

Deliverable I

Carles Capilla Cànovas
Jesús Molina Roldán

Descripció de les dades	4
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
Imputació	21
Imputació variables categòriques	21
Imputació de variables numèriques	22
Multivariant outliers	23
Data quality report	25
Missings	25
Errors	25
Outliers	26
Discretització	27
espeed	27
tlenkm	27
traveltime	28
distHaversine	28
Fare_amount	28

Univariant exploratory analysis (EDA)	29
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
Profiling	35
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
```

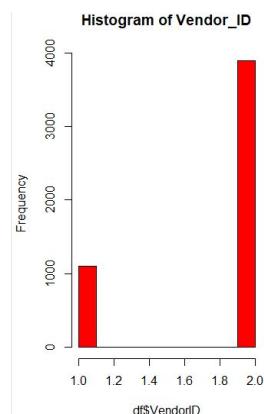
En aquesta primera vista de la mostra podem observar com, per exemple, la variable ehail_fee conté 5000 valors faltants identificats amb NA.

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

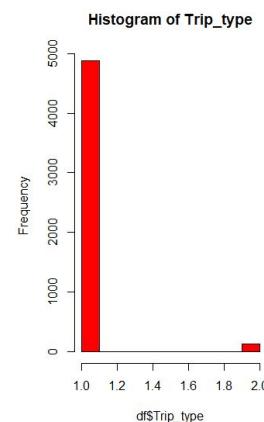


Començant per la variable VendorID on hi ha els valors 1 i 2 depenent de si es tracta de Mobile o Verizon respectivament, el que fem és assignar cada valor a un nivell(valors possibles de la variable) diferent. A més, modificarem el nom de les etiquetes per indicar que es tracta d'un factor de tipus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

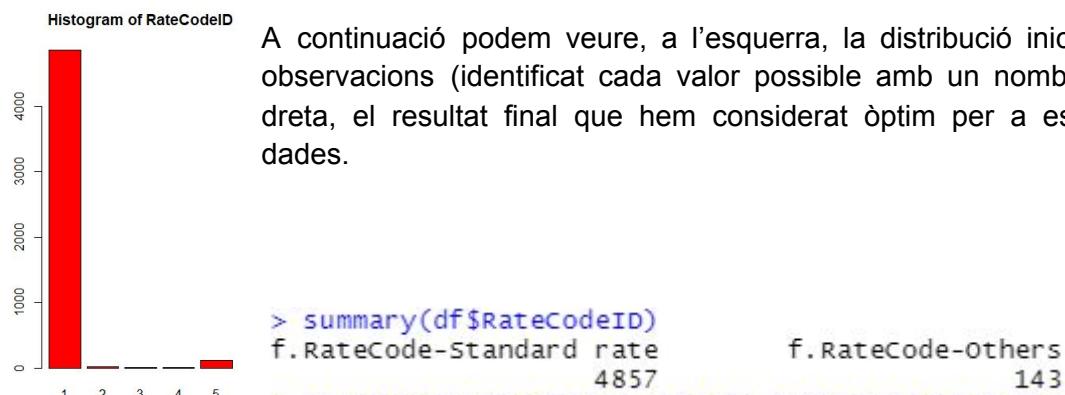
colnames(df)[which(names(df) == "lpep_dropoff_datetime")] <- "lpep_dropoff_datetime"
df_datatime <- t(as.data.frame(strsplit(as.character(df$lpep_dropoff_datetime), " ")))
df$lpep_dropoff_date <- factor(df_datatime[,1])
df$lpep_dropoff_time <- factor(df_datatime[,2])
```

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

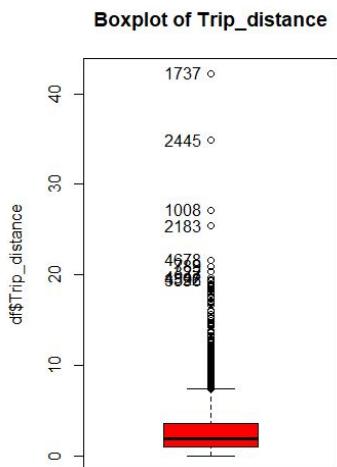


Pel que fa a la variable Passenger_count mitjançant el hist i el boxplot observem com gairebé totes les observacions prenen el valor d'un passatger, tot i això no considerarem els altres valors com a outliers ni errors, ja que pot haver-se reservat un taxi, per exemple, i registrar-se el servei però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

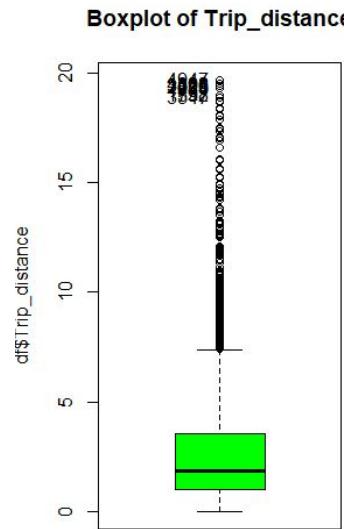
# errors
l <- which(df$Passenger_count<0)
if (length(l)>0) {
  ierrs[]<-ierrs[1]+1
  jerrs["Passenger_count"]<-length(l)
}
df[1,"Passenger_count"]<-NA
```

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[l,"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)
```

```
> 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
```

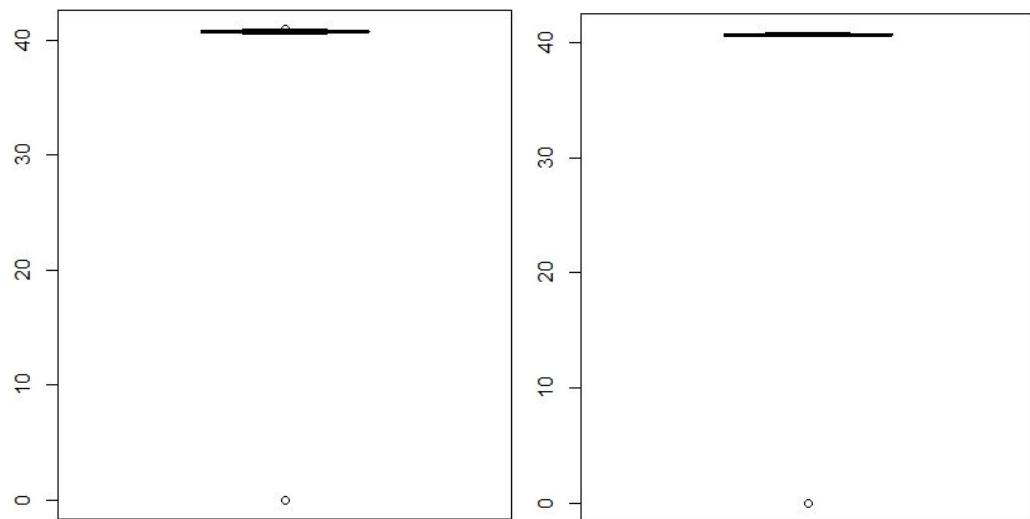
Al boxplot de color verd podem veure com acabaria quedant la distribució final de les observacions i, al summary, veiem els NA apareguts a causa dels 0 i els outliers.

Pickup_latitude i Dropoff_latitude

Pel que fa a la latitud de recollida i la referent a deixar els passatgers al seu destí, podem observar com tindrem alguns valors erronis. Això ho podem afirmar perquè tenint en compte que es tracta d'unes dades referents a la zona de Nova York, no és coherent tenir valors de 0 pel que faria a aquesta latitud veient que la majoria ronda els 40.7.

A més, gràficament podem veure la distribució inicial de les observacions que tot i poder arribar a presentar algun altre outlier més suau un cop eliminades les observacions de 0, aquestes són les més evidents a eliminar.

```
> summary(df$Pickup_latitude)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   0.00 40.69 40.75 40.69 40.80 40.89
> summary(df$Dropoff_latitude)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   0.00 40.70 40.75 40.71 40.79 40.94
```

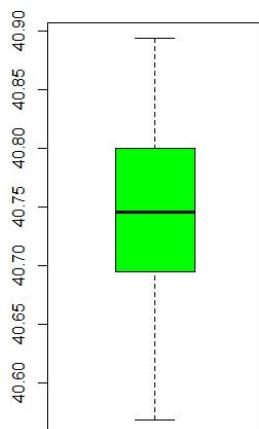


Vist l'error ens disposem a eliminar les observacions que tinguin per valor 0 i posar-les com a NA

```
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
```

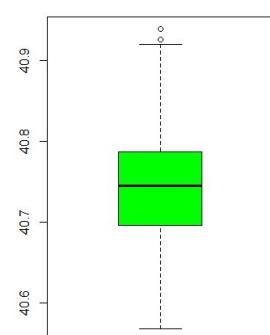
Boxplot of Pickup_latitude



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
```

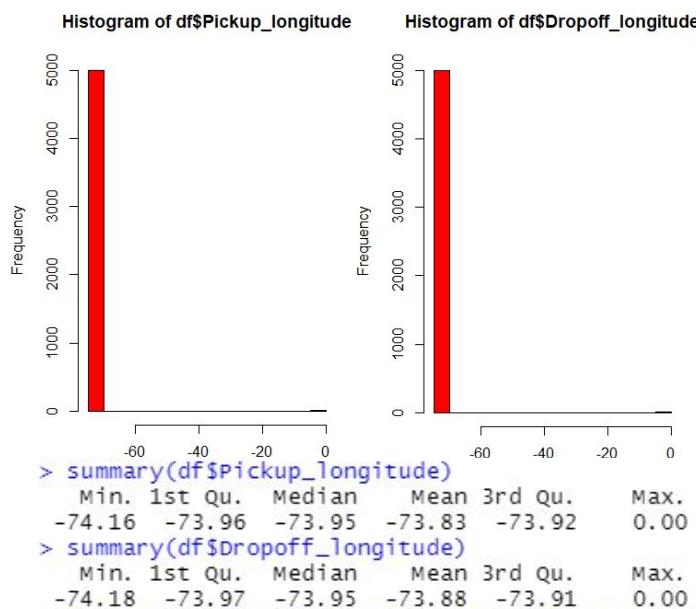
Boxplot of Dropoff_latitude



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

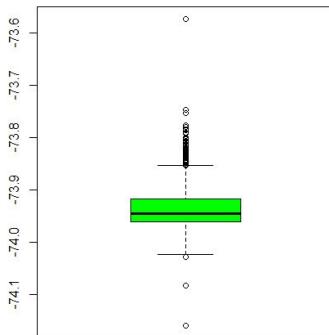


Es presentaria la mateixa situació que pel que feia a les latituds amb els valors de 0 tot i que les longituds rondarien el valor de -74. Aquest cop mostrem visualment les variables amb un hist en comptes d'un boxplot, ja que els boxplots serien semblants als dos anteriors i així mostrem d'un altre manera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

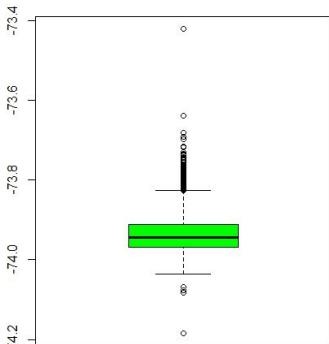
Procedim a eliminar les observacions amb valor 0 i marcar-les com NA de la variable Pickup_longitude.

```
### 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.

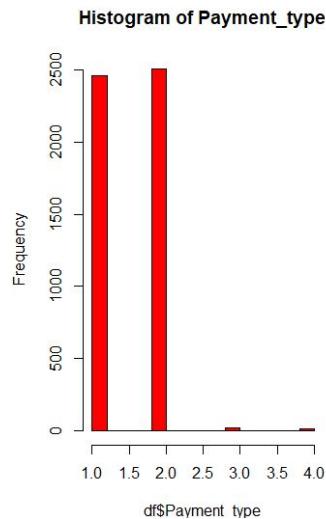
Store_and_fwd_flag

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

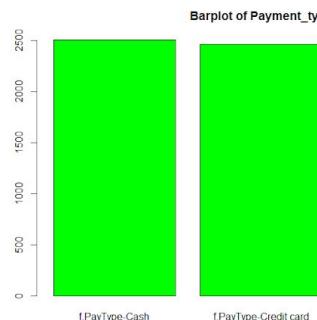


Com podem observar la variable Payment_type és una variable categòrica i com diu la seva definició, té 6 valors possibles. Degut a que en la primera vista que tenim de les observacions de la variable en aquesta llavor només tenim 4 d'aquests 6 possibles valors, procedirem a factoritzar-la ambaquests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



```
> summary(df$Payment_type)
f.PayType-Cash f.PayType-Credit card
2507          2464
f.PayType-Others
29
```

Fare_amount

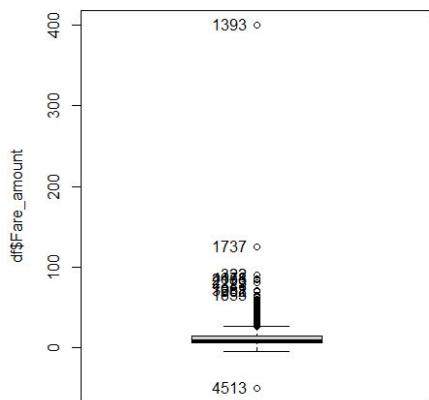
A l'hora de fer un summary de la variable Fare_amount com a primer error que salta a la vista

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

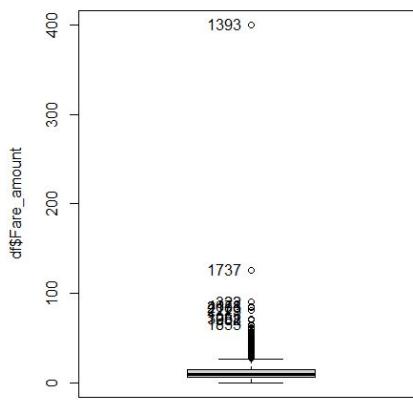
```
> summary(df$Fare_amount)
```

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

Boxplot of Fare_amount



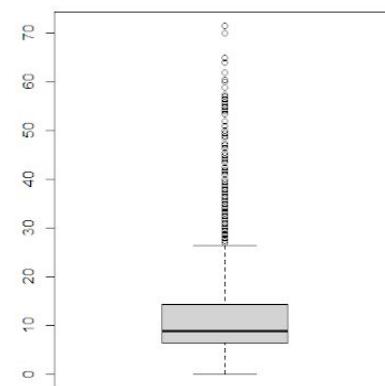
Boxplot of Fare_amount



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

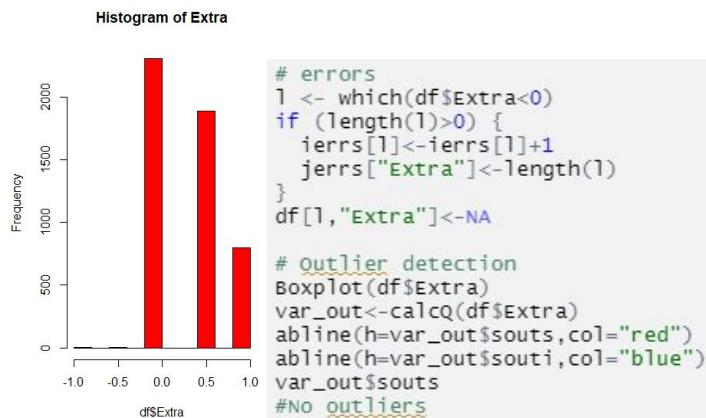


```
> summary(df$Fare_amount)
```

	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.

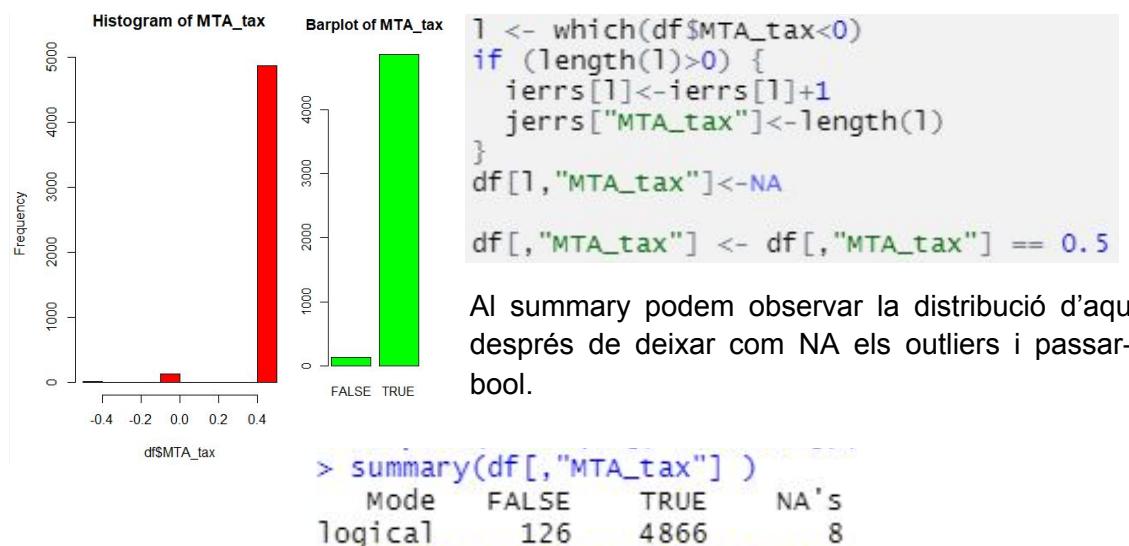


A part de les observacions negatives no trobem d'altres que podem considerar outliers pel que faria a aquesta variable.

```
> 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
```

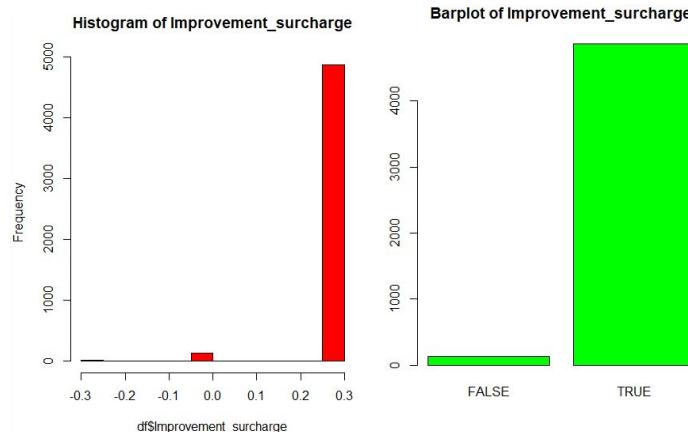
MTA_tax

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.

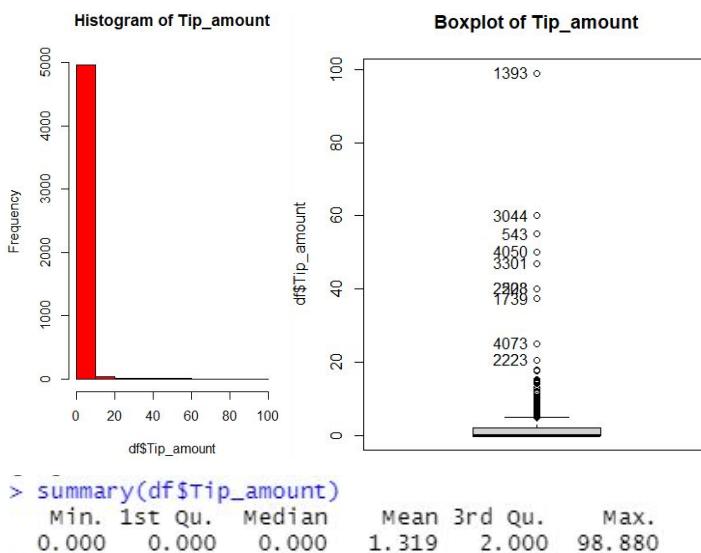


Al Barplot de color verd i al summary podem veure la distribució final que pren la variable.

```
> 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.

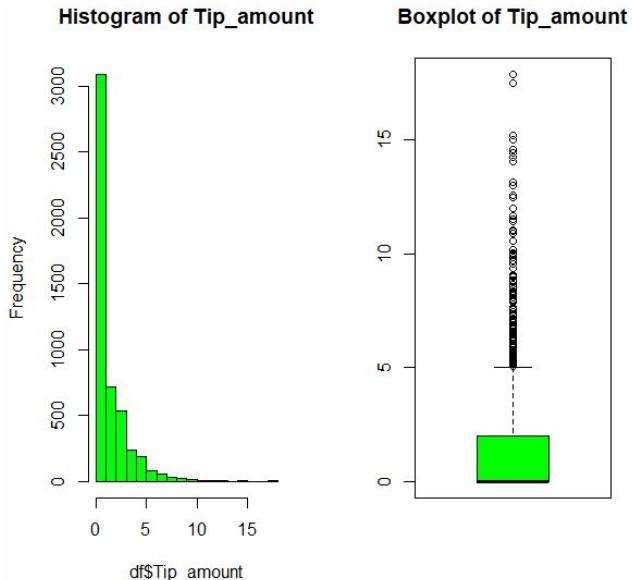
```

# 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

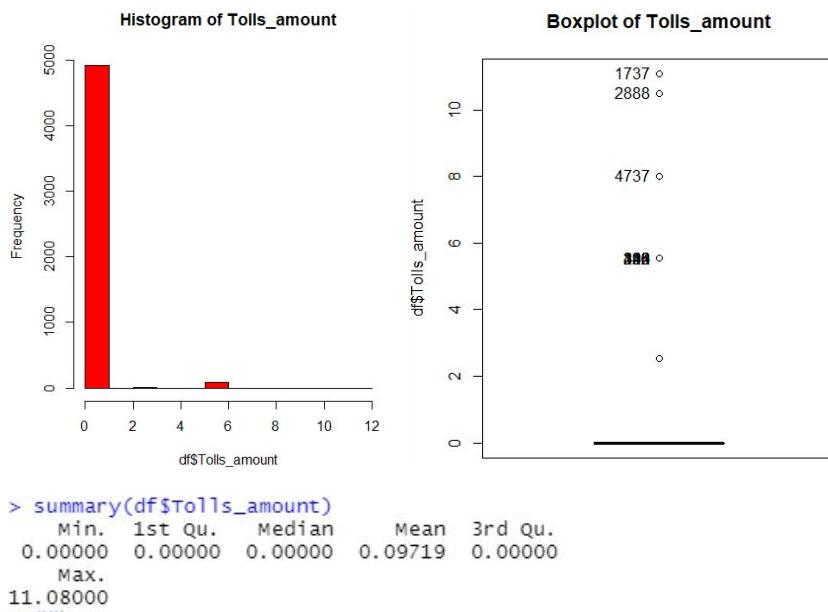
```



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.



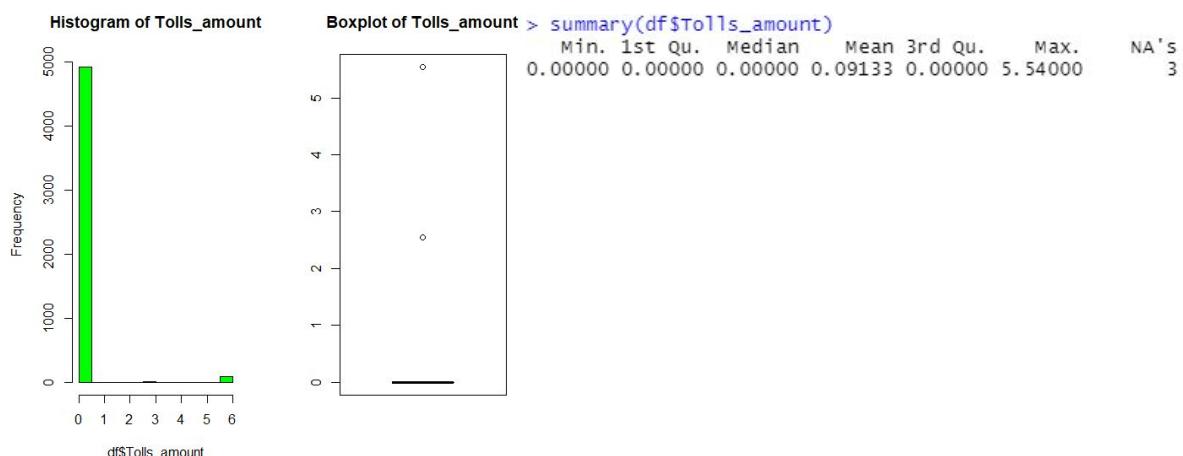
```

# 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

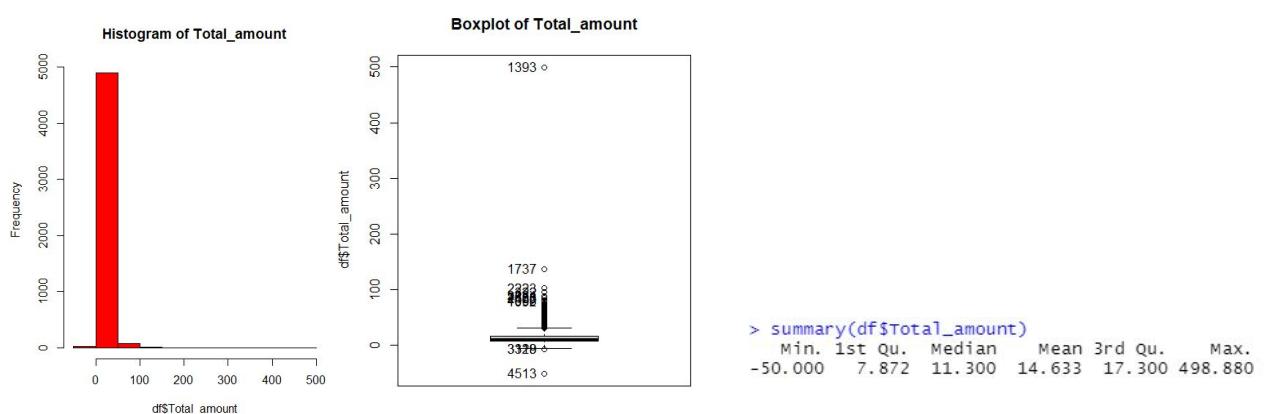
```

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.

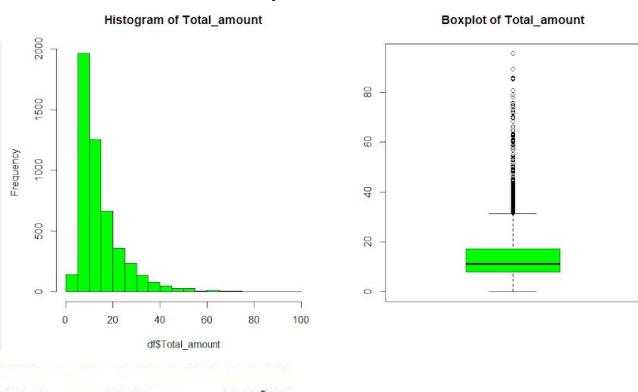
```

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

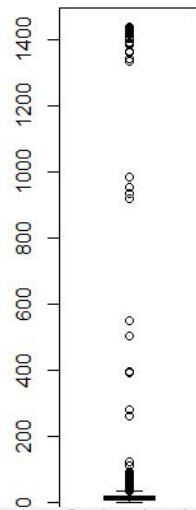
```

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.

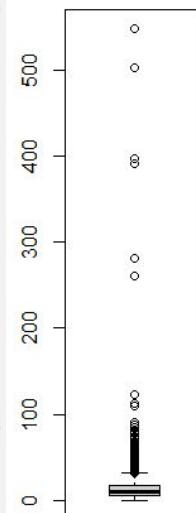
```
# Travel time in min
df$traveltime<-as.numeric(as.POSIXct(df$lpep_dropoff_datetime)) -
as.numeric(as.POSIXct(df$lpep_pickup_datetime))/60

#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
```

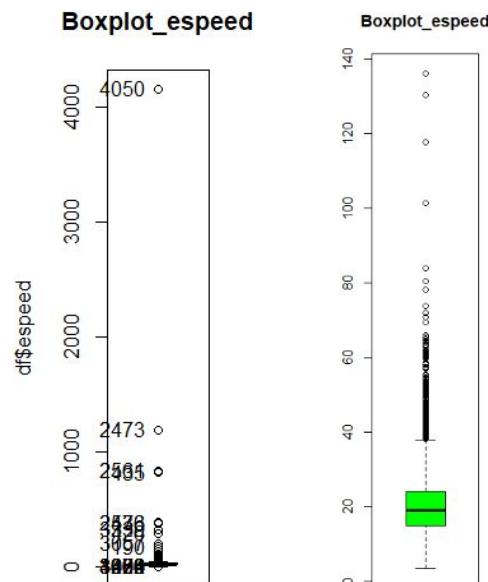
Com podem observar al summary i al primer boxplot, apareixen valors que no tenen cap sentit al voltant de 1000km/h i fins a 4000km/h. És per això que hem decidit comptar com a error i computar com a NA tota velocitat efectiva que superi els 140 km/h, ja que tot i ser una velocitat bastant elevada, podria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

lpep_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 2016-01-31: 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.Extra-1 : 798

      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
                                         2016-01-31: 168
                                         2016-01-09: 186
```

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. : 0.0000      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.: 0.0000      1st Qu.:0.00000
Median :-73.95     Median :40.75     Median :-73.95     Median :40.75     Median : 9.00      Median : 0.0000      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  : 4      NA's  : 15     NA's  :126     NA's  : 10

Tolls_amount    tpep_pickup_time traveltime disthaversine Total_amount
Min. :0.00000  Min. : 0.00000  Min. : 0.0000  Min. : 0.00000  Min. : 0.00000
1st Qu.:0.00000 1st Qu.: 9.000  1st Qu.: 5.933  1st Qu.: 1.242  1st Qu.: 7.88
Median :0.00000  Median :15.000  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.:19.000  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. : 0.00  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

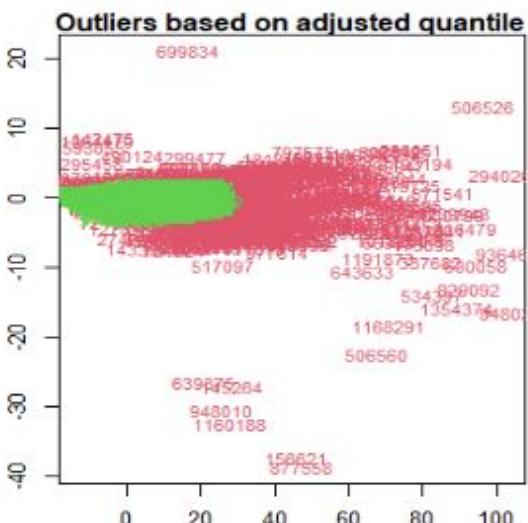
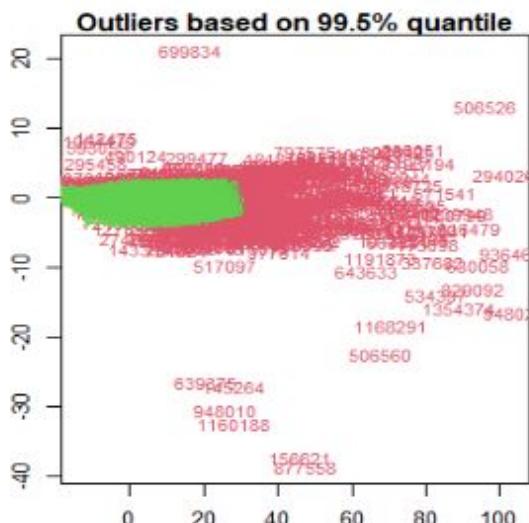
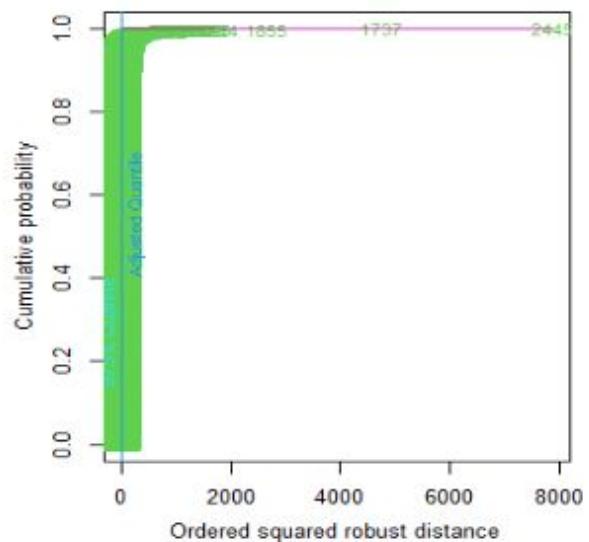
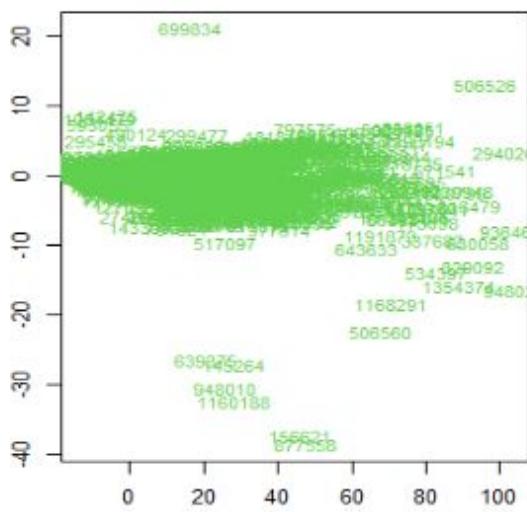
```

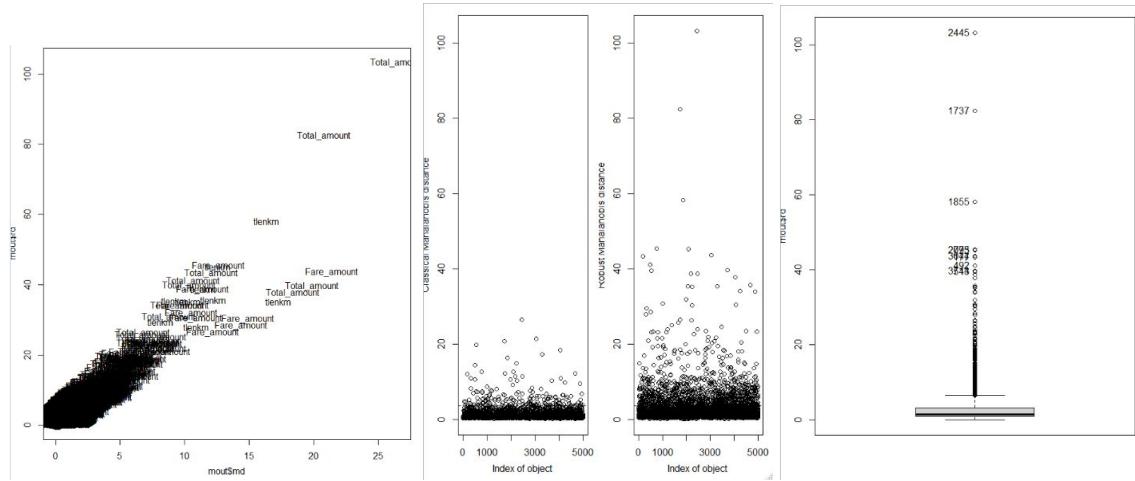
```

## 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"])
``
```

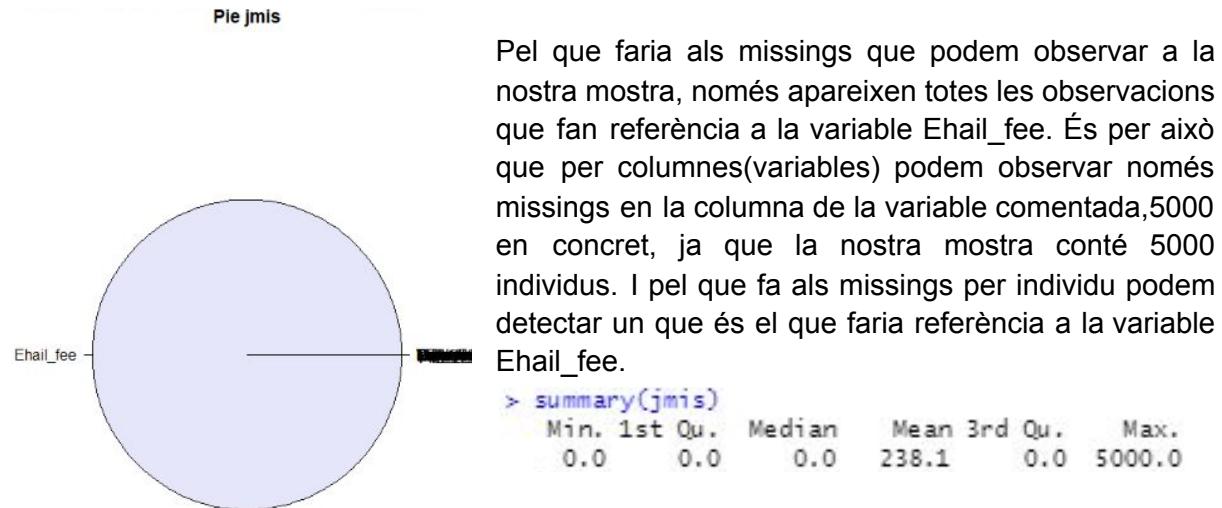




Un cop fet Moutlier, es va crear la variable multiouts. Aquells individus amb distància de Mahanalobis robusta superior a 50 se li va assignar el valor TRUE a la variable multiouts. Altrament, multiouts se li va assignar el valor FALSE.

Data quality report

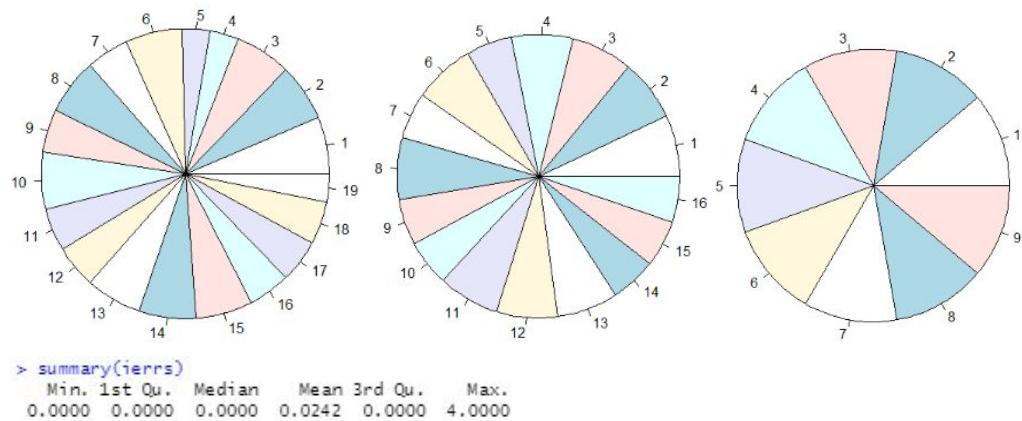
Missings



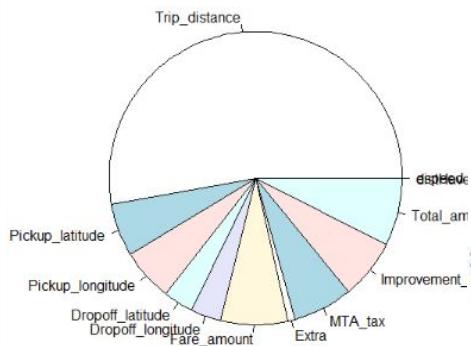
Errors

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.



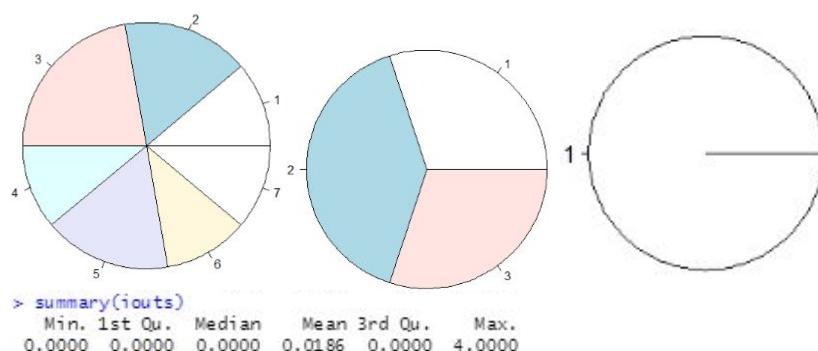
Pie jerrs

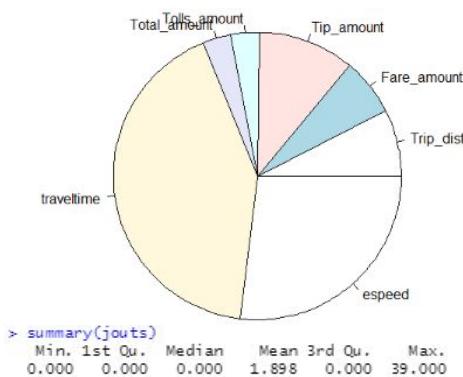


En aquest gràfic en forma de pastís podem veure una comparació entre els errors que presenta la mostra per variable. Podem observar que la que més errors presenta seria 64 errors la variable Trip_distance i, que la mitjana d'errors per variable és de 2.24, un valor molt baix tenint en compte que disposem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

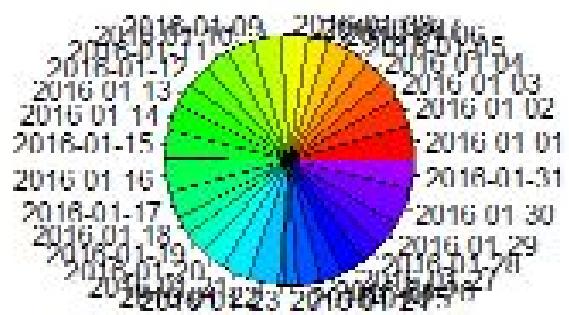
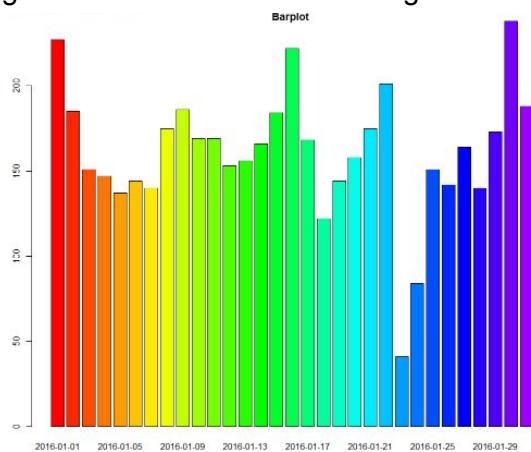
      f.Passenger_count Passenger_count      tlenkm      Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Fare_amount      espeed
f.Passenger_count-1 :4207 Min. :0.0000 Min. : 0.001 Min. :-74.16 Min. :40.57 Min. :74.18 Min. :40.57 Min. : 0.001
f.Passenger_count-2 : 367 1st Qu.:1.0000 1st 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. :0.00000 Min. : 0.000 Min. : 0.000 Min. : 0.000 AnyTip No :2869 Min. : 0.000
1st Qu.: 0.00000 1st Qu.:0.00000 1st Qu.: 9.00 1st Qu.: 5.950 1st Qu.: 1.242 AnyTip Yes:2131 1st Qu.: 7.872
Median : 0.00000 Median :0.00000 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.:0.00000 3rd Qu.: 19.00 3rd Qu.: 16.350 3rd Qu.: 4.126 3rd Qu.:17.300
Max. :17.88000 Max. :5.54000 Max. :168.53 Max. :548.117 Max. :26.106 Max. :95.540
```

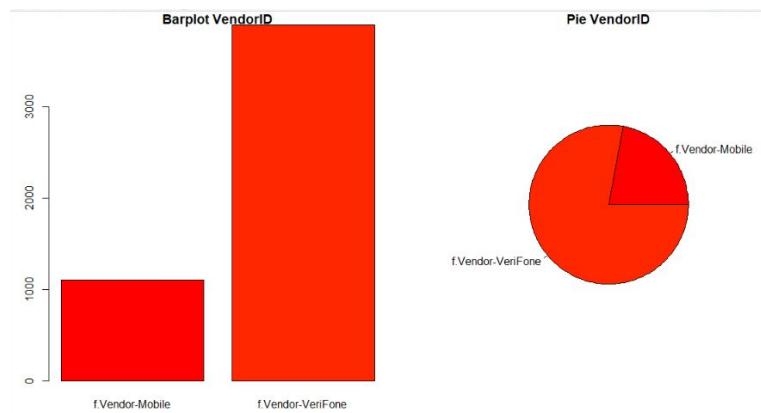
Aquest summary ens permetrà treure algunes conclusions del nostre dataset juntament amb les descripcions gràfiques.

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.

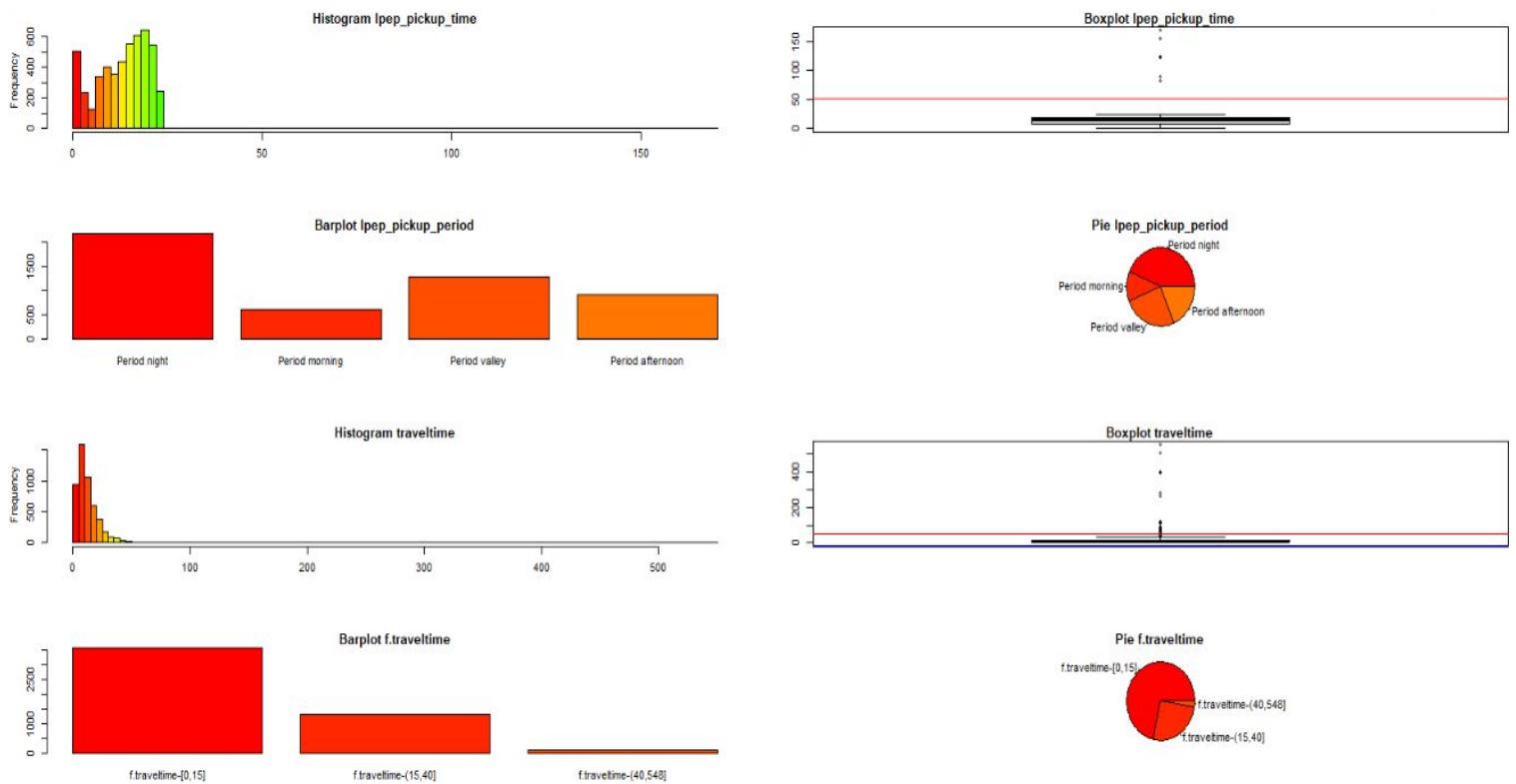


VendorID



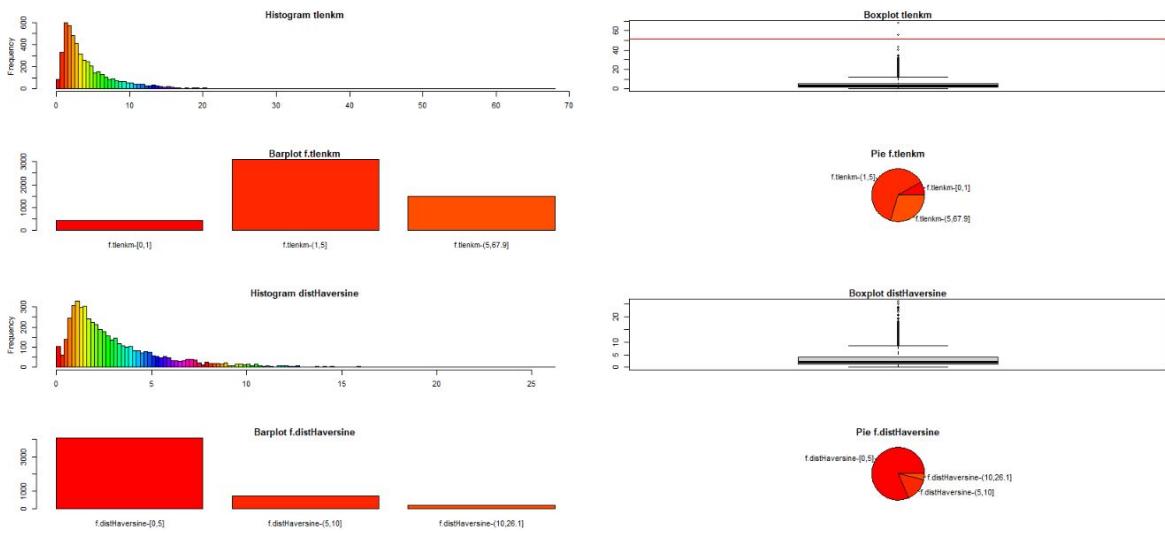
Com podem veure, la gran majoria de tuples del nostre dataset està proporcionat per VeriFone.

lpep_pickup_time, lpep_pickup_period, travelttime, f.travelttime



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.

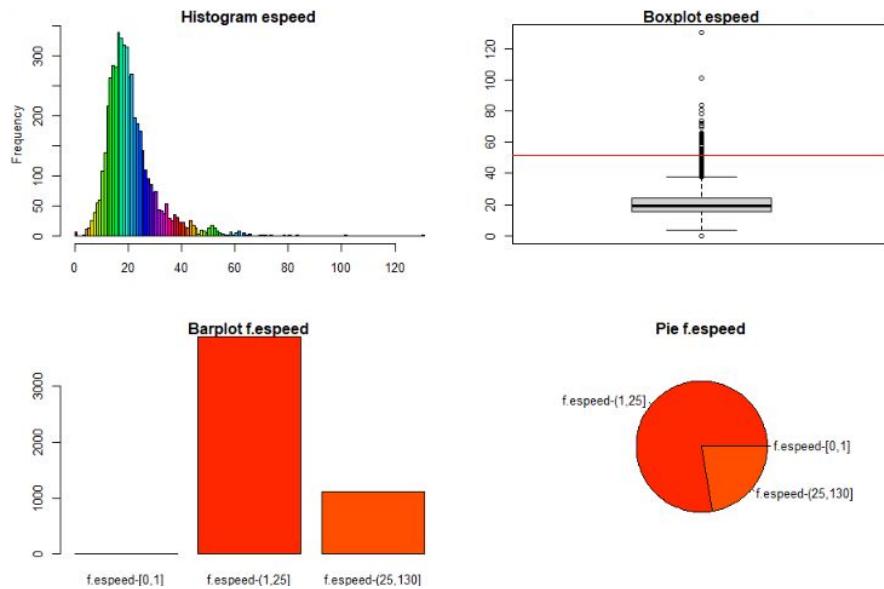
tlenkm, distHaversine



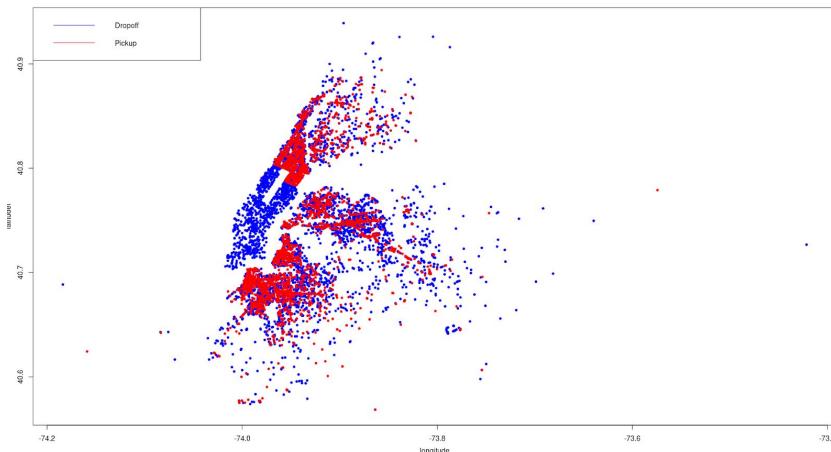
Com hem vist en els gràfics anteriors, la duració dels viatges era d'entre 0 i 15 minuts és per això, que la distància recorreguda pels taxis sigui inferior a 15 km i que la distància entre punts sigui inferior a 5 km. Aquesta informació ens permet dir que els viatges feien trajectes curtes.

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.



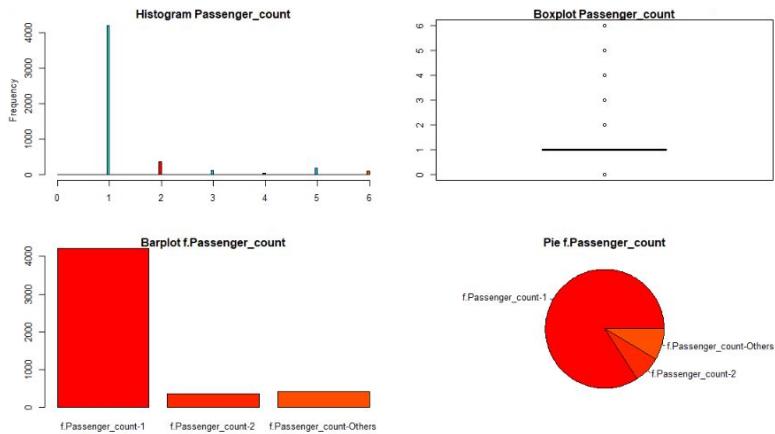
`lpep_pickup_latitude`, `lpep_pickup_longitude`, `lpep_dropoff_latitude`,
`lpep_dropoff_longitude`



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.

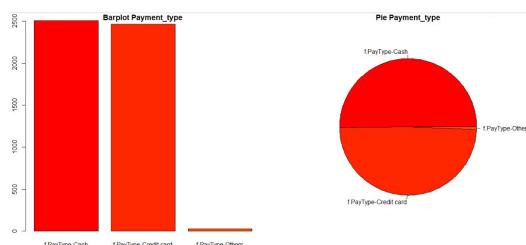
Passenger_count

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

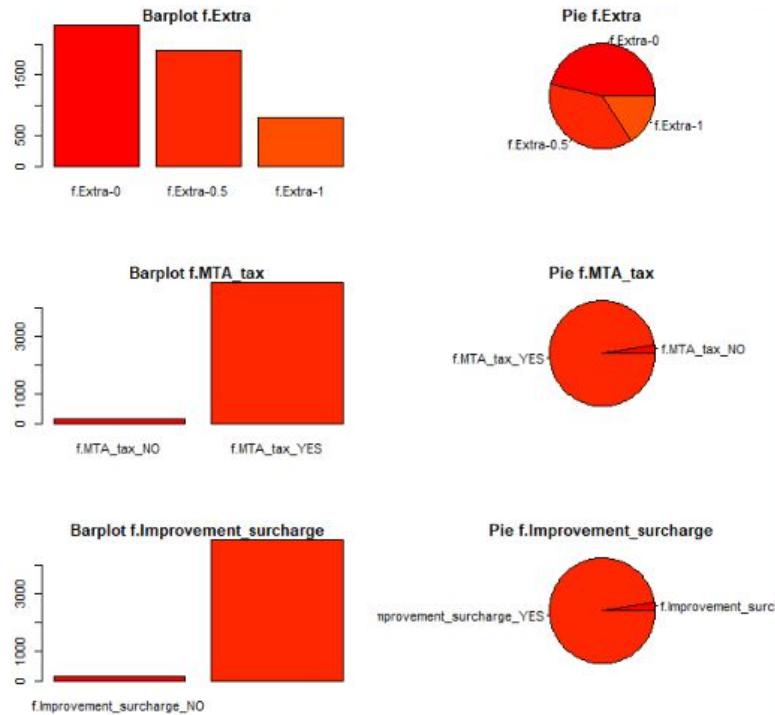


Payment_type

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

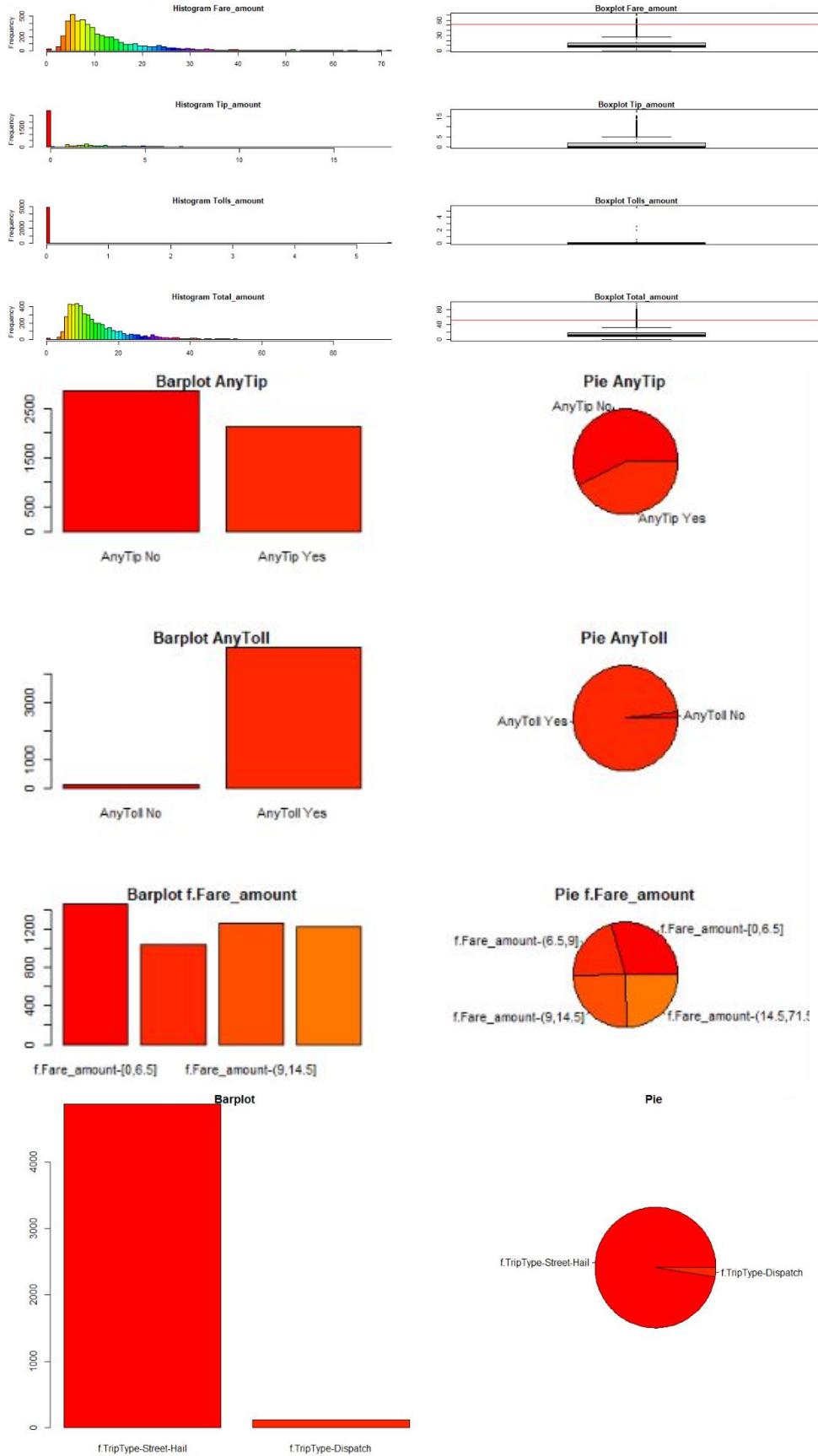


Respecte als extres, podem veure que en la meitat de les ocasions no s'aplica extra per nocturnitat i hores puntes. Aquesta informació no té cap sentit si observem la variable lpep_pickup_period, ja que ens deia que molts dels viatges eran nocturns. Respecte a les variables MTA_tax i Improvement_surcharge, 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 13.00695521 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 3.21886536 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 2.8883275 1.23014878 2.07738644 1.06892443 0.000000e+00
Total_amount   18.399775 17.651990 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 16.22079635 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 9.096866e-07

```

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 43.31551 98.826842
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=lpep_pickup_date[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
lpep_pickup_period=Period valley
Payment_type=f.PayType-Others
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
      Cla/Mod Mod/Cla Global p.value v.test
86.48539 100.000000 49.28 0.000000e+00 Inf
51.68464 35.992492 29.68 4.817935e-17 8.391056
52.30537 32.473817 26.46 1.310886e-16 8.272582
52.72136 30.455185 24.62 2.082215e-16 8.217251
43.30845 98.873766 97.32 8.573293e-10 6.133932
43.22329 98.926969 97.54 1.472744e-09 5.664709
51.170339 17.604294 14.54 9.019685e-08 5.336985
43.170306 16.404005 12.02 2.021134e-02 4.751300
40.87864 25.246363 22.38 2.959330e-05 4.176558
64.36782 2.627074 1.74 4.264542e-05 4.092659
55.67010 5.068043 3.88 2.011444e-04 3.715074
46.11727 27.311122 25.24 3.743941e-03 2.889967
46.23753 23.932426 22.06 6.051906e-03 2.744956
50.42917 5.631159 4.76 1.326216e-02 2.476651
47.28171 13.467855 12.14 1.356961e-02 2.468459
44.68535 39.652745 37.82 2.139829e-02 2.300882
52.72727 2.721727 2.28 3.182030e-02 2.146661
50.28571 4.129517 3.58 3.828862e-02 2.072616
48.28962 4.082591 3.46 3.928834e-02 2.061154
47.68392 8.212107 7.34 4.246028e-02 2.029891
35.82090 3.378695 4.02 4.587826e-02 1.996512
40.71830 44.157677 46.22 1.171496e-02 2.528616
39.85497 19.005161 20.74 8.983861e-03 2.612698
41.30518 76.304294 77.94 6.019593e-03 2.744956
31.27710 3.331769 3.22 2.021134e-02 4.751302
42.23489 97.377126 98.26 4.264542e-05 4.092659
41.05806 74.659784 77.59 3.615147e-05 4.130978
40.83226 58.235574 62.09 2.369828e-06 4.719898
23.77622 1.595495 2.86 2.021134e-02 4.751300
36.68731 22.243078 25.84 4.946405e-07 5.023890
0.00000 0.00000 0.58 9.509930e-08 5.336985
18.69919 1.079305 2.46 1.472744e-08 5.664709
28.56731 5.771938 8.32 1.103123e-08 5.714960
17.91045 1.126232 2.68 8.573293e-10 6.133932
f.Improvement_surcharge=f.Improvement_surcharge_NO
f.distHaversine=f.distHaversine[0,5]
f.Fare_amount=f.Fare_amount[0,15]
33.67347 23.228531 29.49 9.327654e-17 8.313046
38.71601 64.805256 71.34 1.723608e-18 8.774045
0.00000 0.00000 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]        88.94099 3.809999 77.14 1.928576e-10 4.070913
AnyToll=AnyToll_No                  57.72531 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          58.40896 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 11.188112 18.14 4.264542e-02 -2.476539
lpep_pickup_date=2016-01-30         49.57473 4.112082 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.