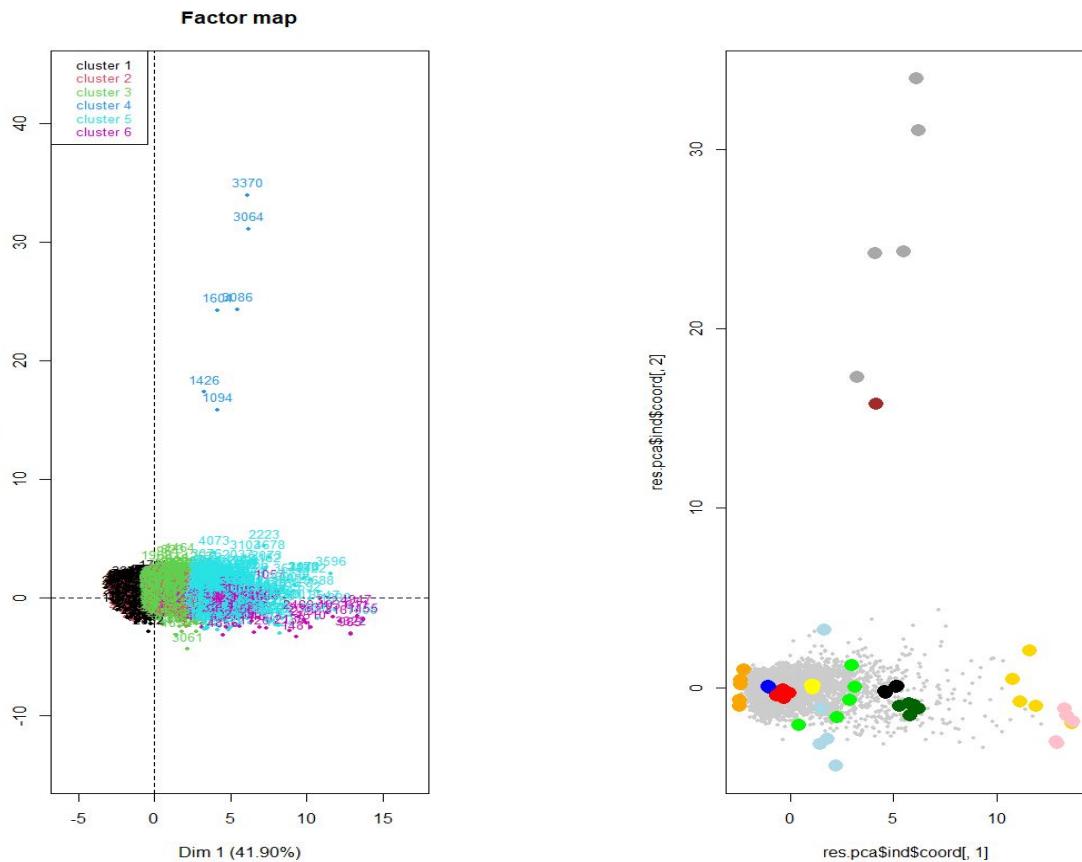
**Cluster 4** Para este cluster, las variables más asociadas son la del tiempo de viaje y la de la hora de recogida. Destacar sobretodo la diferencia entre la media global de 12.97 y la media dentro del cluster de 396.69 del tiempo de trayecto, lo que nos muestra que los trayectos dentro de este cluster son los que a primera vista llevarán más tiempo de todos los demás.

**Cluster 5** Las variables numéricas más asociadas a este cluster serían la distancia del trayecto en km, el distHaversine, el Fare\_amount y el total\_amount, seguidas por la velocidad efectiva y el importe de las propinas. Todas ellas con una media superior dentro del cluster respecto a su media global.

**Cluster 6** La variable numérica más asociada dentro de este cluster sería la de Tolls\_amount seguida del importe total. La media de la primera es de 0.092 mientras que dentro del cluster es de 5.54 siendo este el valor máximo de la misma, lo que da a entender que todas sus observaciones tendrán ese valor para dicha variable lo que da a entender que sean las mismas rutas, o al menos pasen por los mismos peajes durante todo el trayecto.

## Description of the clusters by the individuals

### Characteristic individuals



En el gráfico anterior podemos observar los individuos más representativos de cada uno de los clusters en color azul a la vez que sus más distantes en color naranja.

Para el cluster número 1 vemos como sus individuos más representativos o parangones son los situados a la izquierda del 0 en el eje de las x que corresponde con donde están la mayoría de los individuos. Por otro lado vemos como los elementos más distintivos son los situados a la izquierda del todo del gráfico, lo más separado al centro de gravedad del propio cluster. Podemos ver que sus parangones se identifican por pertenecer al mismo VendorID(Verifone), pagar todos en efectivo, no tener como cierto el Store\_and\_fwd, por tanto tener conexión para guardar el trayecto, pagar la tarifa estandard con un extra de 0 y pagar la tasa MTA a la vez que el sobrecargo de mejora. Por tanto podremos afirmar que la mayoría de los individuos seguirán estas características. Referente a los individuos distintivos vemos como cada uno de ellos muestra alguna o bien varias diferencias con las características mencionadas por lo que resultan distintos a esta mayoría representativa.

Los individuos más representativos del cluster 2 están en la misma nube de puntos mencionada anteriormente para los pertenecientes al cluster 1 pero pintados de rojo, hecho que tiene sentido por la proximidad de este cluster al anterior y su centro de gravedad debido a su dispersión entre los clusters 1 y 3. En cambio, sus individuos más distintivos se sitúan sobre  $x= 2.5$  formando un cuarto de círculo rodeando el extremo derecho de este cluster que parece coincidir con el 3 coloreados de color verde. Para este cluster vemos que las variables con observaciones semejantes varían sutilmente respecto al cluster 1. Estas serían el tipo de VendorID, que es el mismo que para el cluster 1, al Store\_and\_fwd, también falso, el pago de la tarifa estandard así como el pago de la tasa MTA y el improvement surcharge. Además tenemos como el Tolls\_amount en todos es 0 y todos pertenecen al tipo de trayecto Street-Hail. Para los individuos distintivos vemos como predominan los 6 pasajeros en lugar de 5 de los parangones así como que se efectúan más pagos en efectivo que en tarjeta.

Los parangones del cluster 3 aparecen como puntos amarillos a la derecha de la nube de puntos azules y rojos que mencionamos coincidiendo como es lógico con el centro de gravedad de dicho cluster. Sus individuos más distintivos son situados por encima y debajo junto a la nube gris que contiene todos los individuos a la altura de las  $x$  indicada por sus individuos más representativos en este caso pintados de un azul muy claro. Hecho que caracterizaría este cluster respecto a los anteriores según sus parangones sería que estos muestran observaciones de 1 pasajero por trayecto así como la no predominancia de un tipo concreto de VendorID ni de tipo de pago en específico.

Para el cluster 4 observamos solo un individuo como parangón coloreado de marrón siendo este el punto más cercano al eje de las  $y$  del gráfico. Se consideran todos los demás individuos pertenecientes al cluster como individuos distintivos del mismo y podemos verlos de color gris. En este cluster podemos ver trayectos de un solo viajero y nocturnos, de muy larga duración y de importe elevado como más característicos además observamos velocidades efectivas suficientemente altas como para considerar que son trayectos fuera de la ciudad

El cluster 5 tiene representados sus individuos más representativos de color negro, situados donde podemos observar una mayor cantidad de ellos y, de color dorado, al otro extremo(derecho) sus puntos más distintivos, los más lejanos a su centro de gravedad y por tanto, serán menos representativos del cluster. En este caso observamos trayectos más cortos y con velocidades efectivas más controladas suponiendo que se trata de trayectos dentro de la propia ciudad donde también predominan los trayectos nocturnos e individuales.

Para el cluster 6 vemos como los puntos coloreados de rosa, los situados más a la derecha del gráfico, aparecen como sus individuos más distintivos al situarse estos muy alejados de la cantidad principal de estos que forman el cluster. Sus parangones, por otro lado, podemos verlos en verde oscuro cerca del centro de gravedad del cluster 5. En este caso no vemos que predominen absolutamente los trayectos individuales, aún así podemos asumir que son trayectos más situados por el centro de la ciudad debido a las distancias recorridas, su velocidad efectiva así como el tiempo empleado donde estos se distribuyen entre el periodo del medio día y por la noche.

## Hierarchical tree result

### Ratio between within inertias

```
res.hcpc$call$quot[1:res.hcpc$call$nb.clust]  
## [1] 0.7999712 0.7981521 0.7760833 0.7358976 0.8974407 0.9018483
```

Si vemos la relación entre inercias dentro de un mismo cluster vemos que es bastante grande, es decir la inercia intra cluster es elevada.

### Inertia gain

```
res.hcpc$call$inert.gain[1:res.hcpc$call$nb.clust]  
## [1] 1.9240191 1.0252503 0.8276295 0.7327972 0.6707766 0.1916894
```

Si vemos la inertia gain vemos que es bastante alta des del cluster 1 al 5. Si vemos la pérdida entre pasar de 5 cluster a 6 cluster vemos que es bastante bajo con un valor de 0.1916, lo que significa que seguramente la elección de un hierarchical tree con más de 7 cluster no tendría sentido.

### Partition quality

```
(res.hcpc$call$within[1]-res.hcpc$call$within[res.hcpc$call$nb.clust])/res.hcpc$call$within[1]  
## [1] 0.7348677
```

Podemos ver como confirmando que el número de clusters óptimo es 6 obtenido con la llamada a “res.hcpc\$call\$nb.clust” tenemos una calidad de esta división en 6 clusters del 73.49%.

# K-Means: Partitioning in k=6

## Profiling KM

K-means clustering with 6 clusters of sizes 2374, 81, 991, 286, 250, 1018

```
##Within cluster sum of squares by cluster:
```

```
##[1] 1406.1036 1013.3541 811.5994 788.6252 1590.3512 5559.9415
```

```
##(between_SS / total_SS = 68.3 %)
```

Para explicar la clasificación obtenida con el Kmeans hemos decidido coger 6 clusters. Este número de clusters nos deja una explicación del 68.3% con las medidas que podemos ver en la primera línea del output.

## Global association variables numeric

```
##          Eta2      P-value
## Passenger_count 0.444694791 0.000000e+00
## tlenkm          0.603831600 0.000000e+00
## Fare_amount     0.576696172 0.000000e+00
## espeed          0.263490766 0.000000e+00
## Tolls_amount    0.619167093 0.000000e+00
## distHaversine   0.564957345 0.000000e+00
## Total_amount    0.595040329 0.000000e+00
## travelttime     0.233684977 2.464593e-285
## lpep_pickup_time 0.230551028 6.439587e-281
## Tip_amount      0.222218401 2.920709e-269
## Dropoff_longitude 0.021909039 2.966999e-22
## Dropoff_latitude 0.008728726 2.548210e-08
## Pickup_latitude  0.006045956 1.290650e-05
## Pickup_longitude 0.004354320 5.715964e-04
```

Como podemos ver según el p valor de cada una de las variables numéricas, las que están más asociadas globalmente a la muestra son las que hemos marcado en azul, las cuales coinciden con las que habíamos encontrado mediante el hierarchical clustering aunque no con el mismo p valor lo que significa que estas asociaciones han variado.

```
## $`1`
##          v.test Mean in category Overall mean sd in category
## tlenkm      -14.494179  2.1488108  4.58146198  1.14987360
## distHaversine -14.623896  1.6093841  3.21086535  0.88623093
## Total_amount -16.052113  8.5485336 14.54115814  3.66900369
## Fare_amount   -16.658218  6.7237753 11.99993894  2.49596451
## lpep_pickup_time -31.568368  4.6955400 13.63378632  5.06942330
##          Overall sd      p.value
## tlenkm        4.63162829 1.318696e-47
## distHaversine 3.02208211 1.977536e-48
## Total_amount  10.30225573 5.525040e-58
## Fare_amount    8.74051843 2.637671e-62
## lpep_pickup_time 7.81353684 1.003687e-218
##
## $`2`
```

```

##          v.test Mean in category Overall mean sd in category
## espeed      -28.488256   16.6990429 20.97100764  4.78955497
## Total_amount -33.761969    8.9482429 14.54115814  2.82948643
## Fare_amount   -34.220223    7.1904585 11.99993894  2.37583417
## tlenkm       -34.234381    2.0318487 4.58146198  0.90988893
## distHaversine -34.790326    1.5202574 3.21086535  0.72944165
##          Overall sd      p.value
## espeed      9.32574766 1.637649e-178
## Total_amount 10.30225573 7.134769e-250
## Fare_amount   8.74051843 1.209845e-256
## tlenkm       4.63162829 7.448958e-257
## distHaversine 3.02208211 3.406549e-265
##
## $`3`
##          v.test Mean in category Overall mean sd in category
## Passenger_count 46.567793    3.5220126 1.37460000  2.10570696
## Fare_amount     22.254258    20.4714885 11.99993894  13.07651024
## distHaversine  21.227733    6.0048397 3.21086535  4.36243875
## Total_amount    21.207526    24.0567296 14.54115814  14.81389736
## tlenkm        21.203364    8.8585778 4.58146198  6.44443468
## espeed        15.801924    27.3891029 20.97100764  14.77674556
##          Overall sd      p.value
## Passenger_count 1.05880822 0.000000e+00
## Fare_amount     8.74051843 1.025760e-109
## distHaversine  3.02208211 5.295615e-100
## Total_amount    10.30225573 8.138191e-100
## tlenkm        4.63162829 8.890774e-100
## espeed        9.32574766 3.017646e-56
##
## $`4`
##          v.test Mean in category Overall mean sd in category
## distHaversine  8.368848    3.913570e+00 3.21086535  1.11636717
## Fare_amount     8.027703    1.394947e+01 11.99993894  3.33137049
## Total_amount    7.468417    1.667893e+01 14.54115814  4.16121161
## travertime     6.496199    1.607079e+01 13.00095517  5.77794647
## Passenger_count -8.967635   1.110787e+00 1.37460000  0.35730644
##          Overall sd      p.value
## distHaversine  3.02208211 5.818427e-17
## Fare_amount     8.74051843 9.931486e-16
## Total_amount    10.30225573 8.116542e-14
## travertime     17.00805091 8.237461e-11
## Passenger_count 1.05880822 3.029493e-19
##
## $`5`
##          v.test Mean in category Overall mean sd in category
## Tolls_amount    55.633797    3.461877 0.09183507  2.65949782
## tlenkm         34.575806    18.390031 4.58146198  10.96442081
## Total_amount    33.806675    44.572688 14.54115814  18.77537258
## distHaversine  30.937505    11.272706 3.21086535  6.54220495
## Fare_amount     29.587600    34.299155 11.99993894  16.05801588
##          Overall sd      p.value
## Tolls_amount   0.70251226 0.000000e+00
## tlenkm         4.63162829 5.837763e-262
## Total_amount   10.30225573 1.573501e-250
## distHaversine  3.02208211 3.741411e-210

```

```

## Fare_amount     8.74051843 2.157664e-192

##
## $`6` 
##          v.test Mean in category Overall mean sd in category
## Fare_amount    26.523537    21.613820 11.99993894  5.95924104
## tlenkm        25.381329    9.456502  4.58146198  2.16340894
## distHaversine 25.023467    6.346922  3.21086535  2.06401744
## Total_amount   24.648529    25.071766 14.54115814  6.17563347

##          Overall sd      p.value
## Fare_amount    8.74051843 5.188538e-155
## tlenkm         4.63162829 4.054420e-142
## distHaversine 3.02208211 3.395953e-138
## Total_amount   10.30225573 3.816230e-134

res.cat$test.chi2

##          p.value df
## f.tlenkm    0.000000e+00 10
## f.traveltime 0.000000e+00 20
## f.distHaversine 0.000000e+00 10
## AnyToll     0.000000e+00  5
## f.Fare_amount 0.000000e+00 15
## f.Passenger_count 0.000000e+00 10
## f.Total_amount 0.000000e+00 35
## hcpck       0.000000e+00 25
## f.espeed     1.998801e-282 10
## lpep_pickup_period 6.478927e-115 15
## f.Extra      1.929912e-86 10
## AnyTip      2.328424e-42  5
## Payment_type 1.311420e-35 10
## lpep_pickup_date 8.734495e-21 150
## RateCodeID   9.513616e-19  5
## multiouts   7.109065e-10  5
## VendorID    2.397221e-05  5
## Trip_type    1.137969e-04  5
## f.MTA_tax    3.822004e-04  5
## f.Improvement_surcharge 2.412695e-03 5

```

Para el cluster 1 vemos que la variable más caracterizada dentro del propio cluster es la hora de recogida con un p valor muy pequeño en comparación a las demás, la cual observamos que tiene una media inferior a la global por lo que asumimos que contendrá trayectos con recogidas temprano.

Referente al cluster 2 tenemos las variables distHaversine, distancia en kilómetros, importe de la tarifa, importe total y velocidad efectiva como más caracterizadas dentro del cluster. Estas tienen una media inferior dentro del cluster que la que les corresponde al total de observaciones por lo que podremos asumir que se trata de un cluster compuesto por trayectos cortos, lentos y baratos, es decir, trayectos urbanos.

Por lo que hace al cluster 3 vemos caracterizadas, sobretodo, las variables del número de pasajeros, el importe de la tarifa y total, el distHaversine así como la distancia en kilómetros y la velocidad efectiva. En este caso estas tienen una media superior dentro

del cluster que en el total de observaciones por lo que podemos inducir que se trata de trayectos interurbanos que en consecuencia de salir del centro pueden circular a mayor velocidad al evitar atascos.

Para el cluster 4 tendríamos caracterizadas las variables distHaversine así como el importe total del trayecto y su tarifa, el número de pasajeros, estos con una media menor dentro del cluster, y finalmente el tiempo de viaje. En este caso no están tan caracterizadas como en los clusters anteriores por lo que suposiciones que hagamos pueden no ser tan acertadas al ser un cluster más equilibrado. En este caso podríamos llegar a asumir que se trata de trayectos de larga distancia que, en consecuencia, derivan en caros.

Referente al cluster 5 tenemos caracterizados el importe por peajes, el de la tarifa y el total, así como las distancias en km, la distHaversine así como el tiempo de trayecto. Todas estas con una media mayor dentro del cluster muy caracterizada por lo que podemos asumir que este cluster también lo formarán principalmente trayectos interurbanos que pasen por carreteras de pago y cuya tarifa tenga un coste mayor.

Y finalmente, para el cluster 6 vemos caracterizadas con una media superior dentro del cluster las variables referentes a la tarifa del trayecto y la total, y la distHaversine y la distancia en kilómetros. Lo que nos llevaría a asumir que se trata de un cluster formado por trayectos largos.

Si analizamos la variables categóricas que contribuye más en la construcción de Hierarchical clustering, vemos que los factores tlenkm, travelttime y distHaversine tienen una chi2 igual a 0 y por lo tanto son las variables que caracterizan los clustering. En la figura marcada en azul podemos ver las variables categóricas que caracterizan los clustering ordenadas según la probabilidad value (p-value).

## Global association category categoricas

|   | Cla/Mod    | Mod/Cla       | Global | p.value       | v.test     |
|---|------------|---------------|--------|---------------|------------|
| f.distHaversine=f.distHaversine-[0,5]   | 52.9786712 | 100.000000000 | 81.58  | 4.703457e-261 | 34.515457  |
| f.tlenkm=f.tlenkm-(1,5]                 | 61.1935484 | 87.78343360   | 62.00  | 1.757096e-255 | 34.142008  |
| f.travelttime=f.travelttime-[0,10]      | 64.2004773 | 74.68764461   | 50.28  | 3.383813e-206 | 30.641942  |
| f.espeed=f.espeed-(1,25]                | 53.1096774 | 95.23368811   | 77.50  | 2.629290e-174 | 28.146709  |
| f.espeed=f.espeed-(25,130]              | 9.2046470  | 4.76631189    | 22.38  | 1.382287e-172 | -28.005765 |
| f.distHaversine=f.distHaversine-(5,10]  | 0.0000000  | 0.00000000    | 14.54  | 1.215477e-199 | -30.145857 |
| f.Fare_amount=f.Fare_amount-(14.5,71.5] | 0.7311129  | 0.41647385    | 24.62  | 0.000000e+00  | -Inf       |
| f.tlenkm=f.tlenkm-(5,67.9]              | 0.3369272  | 0.23137436    | 29.68  | 0.000000e+00  | -Inf       |

Para el cluster 1 vemos que tenemos sobrerepresentadas las categorías de distancia y tiempo cercanas a 0 e infrarrepresentadas, por otro lado las de tiempo de trayecto prolongado así como su distancia. También vemos sobrerepresentadas las

velocidades efectivas inferiores a 25 km/h que junto al análisis de las variables numéricas de cada cluster nos podría indicar que se trata de trayectos urbanos que suceden por la mañana.

|  | Cla/Mod    | Mod/Cla    | Global | p.value       | v.test    |
|--|------------|------------|--------|---------------|-----------|
| f.Passenger_count=f.Passenger_count-Others | 66.1971831 | 59.1194969 | 8.52   | 3.320893e-217 | 31.457441 |

Las categorías del cluster 2 solo nos ofrecen información destacable sobre que el número de pasajeros suele ser mayor a 2 debido a la sobrerrepresentación de la misma.

|                    | Cla/Mod     | Mod/Cla    | Global | p.value       | v.test    |
|--------------------|-------------|------------|--------|---------------|-----------|
| AnyToll=AnyToll No | 96.55172414 | 64.1221374 | 1.74   | 1.316343e-143 | 25.515807 |

El cluster 3 nos indica que existe una gran cantidad de las observaciones que no han pasado por ningún peaje. Así que podemos definir que la mayoría de trayectos interurbanos que se producen dentro de este cluster siguen rutas por las que no se suele pasar por carreteras de pago.

|   | Cla/Mod     | Mod/Cla    | Global | p.value       | v.test    |
|---|-------------|------------|--------|---------------|-----------|
| f.Fare_amount=f.Fare_amount-(14.5,71.5) | 39.31762794 | 93.4362934 | 24.62  | 3.235575e-282 | 35.898399 |
| f.tlenkm=f.tlenkm-(5,67.9]              | 34.23180593 | 98.0694981 | 29.68  | 1.469359e-280 | 35.792030 |
| f.distHaversine=f.distHaversine-(5,10]  | 54.74552957 | 76.8339768 | 14.54  | 7.087502e-270 | 35.098572 |
| f.Total_amount=f.Total_amount-(20,30]   | 60.00000000 | 68.9189189 | 11.90  | 5.842961e-251 | 33.835940 |

Para el cluster 4 tenemos sobrerrepresentadas las categorías con una tarifa elevada a la vez que trayectos con distancias largas que concuerda con las observaciones anteriores definiendo un cluster de trayectos de largo recorrido.

|                                       | Cla/Mod    | Mod/Cla     | Global | p.value       | v.test    |
|---------------------------------------|------------|-------------|--------|---------------|-----------|
| f.Total_amount=f.Total_amount-(11,18] | 50.4261364 | 69.40371457 | 28.16  | 4.969131e-217 | 31.444640 |
| f.Fare_amount=f.Fare_amount-(9,14.5]  | 50.7923930 | 62.65884653 | 25.24  | 4.679811e-187 | 29.169849 |
| f.traveltime=f.traveltime-(10,20]     | 41.3004214 | 67.05767351 | 33.22  | 5.378902e-139 | 25.096881 |
| f.tlenkm=f.tlenkm-(5,67.9]            | 38.1401617 | 55.32746823 | 29.68  | 5.573471e-84  | 19.416723 |

El cluster 5 por otro lado tiene unos importes totales y tiempos de trayecto más contenidos aún siendo estas distancias en kilómetros considerables por lo que suponemos que se trata de trayectos interurbanos o más bien por las afueras de la ciudad.

|                                  | Cla/Mod    | Mod/Cla    | Global | p.value       | v.test    |
|----------------------------------|------------|------------|--------|---------------|-----------|
| f.traveltime=f.traveltime-[0,10] | 24.1447892 | 87.9710145 | 50.28  | 6.699286e-112 | 22.478754 |
| lpep_pickup_period=Period night  | 24.0366972 | 75.9420290 | 43.60  | 3.892522e-77  | 18.589713 |

En el cluster 6 vemos sobrerepresentados los trayectos con un tiempo muy bajo y nocturnos, que junto al análisis anterior nos mostraría como incierta esta suposición y no nos permitiría llegar a una suposición del tipo de trayectos que caracterizan este cluster.

## Confusion Table

|        | kKM-3 | kKM-6 | kKM-2 | kKM-1 | kKM-5 | kKM-4 |
|--------|-------|-------|-------|-------|-------|-------|
| kHP- 1 | 0     | 661   | 0     | 2161  | 295   | 0     |
| kHP- 2 | 0     | 17    | 263   | 0     | 0     | 6     |
| kHP- 3 | 0     | 12    | 9     | 0     | 728   | 462   |
| kHP- 4 | 44    | 0     | 205   | 0     | 0     | 50    |
| kHP- 5 | 6     | 0     | 0     | 0     | 0     | 0     |
| kHP- 6 | 81    | 0     | 0     | 0     | 0     | 0     |

```
> (2161+263+0+50)/(661+2161+295+17+263+6+12+9+728+462+44+205+50+6+81)
[1] 0.4948
```

A partir de la tabla de confusión, podemos ver que muchos individuos del cluster 1 del kKM y del kHP coinciden. Si observamos las características que coinciden entre el primer cluster de ambos métodos, vemos que los dos clusters tienen 3km de media aproximadamente y 7 dólares de Fare Amount. Si vemos el resultado del kHP la variable duration time es la sexta variable más importante y en kKM la variable duration time es la octava variable más importante.

Si nos fijamos, el cluster 5 tiene sólo 6 individuos en contraposición el cluster 5 tiene aproximadamente 1000, lo que significa que no hay ningún tipo de relación.

Si comparamos la cantidad de miembros que coincide entre clusters con el mismo índice vemos que solo el 50% de miembros coinciden.

## Correspondence Analysis (CA)

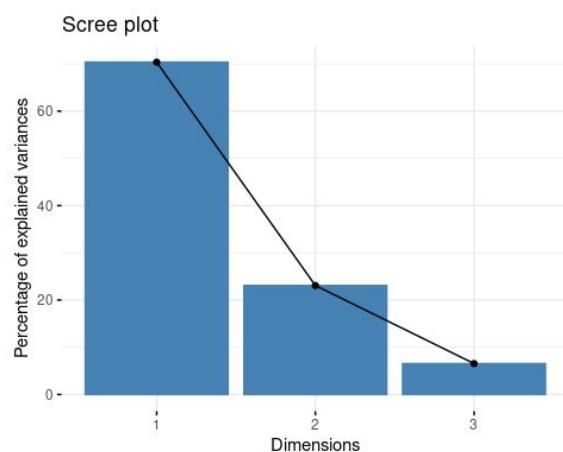
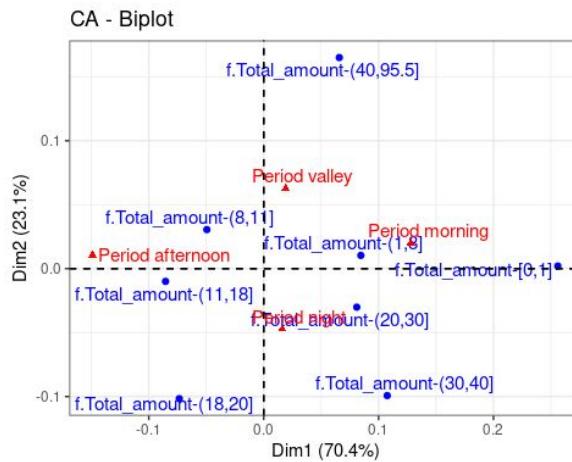
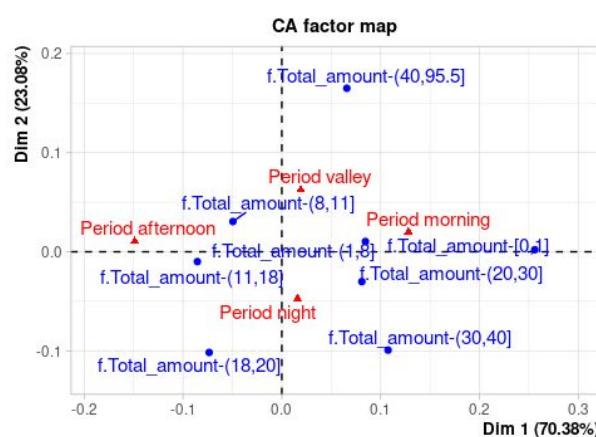
### CA in Total amount and Pick up period

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

```
## data: tt
```

```
## X-squared = 44.632, df = 21, p-value = 0.001935
```

Como podemos ver el p-value es 0.001935, menor que 0.05, por lo tanto es estadísticamente significativa y podríamos rechazar la hipótesis nula: filas y columnas son independientes. Como podemos ver en la tabla de contingencia hay valores inferiores a 5, por lo que la aproximación de Pearson's Chi-squares test puede ser incorrecta.



Viendo la tabla de contingencias, podemos observar como las categorías de importe total de 40-95.5\$, de 30-40\$, DE 18-20\$ así como los trayectos nocturnos, serían los menos presentes en las observaciones al estar más alejados del centro de gravedad(de los ejes).

Podemos observar como los trayectos nocturnos están muy relacionados con costar entre 20\$ y 30\$ y, en cambio, los trayectos que son por la mañana, entre 1\$ y 8\$.

Como conclusión podemos asumir como falsa la suposición de independencia entre categorías de importe total y periodo de recogida.

## CA in Total amount and Travel time

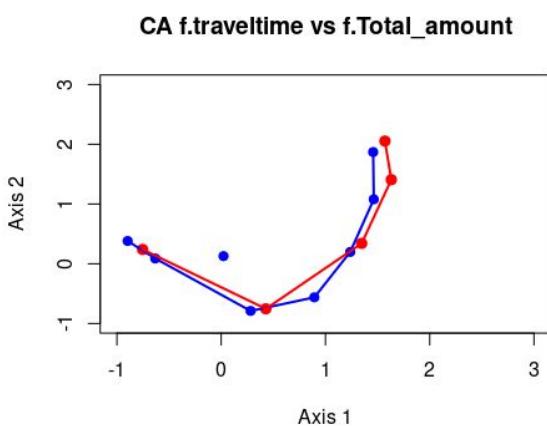
```
chisq.test(tt)
```

```
## X-squared = 6148.5, df = 28, p-value < 2.2e-16
```

Como podemos ver en la tabla de contingencia hay valores inferiores a 5, por lo que la aproximación de Pearson's Chi-squares test puede ser incorrecta. Si observamos el p-value 2.2e-16, es inferior a 0.05 por lo tanto, podríamos rechazar la hipótesis nula.



Si vemos el primer plano factorial vemos como los factores están muy bien representados. Vemos que aquellos viajes que duran entre 40 y 548 minutos tienen un coste entre 40 y 95.5. Los viajes entre 11 y 40 minutos tienen un coste mediano, entre 15 y 50. Y los viajes con coste bajo tienen una duración entre 0 y 15 minutos.



En esta imagen podemos ver perfectamente el efecto Guttman que se forma a partir de aplicar CA. El cual también nos indica la clara relación entre filas y columnas.

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

```
##
```

```
## Call:
```

```
## CA(X = tt)
```

```
##
```

```
## The chi square of independence between the two variables is equal to 6148.456 (p-value = 0 ).
```

```
##
```

```

## Eigenvalues
##           Dim.1  Dim.2  Dim.3  Dim.4
## Variance     0.689  0.389  0.144  0.007
## % of var.  56.013 31.666 11.720  0.601
## Cumulative % of var. 56.013 87.679 99.399 100.000

```

Como podemos ver el estadístico de la chi square es igual a 6148.456 lo que significa que tienen una gran relación las dos variables.

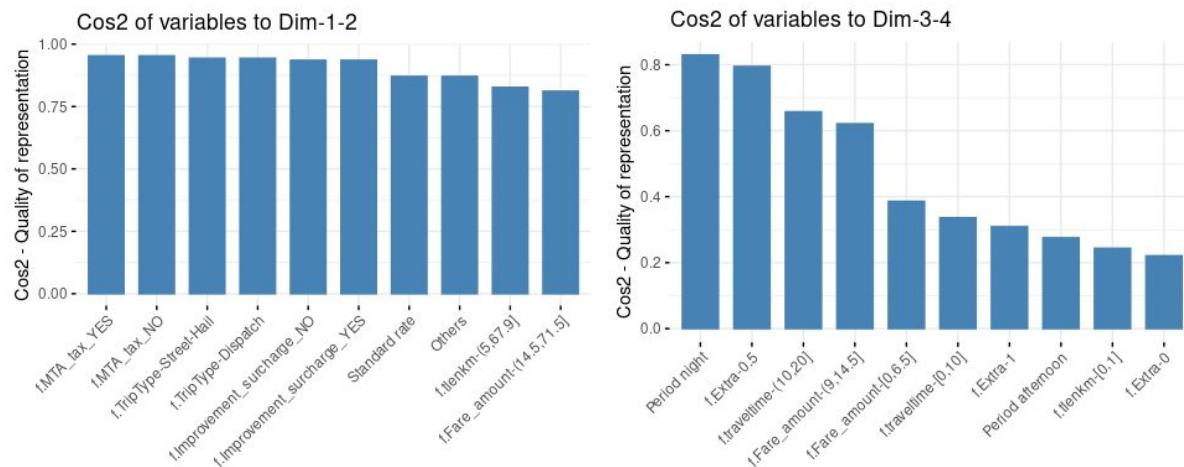
```
mean(res.ca$eig[,1])
```

```
## [1] 0.3074228
```

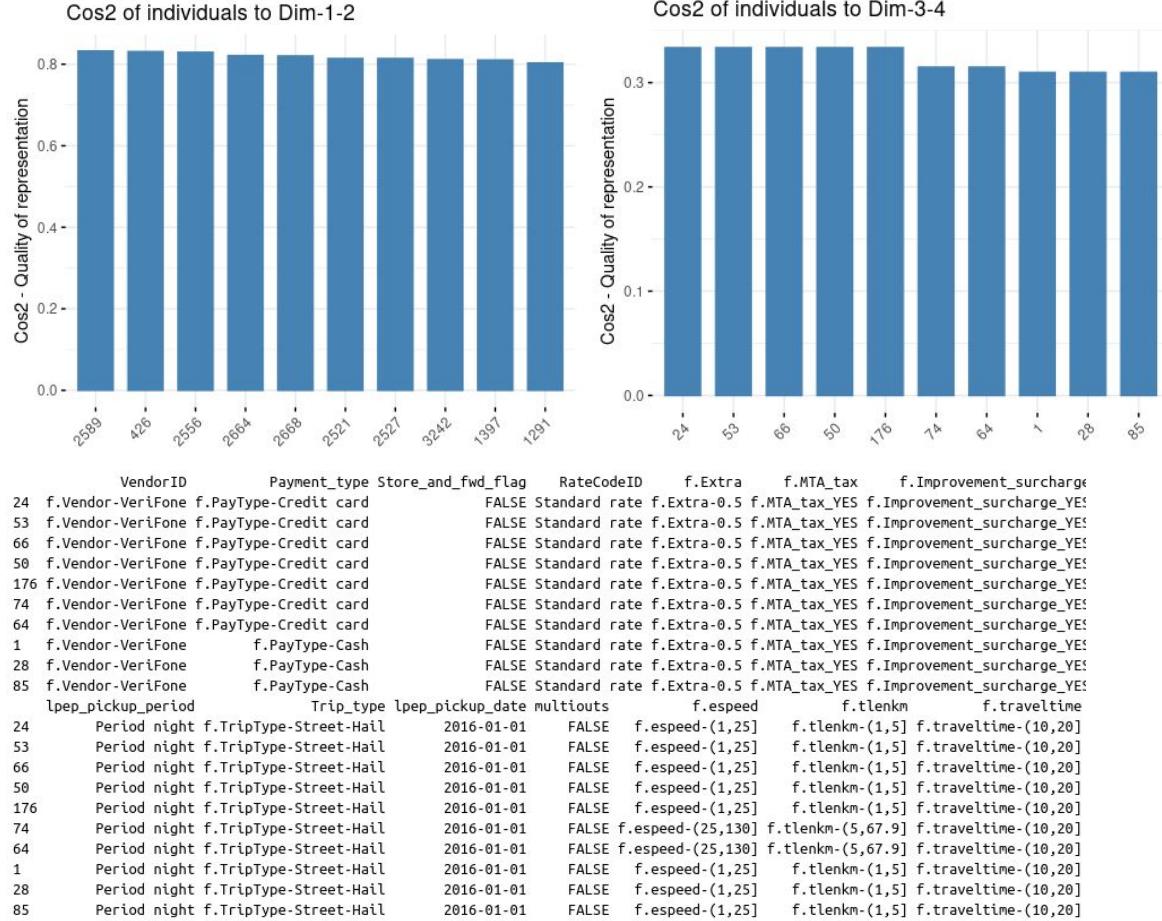
Según el criterio de Kaiser deberíamos coger aquellas dimensiones que tienen un valor propio superior a 0.3074228. Por lo tanto, deberíamos coger hasta la segunda dimensión. Con estas dos dimensiones tendríamos explicado un 87.68% de la muestra.

## MCA analysis

### Quality of representation

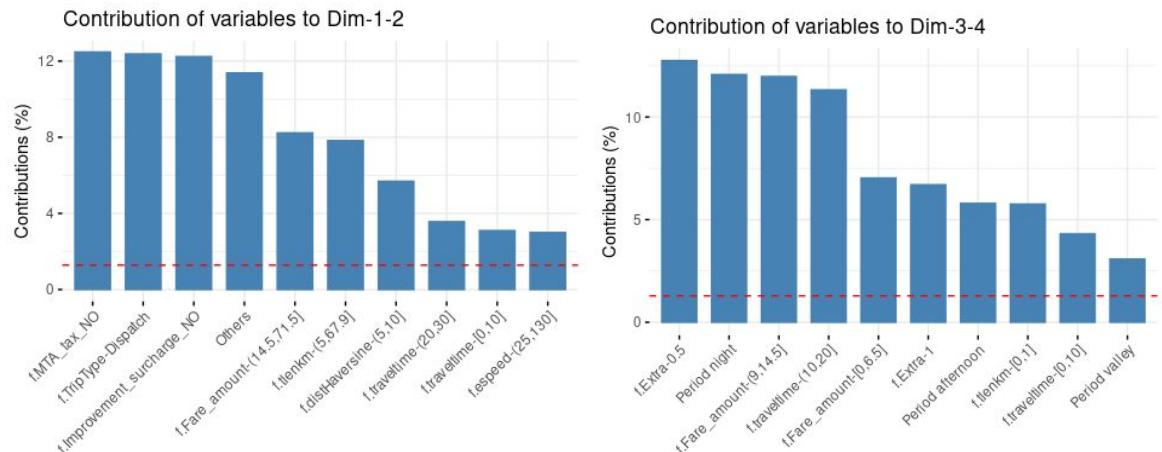


Tras realizar el análisis de correspondencias múltiples, podemos ver que la variables MTA\_tax, TripType y Improvement\_surcharge contribuye en la construcción del primer plano factorial. Podemos ver que los viajes de taxi realizados por la noche son los mejor representados en el segundo plano factorial. También podemos ver que los individuos que han hecho un viaje entre 10 y 20 minutos están bien representados en el segundo plano factorial.

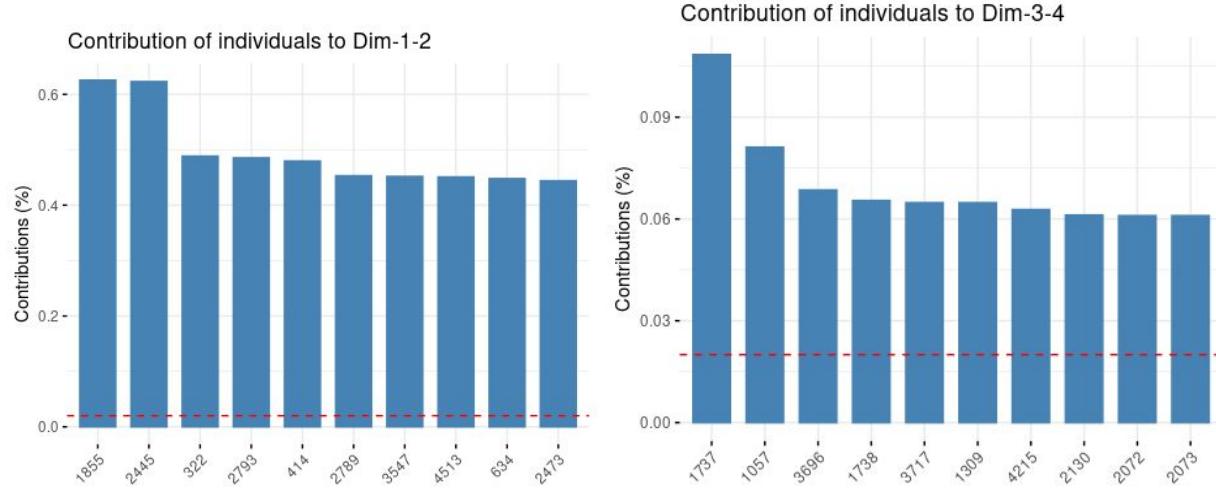


Como podemos ver, los individuos mejor representados en el segundo plano factorial son viajes de taxi realizados por la noche, pagados con extra de 0.5 y con una duración de viaje entre 10 y 20 minutos. Si nos fijamos todas estas características son las mejor representadas en el segundo plano factorial.

## Contribution



Como vemos, los individuos que contribuyen más en el primer plano factorial no han pagado la tasa MTA y no han pagado la tarifa improvement surcharge. En contraposición los individuos que contribuyen más han pagado 0.5 de extra y han hecho el viaje por la noche.



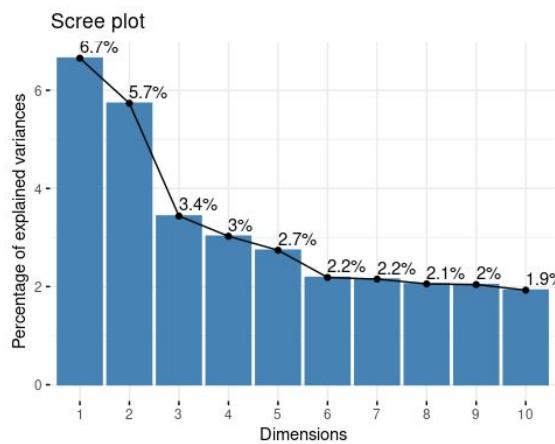
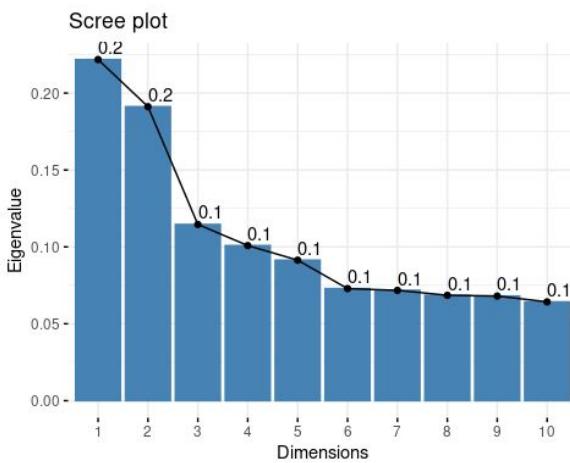
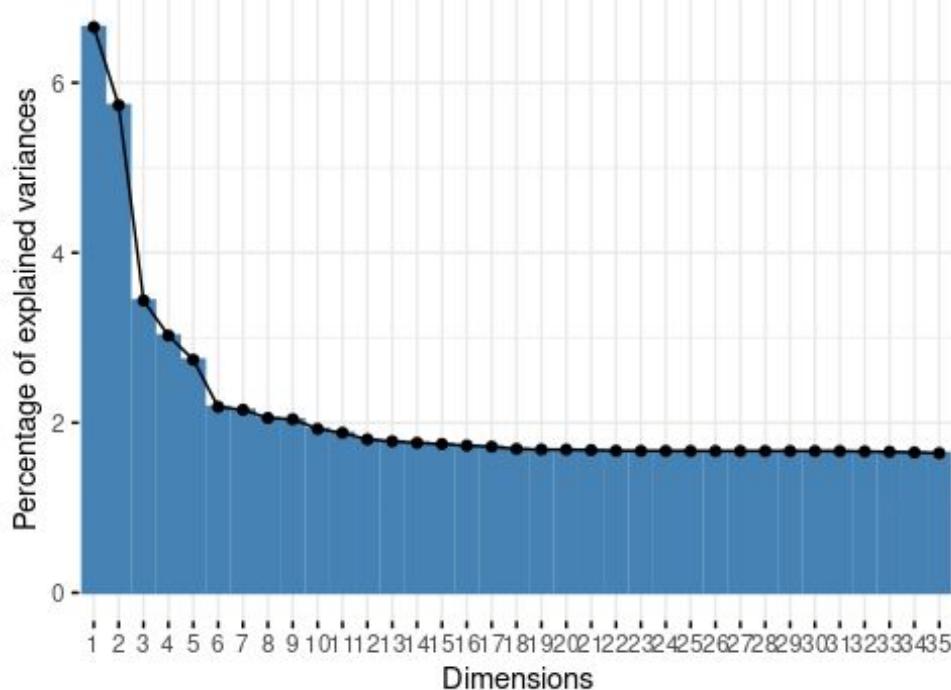
| VendorID           | Payment_type                            | Store_and_fwd_flag     | RateCodeID                | f.Extra                 | f.MTA_tax                | f.Improvement_surcharge                    |
|--------------------|---|------------------------|---------------------------|-------------------------|--------------------------|--|
| 1855               | f.Vendor-VeriFone f.PayType-Credit card | FALSE                  | Others                    | f.Extra-0               | f.MTA_tax_NO             | f.Improvement_surcharge_NO                 |
| 2445               | f.Vendor-VeriFone f.PayType-Cash        | FALSE                  | Others                    | f.Extra-0               | f.MTA_tax_NO             | f.Improvement_surcharge_NO                 |
| 322                | f.Vendor-VeriFone f.PayType-Cash        | FALSE                  | Others                    | f.Extra-0               | f.MTA_tax_NO             | f.Improvement_surcharge_NO                 |
| 1737               | f.Vendor-Mobile f.PayType-Others        | FALSE                  | Standard rate             | f.Extra-0               | f.MTA_tax_YES            | f.Improvement_surcharge_YES                |
| 1057               | f.Vendor-VeriFone f.PayType-Cash        | FALSE                  | Standard rate             | f.Extra-1               | f.MTA_tax_YES            | f.Improvement_surcharge_YES                |
| 3696               | f.Vendor-Mobile f.PayType-Credit card   | FALSE                  | Standard rate             | f.Extra-0.5             | f.MTA_tax_YES            | f.Improvement_surcharge_YES                |
| lpep_pickup_period |   | TripType               | lpep_pickup_date          | multitrips              | f.espeed                 | f.tlenkm f.traveltime                      |
| 1855               | Period morning                          | f.TripType-Dispatch    | 2016-01-12                | TRUE                    | f.espeed-(25,130]        | f.tlenkm-[0,1] f.traveltime-[0,10]         |
| 2445               | Period night                            | f.TripType-Dispatch    | 2016-01-15                | TRUE                    | f.espeed-(25,130]        | f.tlenkm-(5,67.9] f.traveltime-(40,548]    |
| 322                | Period afternoon                        | f.TripType-Dispatch    | 2016-01-02                | FALSE                   | f.espeed-(25,130]        | f.tlenkm-(5,67.9] f.traveltime-(30,40]     |
| 1737               | Period valley                           | f.TripType-Street-Hail | 2016-01-11                | TRUE                    | f.espeed-(25,130]        | f.tlenkm-(5,67.9] f.traveltime-(40,548]    |
| 1057               | Period valley                           | f.TripType-Street-Hail | 2016-01-07                | FALSE                   | f.espeed-(1,25]          | f.tlenkm-(5,67.9] f.traveltime-(40,548]    |
| 3696               | Period night                            | f.TripType-Street-Hail | 2016-01-23                | FALSE                   | f.espeed-(1,25]          | f.tlenkm-[0,1] f.traveltime-[0,10]         |
| f.distHaversine    |   | AnyToll                | f.Fare_amount             | f.Pasenger_count        | f.Total_amount           | Pasenger_count                             |
| 1855               | f.distHaversine-[0,5]                   | AnyToll Yes            | f.Fare_amount-(14.5,71.5] | f.Pasenger_count-2      | f.Total_amount-(40,95.5] | 2  |
| 2445               | f.distHaversine-[0,5]                   | AnyToll Yes            | f.Fare_amount-(9,14.5]    | f.Pasenger_count-Others | f.Total_amount-(8,11]    | 3  |
| 322                | f.distHaversine-(10,26.1]               | AnyToll No             | f.Fare_amount-(14.5,71.5] | f.Pasenger_count-Others | f.Total_amount-(40,95.5] | 4  |
| 1737               | f.distHaversine-(5,10]                  | AnyToll No             | f.Fare_amount-(14.5,71.5] | f.Pasenger_count-1      | f.Total_amount-(40,95.5] | 1  |
| 1057               | f.distHaversine-(10,26.1]               | AnyToll No             | f.Fare_amount-(14.5,71.5] | f.Pasenger_count-1      | f.Total_amount-(40,95.5] | 1  |
| 3696               | f.distHaversine-[0,5]                   | AnyToll Yes            | f.Fare_amount-[0,6.5]     | f.Pasenger_count-Others | f.Total_amount-(1,8]     | 3  |
| tlenkm             |   | Pickup_longitude       | Pickup_latitude           | Dropoff_longitude       | Dropoff_latitude         | Fare_amount espeed Tip_amount Tolls_amount |
| 1855               | 0.0804672                               | -74.08338              | 40.64220                  | -74.08366               | 40.64284                 | 65.00000 48.28032 13.00 0.00000            |
| 2445               | 56.0856400                              | -74.00069              | 40.59997                  | -73.96132               | 40.62992                 | 10.00000 40.04528 0.00 0.00000             |
| 322                | 32.7018701                              | -73.91185              | 40.82726                  | -73.78302               | 40.64885                 | 57.66478 50.33208 0.00 5.54000             |
| 1737               | 67.9143168                              | -73.86628              | 40.87276                  | -73.86414               | 40.82361                 | 61.50166 37.35547 0.00 0.22224             |
| 1057               | 22.0641062                              | -73.86156              | 40.73050                  | -73.95473               | 40.81818                 | 51.00000 20.73369 0.00 5.54000             |
| 3696               | 0.9656064                               | -73.96154              | 40.71416                  | -73.95172               | 40.71512                 | 5.00000 10.89713 1.25 0.00000              |
| lpep_pickup_time   |   | traveltime             | distHaversine             | AnyTip                  | Total_amount             | hcpcd clakM                                |
| 1855               | 9                                       | 0.100000               | 0.07562334                | AnyTip Yes              | 78.00000                 | KHP- 4 kKM-1                               |
| 2445               | 20                                      | 84.03330               | 4.70946300                | AnyTip No               | 10.00000                 | KHP- 4 kKM-3                               |
| 322                | 18                                      | 38.98333               | 22.63957168               | AnyTip No               | 95.54000                 | KHP- 6 kKM-3                               |
| 1737               | 15                                      | 109.08333              | 5.47499787                | AnyTip No               | 70.61709                 | KHP- 4 kKM-3                               |
| 1057               | 16                                      | 63.85000               | 12.52834389               | AnyTip No               | 58.34000                 | KHP- 6 kKM-3                               |
| 3696               | 0                                       | 5.316667               | 0.83537860                | AnyTip Yes              | 7.55000                  | KHP- 1 kKM-4                               |

Como podemos ver los 3 primeros individuos mejor representados en el primer plano factorial no han pagado la tasa MTA y no han pagado la tarifa improvement surcharge. Estas características no se encuentran en los individuos del segundo plano factorial. Los individuos del segundo plano factorial son viajes realizados en Period valley y Period night. También podemos ver que en el segundo plano factorial encontramos que los tres individuos que contribuyen más en el segundo plano factorial han pagado

diferentes extra, por lo que podríamos decir que en el segundo plano factorial se divide los individuos según el tipo de extra pagado y el periodo en el que se realizó el viaje.

## Eigenvalues and dominant axes analysis.

Scree plot

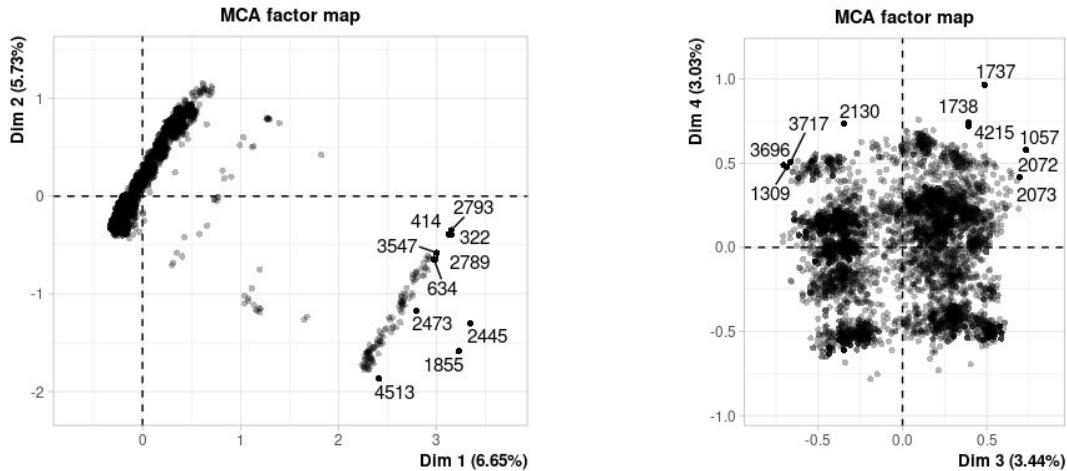


```
mean(res.mca$eig[,1])
## [1] 0.05555556
```

Según el criterio de Kaiser generalizado deberíamos coger aquellas dimensiones cuyo valor propio sea superior a 0.055. Lo que significa que tendríamos que coger hasta la dimensión 35, esta dimensión proporciona un 74.4% de varianza acumulada. En

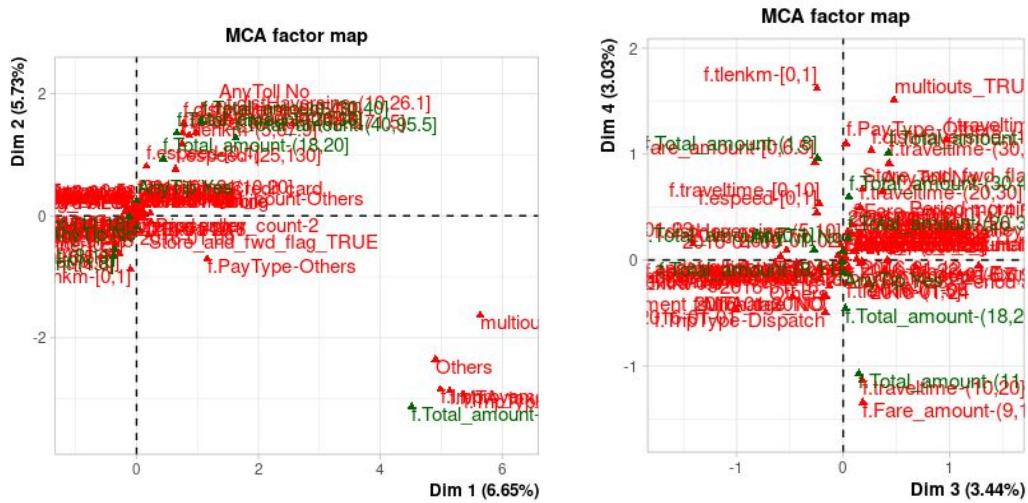
contraposición, si aplicamos la regla de Elbow deberíamos coger hasta la dimensión 6 pero solo nos proporciona un 23.78% de varianza acumulada.

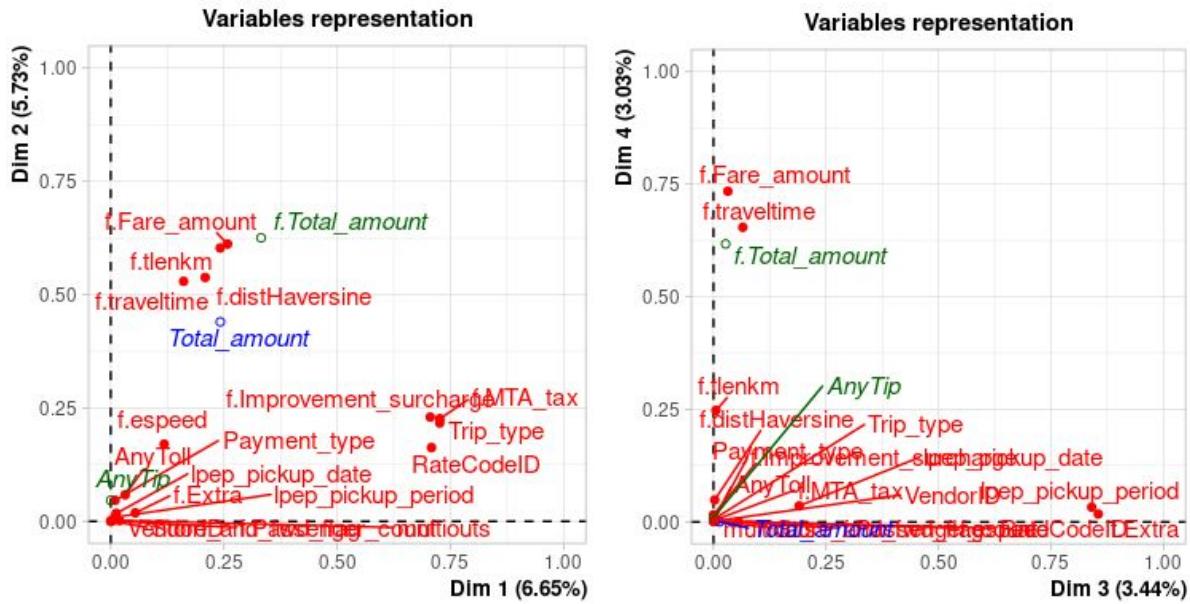
## Individuals



En las imágenes podemos observar los individuos reflejados en los planos factoriales. Si observamos el primer plano nos encontramos que los individuos están separados en 2 conjuntos, pero a pesar de eso los individuos que contribuyen se encuentran en un solo conjunto. En el segundo plano factorial podemos admirar cómo los individuos forman una especie efecto Guttman. También se puede observar que en el segundo plano factorial se forma dos conjuntos pero esta vez, estos dos conjuntos se encuentran más juntos.

## Categorical variables, supplementary numerical variables and supplementary categorical variables

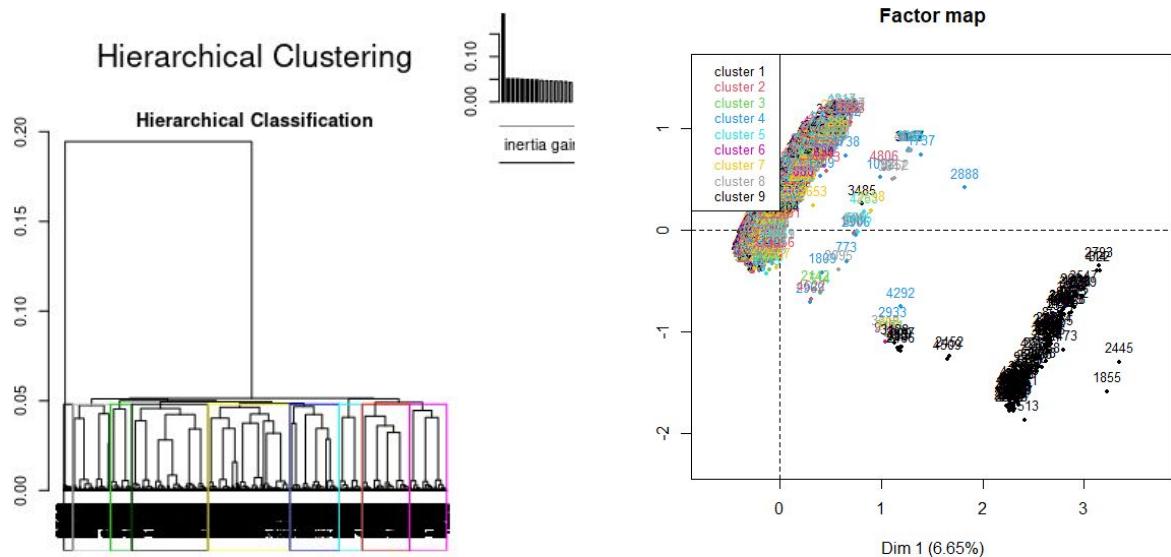


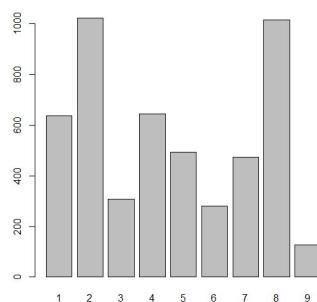


Como podemos ver nos vuelve a aparecer que el total amount está relacionado con el travel time tal como nos apareció haciendo CA. Además podemos observar cómo estas variables están muy relacionadas con las dimensiones 2 y 4 debido a su posición en los ejes.

A parte también podemos observar como la distancia del trayecto está relacionada con estas dos variables, hecho que tiene sentido ya que cuanto más largo sea un trayecto más tiempo hará falta para llevarlo a cabo y, por tanto, más caro será.

## Hierarchical Clustering (from MCA)





En este gráfico podemos observar la distribución de individuos entre cada uno de los clusters donde los clusters 2 y 8 son los que contienen un mayor número respecto a los demás. Y en el Factor map anterior podemos ver como la separación entre clusters no queda nada bien definida a excepción del cluster 9, el cual aparece muy distante del resto.

## Categorical variables which characterizes the clusters

```
res.hcpc$desc.var$test.chi2
```

```
##          p.value df
## RateCodeID      0.000000e+00  8
## f.MTA_tax       0.000000e+00  8
## f.Improvement_surcharge 0.000000e+00  8
## Trip_type       0.000000e+00  8
## lpep_pickup_date 0.000000e+00 240
## f.Total_amount   2.175017e-88 56
## f.Extra         1.081069e-56 16
## Payment_type     2.772605e-12 16
## multiouts        1.917731e-08  8
## f.tlenkm        3.940819e-07 16
## AnyTip          5.725674e-06  8
## f.espeed         1.363908e-05 16
## f.Fare_amount    1.610297e-04 24
## f.distHaversine  3.387161e-03 16
## lpep_pickup_period 1.879570e-02 24
## f.traveltime     2.752741e-02 32
```

Como vemos, las variables categóricas que caracterizan los clusters, según el test de chi cuadrado, serían las anteriores, de las cuales podríamos destacar debido a su p.valor el RateCodeID, las tasas MTA, el tipo de trayecto y la fecha de recogida principalmente.

## Numerical variables which characterizes the clusters

```
$`8`
```

```
v.test Mean in category Overall mean sd in category Overall sd p.value
```

```
Total_amount 2.358116      15.22199   14.54116   10.31011  10.30226 0.01836796
```

```
$`9`
```

```
v.test Mean in category Overall mean sd in category Overall sd p.value
```

```
Total_amount 5.861209      19.81014   14.54116   19.65805  10.30226 4.595097e-09
```

La variable de importe total podemos ver como aparece con una media superior respecto a la global en el cluster 8 así como en el 9, siendo en este más significativa la diferencia.

## Description of each cluster by the categories

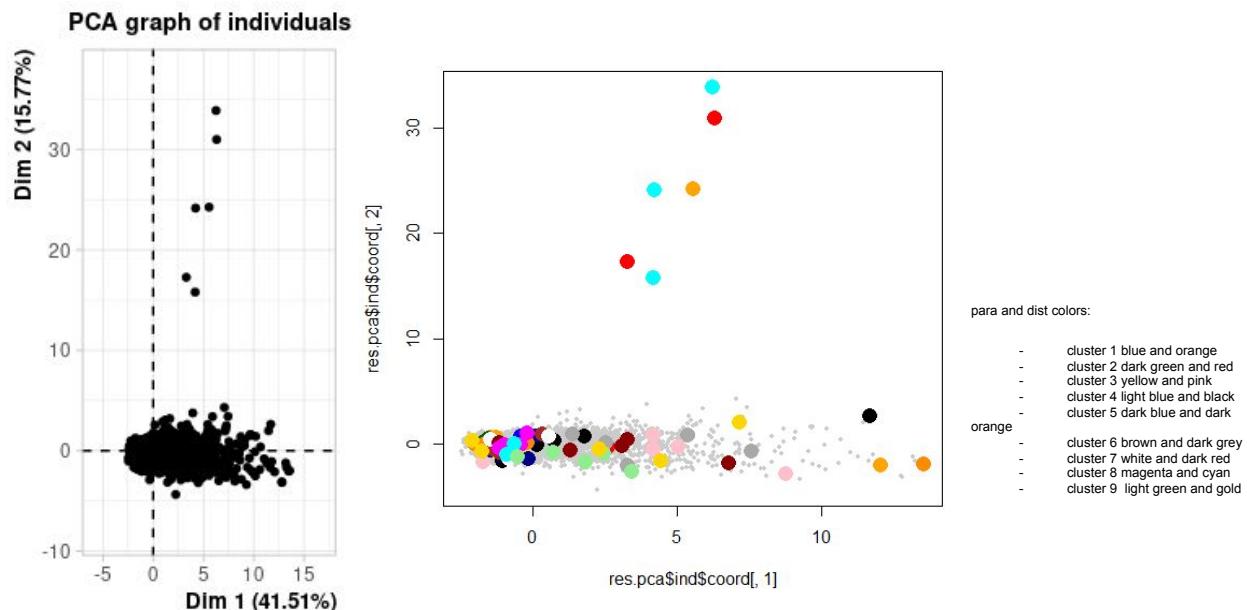
| \$'2'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
|---|-------------|-------------|--------------------|------------|--------|
| lpep_pickup_date=2016-01-22                         | 97.512438   | 30.7692308  | 4.02 7.751839e-180 | 28.595005  |        |
| lpep_pickup_date=2016-01-20                         | 96.835263   | 24.488808   | 3.16 1.770513e-180 | 24.876759  |        |
| lpep_pickup_date=2016-01-12                         | 97.742623   | 30.93948    | 1.54740e-134       | 29.742623  |        |
| lpep_pickup_date=2016-01-28                         | 99.285714   | 21.8210631  | 2.89 3.069388e-129 | 24.186721  |        |
| f.Extra=f.Extra-1                                   | 21.177945   | 26.5306122  | 15.96 2.115516e-13 | 7.341286   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 13.101604   | 100.0000000 | 97.24 5.131088e-09 | 5.842864   |        |
| f.MTA_tax=f.MTA_tax_YES                             | 13.99604    | 100.0000000 | 97.32 8.99570e-09  | 5.742864   |        |
| Trip_type=f.TripType_Street-Hail                    | 13.101604   | 100.0000000 | 97.34 1.482657e-08 | 4.482159   |        |
| RateCodeID=Standard rate                            | 13.073914   | 99.6860283  | 97.14 7.014536e-07 | 4.960561   |        |
| f.Speed=f.Speed-(1,25]                              | 13.341935   | 81.1616954  | 77.50 1.634648e-02 | 2.401896   |        |
| f.distHaversine=f.distHaversine-(10,26,1]           | 8.247423    | 2.5117739   | 3.88 4.746356e-02  | -1.982140  |        |
|   |             |             |                    |            |        |
| \$'2'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-30                         | 98.7394958  | 22.99412916 | 4.76 6.393184e-167 | 27.536661  |        |
| lpep_pickup_date=2016-01-15                         | 96.195652   | 17.31898239 | 3.68 6.455276e-117 | 22.985865  |        |
| lpep_pickup_date=2016-01-10                         | 98.816508   | 16.34950881 | 3.38 6.518664e-117 | 22.985441  |        |
| lpep_pickup_date=2016-01-29                         | 98.7394958  | 22.99412919 | 3.68 6.455276e-117 | 22.985441  |        |
| lpep_pickup_date=2016-01-19                         | 97.5164657  | 17.79647759 | 2.88 4.716456e-96  | 20.795878  |        |
| lpep_pickup_date=2016-01-05                         | 98.5401465  | 13.20939335 | 2.74 3.024228e-93  | 20.483456  |        |
| f.MTA_tax=f.MTA_tax_YES                             | 21.0028771  | 100.0000000 | 97.32 3.091449e-14 | 7.594432   |        |
| f.Total_amount=f.Total_amount-(8,11]                | 20.555504   | 100.0000000 | 97.54 4.109787e-13 | 7.251566   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 10.0000000  | 100.0000000 | 97.54 4.109787e-13 | 7.251566   |        |
| RateCodeID=Standard rate                            | 21.0000000  | 99.80430528 | 97.14 2.949846e-12 | 6.980893   |        |
| f.Fare_amount=f.Fare_amount-[5,6,9]                 | 23.3365477  | 23.67960667 | 20.74 1.019980e-02 | 2.569287   |        |
| f.distHaversine=f.distHaversine-[0,5]               | 21.0835989  | 84.14872798 | 81.58 1.636199e-02 | 2.400739   |        |
| f.Total_amount=f.Total_amount-(8,11]                | 22.5303295  | 25.4600000  | 23.32 4.020000e-02 | 1.334941   |        |
| f.Total_amount=f.Total_amount-(8,11]                | 18.471901   | 23.426145   | 18.42 2.024187e-02 | 0.62808    |        |
| f.TravelTime=f.TravelTime-(5,6,7,9)                 | 18.5389973  | 26.98082348 | 29.68 2.886766e-02 | 2.185289   |        |
| f.distHaversine=f.distHaversine-[5,10]              | 17.3314993  | 12.32876712 | 14.54 2.884340e-02 | 2.276044   |        |
| f.Fare_amount=f.Fare_amount-[14,5,71,5]             | 18.115354   | 21.81996086 | 24.62 1.893066e-02 | -2.346894  |        |
| \$'3'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-14                         | 99.698925   | 27.6827373  | 3.72 1.527403e-156 | 26.655937  |        |
| lpep_pickup_date=2016-01-03                         | 97.5272727  | 20.8000000  | 3.01 1.979503e-154 | 26.478690  |        |
| RateCodeID=Standard rate                            | 97.517429   | 21.461897   | 1.08 7.988389e-154 | 26.478690  |        |
| f.MTA_tax=f.MTA_tax_YES                             | 98.366056   | 16.662519   | 2.44 2.533776e-108 | 22.115266  |        |
| Trip_type=f.TripType_Street-Hail                    | 13.184335   | 100.000000  | 97.54 3.538842e-08 | 5.512444   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 97.566976   | 6.220840    | 0.82 2.877537e-35  | 12.392277  |        |
| lpep_pickup_date=2016-01-23                         | 97.566976   | 17.9600000  | 97.54 3.538842e-08 | 5.512444   |        |
| lpep_pickup_date=2016-01-18                         | 97.566976   | 17.9600000  | 97.54 3.538842e-08 | 5.512444   |        |
| lpep_pickup_date=2016-01-01                         | 97.566976   | 17.9600000  | 97.54 3.538842e-08 | 5.512444   |        |
| Trip_type=f.TripType_Street-Hail                    | 13.184335   | 100.000000  | 97.54 3.538842e-08 | 5.512444   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 99.698925   | 99.943437   | 97.24 8.419961e-06 | 4.454241   |        |
| f.Total_amount=f.Total_amount-(1,8)                 | 13.094583   | 99.911353   | 97.14 1.552561e-03 | 3.164675   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 13.471901   | 81.648523   | 77.94 1.373415e-02 | 2.464142   |        |
| f.Total_amount=f.Total_amount-(30,40)               | 13.922837   | 54.432348   | 50.28 2.412855e-02 | 2.255076   |        |
| f.TravelTime=f.TravelTime-[0,10]                    | 7.906977    | 2.643857    | 4.36 2.050218e-02  | 2.317029   |        |
| VendorID=f.Vendor-Mobile                            | 16.698996   | 18.351477   | 22.06 1.373415e-02 | 2.464142   |        |
| \$'3'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-08                         | 97.713986   | 34.6159846  | 3.59 1.853040e-178 | 28.483924  |        |
| lpep_pickup_date=2016-01-17                         | 97.023810   | 32.9959514  | 3.36 1.075258e-167 | 27.691239  |        |
| lpep_pickup_date=2016-01-27                         | 97.569876   | 32.3886640  | 3.28 9.766056e-166 | 27.437699  |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 10.169424   | 100.0000000 | 97.24 4.715089e-09 | 5.037504   |        |
| f.MTA_tax=f.MTA_tax_YES                             | 97.3296342  | 100.0000000 | 97.32 1.771046e-04 | 7.749617   |        |
| Trip_type=f.TripType_Street-Hail                    | 6.2948534   | 99.6753247  | 97.54 3.737761e-03 | 2.899485   |        |
| f.Total_amount=f.Total_amount-(1,8)                 | 7.7352472   | 31.4935065  | 25.08 3.725716e-02 | 2.621570   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 9.0000000   | 99.9800000  | 97.24 3.245262e-02 | 2.138993   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_NO  | 9.0000000   | 99.9800000  | 97.24 3.245262e-02 | 2.138993   |        |
| f.Total_amount=f.Total_amount-(18,20]               | 14.427861   | 5.8704453   | 4.02 3.579829e-02  | 2.099211   |        |
| \$'5'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-08                         | 97.713986   | 34.6159846  | 3.59 1.853040e-178 | 28.483924  |        |
| lpep_pickup_date=2016-01-17                         | 97.023810   | 32.9959514  | 3.36 1.075258e-167 | 27.691239  |        |
| lpep_pickup_date=2016-01-27                         | 97.569876   | 32.3886640  | 3.28 9.766056e-166 | 27.437699  |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 10.169424   | 100.0000000 | 97.24 4.715089e-09 | 5.037504   |        |
| f.MTA_tax=f.MTA_tax_YES                             | 97.3296342  | 100.0000000 | 97.32 1.771046e-04 | 7.749617   |        |
| Trip_type=f.TripType_Street-Hail                    | 6.2948534   | 99.6753247  | 97.54 3.737761e-03 | 2.899485   |        |
| f.Total_amount=f.Total_amount-(1,8)                 | 7.7352472   | 31.4935065  | 25.08 3.725716e-02 | 2.621570   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 9.0000000   | 99.9800000  | 97.24 3.245262e-02 | 2.138993   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_NO  | 9.0000000   | 99.9800000  | 97.24 3.245262e-02 | 2.138993   |        |
| f.Total_amount=f.Total_amount-(18,20]               | 14.427861   | 5.8704453   | 4.02 3.579829e-02  | 2.099211   |        |
| \$'6'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-26                         | 99.2957746  | 50.3517429  | 3.41 4.922849e-193 | 29.373459  |        |
| lpep_pickup_date=2016-01-06                         | 96.527778   | 49.6428571  | 2.88 1.775953e-183 | 28.886273  |        |
| f.Extra=f.Extra-1                                   | 9.1478697   | 26.0714248  | 15.96 8.253107e-06 | 4.458511   |        |
| RateCodeID=Standard rate                            | 5.7600514   | 100.0000000 | 97.54 3.538842e-08 | 5.512444   |        |
| f.TravelTime=f.TravelTime-[0,10]                    | 5.7431234   | 100.0000000 | 97.54 3.538842e-08 | 5.512444   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 5.7383793   | 99.6428571  | 97.24 3.272953e-03 | 2.940952   |        |
| f.MTA_tax=f.MTA_tax_YES                             | 5.7336621   | 99.6428571  | 97.32 4.047670e-03 | 2.874423   |        |
| \$'7'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-16                         | 97.797356   | 21.87192111 | 4.54 2.303963e-151 | 26.467352  |        |
| lpep_pickup_date=2016-01-02                         | 97.756756   | 21.87192111 | 4.54 2.303963e-151 | 26.467352  |        |
| lpep_pickup_date=2016-01-24                         | 95.4285714  | 16.62328197 | 3.56 5.768208e-127 | 23.119096  |        |
| f.Extra=f.Extra-0                                   | 98.0132459  | 14.58128879 | 3.02 1.156642e-101 | 21.466724  |        |
| Trip_type=f.TripType_Street-Hail                    | 96.7948718  | 14.87684729 | 3.12 5.742471e-101 | 21.331902  |        |
| f.TravelTime=f.TravelTime-[0,10]                    | 97.7789116  | 14.70000000 | 3.01 5.742471e-101 | 20.873131  |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 20.855190   | 100.0000000 | 97.54 3.538842e-08 | 5.512444   |        |
| f.MTA_tax=f.MTA_tax_YES                             | 20.8384710  | 99.90147783 | 97.32 4.165677e-12 | 7.077699   |        |
| RateCodeID=Standard rate                            | 20.7535516  | 99.31034480 | 97.14 1.276323e-01 | 5.282213   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 22.3741240  | 35.00492011 | 29.68 9.911001e-03 | 2.578919   |        |
| \$'8'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| lpep_pickup_date=2016-01-01                         | 97.797356   | 21.87192111 | 4.54 2.303963e-151 | 26.467352  |        |
| lpep_pickup_date=2016-01-31                         | 97.797356   | 21.87192111 | 4.54 2.303963e-151 | 26.467352  |        |
| lpep_pickup_date=2016-01-25                         | 95.4285714  | 16.62328197 | 3.56 5.768208e-127 | 23.119096  |        |
| f.Extra=f.Extra-1                                   | 98.0132459  | 14.58128879 | 3.02 1.156642e-101 | 21.466724  |        |
| lpep_pickup_date=2016-01-13                         | 96.7948718  | 14.87684729 | 3.12 5.742471e-101 | 21.331902  |        |
| lpep_pickup_date=2016-01-04                         | 97.7789116  | 14.70000000 | 3.01 5.742471e-101 | 20.873131  |        |
| Trip_type=f.TripType_Street-Hail                    | 20.855190   | 100.0000000 | 97.54 3.538842e-08 | 5.512444   |        |
| f.TravelTime=f.TravelTime-[0,10]                    | 20.8384710  | 99.90147783 | 97.32 4.165677e-12 | 7.077699   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 20.855190   | 99.90147783 | 97.32 4.165677e-12 | 7.077699   |        |
| f.MTA_tax=f.MTA_tax_YES                             | 20.8384710  | 99.90147783 | 97.32 4.165677e-12 | 7.077699   |        |
| RateCodeID=Standard rate                            | 20.7535516  | 99.31034480 | 97.14 1.276323e-01 | 5.282213   |        |
| f.Improvement_surcharge=f.Improvement_surcharge_YES | 22.3741240  | 35.00492011 | 29.68 9.911001e-03 | 2.578919   |        |
| \$'9'   | Cla/Mod     | Mod/Cla     | Global             | p.value    | v.test |
| f.MTA_tax=f.MTA_tax_NO                              | 95.5228806  | 100.00000   | 2.68 3.453662e-248 | 33.3636857 |        |
| f.Improvement_surcharge=f.Improvement_surcharge_NO  | 92.75362319 | 100.00000   | 2.76 3.342807e-243 | 33.304324  |        |
| Trip_type=f.Dispatch                                | 99.18699187 | 53.31250    | 2.46 1.549633e-236 | 32.849636  |        |
| RateCodeID=Others                                   | 93.1970000  | 93.73600    | 2.04 1.620000e-208 | 30.67558   |        |
| f.Extra=f.Extra-0                                   | 5.45218520  | 29.310000   | 40.22 2.462200e-00 | 1.01502    |        |
| f.Total_amount=f.Total_amount-[0,1]                 | 83.3333333  | 11.71875    | 0.36 4.444113e-22  | 9.660334   |        |

Los clusters del 1 al 8 presentan una sobrerepresentación muy clara de ciertas fechas diferentes según el cluster, constituyendo casi la totalidad de observaciones globales de dichas fechas pertenecientes a cada uno de los clusters.

Por otro lado, en el cluster 9 vemos una sobrerepresentación de trayectos donde no aparece la tasa MTA ni un sobrecargo de mejora, además también lo están el tipo de trayecto Dispatch así como el RateCodeID de otros. A excepción del RatecodeID que supone un 83% del total de observaciones, el resto de categorías superan una

representación del 90% de la totalidad de las observaciones. Es decir, más del 90% de los trayectos que tienen estas categorías, están dentro del cluster 9.

## Description of the clusters by the individuals



Los elementos más distintivos del clustering podemos ver que serían los del cluster 1,2,4,5 y 8 al ser los más alejados de la formación principal de clusters dentro del gráfico.

El del cluster 5 por ejemplo(naranja oscuro) se caracteriza por ser un trayecto con una velocidad efectiva muy alta y una distancia a la par con un tiempo de trayecto muy superior a la media. Por los mismos valores se rige el individuo distintivo del cluster 4(negro) situado más a la derecha.

Los individuos distintivos del cluster 8, por otro lado, vienen caracterizados por tener una velocidad efectiva muy alta, sin ser la distancia de estos muy elevada. Además tenemos como el importe de estos es razonablemente elevado y su RateCodeID pertenece al de otros.

Por lo que hace a los del cluster 1, uno de ellos tiene un tiempo de trayecto excesivamente largo aún teniendo el mismo rango de distancia de trayecto que el otro. Este segundo pese a ser de una duración más reducida parece tener un coste muy superior al primero teniendo los dos el mismo rango de tarifa del trayecto y no haber pasado por ningún peaje. Esto puede deberse a la presencia de una propina, aunque esta debería ser muy alta para marcar una diferencia tan significativa del individuo.

Referente a los individuos a destacar del cluster 2, tenemos dos trayectos cuya duración es muy larga siendo ínfimas sus velocidades efectivas y sus distancias. Tienen un coste final muy similar y los diferencian que uno es efectuado por la noche y otro por el medio día.

## Hierarchical tree result

### Ratio between within inertias

```
res.hcpc$call$t$quot[1:res.hcpc$call$t$nb.clust]  
## [1] 0.9775104 0.9771113 0.9766060 0.9765191 0.9762269 0.9758285 0.9754768  
## [8] 0.9754845 NA
```

Si vemos la relación entre inercias dentro de un mismo cluster vemos que es bastante grande, es decir la inercia intra cluster es elevada.

### Inertia gain

```
res.hcpc$call$t$inert.gain[1:res.hcpc$call$t$nb.clust]  
## [1] 0.19449854 0.05140255 0.05113830 0.05107095 0.05006133 0.04949422 0.04912733  
## [8] 0.04863734 0.04742985
```

Si vemos la inertia gain vemos que es bastante baja des del cluster 1 al 9. La pérdida de pasar de n clusters a n+1 clusters es bastante baja.

### Partition quality

```
(res.hcpc$call$t$within[1]-res.hcpc$call$t$within[res.hcpc$call$t$nb.clust])/res.hcpc$call$t$within[1]  
## [1] 0.2199214
```

Podemos ver que cogiendo 9 cluster obtenemos una calidad del 21.99%, lo que significa que el número de cluster óptimos debería de ser superior a 9. Según hemos observado, para obtener una calidad del 80% deberíamos coger 200 clusters aproximadamente. Estos resultados demuestran que cada cluster divide significa una pequeña población de la muestra y por lo tanto, el clustering no es del todo correcto.

# Deliverable III

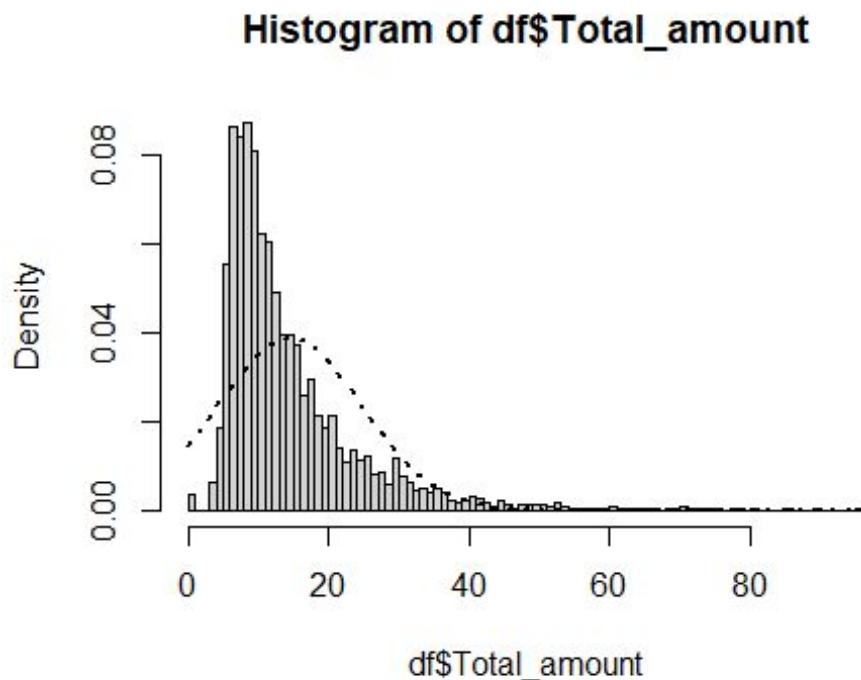
Carles Capilla Cànovas  
Jesús Molina Roldán

|   |           |
|---|-----------|
| <b>Multiple Linear Regression Model</b>                           | <b>3</b>  |
| Multivariate Analysis   | 3         |
| Use explanatory numeric variables                                 | 4         |
| Transformations   | 7         |
| Diagnostics Linear Regression using explanatory numeric variables | 8         |
| Using factors as explanatory variables                            | 22        |
| Clear effects   | 25        |
| Dirty effects   | 27        |
| Interactions  | 28        |
| <b>Binary Regression Model</b>                                    | <b>32</b> |
| Use explanatory numeric variables                                 | 32        |
| Consider factors and interactions                                 | 34        |
| <b>Final Diagnostics</b>  | <b>37</b> |
| Residus   | 37        |
| Observacions potencialment influents                              | 37        |
| Influent data   | 37        |
| Confusion Table   | 40        |

## Multiple Linear Regression Model

En esta sección se construye un modelo lineal para una variable target numérica, Total\_amount.

### Multivariate Analysis



Antes de poder hacer la modelización hay que mirar si la variable respuesta sigue una distribución normal. Para realizar la comprobación utilizamos diferentes indicadores. Si vemos el histograma del Total amount junto a la correcta distribución normal, vemos que los histogramas no se solapan, lo que significa que tenemos que normalizar la variable.

```
shapiro.test(df$Total_amount)
## W = 0.76853, p-value < 2.2e-16
```

Si realizamos el test de normalidad Shapiro-Wilk, observamos que la H<sub>0</sub> puede ser rechazada al mostrar un p-value muy inferior a 0.05. Lo que significa que los datos no siguen una distribución normal.

```
skewness(df$Total_amount)
## [1] 2.485124
```

Si realizamos un test de simetría como Skewness, vemos que nos devuelve un valor diferente a 0. Por lo tanto, los datos son asimétricos y como consecuencia, no siguen una distribución normal. También podemos ver que el valor es superior a 0 por lo tanto los

los datos son right-skewed lo que significa que las observaciones presentan una larga cola de observaciones por la derecha.

```
kurtosis(df$Total_amount)
## [1] 11.95862
```

Si computamos la curtosis de la variable Total amount, observamos que es superior a 3, lo que significa que no sigue una distribución normal. Tras ver todos estos argumentos, vemos que los datos no siguen una distribución normal.

Es por eso que el método más apropiado para calcular la correlación deba ser a partir de Spearman.

```
> round(cor(df[,c("Total_amount",vars_cexp)], method="pearson"),dig=2)
      Total_amount Passenger_count tlenkm Fare_amount espeed Tip_amount Tolls_amount lppep_pickup_time travelttime distHaversine
Total_amount          1.00        0.02    0.89       0.96   0.43      0.62     0.31      -0.04      0.46      0.81
Passenger_count       0.02        1.00        0.02    0.01      0.01      0.02      0.02      -0.02      0.00      0.01
tlenkm                 0.89        0.02        1.00       0.90   0.58      0.45     0.27      -0.06      0.45      0.90
Fare_amount            0.96        0.02        0.90       1.00   0.44      0.48     0.24      -0.05      0.47      0.83
espeed                  0.43        0.01        0.58       0.44   1.00      0.23     0.20      -0.12      -0.01      0.56
Tip_amount              0.62       -0.01        0.45       0.48   0.23      1.00      0.19      -0.02      0.23      0.44
Tolls_amount             0.31        0.02        0.27       0.24   0.20      0.19     1.00      -0.03      0.08      0.27
lppep_pickup_time      -0.04       -0.02       -0.06      -0.05   -0.12      -0.02      -0.03      1.00      0.40      -0.06
travelttime              0.46        0.00        0.45       0.47   -0.01      0.23     0.08      0.40      1.00      0.41
distHaversine            0.81       0.01        0.90       0.83   0.56      0.44     0.27      -0.06      0.41      1.00
```

Si vemos los resultados, observamos que la distancia recorrida, la tarifa abonada, la duración del viaje y la distancia Haversine son las variables más correlacionadas con el target numérico, Total amount.

## Use explanatory numeric variables

Para el modelo inicial podríamos elegir aquellas variables más correlacionadas. A pesar de todo, como tenemos pocas variables explicativas, decidimos coger todas las variables explicativas.

```
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.383185  0.102835 13.451 < 2e-16 ***
## Passenger_count      0.031776  0.025303  1.256  0.20923
## tlenkm                0.224812  0.017958 12.519 < 2e-16 ***
## Fare_amount           0.951982  0.007468 127.483 < 2e-16 ***
## espeed                -0.015837 0.003893 -4.069 4.80e-05 ***
## Tip_amount            1.013289  0.015632 64.821 < 2e-16 ***
## Tolls_amount          1.009723  0.039853 25.336 < 2e-16 ***
## lppep_pickup_time    0.011693  0.003936  2.971  0.00299 **
## travelttime           -0.002567 0.002172 -1.182  0.23744
## distHaversine         -0.147172 0.020475 -7.188 7.54e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.893 on 4990 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9663
## F-statistic: 1.591e+04 on 9 and 4990 DF,  p-value: < 2.2e-16
```

Como podemos ver el modelo tiene una explicatividad del 96'63% de la variabilidad del target. A pesar de todo, hay variables como el passenger count y el travelttime que no son significativas ya que, como podemos ver en el test Anova tienen una proporción de la distribución t superior a 0.05.

```
> vif(m) # Check association between explanatory vars
   Passenger_count      tlenkm      Fare_amount      espeed      Tip_amount      Tolls_amount      lppep_pickup_time      travelttime      distHaversine
1.001878      9.657089      5.946645      1.839423      1.322309      1.094175      1.320458      1.905585      5.344283
```

A partir de la variance inflation factors vemos la asociación entre las variables explicativas. Podemos observar que las variables tlenkm, Fare amount y la distancia Haversine están muy correlacionadas. También podemos ver que la distancia fare amount y la distancia Haversine tienen un valor similar de inflación lo que nos hace creer que hay una gran relación entre ellas.

Tras haber analizado todas las variables que no aportan mucho en el modelo decidimos eliminar las variables passenger count y el travelttime de este ya que eran las dos primeras variables que elimina el vif y las rechazadas por la hipótesis nula. Podríamos eliminar la variable Fare\_amount, ya que es prácticamente la misma que nuestro target, sin embargo decidimos no eliminarla. Aún así esperaremos a ver los resultados que nos muestra el step para acabar de concretar si eliminamos del modelo alguna otra variable.

```
m1 <- step( m, k=log(nrow(df)) )

## Step:  AIC=6439.54
## Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount +
##      distHaversine
##
##              Df Sum of Sq   RSS     AIC
## <none>            17911 6439.5
## - espeed          1      63 17974 6448.7
## - distHaversine  1     192 18103 6484.3
## - tlenkm          1      565 18476 6586.3
## - Tolls_amount    1     2307 20218 7036.7
## - Tip_amount      1     15047 32958 9480.1
## - Fare_amount     1     58799 76710 13704.1

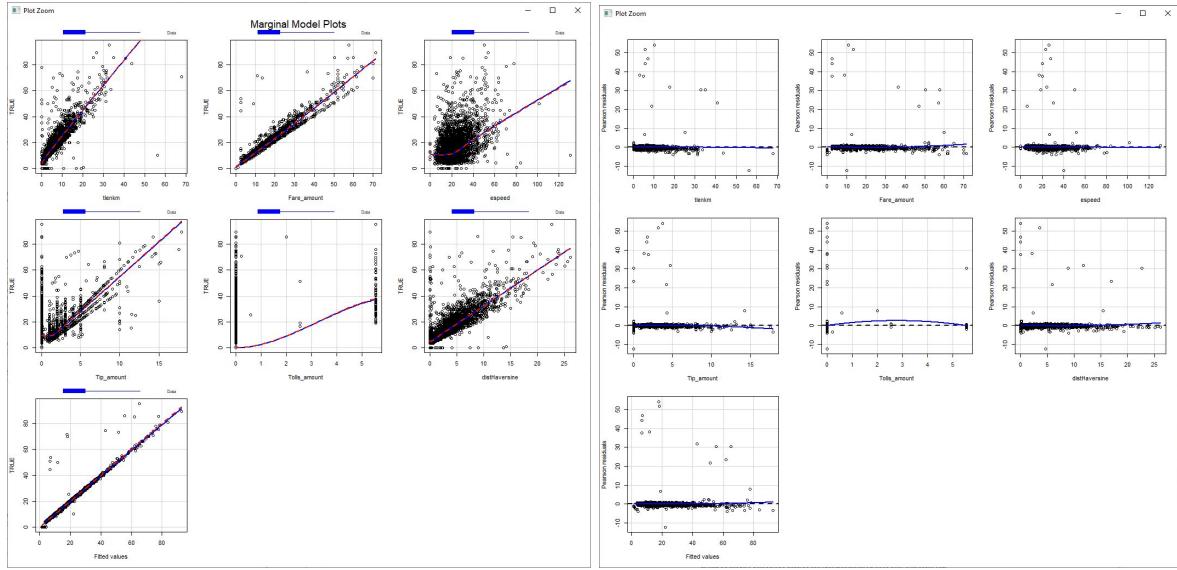
summary(m1)
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.574884   0.079479 19.815 < 2e-16 ***
## tlenkm       0.222321   0.017717 12.548 < 2e-16 ***
## Fare_amount  0.950798   0.007427 128.027 < 2e-16 ***
## espeed      -0.015307   0.003644 -4.200 2.71e-05 ***
## Tip_amount   1.012951   0.015640 64.765 < 2e-16 ***
## Tolls_amount 1.010725   0.039858 25.358 < 2e-16 ***
## distHaversine -0.149401  0.020423 -7.315 2.98e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.894 on 4993 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9662
## F-statistic: 2.382e+04 on 6 and 4993 DF, p-value: < 2.2e-16
```

Como podemos ver que si aplicamos el Stepwise Algorithm Akaike según el Akaike information criterion (AIC), vemos que nos elimina las variables explicativas travelttime, Passenger\_count y lpep\_pickup\_time. La calidad del criterio de AIC se nos queda en 6439.54.

Al realizar un summary del modelo resultante vemos como la explicación de la variabilidad del target, una vez eliminadas las variables del anterior modelo, se mantiene casi igual pasando de 96'63 a 96'62. Consideramos el nuevo modelo m1 ya que obtenemos la misma explicación prácticamente, simplificando en 3 variables el nuevo modelo.

El modelo resultante tendría la siguiente predicción:

$$Y = 1.57 + 0.22tlenkm + 0.95Fare\_amount - 0.015espeed + 1.01Tip\_amount + 1.01Tolls\_amount - 0.15distHaversine$$

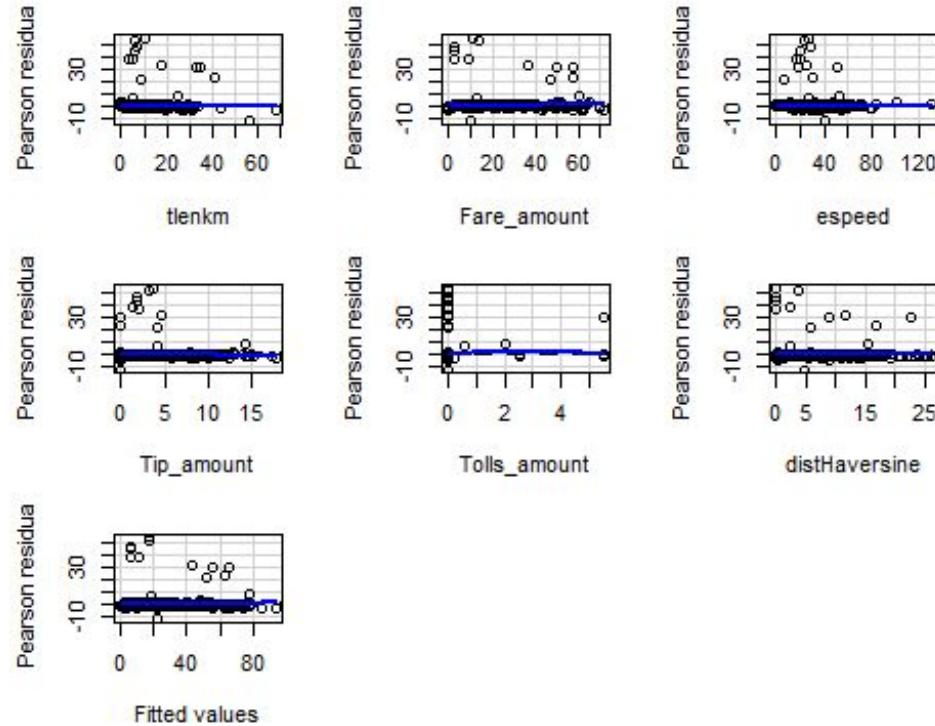


En la imagen superior podemos ver los resultados que obtenemos del modelo de regresión. Como vemos la distancia recorrida en km, el importe de la tarifa y la distancia de Haversine siguen una regresión lineal perfecta. Aun así podemos ver como las observaciones no se distribuyen de una manera homogénea por todo el rango de valores posibles sin llegar, por eso, a mostrar patrones que puedan hacernos considerar tratarlas. A destacar sobretodo el Tolls amount que al ser una variable con valores enteros y más cercana a ser considerada un factor, vemos como no se ajustan sus observaciones al modelo de regresión pero es perfectamente normal debido a sus propiedades.

Al disponer de una explicación de la variabilidad del target por el modelo tan alta y una igualdad entre la predicción de las variables y las observaciones que define el modelo, talvez no tendría demasiado sentido realizar transformaciones para aumentar este Multiple R-squared pero si para reducir el Residual Standard error del modelo.

Si observamos los residuos encontramos algún valor negativo y varios valores fuera de la linea residual. Además encontramos que los residuos no muestran una tendencia normalizada debido a las observaciones de la derecha del gráfico, muy alejadas de lo que sería considerada la normal del target. Una de las modificaciones que hicimos es modificar la escala de la variable target a logarítmica pero no observamos mejoras al respecto.

## Transformations



```
m2 <- lm(Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + poly(Tolls_amount, 2) +
distHaversine, data=df)
summary(m2)

##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##     poly(Tolls_amount, 2) + distHaversine, data = df)
##
## Residuals:
##      Min        1Q        Median       3Q        Max 
## -12.163   -0.394   -0.050    0.184   54.145 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.676645  0.079923 20.978 < 2e-16 ***
## tlenkm      0.220981  0.017712 12.476 < 2e-16 ***
## Fare_amount  0.950796  0.007422 128.111 < 2e-16 ***
## espeed      -0.015660  0.003644 -4.297 1.76e-05 ***
## Tip_amount   1.011293  0.015642 64.653 < 2e-16 ***
## poly(Tolls_amount, 2)1 50.297952  1.978935 25.417 < 2e-16 ***
## poly(Tolls_amount, 2)2 -5.222973  1.898610 -2.751  0.00596 ** 
## distHaversine -0.147326  0.020424 -7.213 6.27e-13 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.893 on 4992 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9663 
## F-statistic: 2.045e+04 on 7 and 4992 DF,  p-value: < 2.2e-16
```

```
##          Test stat Pr(>|Test stat|)  
## tlenkm      -0.6848      0.4934854
```

```

## Fare_amount      3.8600    0.0001148 ***
## espeed         -0.3472    0.7284631
## Tip_amount     -3.4021    0.0006740 ***
## poly(Tolls_amount, 2)  2.9020    0.0037236 **
## distHaversine   1.7407    0.0817347 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m1,m2)

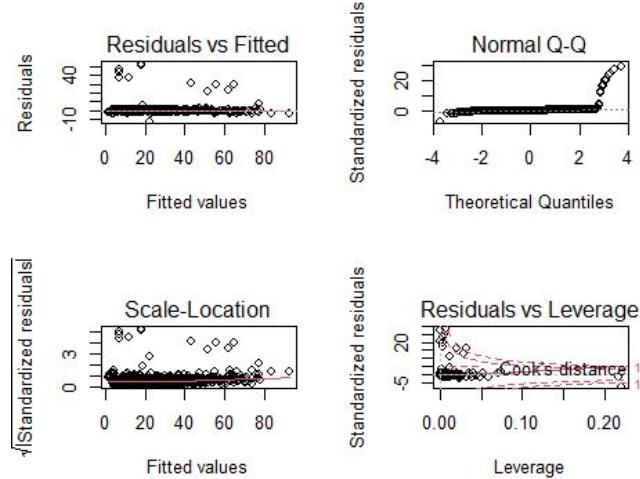
## Analysis of Variance Table
##
## Model 1: Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount +
##           distHaversine
## Model 2: Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + poly(Tolls_amount,
##           2) + distHaversine
##   Res.Df   RSS Df Sum of Sq    F   Pr(>F)
## 1   4993 17911
## 2   4992 17884  1   27.112 7.5677 0.005964 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Al ver los gráficos obtenidos por el residualplots vemos como la única variable que sufre un patrón respecto el smoother sería la de tolls amount así que probaremos a aplicarle alguna transformación.

Una vez usado un polinomio de grado dos para esta transformación vemos como la línea de los residuos se ajusta más al smoother pero, al analizar el summary vemos como la explicación de la variabilidad no varía apenas y al usar el anova(m1,m2) vemos como no debemos considerar esta transformación.

## Diagnostics Linear Regression using explanatory numeric variables



Si observamos los residuos del m1 encontramos algún valor negativo y varios valores fuera de la línea residual. Además encontramos que los residuos no muestran una tendencia normalizada debido a las observaciones de la derecha del gráfico, muy alejadas de lo que sería considerada la normal del target. Una de las modificaciones que hicimos es modificar la escala de la variable target a logarítmica pero no observamos mejoras al respecto.

```

1 <- which(df$Total_amount == 0 )
df[1,'Total_amount'] <- 0.0001
m3 <- lm( log(Total_amount) ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount+ distHaversine,
data=df)
summary(m3)

##
## Call:
## lm(formula = log(Total_amount) ~ tlenkm + Fare_amount + espeed +
##     Tip_amount + Tolls_amount + distHaversine, data = df)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -10.8307 -0.0533  0.0871  0.1761  3.3071 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.562062  0.027989 55.810 <2e-16 ***
## tlenkm      -0.064713  0.006239 -10.372 <2e-16 ***
## Fare_amount  0.083073  0.002615 31.764 <2e-16 ***
## espeed       0.003798  0.001283  2.959  0.0031 **  
## Tip_amount   0.052982  0.005508  9.619 <2e-16 *** 
## Tolls_amount 0.017401  0.014036  1.240  0.2151    
## distHaversine 0.016996  0.007192  2.363  0.0182 *   
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.667 on 4993 degrees of freedom
## Multiple R-squared:  0.4343, Adjusted R-squared:  0.4336 
## F-statistic: 638.8 on 6 and 4993 DF,  p-value: < 2.2e-16

```

Con el objetivo de mejorar el modelo y buscar la normalidad de la variable target Total amount, le aplicamos el logaritmo. Aunque antes de aplicarlo debemos eliminar las observaciones que contengan un 0 y es por eso que les asignamos un valor de 0.0001.

Como podemos ver en el summary las variables espeed, Tolls\_amount y dist\_haversine, debido a su p\_value deberían ser eliminadas del modelo ya que este es muy superior a 0.05, es decir que no aportan información significativa al mismo. Aún así, vemos que el modelo con el logaritmo del importe total solo explica una variabilidad de cerca del 43'43%. Probaremos a eliminar del modelo las variables mencionadas pero esto como mucho mantendrá la explicación de la variabilidad del target.

```

m4 <- lm( log(Total_amount) ~ tlenkm + Fare_amount + espeed + Tip_amount, data=df)
summary(m4)

##
## Call:
## lm(formula = log(Total_amount) ~ tlenkm + Fare_amount + espeed +
##     Tip_amount, data = df)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -10.8336 -0.0544  0.0861  0.1790  2.8938 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.556781  0.027866 55.866 < 2e-16 ***
## tlenkm      -0.056365  0.005284 -10.667 < 2e-16 ***
## Fare_amount  0.083872  0.002594 32.331 < 2e-16 *** 
## espeed       0.004358  0.001267  3.439  0.000588 *** 
## Tip_amount   0.054492  0.005479  9.946 < 2e-16 *** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

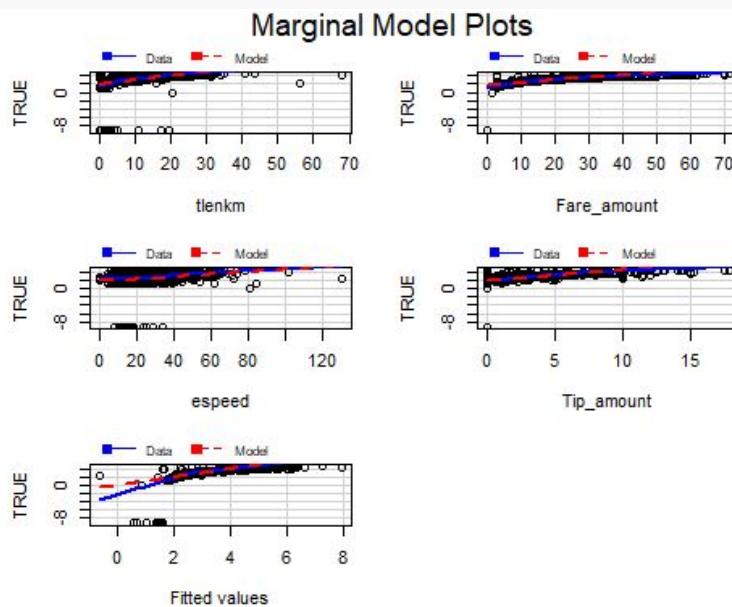
```

## Residual standard error: 0.6673 on 4995 degrees of freedom
## Multiple R-squared:  0.4334, Adjusted R-squared:  0.433
## F-statistic: 955.3 on 4 and 4995 DF,  p-value: < 2.2e-16

Anova(m4)

## Anova Table (Type II tests)
##
## Response: log(Total_amount)
##          Sum Sq Df  F value    Pr(>F)
## tlenkm     50.68  1 113.793 < 2.2e-16 ***
## Fare_amount 465.54  1 1045.325 < 2.2e-16 ***
## espeed      5.27  1   11.828 0.0005882 ***
## Tip_amount   44.05  1   98.918 < 2.2e-16 ***
## Residuals 2224.53 4995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```



Tras realizar las mejoras vemos que el modelo reduce el residual standard error respecto a m1. Aun así seguimos teniendo en el modelo m1 una mejor explicación de la variabilidad del target, lo que resulta en un mejor modelo del mismo.

Además vemos como gracias al plot del marginalModel como todas las predicciones de color azul ya no siguen la distribución del modelo lo que nos hace descartar estas transformaciones.

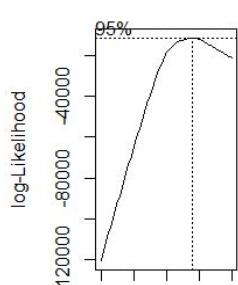
```

BIC(m1,m2,m4)

##      df      BIC
## m1  8 20637.44
## m2  9 20638.39
## m4  6 10191.02

```

Si calculamos Akaike en los modelos, vemos también que m1 es mejor en comparación a los otros.



Si aplicamos una Box-Cox power transformation a nuestros datos, vemos que con el primer modelo tiene un parámetro inferior a 1 pero

casi igual, lo que sugiere que no deberíamos aplicar ninguna transformación, o como mucho probar con una raíz cuadrada ya que ya hemos descartado el logaritmo.

```
m4<-lm(sqrt(Total_amount) ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount+
distHaversine,data=df)
summary(m4)

##
## Call:
## lm(formula = sqrt(Total_amount) ~ tlenkm + Fare_amount + espeed +
##     Tip_amount + Tolls_amount + distHaversine, data = df)
##
## Residuals:
##      Min      1Q  Median      3Q      Max 
## -2.2810 -0.1128  0.0402  0.1506  4.8103 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.1663437  0.0124569 173.908 < 2e-16 ***
## tlenkm      -0.0079799  0.0027769 -2.874  0.00407 **  
## Fare_amount   0.1105970  0.0011640  95.018 < 2e-16 ***
## espeed      -0.0015027  0.0005712 -2.631  0.00854 **  
## Tip_amount    0.1083534  0.0024513  44.202 < 2e-16 ***
## Tolls_amount   0.0703380  0.0062470  11.259 < 2e-16 ***
## distHaversine 0.0216390  0.0032009   6.760 1.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2968 on 4993 degrees of freedom
## Multiple R-squared:  0.9339, Adjusted R-squared:  0.9338 
## F-statistic: 1.175e+04 on 6 and 4993 DF,  p-value: < 2.2e-16

summary(m1)

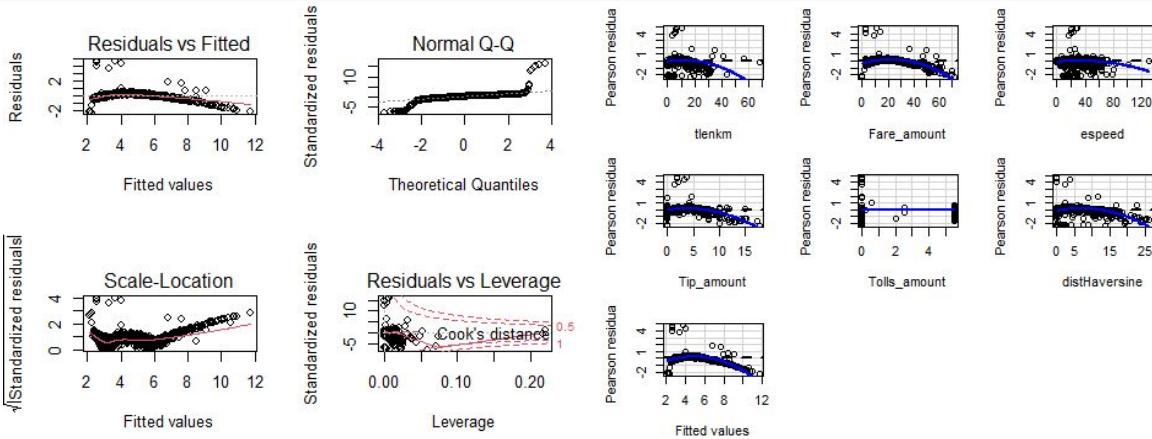
##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##     Tolls_amount + distHaversine, data = df[, c("Total_amount",
##     vars_cexp)])
##
## Residuals:
##      Min      1Q  Median      3Q      Max 
## -12.235 -0.393 -0.055   0.183  54.123 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.574884  0.079479 19.815 < 2e-16 ***
## tlenkm      0.222321  0.017717 12.548 < 2e-16 ***  
## Fare_amount  0.950798  0.007427 128.027 < 2e-16 ***
## espeed      -0.015307  0.003644 -4.200 2.71e-05 ***
## Tip_amount    1.012951  0.015640  64.765 < 2e-16 ***
## Tolls_amount   1.010725  0.039858  25.358 < 2e-16 ***  
## distHaversine -0.149401  0.020423 -7.315 2.98e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.894 on 4993 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9662 
## F-statistic: 2.382e+04 on 6 and 4993 DF,  p-value: < 2.2e-16
```

### Anova(m4)

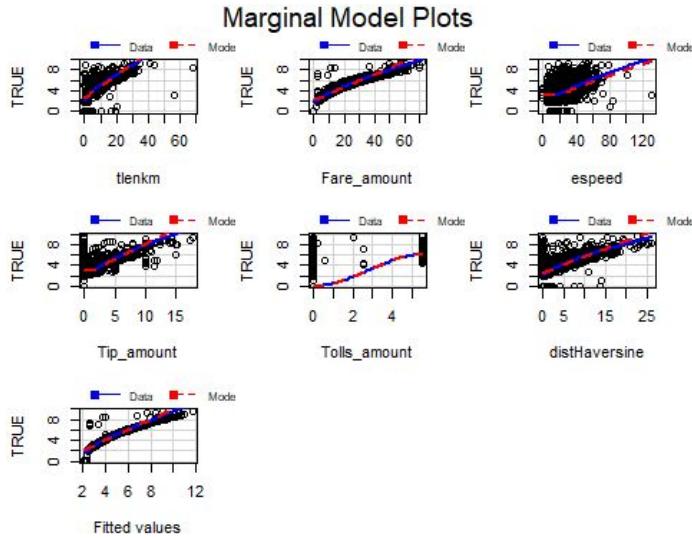
```
## Anova Table (Type II tests)
##
## Response: sqrt>Total_amount
##          Sum Sq Df F value    Pr(>F)
## tlenkm      0.73  1 8.2583 0.004074 **
## Fare_amount 795.57  1 9028.3324 < 2.2e-16 ***
## espeed       0.61  1 6.9212 0.008544 **
## Tip_amount   172.17  1 1953.7944 < 2.2e-16 ***
## Tolls_amount 11.17  1 126.7751 < 2.2e-16 ***
## distHaversine 4.03  1 45.7002 1.536e-11 ***
## Residuals   439.98 4993
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(m4)
```

```
##          tlenkm  Fare_amount      espeed     Tip_amount  Tolls_amount
##         9.385882  5.872915  1.609990  1.321788  1.092833
## distHaversine
##        5.309666
```



```
##          Test stat Pr(>|Test stat|)
## tlenkm      -30.7479 <2e-16 ***
## Fare_amount   -63.0933 <2e-16 ***
## espeed       -8.6098 <2e-16 ***
## Tip_amount    -33.4372 <2e-16 ***
## Tolls_amount    0.8415 0.4001
## distHaversine -40.1881 <2e-16 ***
## Tukey test     -67.4183 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Al probar de aplicar la raíz cuadrada a la variable Total\_amount vemos como la explicación de la variabilidad se reduce a 93.39%, lo que al haber complicado el modelo, nos indicaría no escogerlo como óptimo. Además vemos que para el residualplots las predicciones dejan de ajustarse a sus smoother creando patrones de términos cuadráticos, consecuencia de haber aplicado la raíz en el target.

Una vez aplicado en vif, por eso, vemos cómo Fare Amount y distHaversine están claramente correlacionadas así que probaremos a descartar una de ellas en el siguiente modelo.

```
m5<-lm(Total_amount ~ tlenkm + espeed + Tip_amount + Tolls_amount+ distHaversine,data=df)
summary(m5)

##
## Call:
## lm(formula = Total_amount ~ tlenkm + espeed + Tip_amount + Tolls_amount +
##     distHaversine, data = df)
##
## Residuals:
##   Min     1Q   Median     3Q    Max 
## -88.120 -1.031  -0.427   0.343  59.080 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.90624   0.14008  49.301 < 2e-16 ***
## tlenkm      1.69563   0.02788  60.827 < 2e-16 ***
## espeed     -0.11978   0.00735 -16.297 < 2e-16 ***
## Tip_amount   1.35734   0.03188  42.574 < 2e-16 ***
## Tolls_amount  0.96001   0.08247  11.640 < 2e-16 ***
## distHaversine 0.19326   0.04190   4.613 4.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.919 on 4994 degrees of freedom
## Multiple R-squared:  0.8555, Adjusted R-squared:  0.8553 
## F-statistic:  5911 on 5 and 4994 DF,  p-value: < 2.2e-16

m6<-lm(Total_amount ~ tlenkm +Fare_amount + espeed + Tip_amount + Tolls_amount,data=df)
summary(m6)

##
```

```

## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##     Tolls_amount, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -8.880 -0.404 -0.038  0.162 55.026 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.588970  0.079873 19.894 < 2e-16 ***
## tlenkm      0.153818  0.015119 10.174 < 2e-16 *** 
## Fare_amount  0.943678  0.007401 127.505 < 2e-16 *** 
## espeed      -0.019291  0.003622 -5.326 1.05e-07 *** 
## Tip_amount   1.005206  0.015686 64.081 < 2e-16 *** 
## Tolls_amount 0.995197  0.040010 24.873 < 2e-16 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.904 on 4994 degrees of freedom
## Multiple R-squared:  0.9659, Adjusted R-squared:  0.9659 
## F-statistic: 2.828e+04 on 5 and 4994 DF,  p-value: < 2.2e-16

summary(m1)

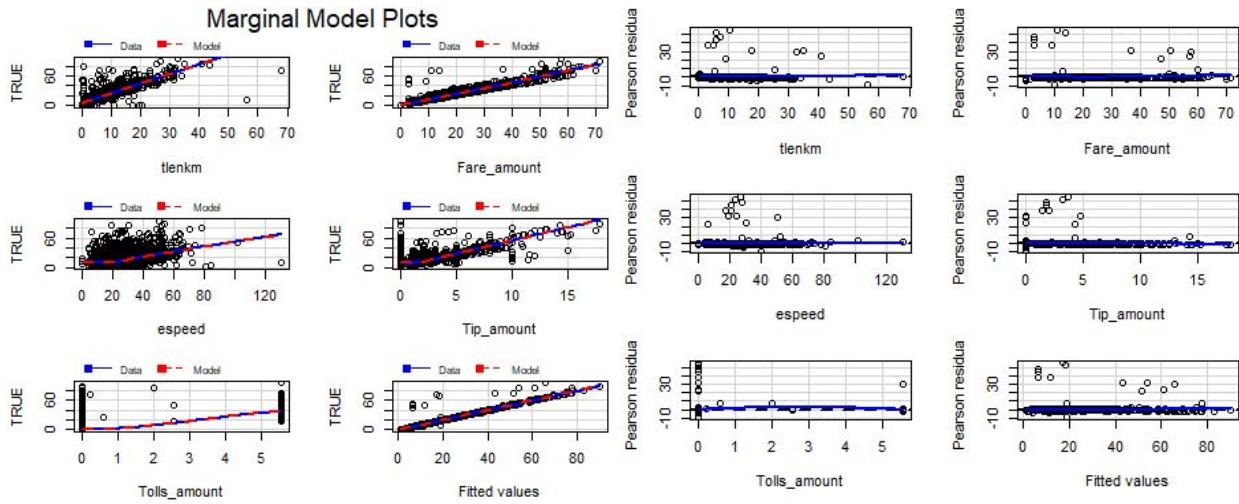
##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##     Tolls_amount + distHaversine, data = df[, c("Total_amount",
##     vars_cexp)])
## 
## Residuals:
##   Min     1Q Median     3Q    Max 
## -12.235 -0.393 -0.055  0.183 54.123 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.574884  0.079479 19.815 < 2e-16 *** 
## tlenkm      0.222321  0.017717 12.548 < 2e-16 *** 
## Fare_amount  0.950798  0.007427 128.027 < 2e-16 *** 
## espeed      -0.015307  0.003644 -4.200 2.71e-05 *** 
## Tip_amount   1.012951  0.015640 64.765 < 2e-16 *** 
## Tolls_amount 1.010725  0.039858 25.358 < 2e-16 *** 
## distHaversine -0.149401  0.020423 -7.315 2.98e-13 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.894 on 4993 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9662 
## F-statistic: 2.382e+04 on 6 and 4993 DF,  p-value: < 2.2e-16

anova(m6,m1)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount
## Model 2: Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount +
##           distHaversine
##   Res.Df   RSS Df Sum of Sq    F   Pr(>F)    
## 1 4994 18103
## 2 4993 17911  1    191.96 53.511 2.98e-13 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

marginalModelPlots(m6)

```

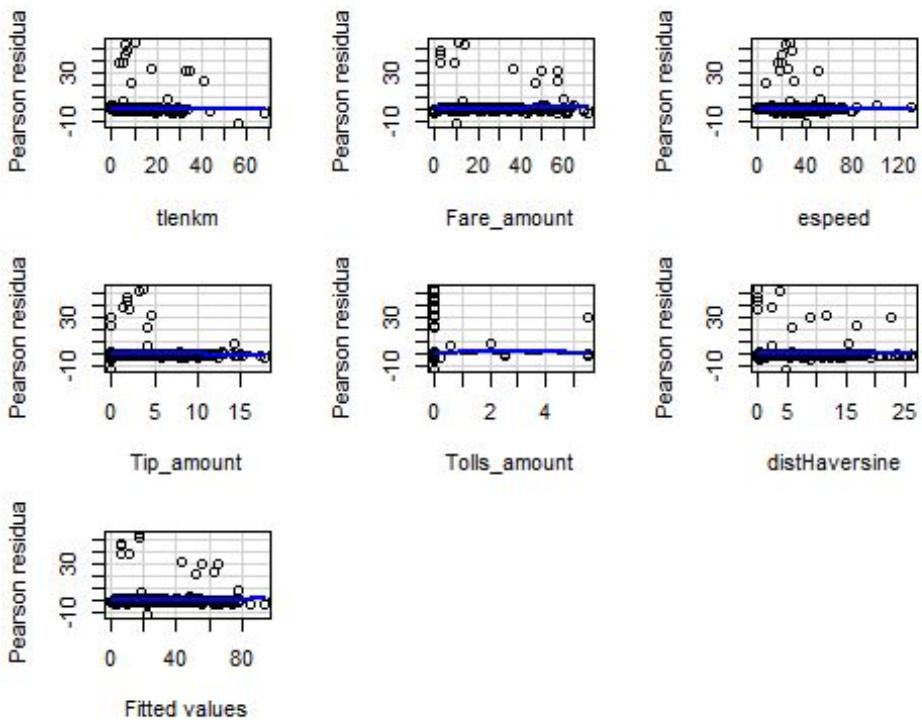


```

##           Test stat Pr(>|Test stat|)
## tlenkm      2.7892  0.0053033 ** 
## Fare_amount  6.0739  1.34e-09 *** 
## espeed     -0.0586  0.9532750  
## Tip_amount   -2.8502  0.0043864 ** 
## Tolls_amount -3.0041  0.0026766 ** 
## Tukey test    3.6569  0.0002553 *** 
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residualPlots(m1)

```



```

##           Test stat Pr(>|Test stat|)
## tlenkm      -0.5294    0.596537
## Fare_amount   3.9819    6.933e-05 ***
## espeed      -0.3207    0.748461
## Tip_amount    -3.1162    0.001842 **
## Tolls_amount   -2.7509    0.005964 **
## distHaversine  2.8748    0.004059 **
## Tukey test     1.8956    0.058014 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Probamos primero a descartar el Fare\_amount debido a que estará muy relacionada con nuestro target, pero vemos como sufrimos un descenso considerable hasta 85.55% de la explicación de la variabilidad por el modelo. En cambio, al quitar el distHaversine solo vemos un descenso de 96.62% a 96.59% reduciendo el uso de una variable respecto al modelo m1. Aplicando anova vemos como es buena opción escoger este nuevo modelo como óptimo.

```

m7<-lm(Total_amount ~ f.tlenkm +Fare_amount + espeed + Tip_amount + Tolls_amount,data=df)
summary(m7)

##
## Call:
## lm(formula = Total_amount ~ f.tlenkm + Fare_amount + espeed +
##     Tip_amount + Tolls_amount, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -1.591 -0.387 -0.037  0.150 55.772 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.212557  0.076090 15.936 <2e-16 ***
## f.tlenkmf.tlenkm-(5,67.9] 0.132659  0.089314  1.485  0.138    
## f.tlenkmf.tlenkm-[0,1]    -0.043945  0.101516 -0.433  0.665    
## Fare_amount              1.003235  0.004947 202.795 <2e-16 ***
## espeed                  -0.004032  0.003369 -1.197  0.231    
## Tip_amount               1.011523  0.015836 63.875 <2e-16 ***
## Tolls_amount              1.027439  0.040313 25.487 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.923 on 4993 degrees of freedom
## Multiple R-squared:  0.9652, Adjusted R-squared:  0.9652 
## F-statistic: 2.308e+04 on 6 and 4993 DF,  p-value: < 2.2e-16

m8<-lm(Total_amount ~ tlenkm +f.Fare_amount + espeed + Tip_amount + Tolls_amount,data=df)
summary(m8)

##
## Call:
## lm(formula = Total_amount ~ tlenkm + f.Fare_amount + espeed +
##     Tip_amount + Tolls_amount, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -72.442 -0.787 -0.118  0.474 53.447 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            12.618792  0.227247 55.53 <2e-16 ***
## tlenkm                 1.386418  0.019843 69.87 <2e-16 ***
## f.Fare_amountf.Fare_amount-(6.5,9] -5.367044  0.204256 -26.28 <2e-16 ***
## f.Fare_amountf.Fare_amount-(9,14.5] -4.022798  0.178615 -22.52 <2e-16 ***
## f.Fare_amountf.Fare_amount-[0,6.5]  -6.515156  0.209623 -31.08 <2e-16 ***

```

```

## espeed                  -0.097683   0.006716  -14.54   <2e-16 ***
## Tip_amount                1.272224   0.029240   43.51   <2e-16 ***
## Tolls_amount               1.066612   0.075498   14.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.589 on 4992 degrees of freedom
## Multiple R-squared:  0.8788, Adjusted R-squared:  0.8787
## F-statistic:  5173 on 7 and 4992 DF,  p-value: < 2.2e-16

m9<-lm(Total_amount ~ f.tlenkm +Fare_amount + f.espeed + Tip_amount + Tolls_amount,data=df)
summary(m9)

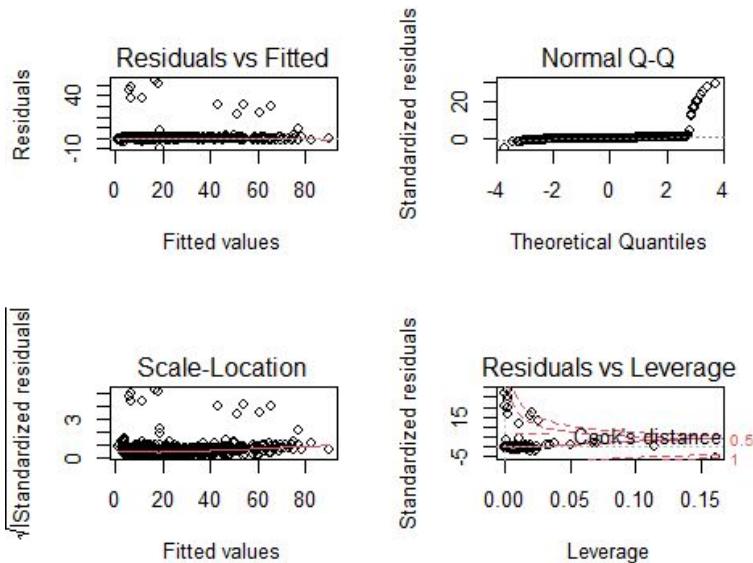
##
## Call:
## lm(formula = Total_amount ~ f.tlenkm + Fare_amount + f.espeed +
##     Tip_amount + Tolls_amount, data = df)
##
## Residuals:
##    Min      1Q Median      3Q      Max
## -1.620 -0.380 -0.039  0.137 55.793
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.148572   0.051685 22.223   <2e-16 ***
## f.tlenkmf.tlenkm-(5,67.9] 0.123461   0.090119  1.370   0.171
## f.tlenkmf.tlenkm-[0,1]   -0.040152   0.101498 -0.396   0.692
## Fare_amount             1.002671   0.004916 203.946   <2e-16 ***
## f.espeedf.espeed-(25,130] -0.048258   0.073715 -0.655   0.513
## f.espeedf.espeed-[0,1]   -0.178280   0.786093 -0.227   0.821
## Tip_amount              1.011469   0.015840  63.856   <2e-16 ***
## Tolls_amount             1.024093   0.040189  25.482   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.924 on 4992 degrees of freedom
## Multiple R-squared:  0.9652, Adjusted R-squared:  0.9651
## F-statistic: 1.977e+04 on 7 and 4992 DF,  p-value: < 2.2e-16

BIC(m6,m7,m8,m9)

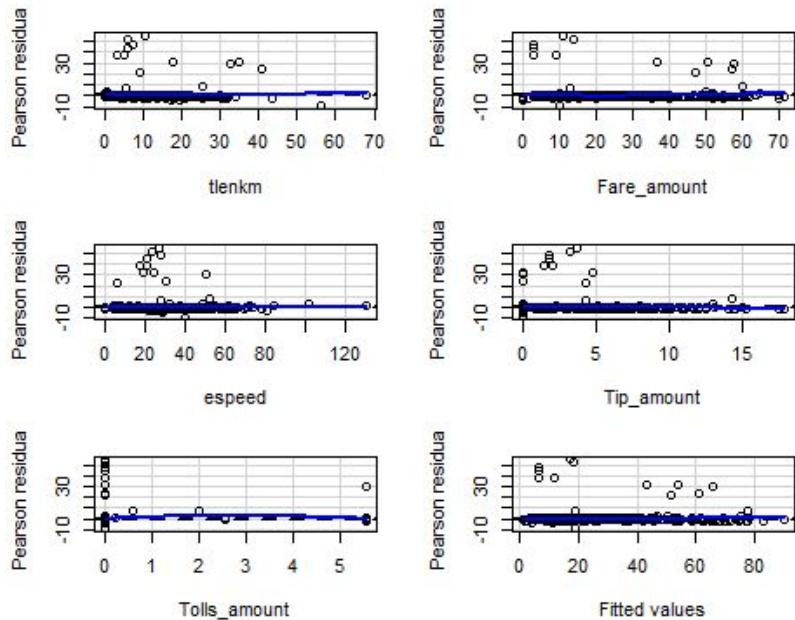
##      df      BIC
## m6  7 20682.23
## m7  8 20790.88
## m8  9 27036.18
## m9  9 20800.36

```

Probamos a cambiar la variable tlenkm por su factor obteniendo una explicación de la variabilidad muy similar, aún así, el método BIC nos indica que sigue siendo mejor el modelo anterior siendo su BIC menor. Y la misma argumentación podemos aplicar para el uso del factor de Fare\_amount aunque este sí que muestra un descenso significativo de la explicación de la variabilidad del target.



```
residualPlots(m6)
```

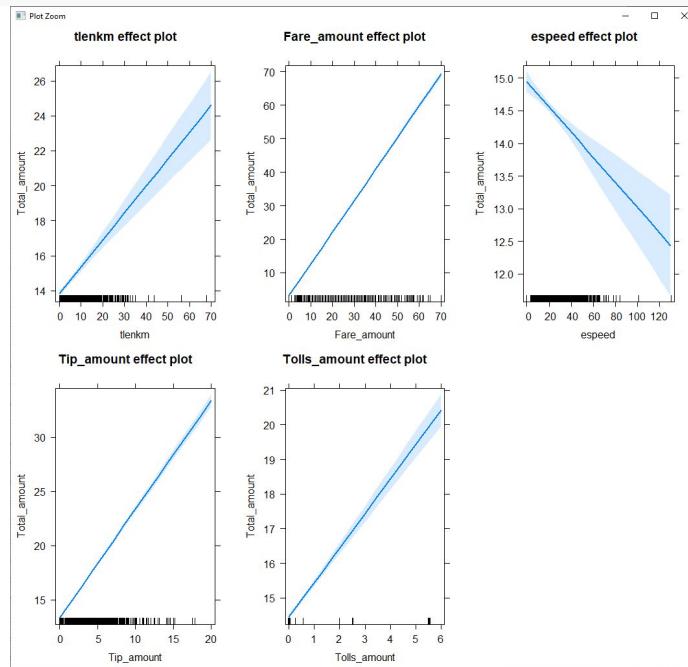


```
##           Test stat Pr(>|Test stat|)
## tlenkm      2.7892  0.0053033 **
## Fare_amount  6.0739  1.34e-09 ***
## espeed     -0.0586  0.9532750
## Tip_amount   -2.8502  0.0043864 **
## Tolls_amount -3.0041  0.0026766 **
## Tukey test    3.6569  0.0002553 ***
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
par(mfrow=c(1,1))
```

Seguimos sin disponer de unos residuos normales, aún así, a excepción de los extremos que distan significativamente, la mayoría de residuos siguen esta distribución. Estos residuos no normales pueden deberse al porcentaje de explicación que le falta a nuestro modelo de la variación de todo el target.

Es cierto que podemos observar un pequeño desajuste respecto a los smoothers del residualPlots pero no son lo suficientemente significativos.

```
library(effects)
plot(allEffects(m6))
```

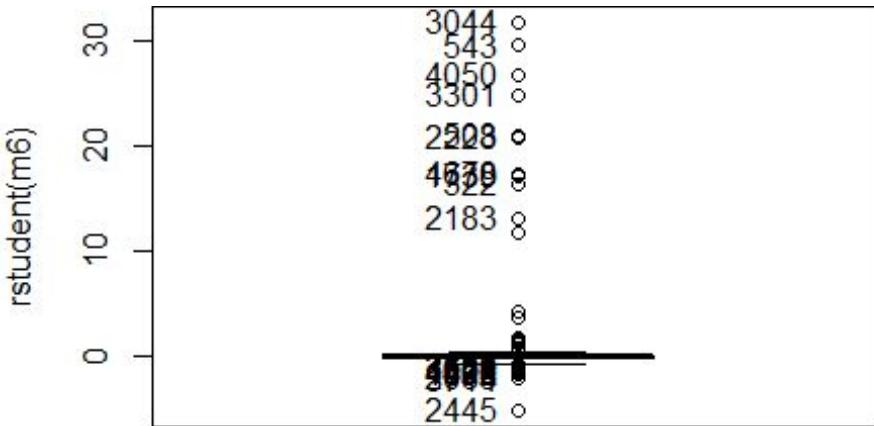


Analizamos como funcionan las variables de nuestro modelo respecto a nuestro target.

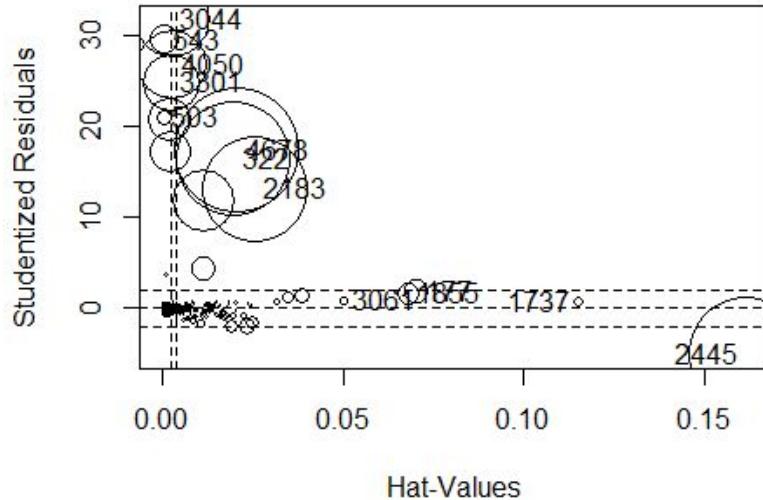
Que a medida que aumenten tanto la distancia en km como el número de peajes así como la tarifa aumente la cuantía total del servicio parece ser coherente por cómo funciona la composición de esta. También podemos llegar a entender que cuanta más propina ofrezca un cliente, más acabará pagando.

La explicación no tan trivial es la de la velocidad efectiva. Se puede justificar esta relación inversa con el hecho de que cuanto más lento vaya el taxi, más tiempo estará produciendo el servicio y por tanto más acabará cobrando. Esto sucede sobretodo en los núcleos urbanos, donde el tiempo predomina frente a la distancia en cuanto al cómputo del precio del transporte.

```
sel1<-Boxplot(rstudent(m6));sel1
```



```
## [1] 2445 3714 2065 492 1674 4492 1026 3547 2789 634 3044 543 4050 3301 503
## [16] 2228 4678 1739 322 2183
influencePlot(m6,id=list(method="noteworthy", n=5))
```



```
##          StudRes      Hat      CookD
## 177  1.7249933 0.0705226674 0.037613327
## 322 16.3094533 0.0196117938 0.842155534
## 503 20.9400558 0.0002536419 0.017047686
## 543 29.5050653 0.0004487365 0.055477416
## 1737 0.5825831 0.1153369866 0.007375861
## 1855 1.5905816 0.0679028642 0.030708213
## 2183 13.0405729 0.0256565037 0.721886847
## 2445 -5.1061262 0.1614370452 0.832385348
## 3044 31.7182308 0.0026453769 0.370229441
## 3061  0.7904553 0.0502540170 0.005510605
## 3301 24.8888840 0.0022600322 0.208090575
## 4050 26.7275138 0.0029229175 0.305398546
## 4678 17.3590485 0.0205975996 0.996306640
```

En el primer boxplot podemos observar las observaciones que consideraríamos outlier con una distribución `rstudent`, que debido a nuestro número de observaciones es equivalente a considerar una normal. Con el segundo plot podemos observar los individuos inusuales y ver si son o no influyentes. Podríamos destacar sobretodo el 2445 al tener un residuo

estandarizado negativo y por otro lado, al parecer bastante influyentes, el 3044,543, 2183 y 4678 así como 322.

## Using factors as explanatory variables

```
m6<-lm(Total_amount ~ tlenkm +Fare_amount + espeed + Tip_amount + Tolls_amount,data=df)
m6a <-lm(Total_amount ~ poly(tlenkm,2) +Fare_amount + espeed + Tip_amount + Tolls_amount,data=df)
m6b <-lm(Total_amount ~ tlenkm +poly(Fare_amount,2) + espeed + Tip_amount + Tolls_amount,data=df)
m6c <-lm(Total_amount ~ tlenkm+ Fare_amount +poly(espeed,2) + Tip_amount + Tolls_amount,data=df)
m6d <-lm(Total_amount ~ tlenkm+ Fare_amount +espeed + poly(Tip_amount,2) + Tolls_amount,data=df)
m6e <-lm(Total_amount ~ tlenkm +Fare_amount +espeed + Tip_amount + poly(Tolls_amount,2),data=df)

m6f<-lm(Total_amount ~ poly(tlenkm,2) +poly(Fare_amount,2) +poly(espeed,2) + poly(Tip_amount,2) +
poly(Tolls_amount,2),data=df)

BIC(m6,m6a,m6b,m6c,m6d,m6e,m6f)

##      df        BIC
## m6     7 20682.23
## m6a    8 20682.96
## m6b    8 20653.94
## m6c    8 20690.74
## m6d    8 20682.62
## m6e    8 20681.72
## m6f   12 20633.85
```

Si intentamos realizar transformaciones en todas aquellas variables que en el modelo del apartado anterior nos da resultado. Observamos que el único modelo que da mejor resultado es aquel con la variable Fare\_amount en la que se aplica un polinomio de segundo grado.

A pesar de todo hemos considerado no oportuno utilizar este modelo debido a la mínima diferencia de BIC así como en cuanto a la explicación de la variabilidad que, si bien aumenta, no consideramos suficiente como para justificar la complejidad añadida al modelo. Al no observar ninguna transformación que mejore sustancialmente el modelo tampoco consideramos hacer una combinación de las mismas.

```
vars_cexp_cat <- c("f.Improvement_surcharge", "f.MTA_tax", "Trip_type", "RateCodeID", "lpep_pickup_period",
"VendorID", "lpep_pickup_date", "f.Extra")

m10<-lm(Total_amount~. ,family="binomial",data=df[,c("Total_amount", vars_cexp_cat)])
```

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

```
vif(m10)

##                  GVIF Df GVIF^(1/(2*Df))
## f.Improvement_surcharge 19.079653  1      4.368026
## f.MTA_tax                22.004858  1      4.690934
## Trip_type                 14.431207  1      3.798843
## RateCodeID                 5.788352  1      2.405899
## lpep_pickup_period       8.756254  3      1.435665
## VendorID                  1.012459  1      1.006210
## lpep_pickup_date         1.339936 30      1.004889
## f.Extra                   9.758503  2      1.767445
```

```

Anova(m10)

## Anova Table (Type II tests)
##
## Response: Total_amount
##          Sum Sq Df F value    Pr(>F)
## f.Improvement_surcharge   137   1 1.4086  0.2353
## f.MTA_tax                 144   1 1.4736  0.2248
## Trip_type                4383   1 44.9457 2.252e-11 ***
## RateCodeID                36962   1 379.0241 < 2.2e-16 ***
## lpep_pickup_period       56    3 0.1918  0.9020
## VendorID                  83    1 0.8461  0.3577
## lpep_pickup_date         2727   30 0.9321  0.5724
## f.Extra                   73    2 0.3756  0.6869
## Residuals                483597 4959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step(m10, k=log(nrow(df)))

## Step:  AIC=22925.57
## Total_amount ~ f.MTA_tax + Trip_type + RateCodeID
##
##          Df Sum of Sq   RSS   AIC
## <none>           486744 22926
## - f.MTA_tax     1      936 487681 22927
## - Trip_type     1      5467 492211 22973
## - RateCodeID     1      39683 526427 23309

##
## Call:
## lm(formula = Total_amount ~ f.MTA_tax + Trip_type + RateCodeID,
##      data = df[, c("Total_amount", vars_cexp_cat)], family = "binomial")
##
## Coefficients:
##             (Intercept)      f.MTA_taxf.MTA_tax_YES
##                   21.25                      8.61
## Trip_typef.TripType-Street-Hail  RateCodeIDStandard rate
##                               24.08                     -39.71

m11<- lm(Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount + Trip_type + RateCodeID
,data=df)
summary(m11)

##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##      Tolls_amount + Trip_type + RateCodeID, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.830 -0.414 -0.059  0.146 55.043 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.703558  0.204604  3.439 0.000589 ***
## tlenkm       0.149000  0.015358  9.702 < 2e-16 ***
## Fare_amount  0.947452  0.007693 123.163 < 2e-16 ***
## espeed      -0.017603  0.003656 -4.815 1.52e-06 ***
## Tip_amount   0.998823  0.015710  63.579 < 2e-16 ***
## Tolls_amount 0.990151  0.040145  24.664 < 2e-16 ***
## Trip_typef.TripType-Street-Hail 1.043899  0.415641  2.512 0.012052 *  
## RateCodeIDStandard rate -0.188495  0.397269 -0.474 0.635180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.9 on 4992 degrees of freedom

```

```

## Multiple R-squared:  0.9661, Adjusted R-squared:  0.966
## F-statistic: 2.029e+04 on 7 and 4992 DF, p-value: < 2.2e-16

vif(m11)

##      tlenkm  Fare_amount      espeed  Tip_amount Tolls_amount  Trip_type
##    7.010194    6.263560    1.610679    1.325573    1.101979    5.743132
##  RateCodeID
##    6.074742

step(m11, k=log(nrow(df)))
## Step: AIC=6468.86
## Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount +
##   Trip_type
##
##           Df Sum of Sq   RSS     AIC
## <none>             18016 6468.9
## - espeed          1      83 18099 6483.3
## - Trip_type       1      87 18103 6484.3
## - tlenkm          1     344 18360 6554.9
## - Tolls_amount    1    2228 20245 7043.4
## - Tip_amount      1    14589 32605 9426.2
## - Fare_amount     1    58537 76553 13693.9

##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##   Tolls_amount + Trip_type, data = df)
##
## Coefficients:
## (Intercept)          tlenkm
##          0.68526        0.14777
## Fare_amount          espeed
##          0.94837       -0.01742
## Tip_amount          Tolls_amount
##          0.99870        0.99214
## Trip_type
## Street-Hail
##          0.86538

```

También intentamos añadir algunas variables categóricas en el modelo. Como resultado nos dio que las variables categóricas Trip\_Type y RateCodeID funcionan bien en el modelo de predicción. Si unimos nuestro mejor modelo del apartado anterior junto con dichas variables nos da que la explicación de la variabilidad del target es del 96.61%.

A pesar de todo hemos considerado no oportuno utilizar este modelo debido a la mínima diferencia de BIC así como en cuanto a la explicación de la variabilidad que, si bien aumenta, no consideramos lo suficiente como para justificar la complejidad añadida al modelo. Al no observar ninguna transformación que mejore sustancialmente el modelo tampoco consideramos hacer una combinación de las mismas.

## Clear effects

```

anova(m6a,m6f)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ poly(tlenkm, 2) + Fare_amount + espeed + Tip_amount +
##   Tolls_amount
## Model 2: Total_amount ~ poly(tlenkm, 2) + poly(Fare_amount, 2) + poly(espeed,
##   2) + poly(Tip_amount, 2) + poly(Tolls_amount, 2)
## Res.Df   RSS Df Sum of Sq   F   Pr(>F)
## 1   4993 18075
## 2   4989 17777  4    298.2 20.922 < 2.2e-16 ***

```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m6b,m6f)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ tlenkm + poly(Fare_amount, 2) + espeed + Tip_amount +
##           Tolls_amount
## Model 2: Total_amount ~ poly(tlenkm, 2) + poly(Fare_amount, 2) + poly(espeed,
##           2) + poly(Tip_amount, 2) + poly(Tolls_amount, 2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4993 17970
## 2 4989 17777  4     193.58 13.582 5.151e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m6c,m6f)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ tlenkm + Fare_amount + poly(espeed, 2) + Tip_amount +
##           Tolls_amount
## Model 2: Total_amount ~ poly(tlenkm, 2) + poly(Fare_amount, 2) + poly(espeed,
##           2) + poly(Tip_amount, 2) + poly(Tolls_amount, 2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4993 18103
## 2 4989 17777  4     326.35 22.897 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m6d,m6f)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ tlenkm + Fare_amount + espeed + poly(Tip_amount,
##           2) + Tolls_amount
## Model 2: Total_amount ~ poly(tlenkm, 2) + poly(Fare_amount, 2) + poly(espeed,
##           2) + poly(Tip_amount, 2) + poly(Tolls_amount, 2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4993 18074
## 2 4989 17777  4     296.95 20.835 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m6e,m6f)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + poly(Tolls_amount,
##           2)
## Model 2: Total_amount ~ poly(tlenkm, 2) + poly(Fare_amount, 2) + poly(espeed,
##           2) + poly(Tip_amount, 2) + poly(Tolls_amount, 2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4993 18070
## 2 4989 17777  4     293.7 20.606 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova(m6f)

## Anova Table (Type II tests)
##
## Response: Total_amount
##             Sum Sq  Df  F value    Pr(>F)
## poly(tlenkm, 2)      400   2  56.073 < 2.2e-16 ***
## poly(Fare_amount, 2) 45256   2 6350.436 < 2.2e-16 ***
## poly(espeed, 2)       138   2  19.407 4.019e-09 ***

```

```

## poly(Tip_amount, 2)    14771     2 2072.675 < 2.2e-16 ***
## poly(Tolls_amount, 2)  2280      2 319.925 < 2.2e-16 ***
## Residuals              17777 4989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Aplicando el método anova para comprobar los efectos limpios producidos por el nuevo modelo, obtenemos que el uso de polinomios de grado dos para todas las variables que forman el modelo son útiles para su construcción. Esto podemos comprobarlo con el método Anova.

Aún así, vemos como las transformaciones que más mejoras aportan, aún siendo estas ínfimas, son las de aplicar el polinomio a las variables Fare\_amount y Tolls\_amount.

## Dirty effects

```

m0<-lm(Total_amount ~ 1,data=df)
anova(m0,m6b)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ 1
## Model 2: Total_amount ~ tlenkm + poly(Fare_amount, 2) + espeed + Tip_amount +
##           Tolls_amount
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4999 530682
## 2 4993 17970  6   512712 23743 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m0,m6c)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ 1
## Model 2: Total_amount ~ tlenkm + Fare_amount + poly(espeed, 2) + Tip_amount +
##           Tolls_amount
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4999 530682
## 2 4993 18103  6   512579 23562 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m0,m6d)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ 1
## Model 2: Total_amount ~ tlenkm + Fare_amount + espeed + poly(Tip_amount,
##           2) + Tolls_amount
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4999 530682
## 2 4993 18074  6   512609 23602 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m0,m6e)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ 1
## Model 2: Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + poly(Tolls_amount,
##           2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)

```

```

## 1 4999 530682
## 2 4993 18070 6 512612 23606 < 2.2e-16 ***
## ...
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(m0,m6f)

## Analysis of Variance Table
##
## Model 1: Total_amount ~ 1
## Model 2: Total_amount ~ poly(tlenkm, 2) + poly(Fare_amount, 2) + poly(espeed,
## 2) + poly(Tip_amount, 2) + poly(Tolls_amount, 2)
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 4999 530682
## 2 4989 17777 10 512906 14395 < 2.2e-16 ***
## ...
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Como podemos ver debido al uso de los efectos sucios, rechazamos la hipótesis nula para todos los modelos que incluían distintos usos del polinomio, por tanto, podemos confirmar que no se trata de modelos equivalentes y necesitamos la aplicación de estos modelos.

Como bien llevamos diciendo durante todo el tratamiento de modelos, este aumento en la complejidad del modelo respecto a las ventajas que nos ofrece no nos parece justificable así que no llegaríamos a adoptar como modelo del target el m6f sinó el m6, el cual no contiene ninguno de los polinomios.

## Interactions

```

m12<- lm(Total_amount ~ (tlenkm +Fare_amount + espeed + Tip_amount + Tolls_amount) + (Trip_type
+RateCodeID)^2 ,data=df) # Interacciones dobles en factores
m12<-step(m12, k=log(nrow(df)))

## Step: AIC=6468.86
## Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount +
##   Trip_type
##
##           Df Sum of Sq   RSS     AIC
## <none>                 18016 6468.9
## - espeed      1       83 18099 6483.3
## - Trip_type   1       87 18103 6484.3
## - tlenkm      1      344 18360 6554.9
## - Tolls_amount 1     2228 20245 7043.4
## - Tip_amount   1     14589 32605 9426.2
## - Fare_amount   1     58537 76553 13693.9

m13<- lm(Total_amount ~ (tlenkm +Fare_amount + espeed + Tip_amount + Tolls_amount) * (Trip_type
+RateCodeID) ,data=df) # Interacciones dobles en factor-numèrica
m13<-step(m13, k=log(nrow(df)))
## Step: AIC=6260.07
## Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount + Tolls_amount +
##   Trip_type + RateCodeID + tlenkm:RateCodeID + Fare_amount:RateCodeID +
##   espeed:RateCodeID + Tolls_amount:Trip_type
##
##           Df Sum of Sq   RSS     AIC
## <none>                 17133 6260.1
## - espeed:RateCodeID   1      50.5 17184 6266.3
## - Tolls_amount:Trip_type 1     285.2 17418 6334.1
## - Fare_amount:RateCodeID 1     450.7 17584 6381.4
## - tlenkm:RateCodeID    1     474.8 17608 6388.2
## - Tip_amount            1    15180.1 32313 9423.9

BIC(m6,m11,m12,m13)

```

```

##      df      BIC
## m6    7 20682.23
## m11   9 20675.05
## m12   8 20666.76
## m13  13 20457.97

summary(m12)

##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##     Tolls_amount + Trip_type, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -7.759 -0.413 -0.061  0.147 55.050 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               0.685261  0.200921  3.411 0.000653 ***
## tlenkm                  0.147766  0.015135  9.763 < 2e-16 ***
## Fare_amount               0.948368  0.007446 127.368 < 2e-16 ***
## espeed                   -0.017415  0.003634 -4.792 1.7e-06 ***
## Tip_amount                0.998699  0.015707  63.585 < 2e-16 ***
## Tolls_amount              0.992137  0.039923  24.851 < 2e-16 ***
## Trip_typef.TripType-Street-Hail 0.865381  0.176620   4.900 9.9e-07 ***
## ---                        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.9 on 4993 degrees of freedom
## Multiple R-squared:  0.9661, Adjusted R-squared:  0.966 
## F-statistic: 2.368e+04 on 6 and 4993 DF,  p-value: < 2.2e-16

summary(m13)

##
## Call:
## lm(formula = Total_amount ~ tlenkm + Fare_amount + espeed + Tip_amount +
##     Tolls_amount + Trip_type + RateCodeID + tlenkm:RateCodeID +
##     Fare_amount:RateCodeID + espeed:RateCodeID + Tolls_amount:Trip_type,
##     data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -10.841 -0.435 -0.053  0.184 53.321 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -0.034628  0.312090 -0.111  
## tlenkm                  0.034592  0.021215  1.631  
## Fare_amount              0.997337  0.011025 90.460  
## espeed                  -0.005694  0.010186 -0.559  
## Tip_amount                1.027393  0.015454 66.479  
## Tolls_amount              2.871734  0.209986 13.676  
## Trip_typef.TripType-Street-Hail 0.952772  0.476165  2.001 
## RateCodeIDStandard rate  1.676628  0.616624  2.719  
## tlenkm:RateCodeIDStandard rate 0.409302  0.034811 11.758  
## Fare_amount:RateCodeIDStandard rate -0.199409  0.017408 -11.455 
## espeed:RateCodeIDStandard rate -0.042593  0.011106 -3.835  
## Tolls_amount:Trip_typef.TripType-Street-Hail -1.939430  0.212852 -9.112 
##                                         Pr(>|t|)    
## (Intercept)                     0.911655  
## tlenkm                         0.103052  
## Fare_amount                      < 2e-16 ***
## espeed                          0.576172  
## Tip_amount                       < 2e-16 ***
## Tolls_amount                      < 2e-16 *** 
## Trip_typef.TripType-Street-Hail 0.045454 * 

```

```

## RateCodeIDStandard rate          0.006570 **
## tlenkm:RateCodeIDStandard rate   < 2e-16 ***
## Fare_amount:RateCodeIDStandard rate   < 2e-16 ***
## espeed:RateCodeIDStandard rate      0.000127 ***
## Tolls_amount:Trip_typef.TripType-Street-Hail < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.853 on 4988 degrees of freedom
## Multiple R-squared:  0.9677, Adjusted R-squared:  0.9676
## F-statistic: 1.359e+04 on 11 and 4988 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))
influencePlot(m6)

##           StudRes       Hat       CookD
## 322  16.3094533 0.0196117938 0.842155534
## 543  29.5050653 0.0004487365 0.055477416
## 1737  0.5825831 0.1153369866 0.007375861
## 2445 -5.1061262 0.1614370452 0.832385348
## 3044 31.7182308 0.0026453769 0.370229441
## 4678 17.3590485 0.0205975996 0.996306640

influencePlot(m12)

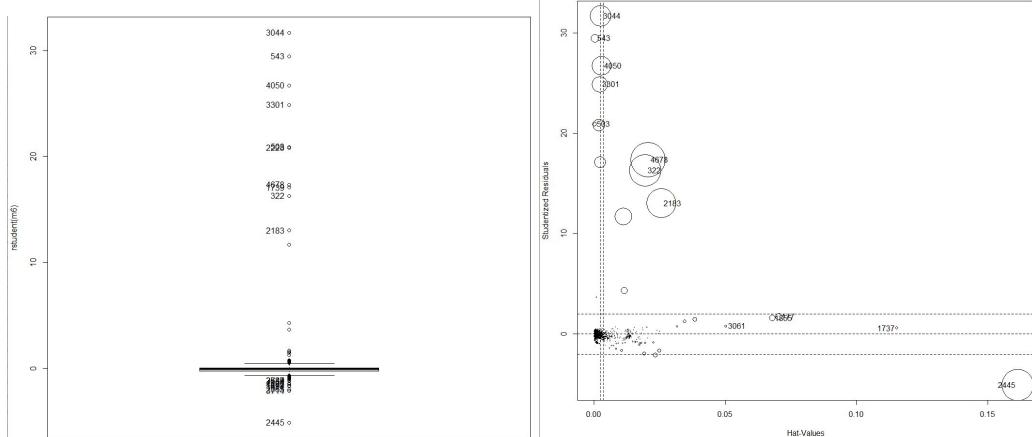
##           StudRes       Hat       CookD
## 322  16.8428511 0.0261692343 1.030679619
## 543  29.5763724 0.0004508224 0.047968509
## 1737  0.6351762 0.1154336530 0.007522191
## 2445 -4.5083574 0.1759470935 0.617572508
## 3044 31.8231469 0.0026518042 0.319854291
## 4678 17.3889190 0.0206030687 0.856973131

influencePlot(m13)

##           StudRes       Hat       CookD
## 322  14.4448459 0.3400346495 8.60068176
## 414  -7.0136487 0.3451641260 2.14005133
## 543  30.4108580 0.0004609966 0.02999020
## 2793 -7.2154961 0.3360629530 2.17380963
## 3044 31.6474445 0.0077899326 0.54579570
## 3061  0.5465404 0.3785873287 0.01516737

111<-Boxplot(rstudent(m6));111

```



```

## [1] 172

112<-which(row.names(df) %in% names(hatvalues(m6)[sel2]));
sel3<-which(cooks.distance(m6)> 0.5 );sel3;length(sel3)

## 322 2183 2445 4678
## 322 2183 2445 4678

## [1] 4

113<-which(row.names(df) %in% names(cooks.distance(m6)[sel3]));113

## [1] 322 2183 2445 4678

111<-Boxplot(rstudent(m13));

## [1] 2793 414 1737 2223 1008 2937 4415 592 1738 1712 3044 543 4050 3301 503
## [16] 2228 1739 4678 322 4073

sel2<-which(hatvalues(m13)>5*length(m13$coefficients)/nrow(df));length(sel2)

## [1] 215

112<-which(row.names(df) %in% names(hatvalues(m13)[sel2]));

sel3<-which(cooks.distance(m13)> 0.5 );sel3;length(sel3)

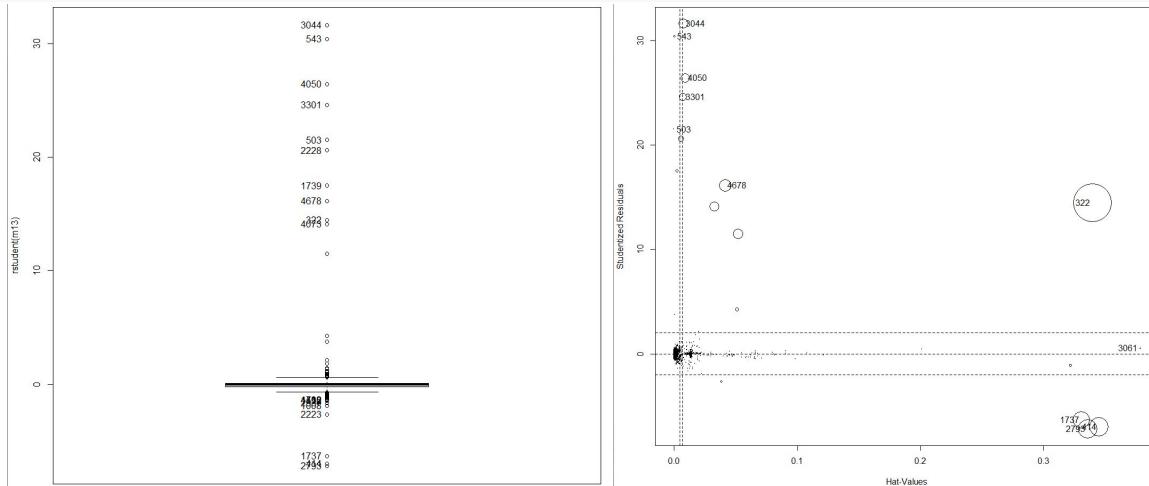
## 322 414 1737 2183 2793 3044 4073 4678
## 322 414 1737 2183 2793 3044 4073 4678

## [1] 8

113<-which(row.names(df) %in% names(cooks.distance(m13)[sel3]));113

## [1] 322 414 1737 2183 2793 3044 4073 4678

```



Para analizar las interacciones que pueden hacer mejorar nuestro modelo, consideramos el modelo m6 y m11 al disponer este último de las variables categóricas necesarias.

Después de aplicar una serie de interacciones entre factores así como entre numéricas y factores, vemos como aparecen los dos modelos con un BIC inferior a nuestro modelo sin interacciones m11, donde el m13 presenta una mejora sustancial en cuanto a su AIC.

Aún así la mejora que obtenemos al aplicar estas interacciones no es lo suficientemente significativa como para justificar el aumento de complejidad del modelo, así que como modelo principal seguiremos usando el m6 ya que como dijimos, el m11 lo descartamos por el mismo motivo producido por las variables extra.

Aquí podemos ver una comparación de las observaciones inusuales tanto de nuestro modelo como del que hemos escogido con interacciones(m13) y vemos como el modelo m13 presenta tal vez menos individuos inusuales aún así acaba presentando el doble(8) de individuos significativos que en el modelo m6 destacando : 322 414 1737 2183 2793 3044 4073 4678. Sobre todo vemos cómo el individuo 322 ha pasado a ser mucho más influyente en este nuevo modelo, aún así no hay una gran diferencia entre los demás individuos influyentes como para considerarlo mejor o peor por ello.

## Binary Regression Model

En esta sección se construye un modelo lineal para una variable target numérica, AnyTip.

Para la realización del modelo, consideramos que la variable Tip\_amount no estuviese como variable explicativa ya que la variable target AnyTip se calculaba a partir del Tip\_amount. Decidimos seleccionar aquellos individuos que no pagaban en efectivo, ya que eran los únicos que tenían registrada la propina. Como consecuencia la variable Payment\_type se elimina como variable explicativa. Para poder realizar el estudio, un 70% de nuestro dataset formó parte del train dataset y el 30% formó parte del test dataset. Inicialmente consideramos como variables explicativas: Passenger\_count, tlenkm, Fare\_amount, espeed, Tolls\_amount, lpep\_pickup\_time, travelttime y distHaversine.

## Use explanatory numeric variables

```
m<-glm(AnyTip~.,family="binomial",data=train_dataset[,c("AnyTip",vars_cexp)])
summary(m)

## Analysis of Deviance Table (Type II tests)
##
## Response: AnyTip
##              Df  Chisq Pr(>Chisq)
## Passenger_count   1 0.3979  0.52816
## tlenkm            1 4.5918  0.03213 *
## Fare_amount        1 0.1895  0.66333
## espeed             1 2.5285  0.11181
## Tolls_amount       1 0.0071  0.93297
## lpep_pickup_time  1 0.0553  0.81411
## travelttime        1 2.6192  0.10558
## distHaversine     1 3.8074  0.05103 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

vif(m)

##   Passenger_count      tlenkm      Fare_amount      espeed
##          1.004448     10.557730     7.221147     3.564562
##   Tolls_amount lpep_pickup_time travelttime distHaversine
##          1.057881      1.008695     6.207482     3.806424
```

Como podemos ver con el Anova las variables tlenkm, es la única variable que rechaza la hipótesis nula y por lo tanto son las variables que tienen más asociación. Podríamos

considerar que la variable distHaversine también está asociada a la variable target ya que su Chisq es muy próximo al nivel de significancia 0.05. Hemos considerado que no hay problemas de multicolinealidad si el valor de VIF es inferior a 11.

Como vemos que hay pocas variables explicativas asociadas con la variable target, intentaremos ampliar la lista de variables explicativas.

```
##           Eta2      P-value
## Tip_amount   0.212639269 1.354923e-92
## Pickup_longitude  0.018844441 8.532558e-09
## Dropoff_longitude  0.010593539 1.648365e-05
## Total_amount    0.008553588 1.093070e-04
## Pickup_latitude   0.005779308 1.482968e-03
## Dropoff_latitude  0.005663466 1.655806e-03
```

Si realizamos un categorical description, vemos que las variables más correlacionadas con la variable target son el Pickup\_longitude , el Dropoff\_longitude y el total amount. Como el Total amount es una variable target, hemos decidido no incluirla como variable explicativa, a pesar de todo, podría incluirse.

```
m2<-glm(AnyTip~tlenkm+Pickup_longitude+ Dropoff_longitude + distHaversine
,family="binomial",data=train_dataset)
summary(m2)

##
## Call:
## glm(formula = AnyTip ~ tlenkm + Pickup_longitude + Dropoff_longitude +
##     distHaversine, family = "binomial", data = train_dataset)
##
## Deviance Residuals:
##    Min      1Q      Median      3Q      Max
## -2.3142  0.4502  0.5073  0.5477  1.4798
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -616.92335  112.61024 -5.478 4.29e-08 ***
## tlenkm       -0.03843   0.02278 -1.687 0.091644 .
## Pickup_longitude  -7.77988  2.32850 -3.341 0.000834 ***
## Dropoff_longitude  -0.58682  1.98806 -0.295 0.767861
## distHaversine     0.09409  0.03898  2.414 0.015783 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1408.6 on 1744 degrees of freedom
## Residual deviance: 1372.2 on 1740 degrees of freedom
## AIC: 1382.2
##
## Number of Fisher Scoring iterations: 4
```

Si cogemos aquellas variables numéricas que no fueron eliminadas por hipótesis del modelo anterior más aquellas variables correlacionadas, vemos que eliminaríamos todas las variables explicativas excepto el Pickup\_longitude y distHaversine. A pesar de todo decidimos quedarnos con un modelo cuyas variables explicativas son Pickup\_longitude, distHaversine y tlenkm.

```
m3<-glm(AnyTip~tlenkm+Pickup_longitude + distHaversine ,family="binomial",data=train_dataset)
summary(m3)

##
## Call:
```

```

## glm(formula = AnyTip ~ tlenkm + Pickup_longitude + distHaversine,
##      family = "binomial", data = train_dataset)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max 
## -2.3445  0.4500  0.5091  0.5485  1.4777 
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      -612.21900  111.60934 -5.485 4.13e-08 ***
## tlenkm          -0.03869   0.02279  -1.698  0.0895 .  
## Pickup_longitude -8.30310   1.50954  -5.500 3.79e-08 *** 
## distHaversine    0.09451   0.03901   2.423  0.0154 *  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1408.6 on 1744 degrees of freedom
## Residual deviance: 1372.3 on 1741 degrees of freedom
## AIC: 1380.3
##
## Number of Fisher Scoring iterations: 4

```

## Consider factors and interactions

```

m41<-glm(AnyTip~f.tlenkm+Pickup_longitude + distHaversine ,family="binomial",data=train_dataset)
m42<-glm(AnyTip~tlenkm+Pickup_longitude + f.distHaversine ,family="binomial",data=train_dataset)
summary(m41)

BIC(m3,m41,m42)

##      df      BIC
## m3     4 1402.171
## m41    5 1411.438
## m42    5 1414.632

```

Si intentamos categorizar todas aquellas variables del modelo del apartado anterior. Observamos que el modelo empeora por lo tanto, no utilizaremos la categorización de las variables numéricas.

```

res.cat <- catdes(df,num.var=which(names(df)=="AnyTip"))

##                               p.value df
## Payment_type            0.000000e+00 2
## f.Total_amount          1.923205e-104 7
## f.Fare_amount           6.129444e-24 3
## f.tlenkm                1.087294e-19 2
## f.traveltime             2.107456e-18 4
## f.distHaversine         7.945740e-11 2
## f.Improvement_surcharge 3.570885e-09 1
## f.MTA_tax                4.542332e-09 1
## Trip_type               5.576360e-08 1
## RateCodeID              3.777248e-06 1
## lpep_pickup_period      5.610059e-06 3
## AnyToll                  3.502094e-05 1
## f.espeed                 1.425728e-04 2
## VendorID                 5.925000e-03 1
## lpep_pickup_date        1.450254e-02 30
## f.Extra                  3.273592e-02 2

m5<-glm(AnyTip~. ,family="binomial",data=train_dataset[,c("AnyTip", vars_cexp_cat)]) 

##                               GVIF Df GVIF^(1/(2*Df))
## f.Improvement_surcharge 2.027487e+07 1     4502.762320
## f.MTA_tax                2.803930e+07 1     5295.215279

```

```

## Trip_type          1.445557e+07  1    3802.048400
## RateCodeID        6.691136e+06  1    2586.722988
## lpep_pickup_period 9.945849e+00  3     1.466472
## VendorID          1.018046e+00  1     1.008983
## lpep_pickup_date   1.375646e+00 30    1.005330
## f.Extra           1.095316e+01  2     1.819218

```

Si vemos la multicolinealidad de las variables más correlacionadas con el target, nos encontramos que las variables f.Improvement\_surcharge, f.MTA\_tax, Trip\_Type y RateCodeID tienen un valor GVIF muy alto. Es por eso que de todas esas variables vamos a quedarnos solo con f.MTA\_tax.

```

m6<-glm(AnyTip~(f.MTA_tax+lpep_pickup_period+VendorID+lpep_pickup_date+f.Extra),family="binomial",data=train_dataset)
vif(m6)

##                                     GVIF Df GVIF^(1/(2*Df))
## f.MTA_tax             1.246385  1     1.116416
## lpep_pickup_period   9.311071  3     1.450441
## VendorID              1.017422  1     1.008674
## lpep_pickup_date     1.393596 30    1.005547
## f.Extra                10.340142 2     1.793212

Anova(m6,test="Wald")

## Analysis of Deviance Table (Type II tests)
##
## Response: AnyTip
##               Df  Chisq Pr(>Chisq)
## f.MTA_tax      1 34.4323  4.413e-09 ***
## lpep_pickup_period 3  9.2760   0.02584 *
## VendorID       1  0.0736   0.78615
## lpep_pickup_date 30 29.6474   0.48381
## f.Extra         2  5.7404   0.05669 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Como podemos observar, las únicas variables categóricas que han pasado el test de la Chisq son f.MTA\_tax y lpep\_pickup\_period. A pesar de todo, la variable f.Extra está cerca del 0.05 por lo tanto la consideraremos para el estudio.

```

m6<-glm(AnyTip~f.MTA_tax+lpep_pickup_period+f.Extra,family="binomial",data=train_dataset)
step(m6, k=log(nrow(df)))

## Step:  AIC=1395.63
## AnyTip ~ f.MTA_tax
##
##               Df Deviance   AIC
## <none>            1378.6 1395.6
## - f.MTA_tax      1   1408.6 1417.1

##
## Call:  glm(formula = AnyTip ~ f.MTA_tax, family = "binomial", data = train_dataset)
##
## Coefficients:
## (Intercept)  f.MTA_taxf.MTA_tax_YES
##                 -0.05407             1.94492
##
## Degrees of Freedom: 1744 Total (i.e. Null);  1743 Residual
## Null Deviance:      1409
## Residual Deviance: 1379  AIC: 1383

summary(m6)

```

```
## AIC: 1380.3
```

Si utilizamos la función step de R nos menciona que deberíamos eliminar la variable f.Extra y lpep\_pickup\_period. A pesar de todo, vamos a mantener todas las variables explicativas. El resultado que nos da es de AIC 1380.3

```
m7<-glm(AnyTip~ (tlenkm+Pickup_longitude+distHaversine)+(f.MTA_tax+lpep_pickup_period+f.Extra),
family="binomial", data=train_dataset)
m71<-glm(AnyTip~ (poly(tlenkm,2)+Pickup_longitude+distHaversine)+(f.MTA_tax+lpep_pickup_period+f.Extra),
family="binomial", data=train_dataset)

Anova(m7,test="Wald")

## Analysis of Deviance Table (Type II tests)
##
## Response: AnyTip
##              Df   Chisq Pr(>Chisq)
## tlenkm          1  0.9072  0.34085
## Pickup_longitude 1 20.3001 6.620e-06 ***
## distHaversine   1  2.8814  0.08961 .
## f.MTA_tax        1 21.8609 2.931e-06 ***
## lpep_pickup_period 3  8.7453  0.03288 *
## f.Extra          2  5.3584  0.06862 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova(m71,test="Wald")

## Analysis of Deviance Table (Type II tests)
##
## Response: AnyTip
##              Df   Chisq Pr(>Chisq)
## poly(tlenkm, 2)  2  7.6596  0.02171 *
## Pickup_longitude 1 20.4303 6.184e-06 ***
## distHaversine   1  0.0260  0.87181
## f.MTA_tax        1 22.4679 2.137e-06 ***
## lpep_pickup_period 3  8.1314  0.04337 *
## f.Extra          2  4.9681  0.08340 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si intentamos unir las variables numéricas explicativas y las variables categóricas explicativas, nos sale que las variables numéricas tlenkm y distHaversine deben eliminarse. Pero si intentamos hacer el polinomio ortogonal de base 2 respecto a la variable tlenkm, nos sale que solo debemos eliminar la variable distHaversine.

```
m8<-glm(AnyTip~ (poly(tlenkm,2)+Pickup_longitude)+(f.MTA_tax+lpep_pickup_period+f.Extra),
family="binomial", data=train_dataset)
m82<-glm(AnyTip ~ (poly(tlenkm,2)+Pickup_longitude)+(f.MTA_tax+lpep_pickup_period+f.Extra)^2,
family="binomial", data=train_dataset) # Interacciones dobles en factores
m83<-glm(AnyTip ~ (poly(tlenkm,2)+Pickup_longitude)*(f.MTA_tax+lpep_pickup_period+f.Extra),
family="binomial", data=train_dataset) # Interacciones dobles en factor-numèrica

BIC(m8,m82,m83)

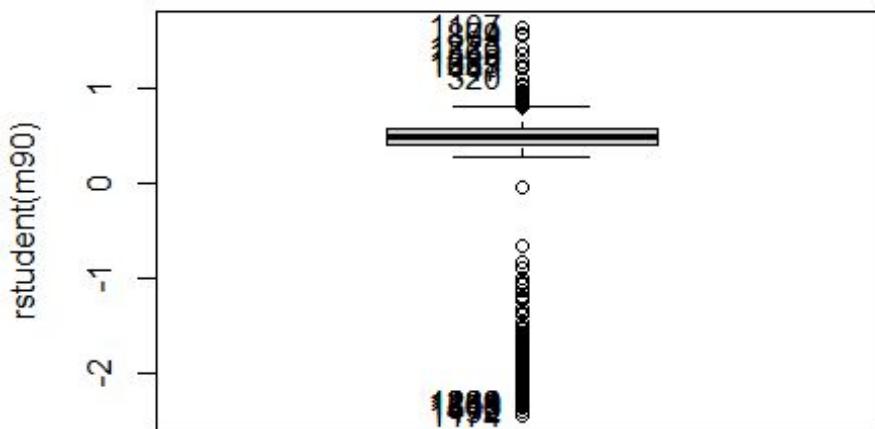
##      df      BIC
## m8  10 1409.216
## m82 18 1453.547
## m83 28 1510.916

m90 <- m8
```

Si intentamos comparar el modelo normal con el modelo con interacción doble en factor y el modelo con interacción doble en factor-numèrica, observamos que los modelos empeoran. Por lo tanto, seguiremos con el modelo normal.

## Final Diagnostics

### Residus



```
## [1] 141 545 683 813 1019 1139 1174 1177 1238 1239 1406 1464 1492
```

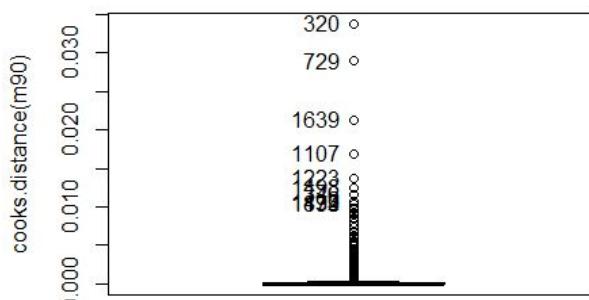
Como podemos ver, en total encontramos 13 individuos que tienen como residuo un valor superior a 2.25.

### Observacions potencialment influents

```
## [1] 47 49 60 75 101 106 139 186 312 320 340 346 363 447 485
## [16] 519 569 577 614 692 699 729 744 757 758 781 785 840 855 869
## [31] 874 961 1064 1079 1107 1132 1194 1210 1223 1367 1436 1442 1503 1569 1589
## [46] 1639 1653 1726
```

En total encontramos que 48 individuos son potencialmente influyentes.

### Influent data



```
## [1] 141 320 346 363 447 545 614 729 813 874 961 1107 1139 1174 1223
## [16] 1238 1239 1406 1464 1492 1498 1639
```

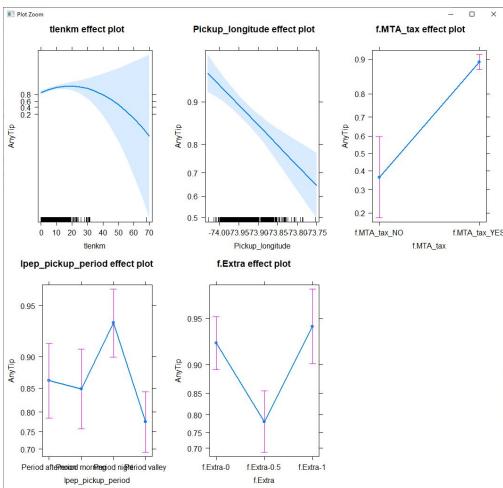
En total encontramos que 22 individuos son influyentes.

```
m10<-glm(AnyTip~ (poly(tlenkm,2)+Pickup_longitude)+(f.MTA_tax+lpep_pickup_period+f.Extra),
family="binomial", data=df1)
m101<-glm(AnyTip ~ (poly(tlenkm,2)+Pickup_longitude)+(f.MTA_tax+lpep_pickup_period+f.Extra)^2,
family="binomial", data=df1) # Interacciones dobles en factores
m102<-glm(AnyTip ~ (poly(tlenkm,2)+Pickup_longitude)*(f.MTA_tax+lpep_pickup_period+f.Extra),
family="binomial", data=df1) # Interacciones dobles en factor-numérica

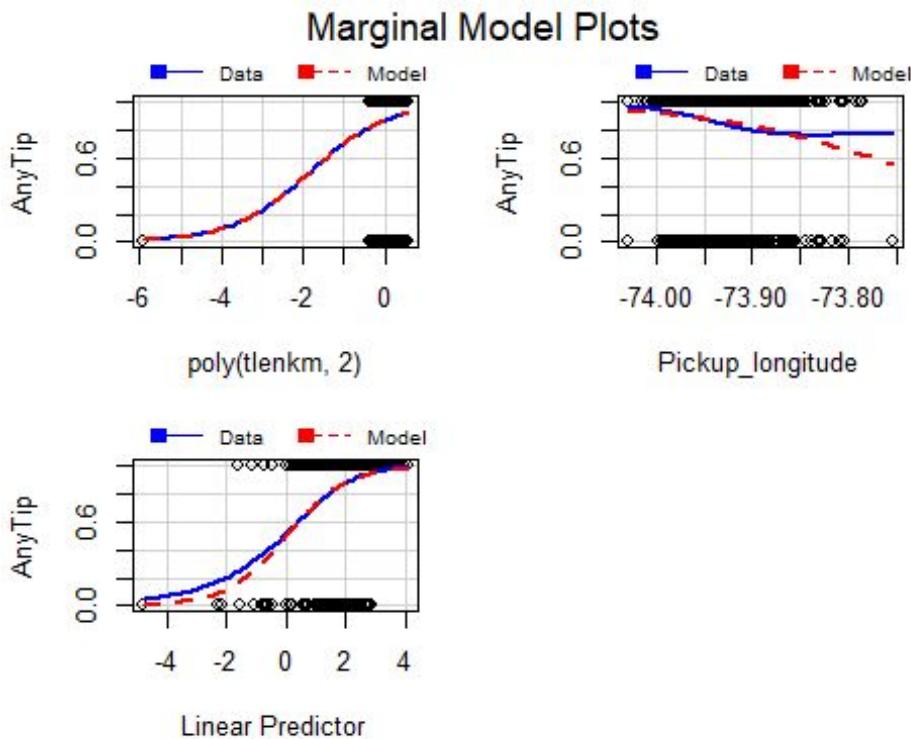
BIC(m10,m101,m102)

##      df      BIC
## m10  10 1328.457
## m101 17 1367.295
## m102 28 1429.212
```

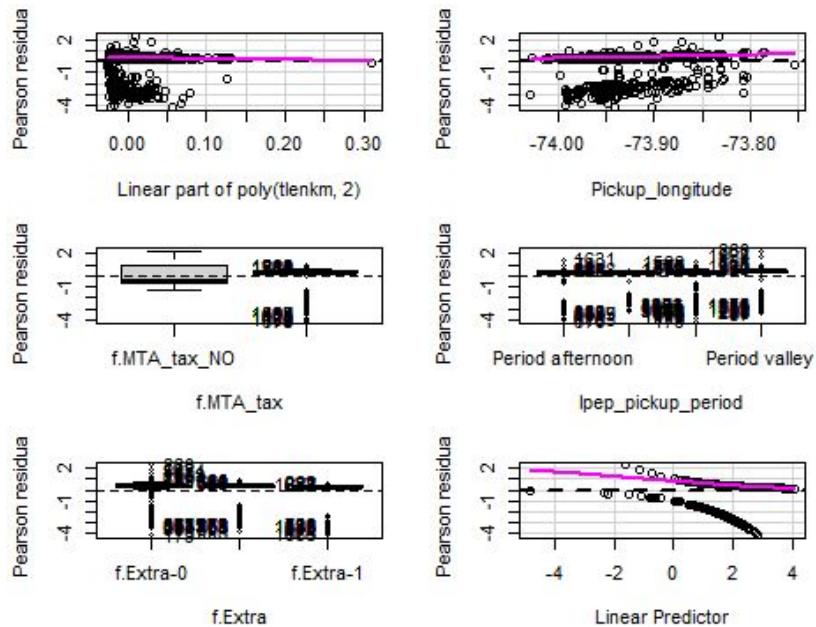
Si volvemos a recalcular el modelo observamos que el modelo cuya fórmula es un sumatorio, sigue siendo el mejor modelo para nuestros datos.



Si intentamos entender el modelo, vemos que si aumenta la distancia recorrida en km o aumenta el valor Pickup\_longitude disminuye la probabilidad de dar propina. Si se paga taxa entonces la probabilidad de dar propina es muy alta y si no se paga taxas entonces la probabilidad de no dar propinas es muy alta.



Si vemos el marginal modelo plot, vemos que la línea de los modelos se ajusta más o menos a la línea de los datos, por cada una de las variables.



Si vemos los residuos nos encontramos que en el plot de Linear Predictor - Pearson residuals vemos como la línea de smoother es inclinado por lo tanto tenemos desajuste en el modelo.

## Confusion Table

```
## AnyTip No AnyTip Yes
## 0.1323273 0.8676727

##
## fit.AnyTip      AnyTip No AnyTip Yes
## Prediction-AnyTip No      13      5
## Prediction-AnyTip Yes     215    1490

100*sum(diag(tt))/sum(tt)

## [1] 87.23157
```

Como podemos ver nuestro modelo tiene una precisión del 87.23157. A pesar de todo, no es un modelo totalmente correcto ya que podemos ver como en nuestro dataset hay más individuos con AnyTip Yes que con AnyTip No, por lo tanto nuestro dataset está desbalanceado. Además podemos ver que nuestro modelo tiene una tendencia a predecir siempre que el individuo da propinas, probablemente causado por el desbalance del dataset.