CASE_STUDY

Katerina Dimitrova, Jose Romero, Sergi ..

March 18, 2018

Introduction

Load requiered packages

Select 5000 samples

```
#Load samples

### Use birthday of 1 member of the group
set.seed(28061963)
nrow(df)
```

```
## [1] 5000
```

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

df<-df[sam,]
#save.image("Taxi5000_raw.RData") # Dont execute again since it will create a new data and the followin
```

Delete unnecessary attributes

```
load("Taxi5000_raw.RData")
table(df$Ehail_fee) ##Delete unnecessary row
```

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

```
df$Ehail_fee<-NULL
table(df$Passanger_count) ##Delete unnecessary row
```

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

```
df$Passanger_count<-NULL

# Now one by one describe vars
names(df)
```

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

```
## [19] "Payment_type"          "Trip_type"
```

Converting numeric variables corresponding to qualitative concepts to factors

VendorID

```
sel<-which(df$VendorID==0.0);length(sel) #No missing Data
```

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

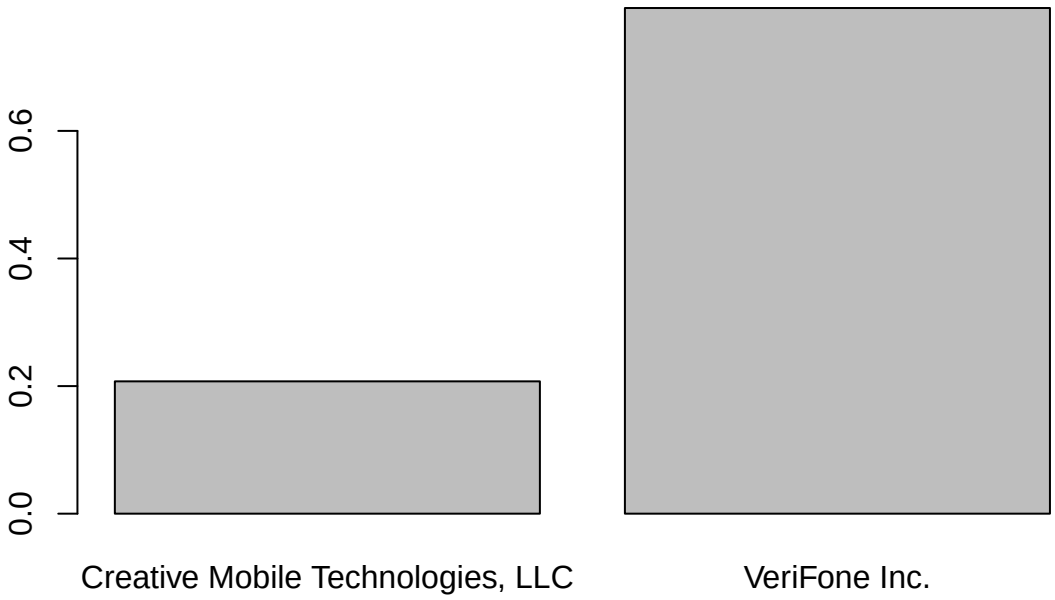
```
df$VendorID<-factor(df$VendorID,labels=c("Creative Mobile Technologies, LLC","VeriFone Inc."))
summary(df$VendorID)
```

```
## Creative Mobile Technologies, LLC                   VeriFone Inc.
##                              1037                            3963
```

```
table(df$VendorID)
```

```
##
## Creative Mobile Technologies, LLC                   VeriFone Inc.
##                              1037                            3963
```

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



RateCodeID, there whas no group ride

```
sel<-which(df$RateCodeID==0.0);length(sel) #No missing Data
```

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

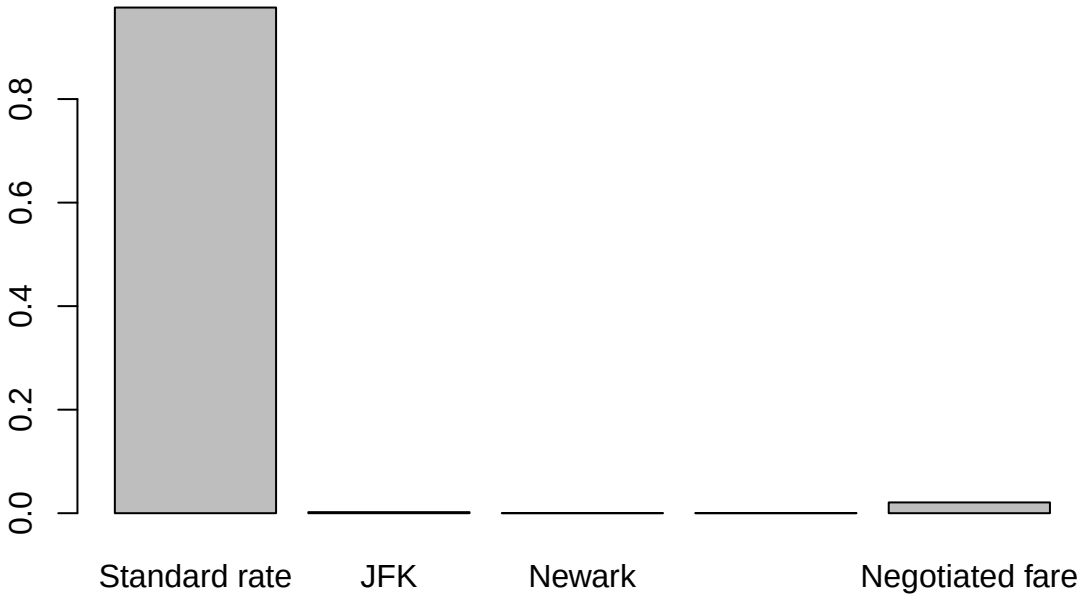
```
df$RateCodeID<-factor(df$RateCodeID,labels=c("Standard rate","JFK","Newark","Nassau or Westchester","Neg
summary(df$RateCodeID)
```

```
##         Standard rate                  JFK                Newark
##                  4884                   10                     1
## Nassau or Westchester      Negotiated fare
##                     1                  104
```

```
table(df$RateCodeID)
```

```
##
##         Standard rate                  JFK                Newark
##                  4884                   10                     1
## Nassau or Westchester      Negotiated fare
##                     1                  104
```

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



Store_and_fwd_flag

```
//first the N and than Y
sel<-which(df$Store_and_fwd_flag==0.0);length(sel) #No missing Data
```

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

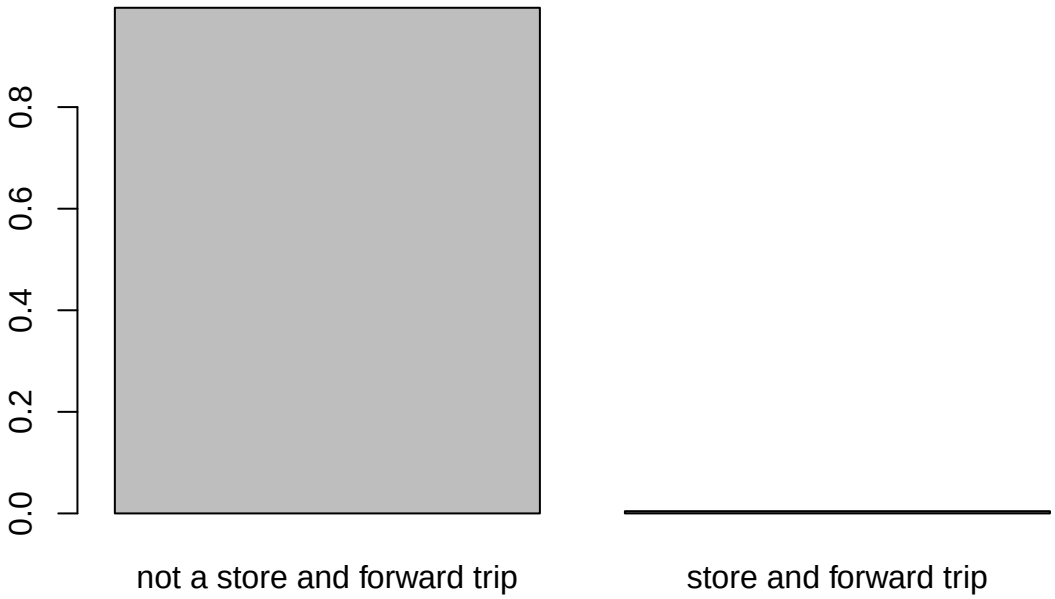
```
df$Store_and_fwd_flag<-factor(df$Store_and_fwd_flag,labels=c("not a store and forward trip","store and f
summary(df$Store_and_fwd_flag)
```

```
## not a store and forward trip       store and forward trip
##                         4978                           22
```

```
table(df$Store_and_fwd_flag)
```

```
##
## not a store and forward trip       store and forward trip
##                         4978                           22
```

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



Payment_type

```
sel<-which(df$Payment_type==0.0);length(sel) #No missing Data
```

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

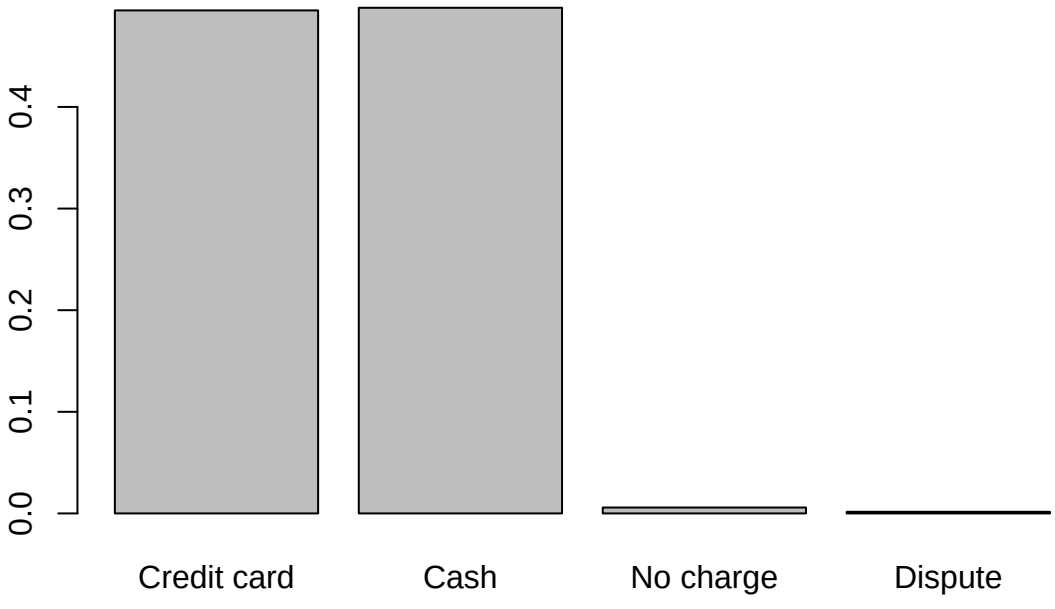
```
df$Payment_type<-factor(df$Payment_type,labels=c("Credit card","Cash", "No charge", "Dispute"))
summary(df$Payment_type)
```

```
## Credit card        Cash   No charge     Dispute
##        2475        2488          29           8
```

```
table(df$Payment_type)
```

```
##
## Credit card        Cash   No charge     Dispute
##        2475        2488          29           8
```

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



Trip_type

```
sel<-which(df$Trip_type==0.0);length(sel) #No missing Data
```

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

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

```
## Street-hail    Dispatch
##        4898         102
```

```
table(df$Trip_type)
```

```
##
## Street-hail    Dispatch
##        4898         102
```

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

Creating additional factors as a discretization

Factorize function:

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

Passenger_count

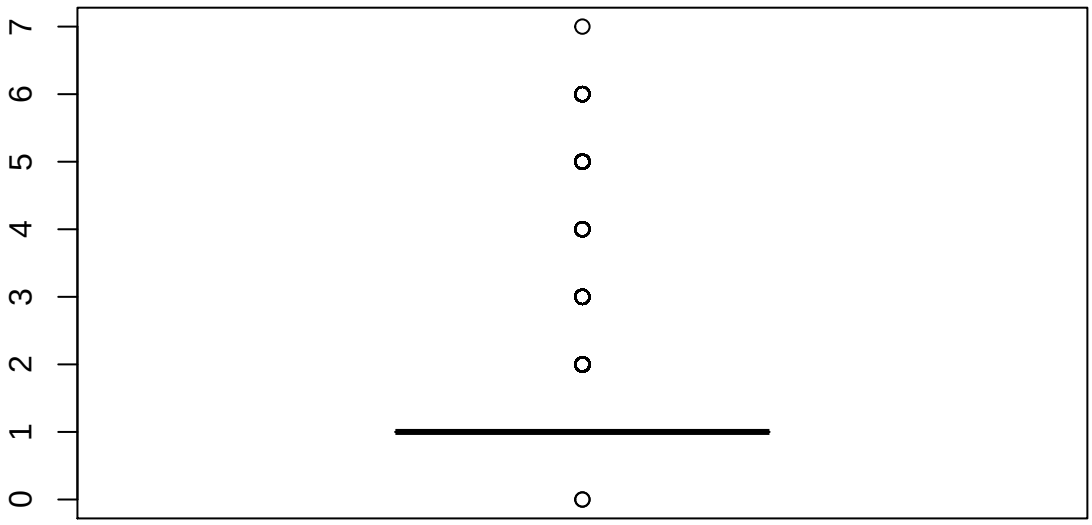
```
df$f.passanger<-factorize(df$Passenger_count)
summary(df$f.passanger)
```

```
## (0,1] (1,7]  NA's
##  4236   762     2
```

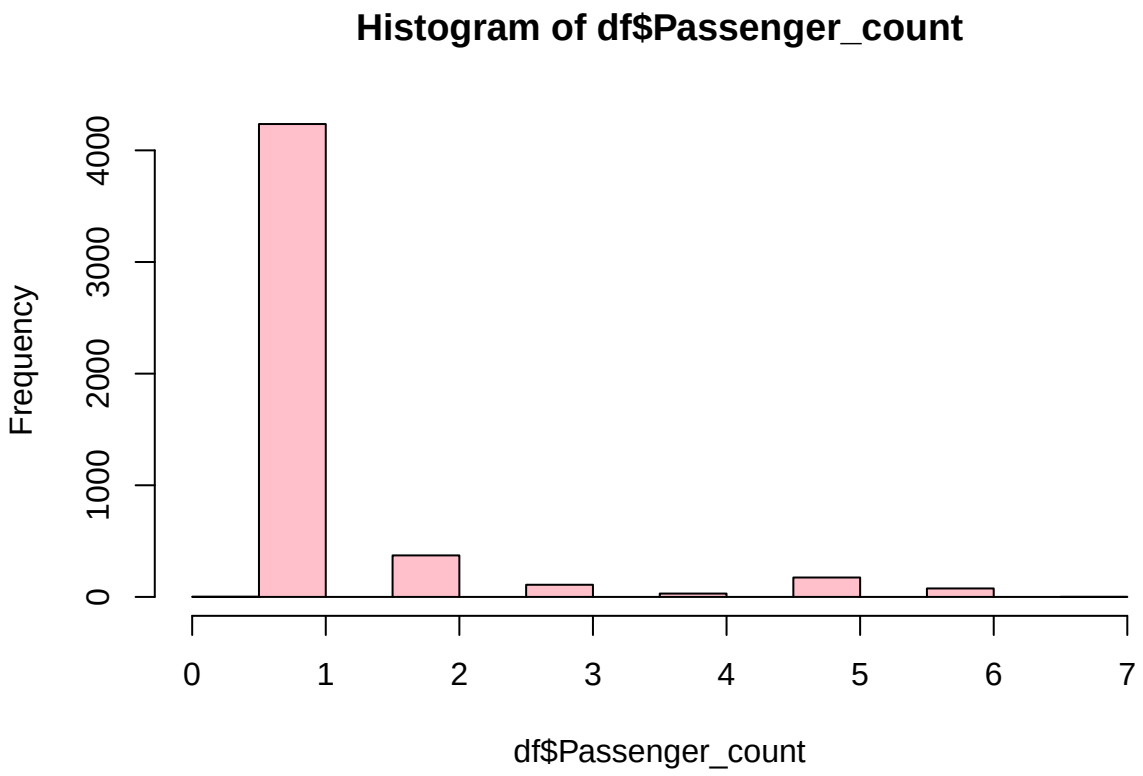
```
sel<-which(df$Passenger_count==0.0);length(sel) #2 missings
```

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

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




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



Trip_distance

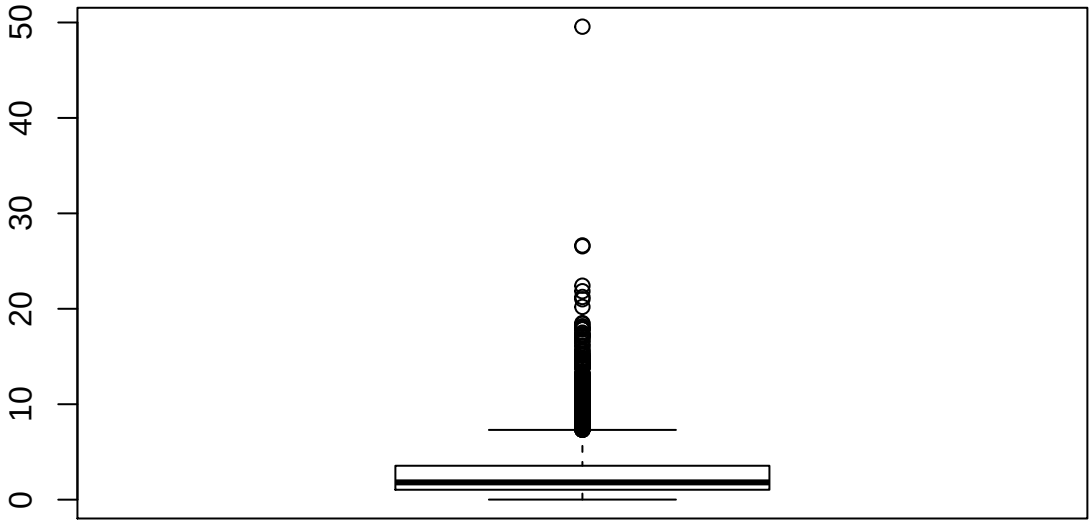
```
df$f.distance<-factorize(df$Trip_distance) # NO VA be
summary(df$distance)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

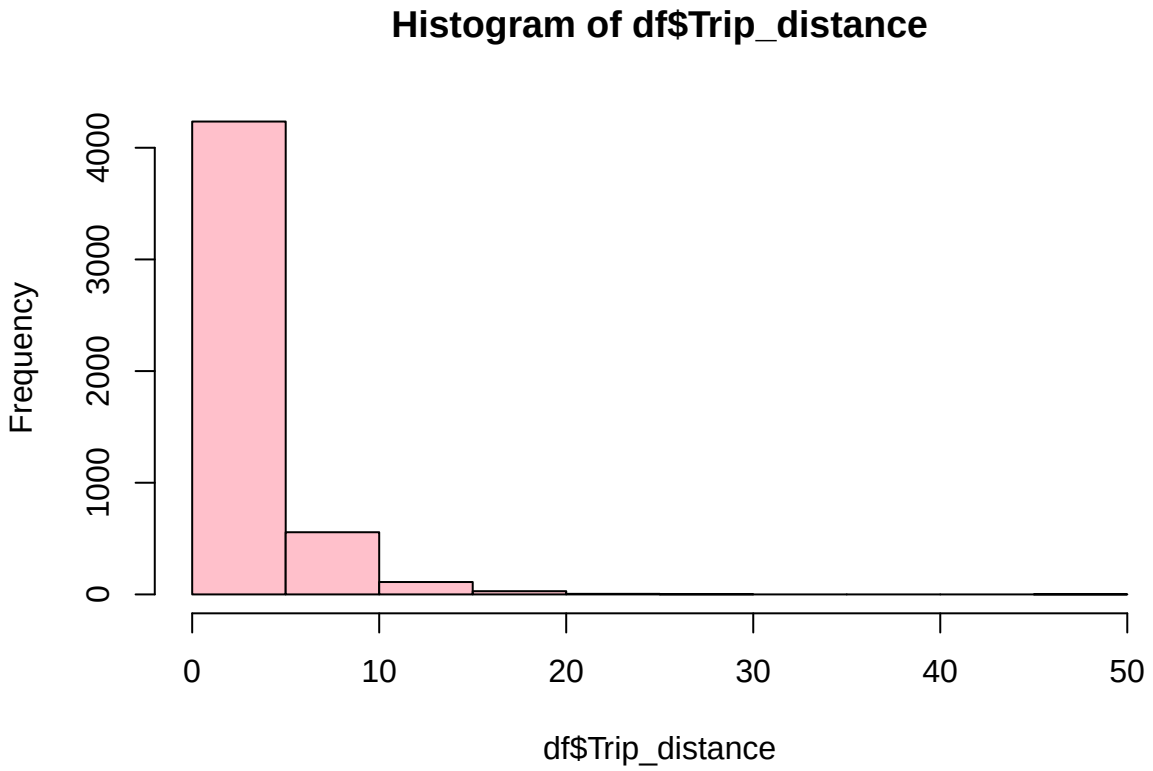
```
sel<-which(df$Trip_distance==0.0);length(sel) #60 missings
```

```
## [1] 60
```

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




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



##

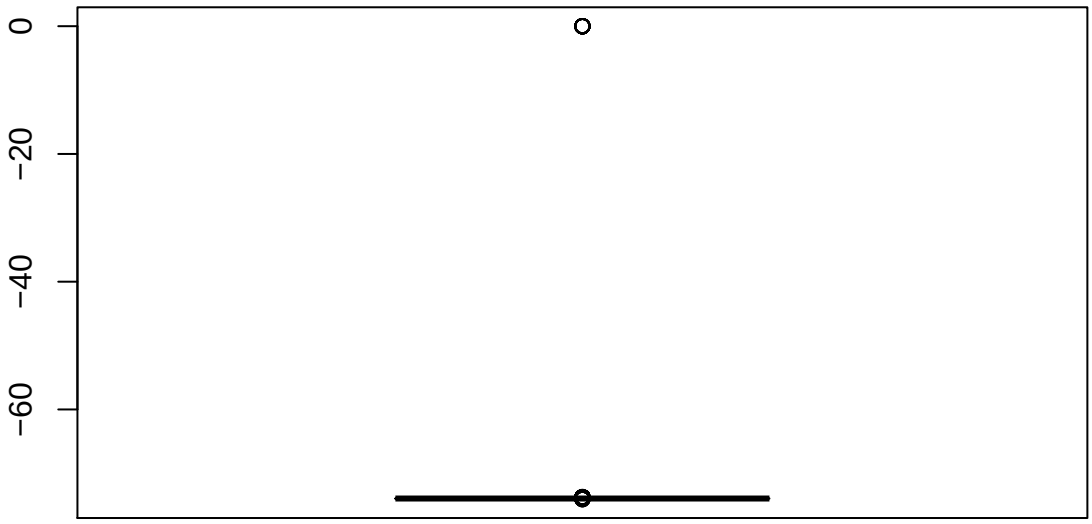
Pickup_longitude

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

```
## (-74.04,-73.96] (-73.96,-73.95] (-73.95,-73.92]      (-73.92,0]
##            1250            1250            1249            1250
##            NA's
##               1
```

```
#How to detect missing values? 0.0 is a possible value?
#sel<-which(df$Pickup_longitude==0.0);length(sel) #11 missings
#df[sel,"Pickup_longitude"]<-NA
```

```
boxplot(df$Pickup_longitude)
```

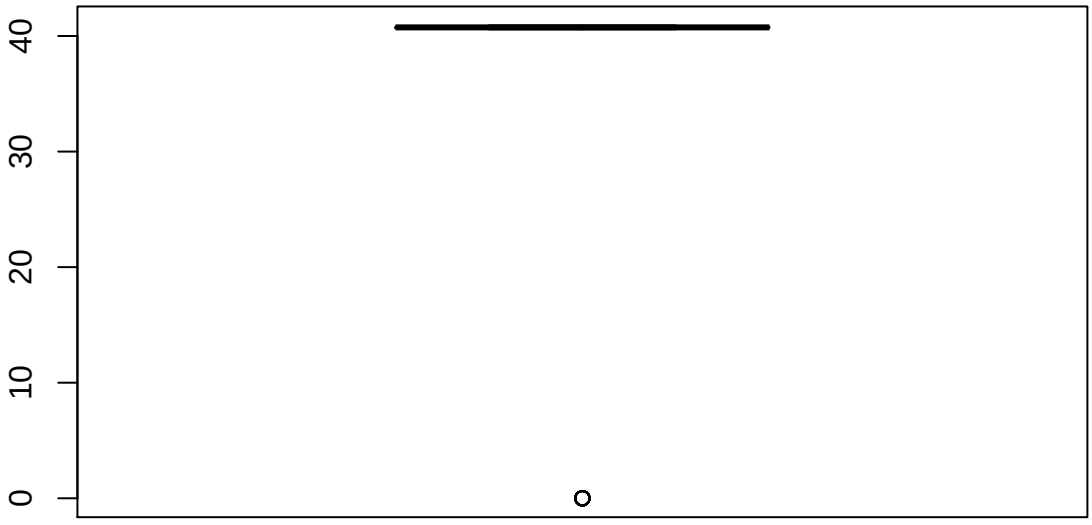


Pickup_latitude

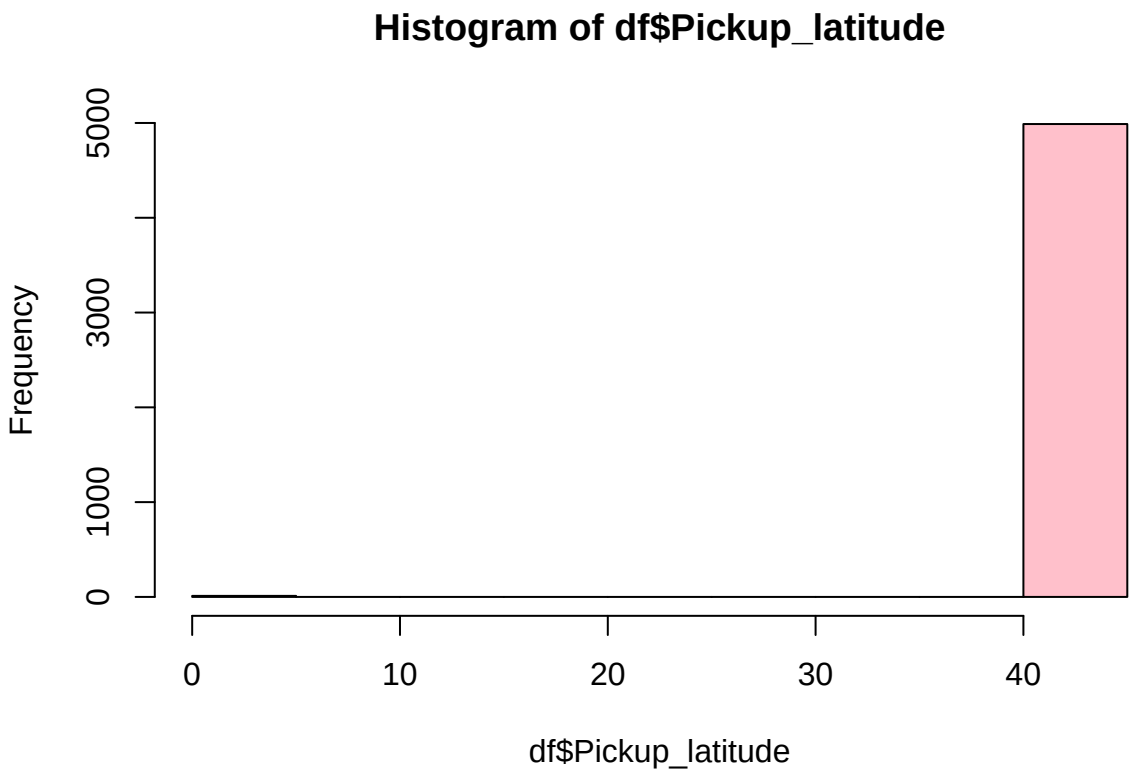
```
df$f.latitude<-factorize(df$Dropoff_latitude)
summary(df$f.latitude) #11 NAs
```

```
##      (0,40.7]  (40.7,40.75] (40.75,40.79] (40.79,40.94]          NA's
##          1246          1250          1250          1250             4
```

```
boxplot(df$Pickup_latitude)
```




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

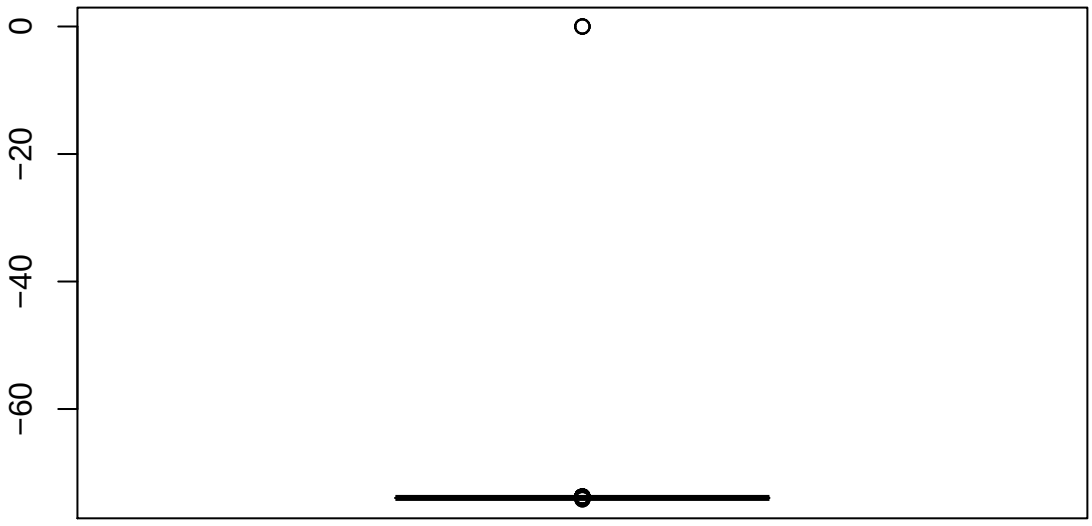


Dropoff_longitude

```
df$f.longtitudeDrop<-factorize(df$Dropoff_longitude)
summary(df$f.longtitudeDrop) # 1 NAs
```

```
## (-74.18,-73.97] (-73.97,-73.95] (-73.95,-73.91]      (-73.91,0]
##            1249            1250            1250            1250
##            NA's
##               1
```

```
boxplot(df$Dropoff_longitude)
```



Dropoff_latitude

```
quantile(df$Dropoff_latitude,seq(0,1,0.1))
```

```
##       0%      10%      20%      30%      40%      50%      60%      70%
##  0.00000 40.67360 40.68850 40.70848 40.72754 40.74601 40.75980 40.77458
##      80%      90%     100%
## 40.80080 40.81875 40.93954
```

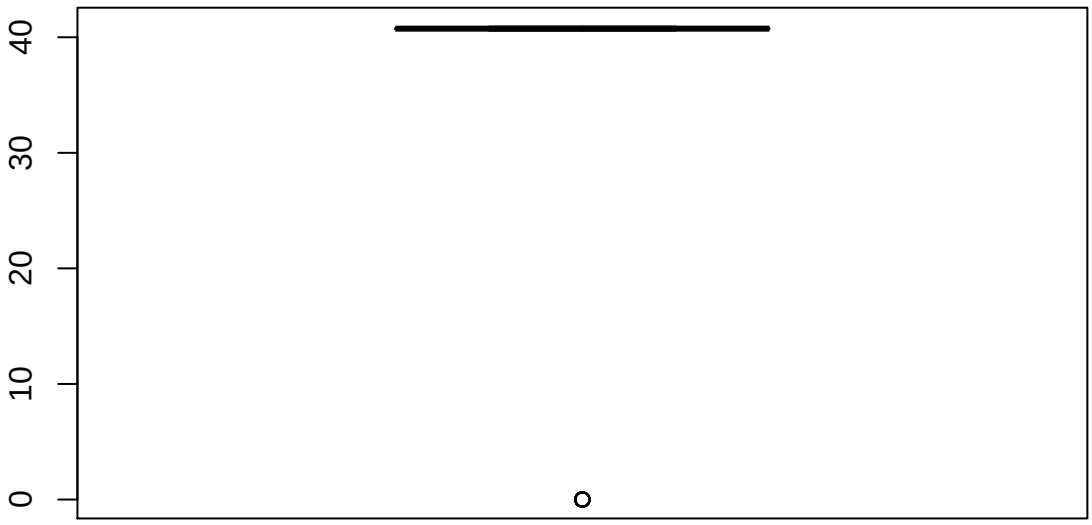
```
pp<-quantile(df$Dropoff_latitude);pp
```

```
##       0%      25%      50%      75%     100%
##  0.00000 40.69549 40.74601 40.78835 40.93954
```

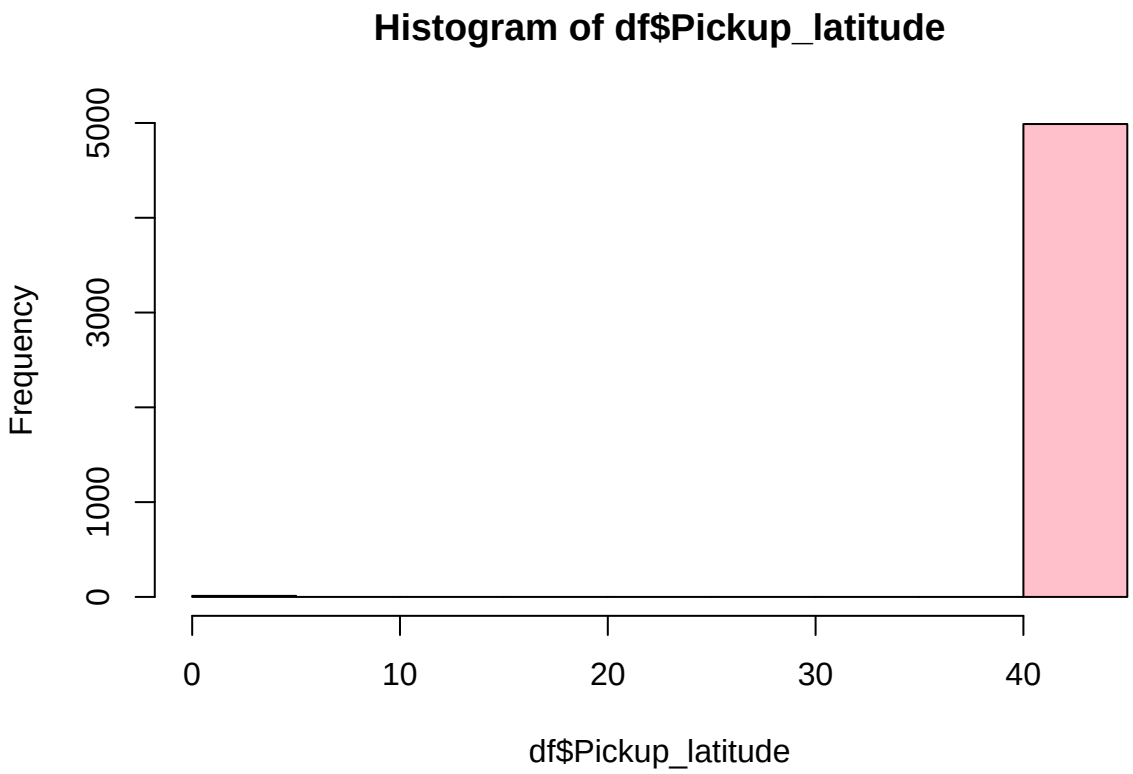
```
df$f.latitudeDrop<-factor(cut(df$Dropoff_latitude,pp))  # NO VA be
summary(df$f.latitudeDrop) # 4 NAs ? Outlier
```

```
##      (0,40.7]  (40.7,40.75] (40.75,40.79] (40.79,40.94]          NA's
##          1246          1250          1250          1250             4
```

```
boxplot(df$Pickup_latitude)
```




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



##

Fare_amount

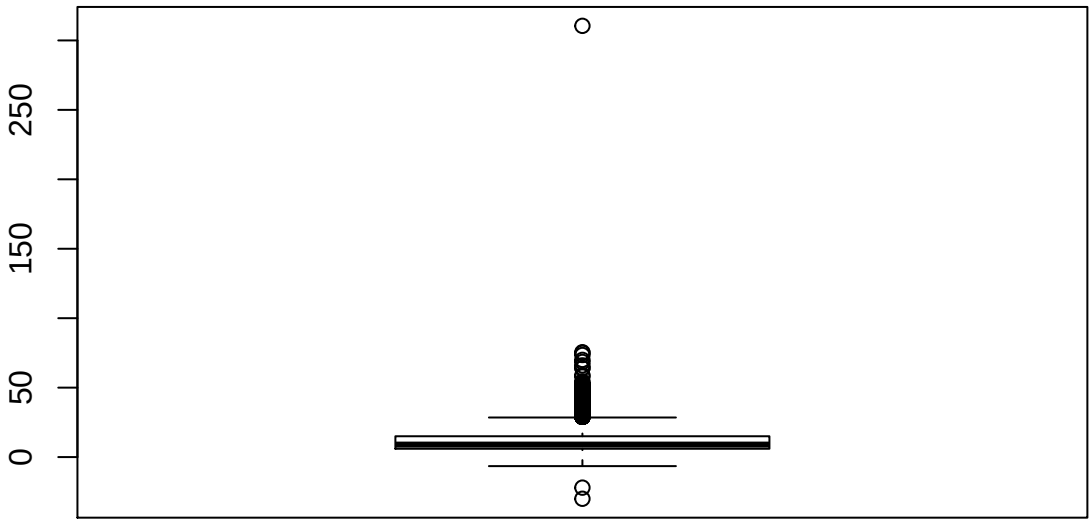
```
df$f.fare_amount<-factorize(df$Fare_amount)
summary(df$f.fare_amount)
```

```
##  (-30,6]    (6,9]   (9,15] (15,310]     NA's
##     1311     1226     1280     1182        1
```

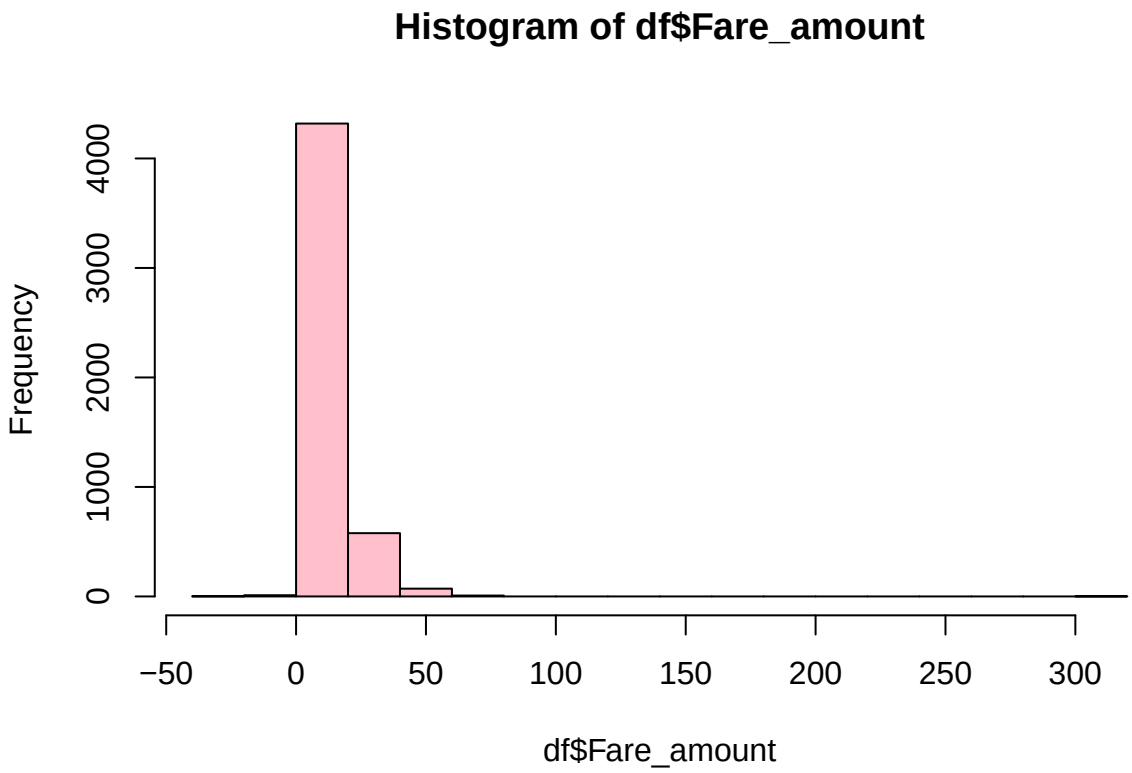
```
sel<-which(df$Fare_amount==0.0);length(sel) #10 missings
```

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

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



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

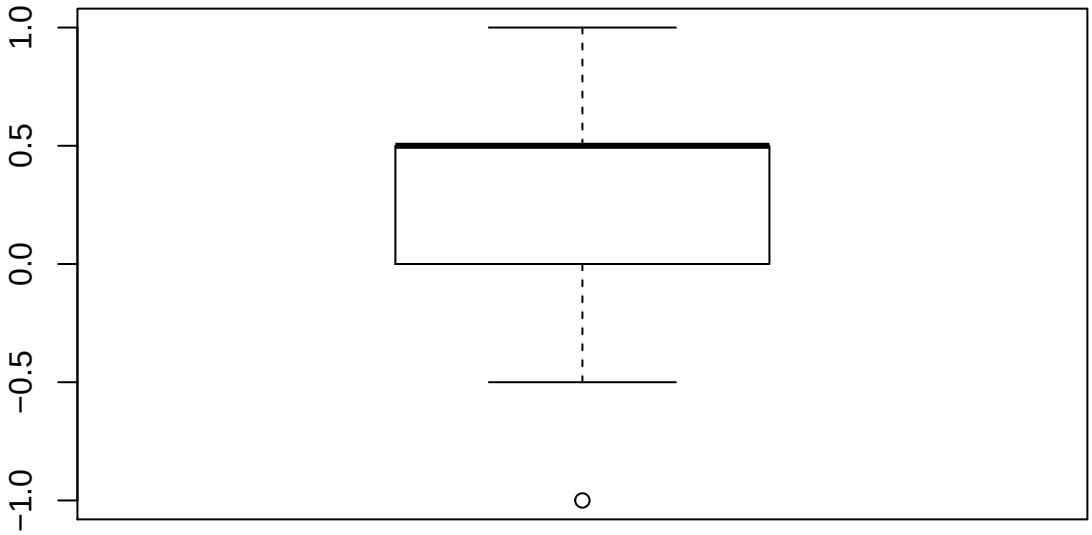


Extra

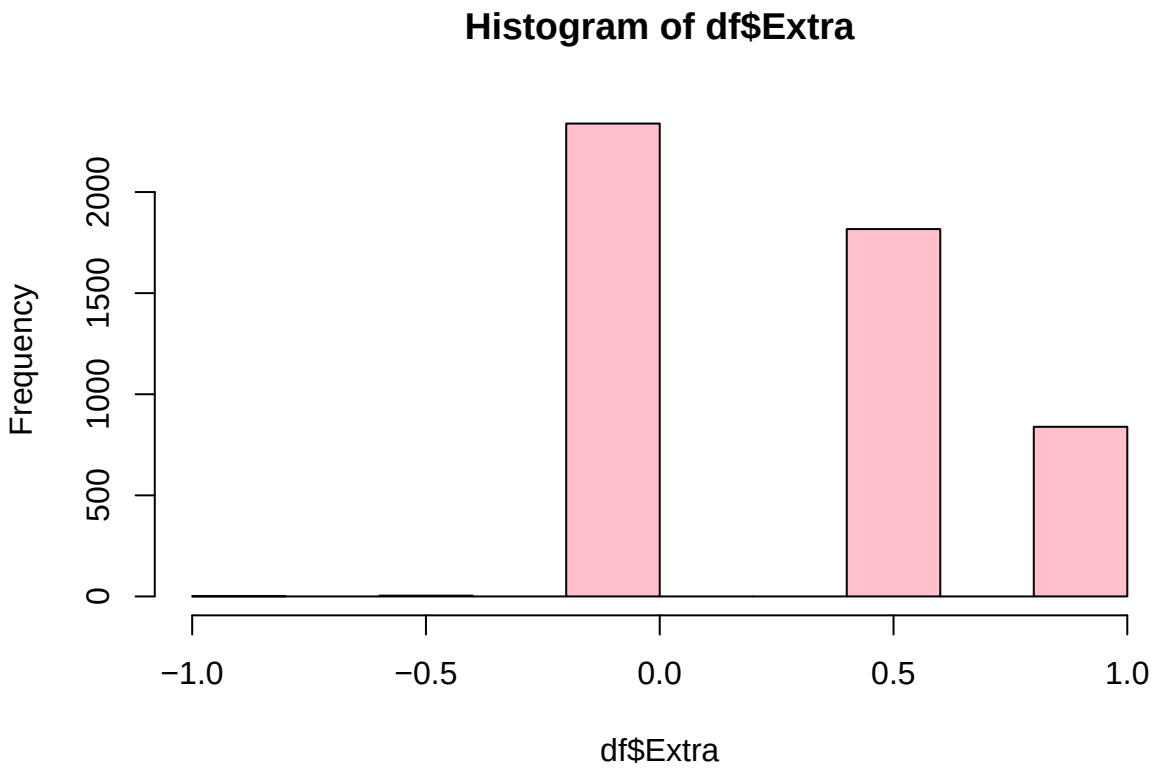
```r
df$f.extra<-factorize(df$Extra)
summary(df$f.extra) #1 NA's
```

```
##  (-1,0] (0,0.5] (0.5,1]    NA's
##    2343    1817     839       1
```

```r
boxplot(df$Extra)
```



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



MTA_tax

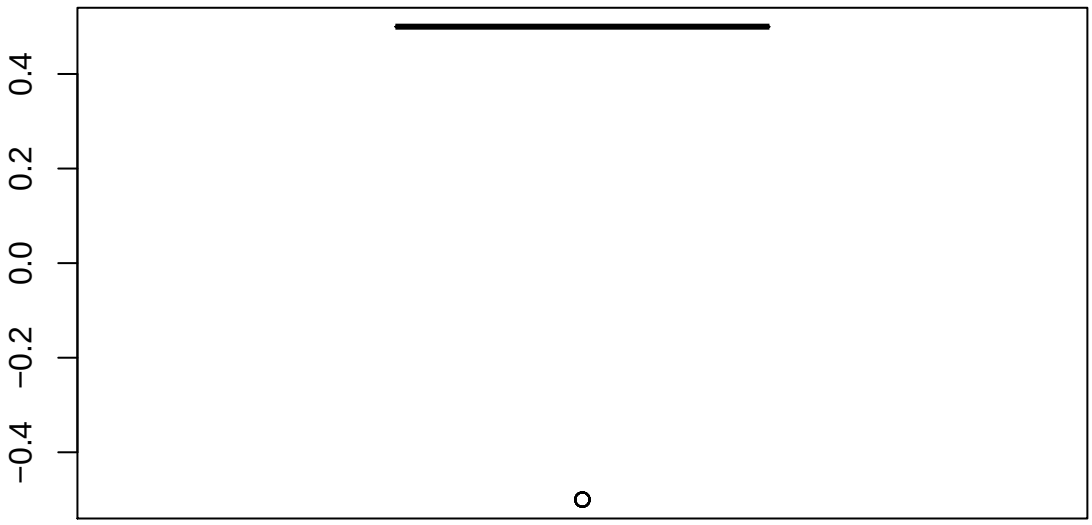
```r
df$f.MTA_tax<-factorize(df$MTA_tax)
summary(df$f.MTA_tax) #11 NA's -> values of -0.5 => Outliers?
```

```
## (-0.5,0.5]       NA's
##       4989         11
```

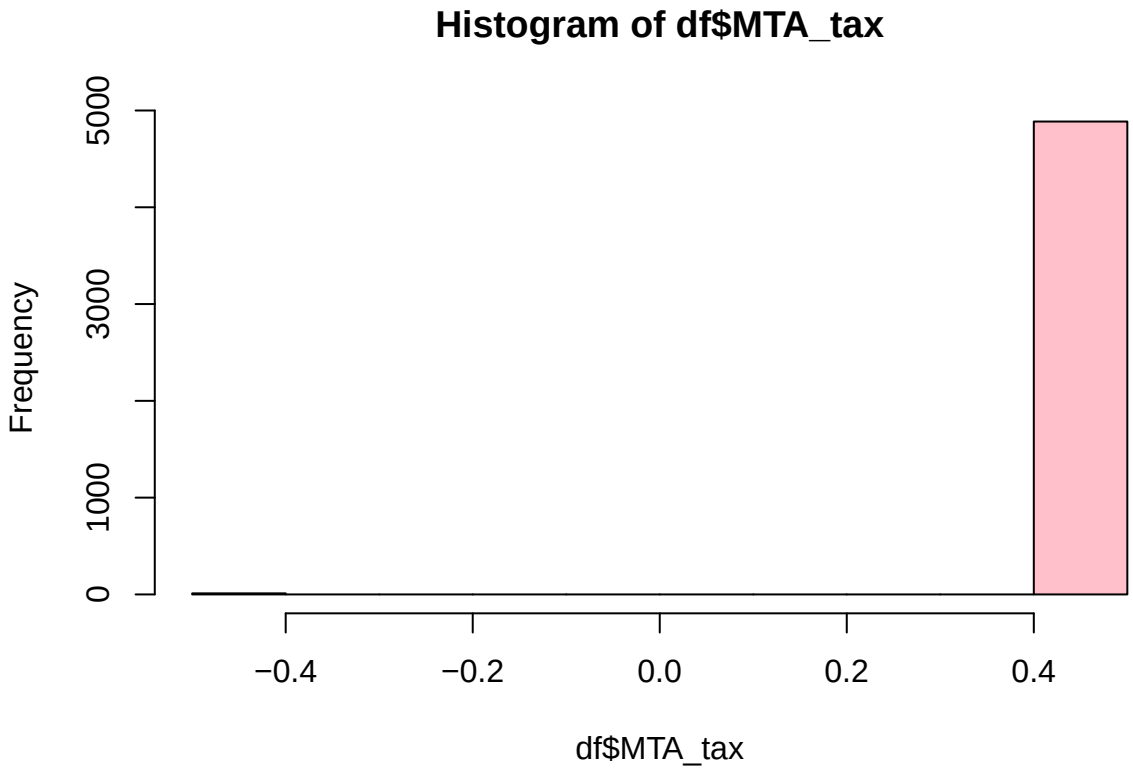
```r
sel<-which(df$MTA_tax==0.0);length(sel) #103 missings
```

```
## [1] 103
```

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




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



Improvement_surcharge

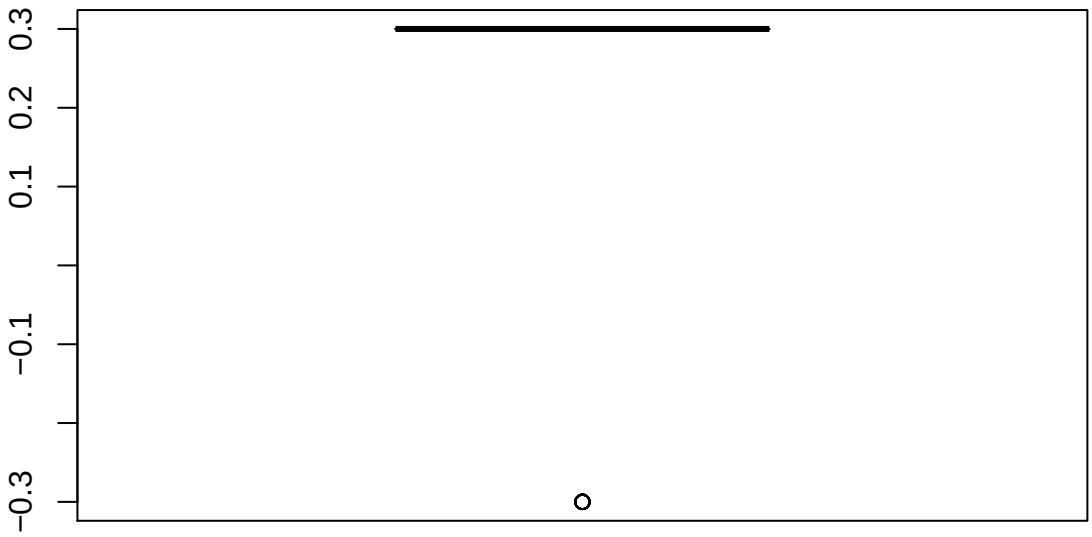
```
df$f.Improvement_surcharge<-factorize(df$improvement_surcharge)
summary(df$f.Improvement_surcharge) #11 NA's -> values of -0.3 => Outliers?
```

```
## (-0.3,0.3]       NA's
##       4989         11
```

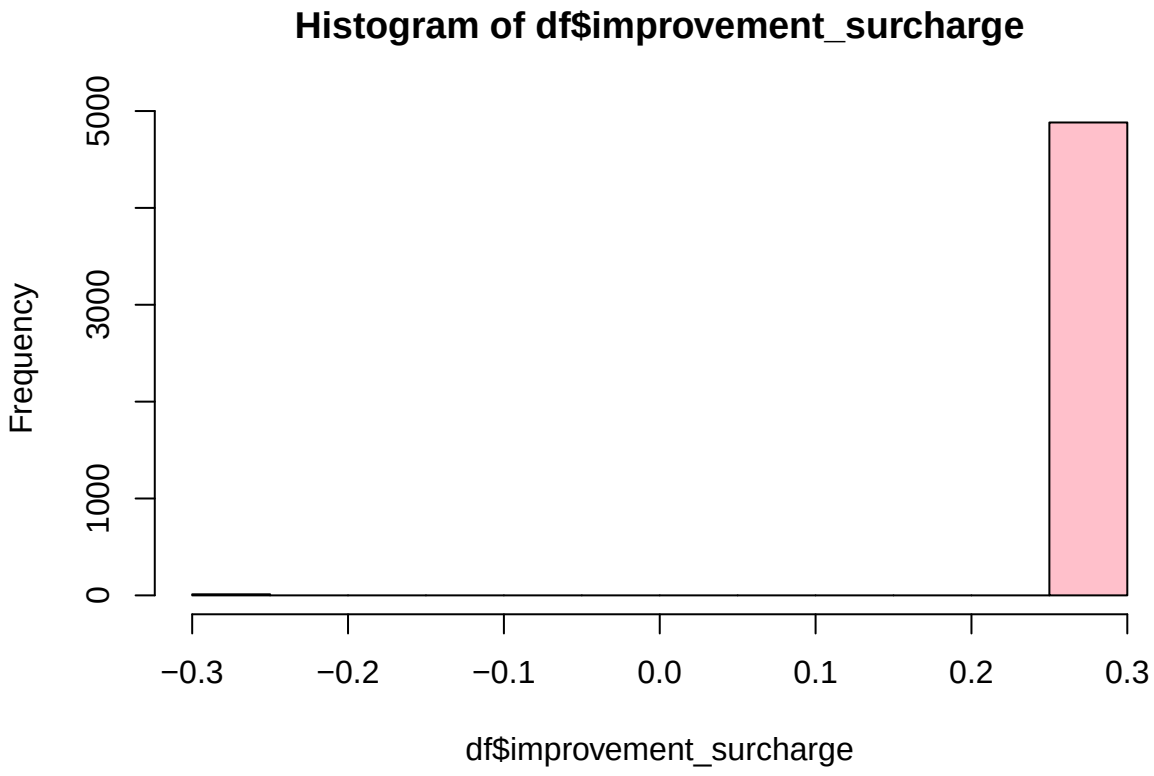
```
sel<-which(df$improvement_surcharge==0.0);length(sel) #107 missings
```

```
## [1] 107
```

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




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

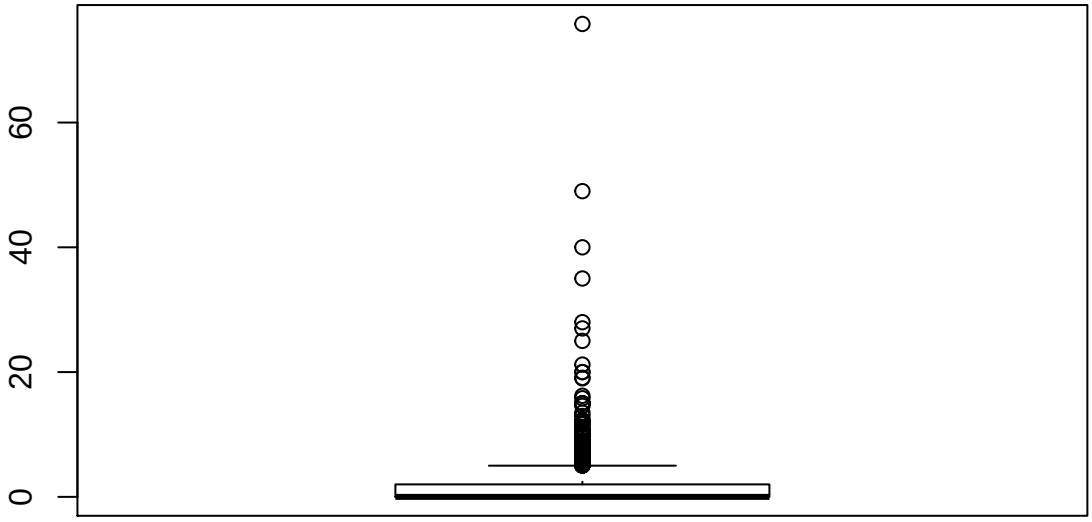


Tip_amount

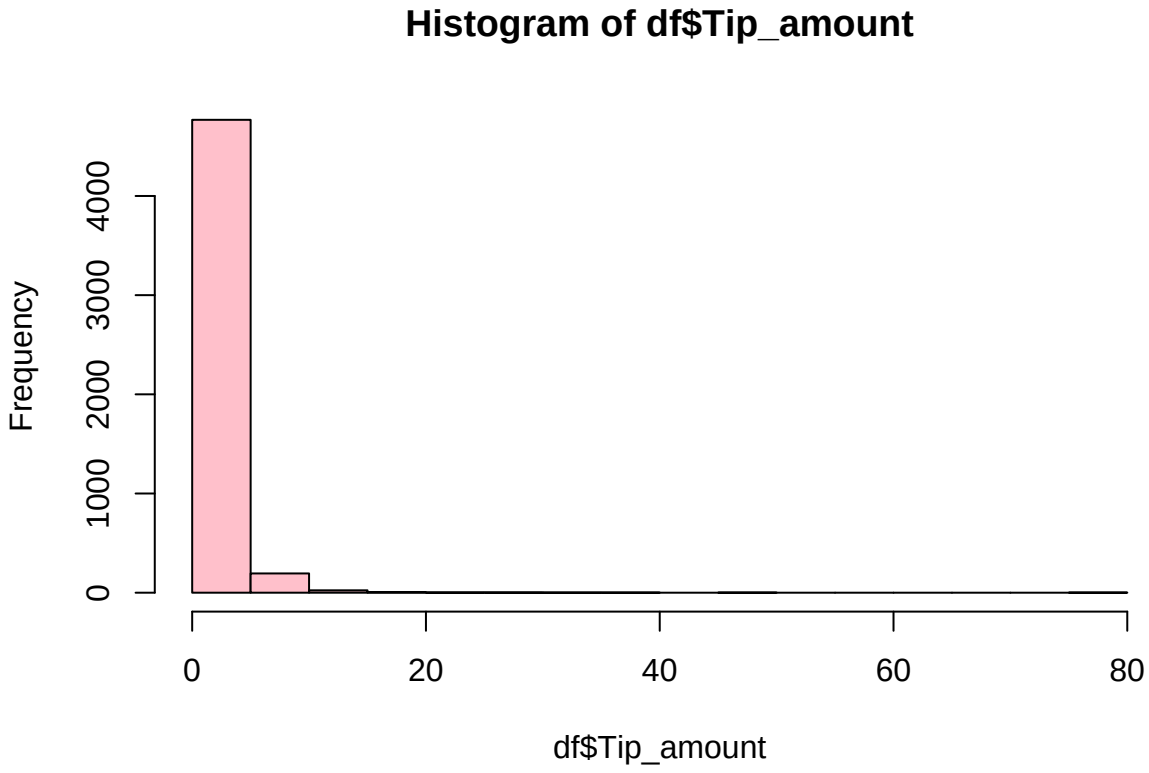
```
df$f.tip_amount<-factorize(df$Tip_amount)
summary(df$f.tip_amount) #2869 NA's
```

```
##    (0,2] (2,75.8]     NA's
##      965     1166     2869
```

```
boxplot(df$Tip_amount)
```




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

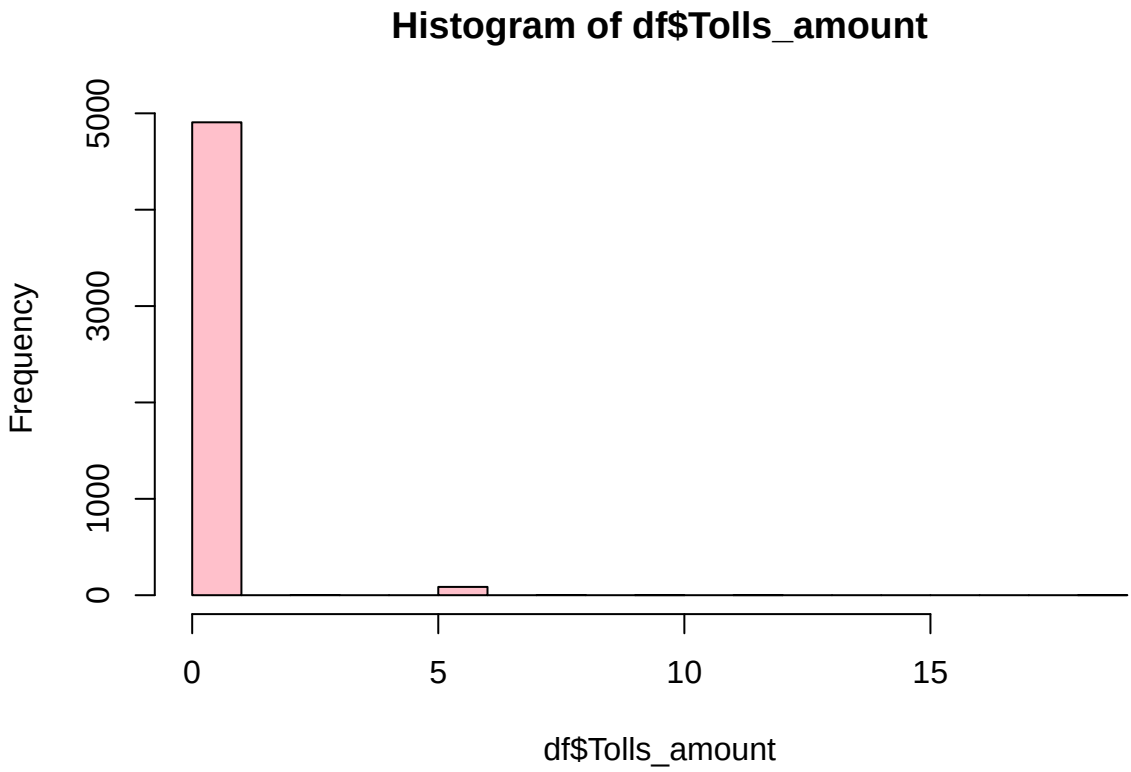


Tolls_amount

```
df$f.toll<-factorize(df$Tolls_amount)
summary(df$f.toll) #4907 NA's, not well factorized
```

```
## (0,18]   NA's
##     93   4907
```

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



Total_amount

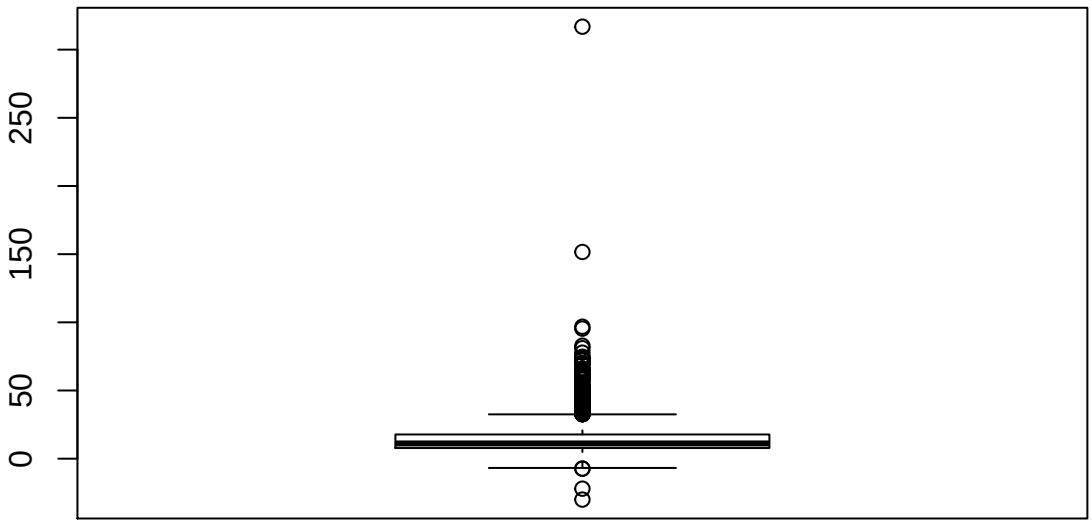
```
df$f.total<-factorize(df$Total_amount)  # NO VA be
summary(df$f.total)
```

```
##   (-30,7.8]  (7.8,11.2] (11.2,17.8]  (17.8,317]        NA's
##        1287        1228        1242        1242           1
```

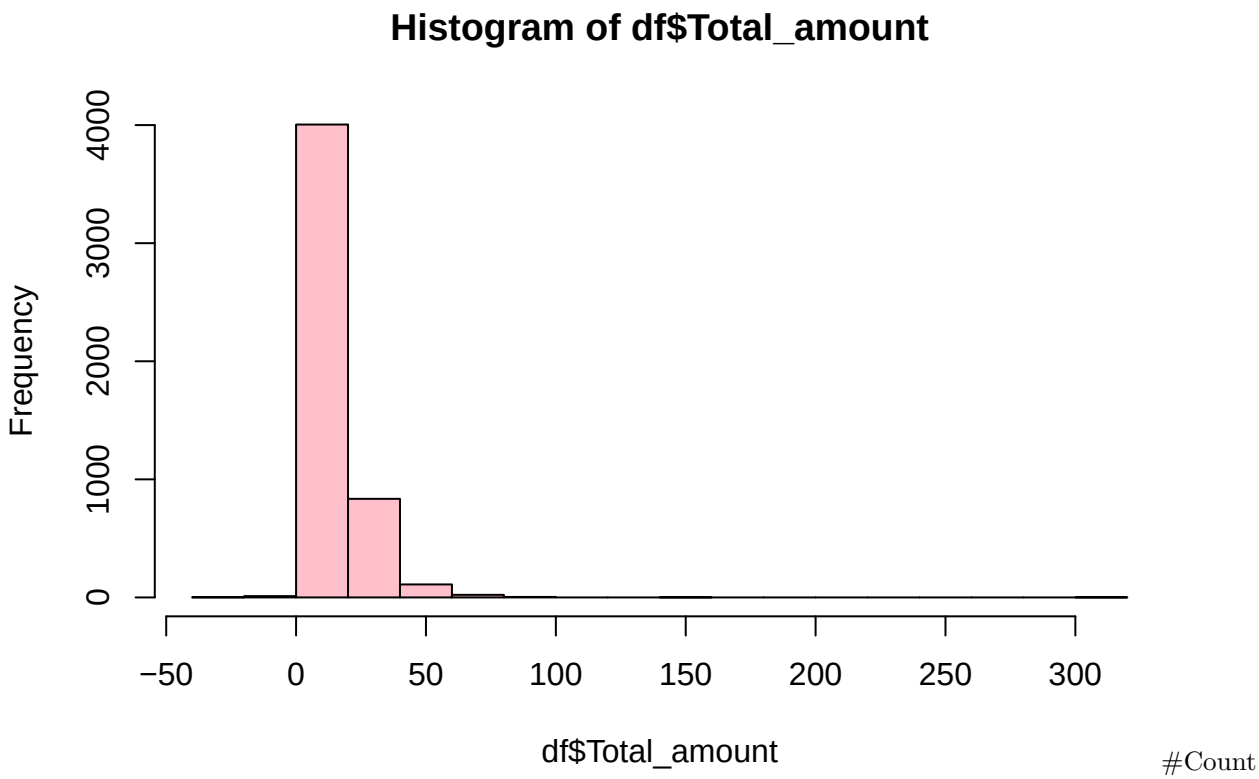
```
sel<-which(df$Total_amount==0.0);length(sel) #9 missings
```

```
## [1] 9
```

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




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



#Count per Variable:

Number of missing values:

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

```
  for (j in 1:ncol(x)) {mis_i <- mis_i + as.numeric(is.na(x[,j])) }
  list(mis_col=mis_x,mis_ind=mis_i) }
mis1<-countNA(df)

attributes(mis1)
```

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

```
mis1$mis_col
```

```
##                          mis_x
## VendorID                     0
## lpep_pickup_datetime         0
## Lpep_dropoff_datetime        0
## Store_and_fwd_flag           0
## RateCodeID                   0
## Pickup_longitude             0
## Pickup_latitude              0
## Dropoff_longitude            0
## Dropoff_latitude             0
## Passenger_count              0
## Trip_distance               60
## Fare_amount                 10
## Extra                        0
## MTA_tax                    103
## Tip_amount                   0
## Tolls_amount                 0
## improvement_surcharge      107
## Total_amount                 9
## Payment_type                 0
## Trip_type                    0
## f.passanger                  2
## Passanger_count           5000
## f.distance                  60
## f.longtitude                 1
## f.latitude                   4
## f.longtitudeDrop             1
## f.latitudeDrop               4
## f.fare_amount                1
## f.extra                      1
## f.MTA_tax                   11
## f.Improvement_surcharge     11
## f.tip_amount              2869
## f.toll                    4907
## f.total                      1
```

```
df$mis_ind <- mis1$mis_ind # new attribute missing values
summary(mis1$mis_ind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.632   3.000   9.000
```

Number of outliers ???

```
outs<-rep(0,ncol(df))
show(outs)
```

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

Multivariant Outlier Detection

#...In process

```
#vars_con<-names(df)[c(6:9,11:18)] #Continuous variables


#install.packages('mvoutlier')
#library(mvoutlier) #not found??
#names(df)
#vars_con # Problems c(5,8,9,10,11,12)
#summary(df[,vars_con])
#vars_con_out<-vars_con[c(1:4)]
#aq.plot(df[,vars_con_out]) # Problems when few numeric values are present in one variable

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

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

# Use common sense
#vars_con
#vars_con_out<-vars_con[c(6,13,16)]
#aq.plot(df[,vars_con_out])  # Problems when missing data are present
#vars_con_out
```

```
#install.packages("car")
#library(car) #not found??
#hist(df$Tip_amount, col="pink")
#catdes(data, 1)
```