Deliverable 2

PCA, CA and Clustering

Júlia Gasull i Claudia Sánchez

# First setups

if(!is.null(dev.list())) dev.off() # Clear plots  
rm(list=ls()) # Clean workspace

## Load Required Packages for this deliverable

We load the necessary packages and set working directory

setwd("~/Documents/uni/FIB-ADEI-LAB/deliverable2")  
filepath<-"~/Documents/uni/FIB-ADEI-LAB/deliverable2"  
#setwd("C:/Users/Claudia Sánchez/Desktop/FIB/TARDOR 2020-2021/ADEI/DELIVERABLE1/FIB-ADEI-LAB/deliverable2")  
#filepath<-"C:/Users/Claudia Sánchez/Desktop/FIB/TARDOR 2020-2021/ADEI/DELIVERABLE1/FIB-ADEI-LAB/deliverable2"  
  
# Load Required Packages  
options(contrasts=c("contr.treatment","contr.treatment"))  
requiredPackages <- c("missMDA","chemometrics","mvoutlier","effects","FactoMineR","car", "factoextra","RColorBrewer","dplyr","ggmap","ggthemes","knitr")  
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]  
if(length(missingPackages)) install.packages(missingPackages)  
lapply(requiredPackages, require, character.only = TRUE)

## Load processed data from first deliverable

load(paste0(filepath,"/Taxi5000\_del1.RData"))

## Clean data

# remove some columns  
#names(df)  
df$lpep\_pickup\_datetime <- NULL  
df$Lpep\_dropoff\_datetime <- NULL  
df$Store\_and\_fwd\_flag <- NULL  
df$Ehail\_fee <- NULL  
df$CashTips <- NULL  
df$Sum\_total\_amount <- NULL  
df$yearGt2015 <- NULL  
  
# imputation  
library(missMDA)  
long\_lat<-names(df)[c(3:6)]  
imp\_long\_lat<-imputePCA(df[,long\_lat])  
df[,long\_lat]<-imp\_long\_lat$completeObs

# Principal Component Analysis (PCA)

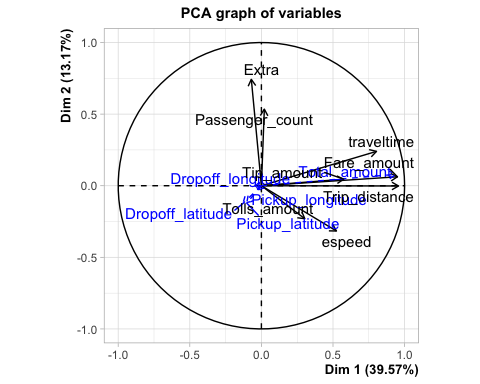
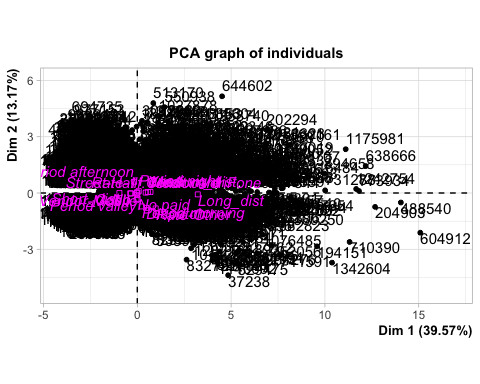
names(df)

## [1] "VendorID" "RateCodeID" "Pickup\_longitude"   
## [4] "Pickup\_latitude" "Dropoff\_longitude" "Dropoff\_latitude"   
## [7] "Passenger\_count" "Trip\_distance" "Fare\_amount"   
## [10] "Extra" "MTA\_tax" "Tip\_amount"   
## [13] "Tolls\_amount" "improvement\_surcharge" "Total\_amount"   
## [16] "Payment\_type" "Trip\_type" "hour"   
## [19] "period" "tlenkm" "traveltime"   
## [22] "espeed" "pickup" "dropoff"   
## [25] "Trip\_distance\_range" "paidTolls" "TipIsGiven"   
## [28] "passenger\_groups"

vars\_res<-names(df)[c(15,27)]   
vars\_quantitatives<-names(df)[c(3:10,12,20:22)]   
vars\_categorical<-names(df)[c(1,2,16:17,19,25,28)]

We have already seen profiling in the previous installment. So now, let’s proceed to look at the main components.

library(FactoMineR)  
res.pca <- PCA(df[,c(1:10,12,13,15:17,19,21,22,25,27)],quanti.sup=c(3:6,13),quali.sup=c(1,2,14:16,19:20))



#plot.PCA(res.pca,choix=c("var"), invisible=c("quanti.sup"))  
#plot.PCA(res.pca,choix=c("var"), invisible=c("var"))  
#plot.PCA(res.pca,choix=c("ind"), invisible=c("ind"))

## Multivariant outliers should be included as supplementary observations

Since the data set we have is pretty good, we considered that we don’t have multivariate outliers

## Eigenvalues and dominant axes analysis

Eigenvalues correspond to the amount of the variation explained by each principal component (PC). Eigenvalues are large for the first PC and small for the subsequent PCs.

### How many axes we have to interpret according to Kaiser?

A PC with an eigenvalue > 1 indicates that the PC accounts for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point to determine the number of PCs to retain, using the Kaiser criteria.

eigenvalues <- res.pca$eig  
head(eigenvalues[, 1:3])

## eigenvalue percentage of variance cumulative percentage of variance  
## comp 1 3.1654602 39.568252 39.56825  
## comp 2 1.0538386 13.172983 52.74124  
## comp 3 1.0394009 12.992511 65.73375  
## comp 4 0.9538540 11.923175 77.65692  
## comp 5 0.8970712 11.213390 88.87031  
## comp 6 0.7211678 9.014597 97.88491

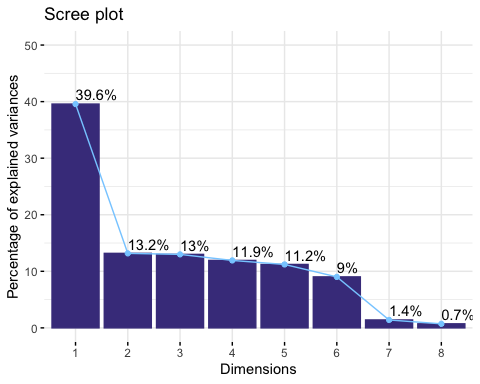
In this case, then, we will use up to dimension 3, and they will explain 65.73% of the total inertia.

### How many axes we have to interpret according to Elbow’s rule?

As a brief definition, we would say that the elbow rule is based on selecting dimensions until the difference in variance of that of the next factorial plane is almost the same as that of the current plane.

So let’s look at exactly where we have this minimal difference:

fviz\_screeplot(  
 res.pca,   
 addlabels=TRUE,   
 ylim=c(0,50),   
 barfill="darkslateblue",   
 barcolor="darkslateblue",  
 linecolor = "skyblue1"  
)

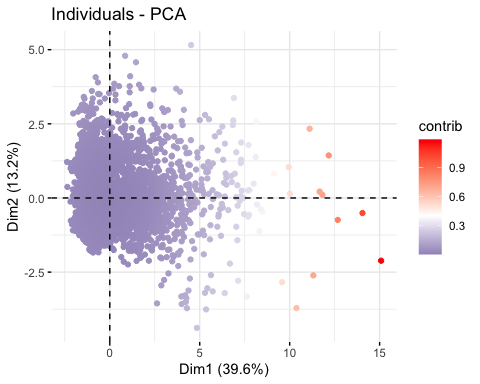


We could say, then, that there is little difference between dimension 3 and 4, or between 5 and 6. Therefore, we could be left with 3 dimensions (as with Kasier) or 5.

## Individuals point of view

### Contribution

# head(res.pca$ind$contrib) # contribition of individuals to the princial components  
fviz\_pca\_ind(res.pca, col.ind="contrib", geom = "point") +  
scale\_color\_gradient2(low="darkslateblue", mid="white",  
 high="red", midpoint=0.40)

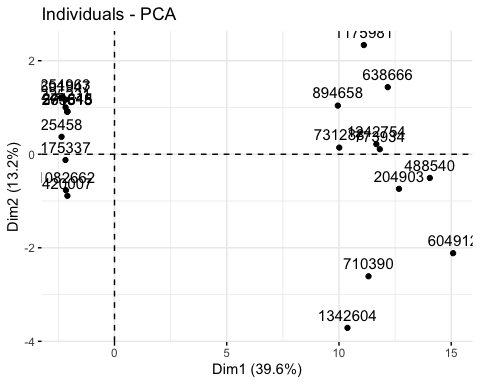


We can see that there are some individuals that are too contributive. So now, let’s try to understand them better with extreme individuals.

### Extreme individuals

#### In dimension 1:

rang<-order(res.pca$ind$coord[,1])  
contrib.extremes<-c(row.names(df)[rang[1]], row.names(df)[rang[length(rang)]])  
  
contrib.extremes<-c(row.names(df)[rang[1:10]], row.names(df)[rang[(length(rang)-10):length(rang)]])  
fviz\_pca\_ind(res.pca, select.ind = list(names=contrib.extremes))



We can now have a look at them:

df[which(row.names(df) %in% row.names(df)[rang[(length(rang)-10):length(rang)]]), 1:28]

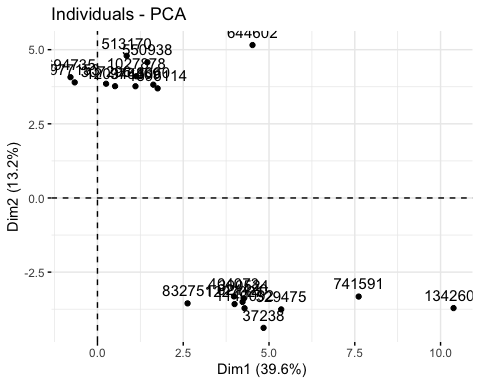
## VendorID RateCodeID Pickup\_longitude Pickup\_latitude  
## 204903 f.Vendor-Mobile Rate-1 -73.98677 40.70252  
## 488540 f.Vendor-VeriFone Rate-1 -73.91121 40.75299  
## 604912 f.Vendor-VeriFone Rate-1 -73.81548 40.62804  
## 638666 f.Vendor-VeriFone Rate-Other -73.80701 40.69907  
## 710390 f.Vendor-VeriFone Rate-1 -73.93688 40.81975  
## 731288 f.Vendor-VeriFone Rate-1 -73.94330 40.63695  
## 773934 f.Vendor-VeriFone Rate-1 -73.95317 40.81768  
## 894658 f.Vendor-Mobile Rate-1 -73.94506 40.79953  
## 1175981 f.Vendor-VeriFone Rate-1 -73.92376 40.76116  
## 1242754 f.Vendor-VeriFone Rate-1 -73.96619 40.58548  
## 1342604 f.Vendor-Mobile Rate-Other -73.94370 40.81538  
## Dropoff\_longitude Dropoff\_latitude Passenger\_count Trip\_distance  
## 204903 -73.97940 40.64393 1 27.00000  
## 488540 -73.91345 40.75084 1 30.00000  
## 604912 -73.99866 40.59183 1 27.33295  
## 638666 -73.81952 40.71432 1 18.21000  
## 710390 -73.84977 40.67285 1 19.00000  
## 731288 -73.86108 40.83635 6 19.94000  
## 773934 -73.95087 40.72394 1 24.92000  
## 894658 -73.94336 40.71036 1 25.70000  
## 1175981 -73.90582 40.76783 5 27.76064  
## 1242754 -73.87349 40.77394 1 22.46000  
## 1342604 -73.94130 40.64498 1 18.30000  
## Fare\_amount Extra MTA\_tax Tip\_amount Tolls\_amount improvement\_surcharge  
## 204903 60.00000 0.0 Yes 14.35 0.000000 Yes  
## 488540 60.00000 0.0 Yes 17.00 0.000000 Yes  
## 604912 60.00000 0.5 Yes 17.00 5.540000 Yes  
## 638666 60.00000 1.0 Yes 17.00 3.020141 Yes  
## 710390 50.50000 0.5 Yes 11.47 5.540000 Yes  
## 731288 48.79243 0.0 Yes 0.00 5.540000 Yes  
## 773934 60.00000 0.5 Yes 13.36 0.000000 Yes  
## 894658 60.00000 1.0 Yes 0.00 0.000000 Yes  
## 1175981 60.00000 0.5 Yes 0.00 0.000000 Yes  
## 1242754 60.00000 0.0 Yes 12.86 0.000000 Yes  
## 1342604 52.00000 0.0 Yes 6.00 5.540000 Yes  
## Total\_amount Payment\_type Trip\_type hour period tlenkm  
## 204903 86.15 Credit card Street-Hail 7 Period night 43.45229  
## 488540 128.76 Credit card Street-Hail 6 Period night 48.28000  
## 604912 108.41 Credit card Street-Hail 20 Period afternoon 48.28000  
## 638666 111.05 Credit card Street-Hail 16 Period valley 29.30615  
## 710390 68.81 Credit card Street-Hail 23 Period night 30.57754  
## 731288 68.84 Credit card Street-Hail 10 Period morning 32.09032  
## 773934 80.16 Credit card Street-Hail 0 Period night 40.10485  
## 894658 72.80 Cash Street-Hail 18 Period afternoon 41.36014  
## 1175981 116.30 Cash Street-Hail 23 Period night 48.28000  
## 1242754 77.16 Credit card Street-Hail 14 Period valley 36.14587  
## 1342604 64.34 Credit card Street-Hail 6 Period night 29.45100  
## traveltime espeed pickup dropoff Trip\_distance\_range paidTolls  
## 204903 41.71667 55.00000 07 08 Long\_dist No  
## 488540 49.00000 55.00000 06 07 Short\_dist No  
## 604912 43.18333 55.00000 20 21 Short\_dist Yes  
## 638666 60.00000 25.41608 16 17 Long\_dist <NA>  
## 710390 30.53333 55.00000 23 00 Long\_dist Yes  
## 731288 60.00000 31.56425 10 11 Long\_dist Yes  
## 773934 36.73333 55.00000 00 01 Long\_dist No  
## 894658 46.28333 53.61776 18 19 Long\_dist No  
## 1175981 60.00000 55.00000 23 00 Short\_dist No  
## 1242754 57.71667 37.57584 14 15 Long\_dist No  
## 1342604 30.75000 55.00000 06 06 Long\_dist Yes  
## TipIsGiven passenger\_groups  
## 204903 Yes Single  
## 488540 Yes Single  
## 604912 Yes Single  
## 638666 Yes Single  
## 710390 Yes Single  
## 731288 No Group  
## 773934 Yes Single  
## 894658 No Single  
## 1175981 No Group  
## 1242754 Yes Single  
## 1342604 Yes Single

df[which(row.names(df) %in% row.names(df)[rang[1:10]]),1:28]

## VendorID RateCodeID Pickup\_longitude Pickup\_latitude  
## 25458 f.Vendor-VeriFone Rate-1 -73.89600 40.85568  
## 175337 f.Vendor-Mobile Rate-1 -73.85332 40.72649  
## 225231 f.Vendor-VeriFone Rate-1 -73.94785 40.80964  
## 263515 f.Vendor-VeriFone Rate-1 -73.95492 40.82026  
## 274645 f.Vendor-Mobile Rate-1 -73.94057 40.62366  
## 420007 f.Vendor-Mobile Rate-1 -73.89059 40.74692  
## 591818 f.Vendor-VeriFone Rate-1 -73.97880 40.68356  
## 691947 f.Vendor-VeriFone Rate-1 -73.80762 40.70077  
## 1082662 f.Vendor-VeriFone Rate-1 -73.93958 40.81605  
## 1254963 f.Vendor-VeriFone Rate-1 -73.99031 40.69246  
## Dropoff\_longitude Dropoff\_latitude Passenger\_count Trip\_distance  
## 25458 -73.89645 40.85497 1 0.05000000  
## 175337 -73.85199 40.72478 2 0.10000000  
## 225231 -73.94830 40.80927 1 0.04000000  
## 263515 -73.95686 40.81767 1 0.03813833  
## 274645 -73.94056 40.62366 1 0.03807637  
## 420007 -73.89084 40.74857 1 0.10000000  
## 591818 -73.97880 40.68356 1 0.03810496  
## 691947 -73.80876 40.69843 1 0.16000000  
## 1082662 -73.94041 40.81475 1 0.09000000  
## 1254963 -73.99083 40.69273 1 0.03000000  
## Fare\_amount Extra MTA\_tax Tip\_amount Tolls\_amount improvement\_surcharge  
## 25458 3.0 0.5 Yes 0 0 Yes  
## 175337 3.5 0.0 Yes 0 0 Yes  
## 225231 2.5 1.0 Yes 0 0 Yes  
## 263515 2.5 1.0 Yes 0 0 Yes  
## 274645 2.5 1.0 Yes 0 0 Yes  
## 420007 2.5 0.0 Yes 0 0 Yes  
## 591818 2.5 1.0 Yes 0 0 Yes  
## 691947 3.0 1.0 Yes 0 0 Yes  
## 1082662 3.0 0.0 Yes 0 0 Yes  
## 1254963 2.5 1.0 Yes 0 0 Yes  
## Total\_amount Payment\_type Trip\_type hour period tlenkm  
## 25458 4.3 Cash Street-Hail 4 Period night 0.08046720  
## 175337 4.3 Cash Street-Hail 14 Period valley 0.16093440  
## 225231 4.3 Cash Street-Hail 17 Period afternoon 0.06437376  
## 263515 4.3 Cash Street-Hail 16 Period valley 0.00000000  
## 274645 4.3 No paid Street-Hail 19 Period afternoon 0.00000000  
## 420007 3.3 Cash Street-Hail 19 Period afternoon 0.16093440  
## 591818 4.3 Credit card Street-Hail 16 Period valley 0.00000000  
## 691947 4.8 Cash Street-Hail 18 Period afternoon 0.25749504  
## 1082662 3.8 Cash Street-Hail 19 Period afternoon 0.14484096  
## 1254963 4.3 Cash Street-Hail 18 Period afternoon 0.04828032  
## traveltime espeed pickup dropoff Trip\_distance\_range paidTolls  
## 25458 1.3500000 3.576320 04 04 Short\_dist No  
## 175337 2.1333333 4.526280 14 14 Short\_dist No  
## 225231 0.3000000 12.874752 17 17 Short\_dist No  
## 263515 0.0500000 15.398313 16 16 Short\_dist No  
## 274645 0.2666667 15.382913 19 19 Short\_dist No  
## 420007 0.8833333 10.931393 19 19 Short\_dist No  
## 591818 0.1666667 15.390021 16 16 Short\_dist No  
## 691947 1.6833333 9.178041 18 19 Short\_dist No  
## 1082662 1.1166667 7.782499 19 19 Short\_dist No  
## 1254963 0.4166667 6.952366 18 18 Short\_dist No  
## TipIsGiven passenger\_groups  
## 25458 No Single  
## 175337 No Couple  
## 225231 No Single  
## 263515 No Single  
## 274645 No Single  
## 420007 No Single  
## 591818 No Single  
## 691947 No Single  
## 1082662 No Single  
## 1254963 No Single

#### In dimension 2:

rang<-order(res.pca$ind$coord[,2])  
contrib.extremes<-c(row.names(df)[rang[1]], row.names(df)[rang[length(rang)]])  
  
contrib.extremes<-c(row.names(df)[rang[1:10]], row.names(df)[rang[(length(rang)-10):length(rang)]])  
fviz\_pca\_ind(res.pca, select.ind = list(names=contrib.extremes))



We can now have a look at them:

df[which(row.names(df) %in% row.names(df)[rang[(length(rang)-10):length(rang)]]), 1:28]

## VendorID RateCodeID Pickup\_longitude Pickup\_latitude  
## 3060 f.Vendor-VeriFone Rate-1 -73.86355 40.73727  
## 307296 f.Vendor-VeriFone Rate-1 -73.95361 40.78796  
## 513170 f.Vendor-VeriFone Rate-1 -73.91908 40.75881  
## 550938 f.Vendor-VeriFone Rate-1 -73.93481 40.74301  
## 644602 f.Vendor-VeriFone Rate-1 -73.92159 40.76666  
## 694735 f.Vendor-VeriFone Rate-1 -73.98262 40.66566  
## 976469 f.Vendor-VeriFone Rate-1 -73.96669 40.80442  
## 977153 f.Vendor-VeriFone Rate-1 -73.89025 40.74623  
## 1027878 f.Vendor-VeriFone Rate-1 -73.96809 40.63953  
## 1203448 f.Vendor-VeriFone Rate-1 -73.97668 40.68291  
## 1396114 f.Vendor-VeriFone Rate-1 -73.96153 40.71631  
## Dropoff\_longitude Dropoff\_latitude Passenger\_count Trip\_distance  
## 3060 -73.91945 40.74348 5 3.05  
## 307296 -73.96581 40.76854 5 1.68  
## 513170 -73.90479 40.77545 5 1.47  
## 550938 -73.96293 40.75823 6 2.87  
## 644602 -73.98792 40.73801 6 6.26  
## 694735 -73.97092 40.67282 6 0.97  
## 976469 -73.96804 40.76556 5 3.45  
## 977153 -73.92136 40.75252 6 1.81  
## 1027878 -73.98267 40.67964 6 3.58  
## 1203448 -73.93872 40.69656 5 3.11  
## 1396114 -73.98534 40.72356 6 2.49  
## Fare\_amount Extra MTA\_tax Tip\_amount Tolls\_amount improvement\_surcharge  
## 3060 14.0 0.5 Yes 0.00 0 Yes  
## 307296 14.0 1.0 Yes 3.16 0 Yes  
## 513170 8.0 1.0 Yes 0.00 0 Yes  
## 550938 19.0 1.0 Yes 4.16 0 Yes  
## 644602 32.5 1.0 Yes 6.86 0 Yes  
## 694735 9.0 1.0 Yes 2.16 0 Yes  
## 976469 18.0 1.0 Yes 2.50 0 Yes  
## 977153 10.5 1.0 Yes 0.00 0 Yes  
## 1027878 16.0 1.0 Yes 3.56 0 Yes  
## 1203448 17.0 1.0 Yes 0.00 0 Yes  
## 1396114 19.0 0.5 Yes 6.09 0 Yes  
## Total\_amount Payment\_type Trip\_type hour period tlenkm  
## 3060 15.30 Cash Street-Hail 0 Period night 4.908499  
## 307296 18.96 Credit card Street-Hail 16 Period valley 2.703698  
## 513170 9.80 Cash Street-Hail 18 Period afternoon 2.365736  
## 550938 24.96 Credit card Street-Hail 17 Period afternoon 4.618817  
## 644602 41.16 Credit card Street-Hail 18 Period afternoon 10.074493  
## 694735 12.96 Credit card Street-Hail 19 Period afternoon 1.561064  
## 976469 22.30 Credit card Street-Hail 16 Period valley 5.552237  
## 977153 12.30 Cash Street-Hail 17 Period afternoon 2.912913  
## 1027878 21.36 Credit card Street-Hail 16 Period valley 5.761452  
## 1203448 18.80 Credit card Street-Hail 17 Period afternoon 5.005060  
## 1396114 26.39 Credit card Street-Hail 0 Period night 4.007267  
## traveltime espeed pickup dropoff Trip\_distance\_range paidTolls  
## 3060 60.00000 3.864960 00 01 Medium\_dist No  
## 307296 21.35000 7.598214 16 16 Short\_dist No  
## 513170 60.00000 3.000000 18 18 Short\_dist No  
## 550938 30.50000 9.086198 17 17 Medium\_dist No  
## 644602 52.20000 11.579878 18 19 Long\_dist No  
## 694735 12.08333 7.751489 19 19 Short\_dist No  
## 976469 25.50000 13.064087 16 17 Medium\_dist No  
## 977153 13.81667 12.649560 17 18 Short\_dist No  
## 1027878 21.98333 15.724962 16 16 Medium\_dist No  
## 1203448 26.13333 11.491209 17 18 Medium\_dist No  
## 1396114 31.03333 7.747669 00 00 Short\_dist No  
## TipIsGiven passenger\_groups  
## 3060 No Group  
## 307296 Yes Group  
## 513170 No Group  
## 550938 Yes Group  
## 644602 Yes Group  
## 694735 Yes Group  
## 976469 Yes Group  
## 977153 No Group  
## 1027878 Yes Group  
## 1203448 No Group  
## 1396114 Yes Group

df[which(row.names(df) %in% row.names(df)[rang[1:10]]),1:28]

## VendorID RateCodeID Pickup\_longitude Pickup\_latitude  
## 37238 f.Vendor-VeriFone Rate-1 -73.94037 40.79722  
## 300524 f.Vendor-VeriFone Rate-1 -73.95204 40.79805  
## 404073 f.Vendor-VeriFone Rate-1 -73.92345 40.80943  
## 529475 f.Vendor-VeriFone Rate-1 -73.95724 40.81275  
## 621420 f.Vendor-VeriFone Rate-1 -73.93903 40.81678  
## 741591 f.Vendor-VeriFone Rate-1 -73.89080 40.74696  
## 832751 f.Vendor-VeriFone Rate-1 -73.98846 40.67025  
## 1140092 f.Vendor-Mobile Rate-1 -73.91059 40.76953  
## 1227021 f.Vendor-VeriFone Rate-1 -73.89172 40.74702  
## 1342604 f.Vendor-Mobile Rate-Other -73.94370 40.81538  
## Dropoff\_longitude Dropoff\_latitude Passenger\_count Trip\_distance  
## 37238 -73.87116 40.77416 1 6.29  
## 300524 -73.87309 40.77436 2 7.44  
## 404073 -73.87628 40.76842 1 6.70  
## 529475 -73.86170 40.76838 1 7.85  
## 621420 -73.87211 40.77211 1 7.33  
## 741591 -74.01478 40.71557 1 11.47  
## 832751 -74.01384 40.71449 1 3.66  
## 1140092 -73.86433 40.84798 1 7.50  
## 1227021 -73.91472 40.80377 1 6.62  
## 1342604 -73.94130 40.64498 1 18.30  
## Fare\_amount Extra MTA\_tax Tip\_amount Tolls\_amount improvement\_surcharge  
## 37238 19.0 0.0 Yes 5.07 5.54 Yes  
## 300524 22.5 0.0 Yes 0.00 5.54 Yes  
## 404073 23.5 0.0 Yes 0.00 5.54 Yes  
## 529475 24.0 0.0 Yes 5.00 5.54 Yes  
## 621420 24.0 0.0 Yes 0.00 5.54 Yes  
## 741591 34.0 0.0 Yes 8.07 5.54 Yes  
## 832751 13.5 0.0 Yes 2.00 5.54 Yes  
## 1140092 23.5 0.0 Yes 0.00 5.54 Yes  
## 1227021 19.5 0.5 Yes 0.00 5.54 Yes  
## 1342604 52.0 0.0 Yes 6.00 5.54 Yes  
## Total\_amount Payment\_type Trip\_type hour period tlenkm  
## 37238 30.41 Credit card Street-Hail 9 Period morning 10.122774  
## 300524 28.84 Credit card Street-Hail 13 Period valley 11.973519  
## 404073 29.84 Credit card Street-Hail 14 Period valley 10.782605  
## 529475 35.34 Credit card Street-Hail 6 Period night 12.633350  
## 621420 30.34 Cash Street-Hail 8 Period morning 11.796492  
## 741591 48.41 Credit card Street-Hail 15 Period valley 18.459176  
## 832751 21.84 Credit card Street-Hail 9 Period morning 5.890199  
## 1140092 29.84 Cash Street-Hail 8 Period morning 12.070080  
## 1227021 26.34 Cash Street-Hail 5 Period night 10.653857  
## 1342604 64.34 Credit card Street-Hail 6 Period night 29.450995  
## traveltime espeed pickup dropoff Trip\_distance\_range paidTolls  
## 37238 11.30000 53.74924 09 09 Long\_dist Yes  
## 300524 17.48333 41.09120 13 13 Long\_dist Yes  
## 404073 22.56667 28.66867 14 14 Long\_dist Yes  
## 529475 18.20000 41.64841 06 07 Long\_dist Yes  
## 621420 21.33333 33.17763 08 09 Long\_dist Yes  
## 741591 27.78333 39.86385 15 15 Long\_dist Yes  
## 832751 12.60000 28.04857 09 09 Medium\_dist Yes  
## 1140092 19.23333 37.65363 08 09 Long\_dist Yes  
## 1227021 10.46667 55.00000 05 05 Long\_dist Yes  
## 1342604 30.75000 55.00000 06 06 Long\_dist Yes  
## TipIsGiven passenger\_groups  
## 37238 Yes Single  
## 300524 No Couple  
## 404073 No Single  
## 529475 Yes Single  
## 621420 No Single  
## 741591 Yes Single  
## 832751 Yes Single  
## 1140092 No Single  
## 1227021 No Single  
## 1342604 Yes Single

### Detection of multivariant outliers and influent data.

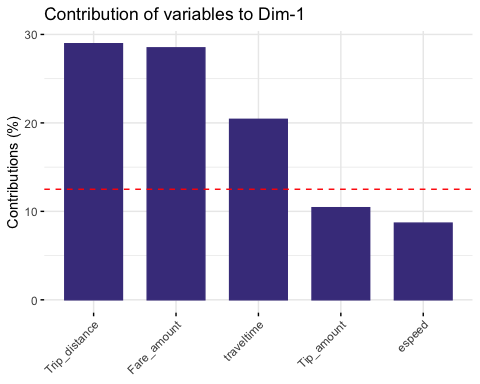
Since we’ve commented before that we don’t consider multivariate outliers, no action should be taken here.

## Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables

res.des <- dimdesc(res.pca)

### First dimension

fviz\_contrib( # contributions of variables to PC1  
 res.pca,   
 fill = "darkslateblue",  
 color = "darkslateblue",  
 choice = "var",   
 axes = 1,   
 top = 5)



res.des$Dim.1

## $quanti  
## correlation p.value  
## Trip\_distance 0.95730706 0.000000e+00  
## Fare\_amount 0.94960484 0.000000e+00  
## Total\_amount 0.93942001 0.000000e+00  
## traveltime 0.80368337 0.000000e+00  
## Tip\_amount 0.57415837 0.000000e+00  
## espeed 0.52394674 0.000000e+00  
## Tolls\_amount 0.30300105 9.013310e-99  
## Pickup\_longitude -0.03125024 3.360908e-02  
## Dropoff\_longitude -0.05426961 2.227979e-04  
## Extra -0.07041780 1.646768e-06  
## Pickup\_latitude -0.10228377 3.148028e-12  
## Dropoff\_latitude -0.12894697 1.345881e-18  
##   
## $quali  
## R2 p.value  
## Trip\_distance\_range 0.691017128 0.000000e+00  
## TipIsGiven 0.060653567 7.774385e-65  
## Payment\_type 0.053034123 2.149327e-55  
## RateCodeID 0.008583339 2.769847e-10  
## period 0.005169311 2.569159e-05  
## Trip\_type 0.001738152 4.580306e-03  
##   
## $category  
## Estimate p.value  
## Trip\_distance\_range=Long\_dist 2.23397417 0.000000e+00  
## TipIsGiven=Yes 0.45216207 7.774385e-65  
## Payment\_type=Credit card 0.41968655 2.271313e-56  
## RateCodeID=Rate-Other 0.50422625 2.769847e-10  
## period=Period morning 0.20884328 1.137211e-03  
## Trip\_type=Dispatch 0.24121859 4.580306e-03  
## period=Period night 0.05154686 3.047979e-02  
## Trip\_type=Street-Hail -0.24121859 4.580306e-03  
## period=Period afternoon -0.19586260 1.290974e-04  
## RateCodeID=Rate-1 -0.50422625 2.769847e-10  
## Trip\_distance\_range=Medium\_dist -0.28824012 2.452911e-45  
## Payment\_type=Cash -0.40559005 2.694846e-56  
## TipIsGiven=No -0.45216207 7.774385e-65  
## Trip\_distance\_range=Short\_dist -1.94573405 0.000000e+00  
##   
## attr(,"class")  
## [1] "condes" "list "

In the first dimension we see that for the **quantitative** variables the most positively related, from more to less, are:

* Trip\_distance (0.95)
* Fare\_amount (0.94)
* Total\_amount (0.93)
* traveltime (0.80)

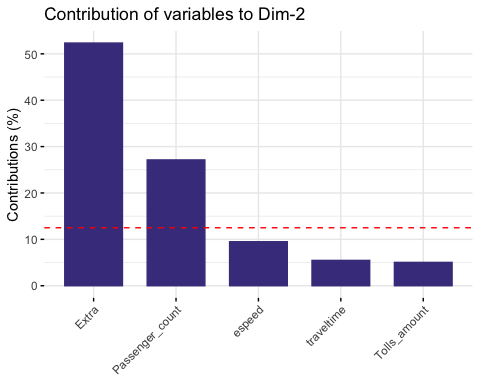
If we take look at the **qualitatives** ones, we that the most related is

* Trip\_distance\_range (0.69)

Finally, if we take a look at the **categories** we see that for the Trip\_distance\_range category long distance trips show a mean 2.23 units over the global mean and short distance ones show a mean -1.94 units under the global mean, so we can reject the H0 done in the t.Student test.

### Second dimension

fviz\_contrib( # contributions of variables to PC1  
 res.pca,   
 fill = "darkslateblue",  
 color = "darkslateblue",  
 choice = "var",   
 axes = 2,   
 top = 5)



res.des$Dim.2

## $quanti  
## correlation p.value  
## Extra 0.74258866 0.000000e+00  
## Passenger\_count 0.53463310 0.000000e+00  
## traveltime 0.23990250 1.615918e-61  
## Total\_amount 0.07947291 6.278874e-08  
## Fare\_amount 0.06251197 2.105822e-05  
## Tip\_amount 0.04580469 1.838358e-03  
## Pickup\_latitude -0.12147081 1.155632e-16  
## Dropoff\_latitude -0.12411309 2.469588e-17  
## Tolls\_amount -0.23032359 1.024002e-56  
## espeed -0.31615982 7.834681e-108  
##   
## $quali  
## R2 p.value  
## period 0.184068800 2.143099e-203  
## RateCodeID 0.018119629 3.862505e-20  
## Trip\_type 0.014819256 9.922508e-17  
## VendorID 0.002425023 8.098907e-04  
## TipIsGiven 0.001332968 1.304433e-02  
## Trip\_distance\_range 0.001446882 3.527015e-02  
##   
## $category  
## Estimate p.value  
## period=Period afternoon 0.69741738 6.273330e-126  
## RateCodeID=Rate-1 0.42270813 3.862505e-20  
## Trip\_type=Street-Hail 0.40639535 9.922508e-17  
## period=Period night 0.19868760 1.141234e-06  
## VendorID=f.Vendor-VeriFone 0.06200633 8.098907e-04  
## TipIsGiven=Yes 0.03867626 1.304433e-02  
## Trip\_distance\_range=Medium\_dist 0.06499883 4.081973e-02  
## Trip\_distance\_range=Long\_dist -0.06734957 4.739997e-02  
## TipIsGiven=No -0.03867626 1.304433e-02  
## VendorID=f.Vendor-Mobile -0.06200633 8.098907e-04  
## Trip\_type=Dispatch -0.40639535 9.922508e-17  
## RateCodeID=Rate-Other -0.42270813 3.862505e-20  
## period=Period valley -0.28051232 5.465420e-55  
## period=Period morning -0.61559267 5.765919e-69  
##   
## attr(,"class")  
## [1] "condes" "list "

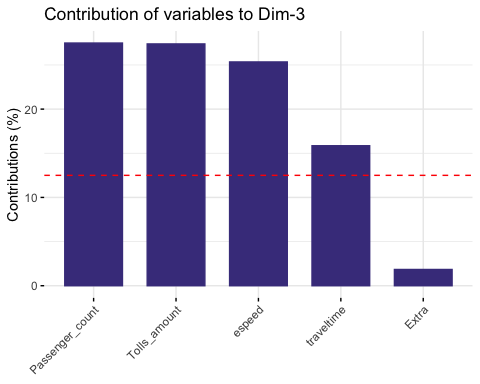
For the second dimension we see that or the **quantitative** variables Extra and Passenger\_count are the most positively related ones with 0.74 and 0.53 respectively.

If we see the **qualitative** variables we notice that period is the most related with 0.18 even though it is not a very remarkable data.

And we see that for this **category**, period afternoon mean is 0.69 units over the global mean and period morning mean, on the contrary, is -0.61 units under the global mean, so we can reject the H0 done in the t.Student test.

### Third dimension

fviz\_contrib( # contributions of variables to PC1  
 res.pca,   
 fill = "darkslateblue",  
 color = "darkslateblue",  
 choice = "var",   
 axes = 3,   
 top = 5)



res.des$Dim.3

## $quanti  
## correlation p.value  
## Passenger\_count 0.53445793 0.000000e+00  
## Tolls\_amount 0.53348146 0.000000e+00  
## espeed 0.51322530 3.958881e-309  
## Extra 0.13832221 3.460374e-21  
## Dropoff\_longitude 0.08626112 4.241523e-09  
## Pickup\_longitude 0.07649050 1.919027e-07  
## Tip\_amount 0.05620014 1.317391e-04  
## Dropoff\_latitude 0.04007164 6.431426e-03  
## Pickup\_latitude 0.03744970 1.088064e-02  
## Total\_amount -0.06349286 1.558600e-05  
## Fare\_amount -0.13644926 1.178290e-20  
## traveltime -0.40591753 6.233710e-183  
##   
## $quali  
## R2 p.value  
## period 0.035886226 2.283135e-36  
## Trip\_distance\_range 0.007909240 1.080799e-08  
## TipIsGiven 0.004524510 4.707055e-06  
## Payment\_type 0.003949701 1.070864e-04  
## VendorID 0.001086215 2.503325e-02  
##   
## $category  
## Estimate p.value  
## period=Period night 0.282886526 4.247490e-30  
## TipIsGiven=Yes 0.070766034 4.707055e-06  
## Payment\_type=Credit card 0.121518708 2.298510e-05  
## Trip\_distance\_range=Short\_dist 0.064024746 1.353427e-04  
## VendorID=f.Vendor-VeriFone 0.041213596 2.503325e-02  
## VendorID=f.Vendor-Mobile -0.041213596 2.503325e-02  
## Payment\_type=Cash -0.004578138 4.465703e-05  
## TipIsGiven=No -0.070766034 4.707055e-06  
## Trip\_distance\_range=Medium\_dist -0.152026208 1.617657e-09  
## period=Period morning -0.205703946 2.492716e-10  
## period=Period valley -0.144508011 4.079781e-16  
##   
## attr(,"class")  
## [1] "condes" "list "

For the last dimension we took into account, the third one, we see that the most related **quantitative** variables are:

* Passenger\_count (0.53)
* Tolls\_amount (0.53)
* espeed (0.51),

For the inversely related one, we also see that traveltime time (-0.40).

For the **quanlitatives**, we see that period is the category that is more related with 0.36, even though it is not a big relation.

And we see that for this **category**, period afternoon mean is 0.28 units over the global mean and period valley mean, on the contrary, is -0.14 units under the global mean, hough it is not either a big relation.

**We can conclude, then, that the first dimension is the one with the biggest correlations.**

## Perform a PCA taking into account also supplementary variables the supplementary variables can be quantitative and/or categorical

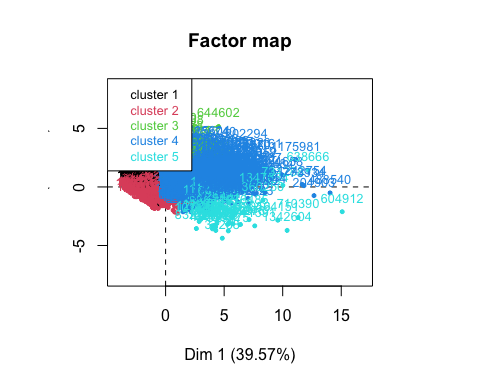
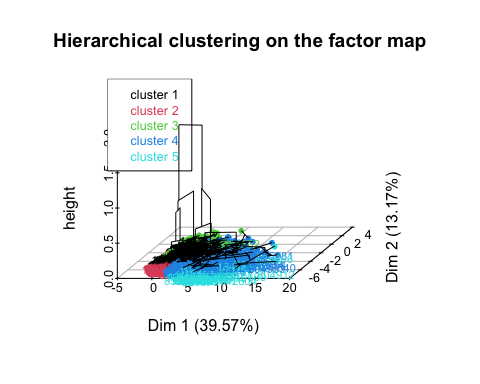
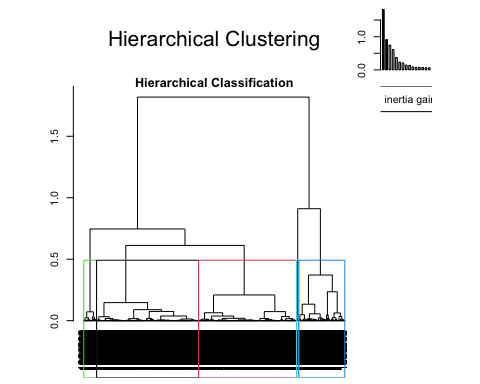
We want to take analyze the supplementary factor **kind of rate**, so we want to add lines that join the categories of this factor for the first factorial plane. With the following plot we can see it.

plot(res.pca$ind$coord[,1],res.pca$ind$coord[,2],pch=19,col="grey30") # draw all the individuals in grey  
points(res.pca$quali.sup$coord[,1],res.pca$quali.sup$coord[,2],pch=15,col="cadetblue1") # points associated with the categories gravitatorial centers  
lines(res.pca$quali.sup$coord[3:4,1],res.pca$quali.sup$coord[3:4,2],lwd=2,lty=2,col="coral") # draw a line that joins the categories that we want to take a look at  
text(res.pca$quali.sup$coord[,1],res.pca$quali.sup$coord[,2],labels=names(res.pca$quali.sup$coord[,1]),col="cadetblue1",cex=0.5) #add the names of the different categories

## 

# Hierarchical Clustering

res.hcpc <- HCPC(res.pca,nb.clust = 5, order = TRUE)



*Note*: If we chose the default number of cluster it would be 3, as we can guess from the inertia reduction plot, that follows the Elbow’s rule (number of black lines plus 1). In our case, due to the amount of data we have, the reason why we chose 5 as the number of clusters is because, after trying different numbers, we thought it was the best way to distribute the data.

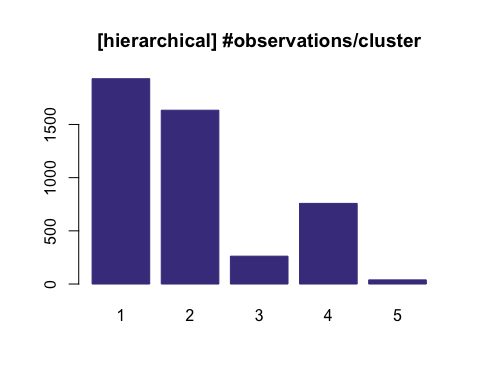
## Description of clusters

Number of observations in each cluster:

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

##   
## 1 2 3 4 5   
## 1930 1634 262 758 39

barplot(table(res.hcpc$data.clust$clust), col="darkslateblue", border="darkslateblue", main="[hierarchical] #observations/cluster")



## Interpret the results of the classification

### The description of the clusters by the variables

names(res.hcpc$desc.var)

## [1] "test.chi2" "category" "quanti.var" "quanti" "call"

res.hcpc$desc.var$test.chi2 # categorical variables which characterizes the clusters

## p.value df  
## period 0.000000e+00 12  
## Trip\_distance\_range 0.000000e+00 8  
## TipIsGiven 4.279197e-36 4  
## Payment\_type 1.274689e-28 8  
## RateCodeID 4.483773e-23 4  
## Trip\_type 1.609776e-21 4  
## VendorID 2.096463e-08 4

We start wit the description of the categorical variables that characterize the clusters, so in this output we do not have dimensions because it is the total association. We can see the intensity of the variables, in our case the variables that affect more to the clustering are **period** and **Trip\_distance\_range** because are the one with the smallest p.value. The variables associated to the clusters are the ones that appear on the output.

Next, we want to see for each cluster which are the categories that characterize them.The clusters that contain more individuals are the first, the second and the fourth one. Cluster number 4 has less individuals. We proceed to analyze them.

res.hcpc$desc.var$category # description of each cluster by the categories

## $`1`  
## Cla/Mod Mod/Cla Global p.value  
## period=Period night 64.0682095 54.50777202 35.518062 7.770495e-116  
## Trip\_distance\_range=Short\_dist 50.7065949 78.08290155 64.287259 1.280121e-63  
## period=Period afternoon 60.8142494 37.15025907 25.502920 6.952752e-53  
## RateCodeID=Rate-1 42.9048043 99.94818653 97.252866 4.277657e-29  
## Trip\_type=Street-Hail 42.7843050 100.00000000 97.577331 1.936966e-27  
## Payment\_type=Cash 44.0128154 56.94300518 54.012546 7.116030e-04  
## TipIsGiven=No 43.6502429 65.18134715 62.340472 7.289207e-04  
## Payment\_type=Credit card 39.0744275 42.43523316 45.338525 7.859632e-04  
## TipIsGiven=Yes 38.5985066 34.81865285 37.659528 7.289207e-04  
## Trip\_type=Dispatch 0.0000000 0.00000000 2.422669 1.936966e-27  
## RateCodeID=Rate-Other 0.7874016 0.05181347 2.747134 4.277657e-29  
## period=Period morning 0.7380074 0.20725389 11.723989 1.260284e-129  
## period=Period valley 12.4603175 8.13471503 27.255029 2.922636e-150  
## Trip\_distance\_range=Long\_dist 0.4511278 0.15544041 14.384599 2.585616e-166  
## v.test  
## period=Period night 22.877574  
## Trip\_distance\_range=Short\_dist 16.838228  
## period=Period afternoon 15.306182  
## RateCodeID=Rate-1 11.195750  
## Trip\_type=Street-Hail 10.852664  
## Payment\_type=Cash 3.385069  
## TipIsGiven=No 3.378464  
## Payment\_type=Credit card -3.357691  
## TipIsGiven=Yes -3.378464  
## Trip\_type=Dispatch -10.852664  
## RateCodeID=Rate-Other -11.195750  
## period=Period morning -24.223432  
## period=Period valley -26.108457  
## Trip\_distance\_range=Long\_dist -27.485937  
##   
## $`2`  
## Cla/Mod Mod/Cla Global p.value  
## period=Period valley 66.587302 51.346389 27.255029 7.063369e-159  
## period=Period morning 74.723247 24.785802 11.723989 1.245802e-88  
## Trip\_distance\_range=Short\_dist 42.698520 77.662179 64.287259 1.943824e-46  
## Trip\_type=Dispatch 73.214286 5.018360 2.422669 1.854170e-16  
## RateCodeID=Rate-Other 66.141732 5.140759 2.747134 1.024771e-12  
## TipIsGiven=No 38.965996 68.727050 62.340472 2.645583e-11  
## Payment\_type=Cash 39.006808 59.608323 54.012546 1.570437e-08  
## Payment\_type=Credit card 30.963740 39.718482 45.338525 1.300378e-08  
## TipIsGiven=Yes 29.350948 31.272950 37.659528 2.645583e-11  
## RateCodeID=Rate-1 34.475089 94.859241 97.252866 1.024771e-12  
## Trip\_type=Street-Hail 34.404788 94.981640 97.577331 1.854170e-16  
## period=Period afternoon 18.999152 13.708690 25.502920 5.030711e-45  
## Trip\_distance\_range=Long\_dist 3.157895 1.285190 14.384599 1.831233e-103  
## period=Period night 10.109622 10.159119 35.518062 2.015359e-175  
## v.test  
## period=Period valley 26.856598  
## period=Period morning 19.959245  
## Trip\_distance\_range=Short\_dist 14.308236  
## Trip\_type=Dispatch 8.231155  
## RateCodeID=Rate-Other 7.127138  
## TipIsGiven=No 6.665059  
## Payment\_type=Cash 5.653685  
## Payment\_type=Credit card -5.686015  
## TipIsGiven=Yes -6.665059  
## RateCodeID=Rate-1 -7.127138  
## Trip\_type=Street-Hail -8.231155  
## period=Period afternoon -14.080144  
## Trip\_distance\_range=Long\_dist -21.599106  
## period=Period night -28.237702  
##   
## $`3`  
## Cla/Mod Mod/Cla Global p.value v.test  
## VendorID=f.Vendor-VeriFone 6.767123 94.2748092 78.953061 1.557606e-12 7.069261  
## period=Period night 6.942753 43.5114504 35.518062 6.033525e-03 2.745954  
## RateCodeID=Rate-1 5.782918 99.2366412 97.252866 2.625621e-02 2.222401  
## RateCodeID=Rate-Other 1.574803 0.7633588 2.747134 2.625621e-02 -2.222401  
## period=Period valley 4.365079 20.9923664 27.255029 1.697607e-02 -2.387226  
## period=Period morning 2.767528 5.7251908 11.723989 8.241798e-04 -3.344544  
## VendorID=f.Vendor-Mobile 1.541624 5.7251908 21.046939 1.557606e-12 -7.069261  
##   
## $`4`  
## Cla/Mod Mod/Cla Global p.value  
## Trip\_distance\_range=Long\_dist 87.5187970 76.781003 14.384599 0.000000e+00  
## TipIsGiven=Yes 24.6984492 56.728232 37.659528 2.002989e-31  
## Payment\_type=Credit card 22.8530534 63.192612 45.338525 3.776109e-27  
## RateCodeID=Rate-Other 28.3464567 4.749340 2.747134 6.121937e-04  
## period=Period night 18.2095006 39.445910 35.518062 1.401893e-02  
## Trip\_type=Dispatch 25.0000000 3.693931 2.422669 1.829357e-02  
## period=Period morning 19.7416974 14.116095 11.723989 2.804593e-02  
## VendorID=f.Vendor-Mobile 18.4994861 23.746702 21.046939 4.833228e-02  
## VendorID=f.Vendor-VeriFone 15.8356164 76.253298 78.953061 4.833228e-02  
## Trip\_type=Street-Hail 16.1826646 96.306069 97.577331 1.829357e-02  
## RateCodeID=Rate-1 16.0587189 95.250660 97.252866 6.121937e-04  
## period=Period afternoon 12.9770992 20.184697 25.502920 1.834710e-04  
## Payment\_type=Cash 10.8930717 35.883905 54.012546 5.912321e-28  
## TipIsGiven=No 11.3809854 43.271768 62.340472 2.002989e-31  
## Trip\_distance\_range=Short\_dist 0.4710633 1.846966 64.287259 0.000000e+00  
## v.test  
## Trip\_distance\_range=Long\_dist Inf  
## TipIsGiven=Yes 11.661577  
## Payment\_type=Credit card 10.791491  
## RateCodeID=Rate-Other 3.426154  
## period=Period night 2.456778  
## Trip\_type=Dispatch 2.359622  
## period=Period morning 2.196643  
## VendorID=f.Vendor-Mobile 1.974435  
## VendorID=f.Vendor-VeriFone -1.974435  
## Trip\_type=Street-Hail -2.359622  
## RateCodeID=Rate-1 -3.426154  
## period=Period afternoon -3.740751  
## Payment\_type=Cash -10.960574  
## TipIsGiven=No -11.661577  
## Trip\_distance\_range=Short\_dist -Inf  
##   
## $`5`  
## Cla/Mod Mod/Cla Global p.value  
## Trip\_distance\_range=Long\_dist 4.51127820 76.923077 14.384599 1.878553e-18  
## Payment\_type=Credit card 1.52671756 82.051282 45.338525 2.937287e-06  
## TipIsGiven=Yes 1.60827111 71.794872 37.659528 1.783365e-05  
## period=Period morning 2.02952030 28.205128 11.723989 5.186239e-03  
## RateCodeID=Rate-Other 3.14960630 10.256410 2.747134 2.519752e-02  
## RateCodeID=Rate-1 0.77846975 89.743590 97.252866 2.519752e-02  
## TipIsGiven=No 0.38167939 28.205128 62.340472 1.783365e-05  
## Payment\_type=Cash 0.28033640 17.948718 54.012546 4.309549e-06  
## Trip\_distance\_range=Short\_dist 0.03364738 2.564103 64.287259 2.003816e-16  
## v.test  
## Trip\_distance\_range=Long\_dist 8.764351  
## Payment\_type=Credit card 4.675157  
## TipIsGiven=Yes 4.290419  
## period=Period morning 2.795233  
## RateCodeID=Rate-Other 2.238361  
## RateCodeID=Rate-1 -2.238361  
## TipIsGiven=No -4.290419  
## Payment\_type=Cash -4.595866  
## Trip\_distance\_range=Short\_dist -8.221854

* Cluster 1
  + The first thing we can notice from this cluster is that **Trip\_type=Street-Hail** that intervents in the 97.58% from the sample, in this cluster is the 100% of the observations, which means that all the observations in this cluster have this type of trip. We have 42.78% from the Trip\_type=Street-Hail observations in this cluster. As we can see and expect, from the other trip\_type that we have in this cluster is that **Trip\_type=Dispatch** that intervents in the 2.42% from the sample, in this cluster is not represented, we get 0% of the observations. Then, we can notice is the kind of rate. We can see that **RateCodeID=Rate-1**, the one that represents the standard rate, and means the 97.25% of our sample, in this cluster is the 99.95% of the observations, almost every observation from this cluster is a standard rate trip. In this cluster we have 42.90% of the observations from this category. In the other hand, we have the kind of rate, that contains the other options, represents the 2.75% of our sample, in this cluster is the 0.05% of the observations. In this cluster, we have the 0.79% of the observations from this category.
* Cluster 2
  + The first thing we can notice from this cluster is that **RateCodeID=Rate-1** (standard rate) and **Trip\_type=Street-Hail** are the most represented in the cluster. We have 94.98% of the observations in the cluster that represent street-hail trips, and we also have 94.86% of the observations in the cluster that represent the standard rate trips. We have 74.72% of the morning period trips of the observations in the sample represented in this cluster, 73.21% of the dispatch type trips of the observations in the sample represented in this cluster, 66.59% of the valley period trips of the observations in the sample represented in this cluster, we also have the 66.14% of the other kind of rates f the observations in the sample represented in this cluster. In the other hand, we only have 3.16% of the long distance trips in the sample represented in this cluster and this category only means the 1.29% of the observations in the cluster of this category. We have 10.11% of the night period trips in the sample represented in this cluster and we have almost 19% of the afternoon period trips in the sample represented in this cluster.
* Cluster 3
  + The first thing we can notice from this cluster is that almost every observation is from standard rate kind. We can see that 99.24% of the observations in the cluster are **RateCodeID=Rate-1**, and the cluster contains the 5.78% of the observations in the sample of this kind. The rest of observations in the cluster are from **RateCodeID=Rate-Other** kind.The next thing we can notice from this cluster is that, also, almost every observation is from Verfione kind of vendor. We have the 94.27% of the observations in this cluster of **VendorID=f.Vendor-VeriFone** category. This categories represents the 78.95% from our sample, and the cluster contains the 6.77% of obervations of this kind. For the other kind of vendor, **VendorID=f.Vendor-Mobile**, that represents the 21.05% of our sample, we have that in this cluster, 5.73% of the observations are from this vendor, and the cluster contains 1.54% of observations of this kind. If we take a look at the period categories, we see that **period=Period night** represents 43.51% of the observations in the cluster, and we have the 6.94% of the observations of this kind from the sample. In this cluster the night period is over represented because this kind of period represents the 35.52% of observations from our sample. For the **period=Period valley**, we have 20.99% of the observations in the cluster of this kind of period. We have in this cluster 4.37% of the observations of this kind from our sample. The last kind of period that we have in this cluster is the moring one, that represents the 5.73% of the observations in the cluster and we have 2.77% of the observations from the sample of this kind in this cluster.
* Cluster 4
  + In this cluster, we can see that the category more represented is **Trip\_type=Street-Hail** with 96.31% of the observations in the cluster. We get 16.18% of the observations of this kind from the sample in the cluster. Another category that is very represented is the standard rate, **RateCodeID=Rate-1**, with 95.25% of the observations in the cluster. From the sample, we get in this cluster, 16.06% of the observations of this kind. We can notice that we have 87.52% of long distance trip observations from the sample in this cluster. We can see that this category is over represented in this cluster because this category represents the 14.38% of the sample, and 76.78% of the observations in the cluster are of this category. In the other hand, we can see that short distance trips that represents 1.85% of the observations in the cluster and we only got 0.47% of the observations of this kind from the sample.
* Cluster 5
  + This cluster is the smallest one, we only have 39 observations from the sample. We can see in this cluster is that the **RateCodeID=Rate-1** represents the 89.75% of the observations in this cluster. In this cluster we only have 0.78% of the observations from the sample of this kind. The rest 10.25% are the **RateCodeID=Rate-Other** observtions in the cluster. In this case, we have a 3.15% of the observations from the sample of this kind in this cluster. Then we have that 82.05% of the observations in the cluster that paid credit card, and we got 1.53% of the observations from sample sample of this kind this cluster. The other 17.95% of the observations in the cluster paid in cash, and we got less representation from the sample in this cluster for this category, we only got 0.28% of the observations from the sample.

We now proceed to see the quantitative variables that characterizes the clusters.

res.hcpc$desc.var$quanti.var # quantitative variables which characterizes the clusters

## Eta2 P-value  
## Passenger\_count 0.781083003 0.000000e+00  
## Trip\_distance 0.578106343 0.000000e+00  
## Fare\_amount 0.575439601 0.000000e+00  
## Extra 0.632538094 0.000000e+00  
## Tolls\_amount 0.981954788 0.000000e+00  
## Total\_amount 0.539522699 0.000000e+00  
## traveltime 0.419905351 0.000000e+00  
## espeed 0.205381252 1.391829e-228  
## Tip\_amount 0.202596695 4.421382e-225  
## Dropoff\_latitude 0.018549311 7.346910e-18  
## Pickup\_latitude 0.016472560 8.618675e-16  
## Dropoff\_longitude 0.009820162 3.006725e-09  
## Pickup\_longitude 0.004646807 2.504182e-04

We can see in the output that all the variables that appear are slightly over represented in the clusters. We can notice that the greatest represented is the Total\_amount with 0.98 units over the global mean, we can also remark the Passenger\_count with 0.78 units over the mean and the Extra variable with 0.63 units over the mean. The least over represented are the Pickup\_longitude with 0.004 units over the mean, the Dropoff\_longitude with 0.01 units over the mean, the Pickup\_latitude with 0.016 units over the mean and the Dropoff\_latitude with 0.02 units over the total mean.

We want to know now which variables are associated with the quantitative variables.

res.hcpc$desc.var$quanti # description of each cluster by the quantitative variables

## $`1`  
## v.test Mean in category Overall mean sd in category  
## Extra 48.725143 0.6626943 0.35226044 0.23425993  
## Dropoff\_longitude 5.981195 -73.9299781 -73.93460830 0.04395684  
## Pickup\_longitude 3.321671 -73.9325877 -73.93496823 0.04237046  
## Dropoff\_latitude -4.282820 40.7409033 40.74500568 0.05287830  
## Pickup\_latitude -4.735737 40.7422169 40.74676502 0.05237977  
## Tolls\_amount -5.433312 0.0000000 0.04769564 0.00000000  
## espeed -8.810257 19.0031003 20.33575305 6.29787224  
## Tip\_amount -10.443222 0.6893179 1.02203842 1.08615941  
## Passenger\_count -12.789408 1.1409326 1.37107208 0.41827819  
## Total\_amount -18.789110 10.6471503 13.92640493 4.50875619  
## traveltime -19.049278 9.1670035 12.48732425 5.94179824  
## Trip\_distance -20.757190 1.7205850 2.72449524 1.03949364  
## Fare\_amount -22.244878 8.4204663 11.61104706 3.53352131  
## Overall sd p.value  
## Extra 0.36668354 0.000000e+00  
## Dropoff\_longitude 0.04455396 2.215059e-09  
## Pickup\_longitude 0.04124656 8.948012e-04  
## Dropoff\_latitude 0.05512875 1.845399e-05  
## Pickup\_latitude 0.05527371 2.182601e-06  
## Tolls\_amount 0.50523041 5.531755e-08  
## espeed 8.70570362 1.248593e-18  
## Tip\_amount 1.83366715 1.573775e-25  
## Passenger\_count 1.03565723 1.878993e-37  
## Total\_amount 10.04487145 9.272116e-79  
## traveltime 10.03175633 6.661465e-81  
## Trip\_distance 2.78356770 1.055625e-95  
## Fare\_amount 8.25496368 1.264366e-109  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category  
## Dropoff\_latitude 8.827382 40.7546869 40.74500568 0.05701522  
## Pickup\_latitude 8.406078 40.7560085 40.74676502 0.05684751  
## Dropoff\_longitude -2.581594 -73.9368965 -73.93460830 0.04060069  
## Tolls\_amount -4.745339 0.0000000 0.04769564 0.00000000  
## Tip\_amount -11.980225 0.5850122 1.02203842 0.99664574  
## Passenger\_count -12.679469 1.1098324 1.37107208 0.37470104  
## espeed -13.935697 17.9222129 20.33575305 6.35570993  
## traveltime -14.229130 9.6475928 12.48732425 6.01107875  
## Fare\_amount -16.360397 8.9242741 11.61104706 4.11025949  
## Trip\_distance -17.849175 1.7360744 2.72449524 1.07373082  
## Total\_amount -18.266469 10.2761689 13.92640493 4.94499736  
## Extra -48.289253 0.0000000 0.35226044 0.00000000  
## Overall sd p.value  
## Dropoff\_latitude 0.05512875 1.071545e-18  
## Pickup\_latitude 0.05527371 4.239492e-17  
## Dropoff\_longitude 0.04455396 9.834518e-03  
## Tolls\_amount 0.50523041 2.081575e-06  
## Tip\_amount 1.83366715 4.510961e-33  
## Passenger\_count 1.03565723 7.685081e-37  
## espeed 8.70570362 3.844308e-44  
## traveltime 10.03175633 6.042928e-46  
## Fare\_amount 8.25496368 3.667285e-60  
## Trip\_distance 2.78356770 2.933368e-71  
## Total\_amount 10.04487145 1.530386e-74  
## Extra 0.36668354 0.000000e+00  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category  
## Passenger\_count 59.986235 5.0992366 1.3710721 0.6863440  
## Extra 3.765260 0.4351145 0.3522604 0.3543457  
## Total\_amount -2.537392 12.3968702 13.9264049 6.8282336  
## Fare\_amount -2.616552 10.3148473 11.6110471 6.3920807  
## Trip\_distance -2.945418 2.2324828 2.7244952 1.8662661  
## Overall sd p.value  
## Passenger\_count 1.0356572 0.0000000000  
## Extra 0.3666835 0.0001663758  
## Total\_amount 10.0448715 0.0111681899  
## Fare\_amount 8.2549637 0.0088822891  
## Trip\_distance 2.7835677 0.0032251885  
##   
## $`4`  
## v.test Mean in category Overall mean sd in category  
## Trip\_distance 49.106302 7.26458247 2.72449524 3.47580089  
## Fare\_amount 49.067121 25.06441195 11.61104706 9.24177619  
## Total\_amount 45.821920 29.21412929 13.92640493 11.86369386  
## traveltime 42.874587 26.77304310 12.48732425 12.32002615  
## espeed 28.378179 28.54141415 20.33575305 12.17319710  
## Tip\_amount 27.211285 2.67931398 1.02203842 3.09282254  
## Tolls\_amount -2.295339 0.00917784 0.04769564 0.14117624  
## Pickup\_longitude -3.443125 -73.93968523 -73.93496823 0.04283372  
## Pickup\_latitude -4.158084 40.73913128 40.74676502 0.05714529  
## Passenger\_count -4.305896 1.22295515 1.37107208 0.65713115  
## Extra -4.496790 0.29749340 0.35226044 0.33420886  
## Dropoff\_longitude -4.799514 -73.94171076 -73.93460830 0.05184553  
## Dropoff\_latitude -5.180004 40.73552077 40.74500568 0.05408675  
## Overall sd p.value  
## Trip\_distance 2.78356770 0.000000e+00  
## Fare\_amount 8.25496368 0.000000e+00  
## Total\_amount 10.04487145 0.000000e+00  
## traveltime 10.03175633 0.000000e+00  
## espeed 8.70570362 3.759899e-177  
## Tip\_amount 1.83366715 4.775939e-163  
## Tolls\_amount 0.50523041 2.171371e-02  
## Pickup\_longitude 0.04124656 5.750332e-04  
## Pickup\_latitude 0.05527371 3.209275e-05  
## Passenger\_count 1.03565723 1.663115e-05  
## Extra 0.36668354 6.898701e-06  
## Dropoff\_longitude 0.04455396 1.590515e-06  
## Dropoff\_latitude 0.05512875 2.218809e-07  
##   
## $`5`  
## v.test Mean in category Overall mean sd in category  
## Tolls\_amount 67.367546 5.475388 0.04769564 0.39829372  
## Total\_amount 17.705432 42.287692 13.92640493 20.69332947  
## Trip\_distance 13.871930 8.882127 2.72449524 5.24509423  
## Fare\_amount 13.439098 29.302370 11.61104706 13.01003029  
## Tip\_amount 12.655167 4.722564 1.02203842 4.52414418  
## espeed 10.141705 34.415339 20.33575305 11.95705914  
## traveltime 7.719334 24.836325 12.48732425 11.22620743  
## Pickup\_longitude 1.961840 -73.922064 -73.93496823 0.04269607  
## Overall sd p.value  
## Tolls\_amount 0.50523041 0.000000e+00  
## Total\_amount 10.04487145 3.807483e-70  
## Trip\_distance 2.78356770 9.372098e-44  
## Fare\_amount 8.25496368 3.567598e-41  
## Tip\_amount 1.83366715 1.047523e-36  
## espeed 8.70570362 3.607463e-24  
## traveltime 10.03175633 1.169396e-14  
## Pickup\_longitude 0.04124656 4.978116e-02

* Cluster 1
  + For this cluster, we can see that the **traveltime** is around 3 units under the overall mean, the **Fare\_amount** as well and the **Total\_amount** too. We can also see that the **Trip\_distance** is 1 unit under the overall mean and the **espeed** as well. We see that the only variable that is over the overall mean is the variable **Extra** with less than 0.3 units over it.
* Cluster 2
  + For the second cluster, happens something similar as with the first one. We see that the **Total\_amount** is around 3.7 units under the overall mean, **espeed** around 2 units under as well, **Tip\_amount** around 0.5 under the overall mean too, **traveltime** and **Fare\_amount** around 3 units under the overall mean as well, **Trip\_distance** around 1 unit under the mean. In this clusters the only variables ver the overall mean are **Dropoff\_latitude** and **Pickup\_latitude** but they are not remarkable since the increase is super light.
* Cluster 3
  + In this cluster we can see that the most remarkable variable is **Passenger\_count** with almost 4 units over the overall mean, then we also have **Total\_amount** with 0.1 units over the meant. In the other hand, we have **Total\_amount** and **Fare\_amount** with around 1 unit under the overall mean. **Trip\_distance** is around 0.5 units under the overall mean.
* Cluster 4
  + In this cluster we can see clearly the most remarkable vairables. We have 5 variables cleary over the overall mean. These are: **Total\_amount** with 26 units over the mean, **Fare\_amount** and **traveltime** with 14 units over the mean, **espeed** with 8 units over the mean and **Trip\_distance** with 5 units over the overall mean.
* Cluster 5
  + In this cluster every variable is over the overall mean. Every variable except **Pickup\_longitude** are remarkably over the overall mean. Firstly, we have the **Total\_amount** around 30 units over, then we have **Fare\_amount** 18 units over, **espeed** 14 units over, **traveltime** 12 units over, **Trip\_distance** 6 units over, **Tolls\_amount** 5 units over and **Tip\_amount** 3.7 units over the overall mean.

### The description of the clusters by the individuals

res.hcpc$desc.ind$para # representative individuals of each cluster

## Cluster: 1  
## 697423 442213 365332 655407 945065   
## 0.4551377 0.4585094 0.4624702 0.4675288 0.4733316   
## ------------------------------------------------------------   
## Cluster: 2  
## 665209 677545 343231 743541 473945   
## 0.1500605 0.1502214 0.1520744 0.1533864 0.1668652   
## ------------------------------------------------------------   
## Cluster: 3  
## 952205 21675 1090746 607516 1397283   
## 0.2651094 0.3722646 0.5401477 0.5498816 0.5620526   
## ------------------------------------------------------------   
## Cluster: 4  
## 1040597 1272173 10891 1445033 693126   
## 0.5534480 0.6419473 0.6769121 0.7137618 0.7296941   
## ------------------------------------------------------------   
## Cluster: 5  
## 1261276 1016299 327762 1010826 529475   
## 1.151077 1.224596 1.305726 1.472585 1.482492

What we obtain are the more representative individuals,paragons, for each cluster. We get the rownames of each paragon in every single cluster.

res.hcpc$desc.ind$dist # individuals distant from each cluster

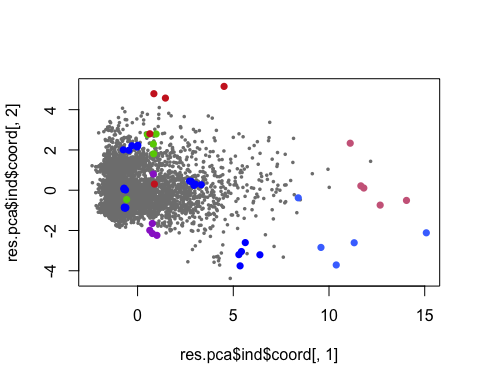
## Cluster: 1  
## 886530 642379 71268 1393691 560933   
## 4.878069 4.760057 4.577272 4.506090 4.465229   
## ------------------------------------------------------------   
## Cluster: 2  
## 36606 533937 535041 829742 1418974   
## 4.641497 4.283722 4.264553 4.177470 3.770009   
## ------------------------------------------------------------   
## Cluster: 3  
## 169380 644602 513170 550938 871576   
## 6.214858 6.161465 5.875364 5.669044 5.651629   
## ------------------------------------------------------------   
## Cluster: 4  
## 488540 204903 773934 1242754 1175981   
## 13.32453 12.61924 12.27617 12.27616 11.95419   
## ------------------------------------------------------------   
## Cluster: 5  
## 604912 710390 194151 1347654 1342604   
## 15.93179 13.33560 12.81720 12.39681 12.21009

What we obtain are those individuals of each cluster that that far away in the same cluster from the rest of the individuals. We also obtain the rownames of each individual with the bigger distance respect the other ones in the cluster.

#### Examine the values of individuals that characterize classes

We get the grpahical representation for the individuals that characterize classes (para and dist).

# characteristic individuals  
para1<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[1]]))  
dist1<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[1]]))  
para2<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[2]]))  
dist2<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[2]]))  
para3<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[3]]))  
dist3<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[3]]))  
para4<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[4]]))  
dist4<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[4]]))  
para5<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$para[[5]]))  
dist5<-which(rownames(res.pca$ind$coord)%in%names(res.hcpc$desc.ind$dist[[5]]))  
  
plot(res.pca$ind$coord[,1],res.pca$ind$coord[,2],col="grey50",cex=0.5,pch=16)  
points(res.pca$ind$coord[para1,1],res.pca$ind$coord[para1,2],col="blue",cex=1,pch=16)  
points(res.pca$ind$coord[dist1,1],res.pca$ind$coord[dist1,2],col="chartreuse3",cex=1,pch=16)  
points(res.pca$ind$coord[para2,1],res.pca$ind$coord[para2,2],col="blue",cex=1,pch=16)  
points(res.pca$ind$coord[dist2,1],res.pca$ind$coord[dist2,2],col="darkorchid3",cex=1,pch=16)  
points(res.pca$ind$coord[para3,1],res.pca$ind$coord[para3,2],col="blue",cex=1,pch=16)  
points(res.pca$ind$coord[dist3,1],res.pca$ind$coord[dist3,2],col="firebrick3",cex=1,pch=16)  
points(res.pca$ind$coord[para4,1],res.pca$ind$coord[para4,2],col="blue",cex=1,pch=16)  
points(res.pca$ind$coord[dist4,1],res.pca$ind$coord[dist4,2],col="palevioletred3",cex=1,pch=16)  
points(res.pca$ind$coord[para5,1],res.pca$ind$coord[para5,2],col="blue",cex=1,pch=16)  
points(res.pca$ind$coord[dist5,1],res.pca$ind$coord[dist5,2],col="royalblue1",cex=1,pch=16)



### Partition quality

We are going to evaluate the partition quality.

#### Gain in inertia (in %)

# ( between sum of squares / total sum of squares ) \* 100  
((res.hcpc$call$t$within[1]-res.hcpc$call$t$within[5])/res.hcpc$call$t$within[1])\*100

## [1] 57.49171

The quality of this reduction if of 57.49%.

In case we wanted to achieve an 80% of the clustering representativity we would need 18 clusters.

((res.hcpc$call$t$within[1]-res.hcpc$call$t$within[18])/res.hcpc$call$t$within[1])\*100

## [1] 80.59951

### Save the results into dataframe

res.hcpc$call$t$inert.gain[1:5]

## [1] 1.8187697 0.9105858 0.7460223 0.6120673 0.3712993

df$hcpck<-res.hcpc$data.clust$clust

# K-Means Classification

## Description of clusters

res.pca <- PCA(df[,c(1:10,12,13,15:17,19,21,22,25,27)],quanti.sup=c(3:6,13),quali.sup=c(1,2,14:16,19:20),ncp=5,graph=FALSE)  
ppcc<-res.pca$ind$coord[,1:3] # 3 components principals (kaiser)  
dim(ppcc)

## [1] 4623 3

### Optimal number of clusters

library("factoextra")  
#fviz\_nbclust(ppcc, kmeans, method = "gap\_stat") !!!!Descomentar pel deliverable, triga molt.

According to the previous plot, the optimal number of clusters per k-means is 1, so we guess maybe something is wrong or missing.

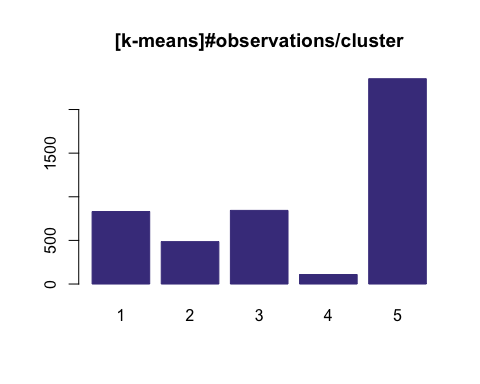
## Classification

dist<-dist(ppcc) # coordenates are real - Euclidean metric  
kc<-kmeans(dist, 5, iter.max=30, trace=TRUE) #caclulate the distances, it turns into a matrix

## KMNS(\*, k=5): iter= 1, indx=3  
## QTRAN(): istep=4623, icoun=0  
## QTRAN(): istep=9246, icoun=52  
## QTRAN(): istep=13869, icoun=6  
## QTRAN(): istep=18492, icoun=13  
## QTRAN(): istep=23115, icoun=1  
## QTRAN(): istep=27738, icoun=9  
## QTRAN(): istep=32361, icoun=27  
## QTRAN(): istep=36984, icoun=7  
## QTRAN(): istep=41607, icoun=49  
## QTRAN(): istep=46230, icoun=1  
## QTRAN(): istep=50853, icoun=6  
## QTRAN(): istep=55476, icoun=2  
## QTRAN(): istep=60099, icoun=777  
## KMNS(\*, k=5): iter= 2, indx=3  
## QTRAN(): istep=4623, icoun=25  
## QTRAN(): istep=9246, icoun=1  
## QTRAN(): istep=13869, icoun=5  
## QTRAN(): istep=18492, icoun=21  
## QTRAN(): istep=23115, icoun=226  
## QTRAN(): istep=27738, icoun=926  
## QTRAN(): istep=32361, icoun=3  
## QTRAN(): istep=36984, icoun=483  
## QTRAN(): istep=41607, icoun=4591  
## KMNS(\*, k=5): iter= 3, indx=3  
## QTRAN(): istep=4623, icoun=225  
## QTRAN(): istep=9246, icoun=690  
## QTRAN(): istep=13869, icoun=3645  
## KMNS(\*, k=5): iter= 4, indx=4623

We see from the output that in 4 iterations it has converged. We now procceed to save in the data frame the number of clusters.

df$claKM<-0  
df$claKM<-kc$cluster  
df$claKM<-factor(df$claKM)  
barplot(table(df$claKM),col="darkslateblue",border="darkslateblue",main="[k-means]#observations/cluster")



### Gain in inertia (in %)

The american school does the partition quality evaluation in 5 clusters is done very fast, and after executing the following chunk we get an explicability of the 77.99%

100\*(kc$betweenss/kc$totss)

## [1] 79.40953

### k-means clusters characteristics

If we want to know the characteristics of each cluster, as we did with the hierarchical, we need to execute a catdes to obtain these characteristics. In the following output we get them.

dim(df)

## [1] 4623 30

res.cat <-catdes(df,30)  
res.cat

##   
## Link between the cluster variable and the categorical variables (chi-square test)  
## =================================================================================  
## p.value df  
## Trip\_distance\_range 0.000000e+00 8  
## paidTolls 0.000000e+00 8  
## hcpck 0.000000e+00 16  
## pickup 1.114117e-215 92  
## dropoff 4.738913e-206 92  
## passenger\_groups 1.560774e-177 8  
## period 8.756108e-127 12  
## TipIsGiven 4.163217e-45 4  
## Payment\_type 4.711245e-34 8  
## RateCodeID 5.628907e-08 4  
## MTA\_tax 4.996468e-06 4  
## improvement\_surcharge 3.086294e-05 4  
## Trip\_type 4.421007e-05 4  
##   
## Description of each cluster by the categories  
## =============================================  
## $`1`  
## Cla/Mod Mod/Cla Global p.value  
## Trip\_distance\_range=Medium\_dist 59.736308 70.8784597 21.328142 9.830260e-273  
## hcpck=4 42.612137 38.8688327 16.396279 5.795841e-70  
## Trip\_distance\_range=Long\_dist 30.827068 24.6690734 14.384599 1.510254e-18  
## TipIsGiven=Yes 23.721999 49.6991576 37.659528 5.633107e-15  
## Payment\_type=Credit card 22.185115 55.9566787 45.338525 1.283731e-11  
## paidTolls=No 18.138112 99.8796631 98.983344 1.064902e-03  
## passenger\_groups=Single 18.738739 87.6052948 84.036340 1.519643e-03  
## pickup=10 26.701571 6.1371841 4.131516 2.257993e-03  
## period=Period night 20.219245 39.9518652 35.518062 3.394570e-03  
## pickup=06 33.333333 2.2864019 1.232966 5.171005e-03  
## dropoff=11 25.396825 5.7761733 4.088254 9.253274e-03  
## VendorID=f.Vendor-Mobile 20.452210 23.9470517 21.046939 2.509928e-02  
## dropoff=21 22.932331 7.3405535 5.753839 3.468393e-02  
## VendorID=f.Vendor-VeriFone 17.315068 76.0529483 78.953061 2.509928e-02  
## pickup=17 12.749004 3.8507822 5.429375 2.247023e-02  
## dropoff=17 12.648221 3.8507822 5.472637 1.935138e-02  
## hcpck=2 15.973072 31.4079422 35.345014 8.402245e-03  
## pickup=16 12.323944 4.2117930 6.143197 8.066705e-03  
## hcpck=5 0.000000 0.0000000 0.843608 4.250924e-04  
## paidTolls=Yes 0.000000 0.0000000 0.865239 3.480305e-04  
## dropoff=18 10.289389 3.8507822 6.727233 1.117498e-04  
## pickup=18 10.191083 3.8507822 6.792126 8.236661e-05  
## period=Period afternoon 13.910093 19.7352587 25.502920 1.740057e-05  
## passenger\_groups=Group 7.848101 3.7304452 8.544235 2.569581e-09  
## Payment\_type=Cash 14.457349 43.4416366 54.012546 1.612540e-11  
## hcpck=3 3.053435 0.9626955 5.667316 3.045013e-14  
## TipIsGiven=No 14.503817 50.3008424 62.340472 5.633107e-15  
## hcpck=1 12.383420 28.7605295 41.747783 1.603615e-17  
## Trip\_distance\_range=Short\_dist 1.244953 4.4524669 64.287259 0.000000e+00  
## v.test  
## Trip\_distance\_range=Medium\_dist 35.285413  
## hcpck=4 17.681760  
## Trip\_distance\_range=Long\_dist 8.788904  
## TipIsGiven=Yes 7.811903  
## Payment\_type=Credit card 6.770461  
## paidTolls=No 3.272794  
## passenger\_groups=Single 3.170906  
## pickup=10 3.054017  
## period=Period night 2.929547  
## pickup=06 2.796183  
## dropoff=11 2.602552  
## VendorID=f.Vendor-Mobile 2.239871  
## dropoff=21 2.112029  
## VendorID=f.Vendor-VeriFone -2.239871  
## pickup=17 -2.282324  
## dropoff=17 -2.338692  
## hcpck=2 -2.635464  
## pickup=16 -2.649265  
## hcpck=5 -3.523995  
## paidTolls=Yes -3.576646  
## dropoff=18 -3.863553  
## pickup=18 -3.937408  
## period=Period afternoon -4.295875  
## passenger\_groups=Group -5.956970  
## Payment\_type=Cash -6.737393  
## hcpck=3 -7.596392  
## TipIsGiven=No -7.811903  
## hcpck=1 -8.519415  
## Trip\_distance\_range=Short\_dist -Inf  
##   
## $`2`  
## Cla/Mod Mod/Cla Global p.value  
## hcpck=4 48.2849604 75.3086420 16.396279 1.021557e-216  
## Trip\_distance\_range=Long\_dist 51.4285714 70.3703704 14.384599 2.789031e-208  
## hcpck=3 42.3664122 22.8395062 5.667316 4.166709e-44  
## passenger\_groups=Group 34.1772152 27.7777778 8.544235 1.987956e-41  
## TipIsGiven=Yes 15.6232051 55.9670782 37.659528 5.218460e-18  
## Payment\_type=Credit card 14.5038168 62.5514403 45.338525 8.678931e-16  
## RateCodeID=Rate-Other 19.6850394 5.1440329 2.747134 1.867120e-03  
## dropoff=17 16.6007905 8.6419753 5.472637 2.315914e-03  
## MTA\_tax=No 19.3277311 4.7325103 2.574086 3.719873e-03  
## Trip\_type=Dispatch 17.8571429 4.1152263 2.422669 1.738247e-02  
## improvement\_surcharge=No 16.9491525 4.1152263 2.552455 3.059642e-02  
## pickup=01 15.4320988 5.1440329 3.504218 4.788939e-02  
## dropoff=12 5.9523810 2.0576132 3.634004 3.970823e-02  
## improvement\_surcharge=Yes 10.3440622 95.8847737 97.447545 3.059642e-02  
## period=Period valley 8.8888889 23.0452675 27.255029 2.592855e-02  
## Trip\_type=Street-Hail 10.3303037 95.8847737 97.577331 1.738247e-02  
## hcpck=5 0.0000000 0.0000000 0.843608 1.289666e-02  
## MTA\_tax=Yes 10.2797513 95.2674897 97.425914 3.719873e-03  
## pickup=12 4.4444444 1.6460905 3.893576 3.252368e-03  
## RateCodeID=Rate-1 10.2535587 94.8559671 97.252866 1.867120e-03  
## pickup=20 5.5555556 3.4979424 6.619079 1.788241e-03  
## dropoff=14 4.5662100 2.0576132 4.737184 1.391931e-03  
## Trip\_distance\_range=Medium\_dist 6.9979716 14.1975309 21.328142 2.510501e-05  
## Payment\_type=Cash 7.1285543 36.6255144 54.012546 4.368131e-16  
## TipIsGiven=No 7.4253990 44.0329218 62.340472 5.218460e-18  
## passenger\_groups=Single 8.3397683 66.6666667 84.036340 6.471313e-24  
## hcpck=2 0.1223990 0.4115226 35.345014 1.010014e-94  
## hcpck=1 0.3626943 1.4403292 41.747783 6.925515e-109  
## Trip\_distance\_range=Short\_dist 2.5235532 15.4320988 64.287259 1.499111e-122  
## v.test  
## hcpck=4 31.421736  
## Trip\_distance\_range=Long\_dist 30.797978  
## hcpck=3 13.929946  
## passenger\_groups=Group 13.482306  
## TipIsGiven=Yes 8.648492  
## Payment\_type=Credit card 8.044230  
## RateCodeID=Rate-Other 3.110593  
## dropoff=17 3.046410  
## MTA\_tax=No 2.900989  
## Trip\_type=Dispatch 2.378516  
## improvement\_surcharge=No 2.162282  
## pickup=01 1.978349  
## dropoff=12 -2.056771  
## improvement\_surcharge=Yes -2.162282  
## period=Period valley -2.227280  
## Trip\_type=Street-Hail -2.378516  
## hcpck=5 -2.486610  
## MTA\_tax=Yes -2.900989  
## pickup=12 -2.942821  
## RateCodeID=Rate-1 -3.110593  
## pickup=20 -3.123319  
## dropoff=14 -3.196319  
## Trip\_distance\_range=Medium\_dist -4.213854  
## Payment\_type=Cash -8.127894  
## TipIsGiven=No -8.648492  
## passenger\_groups=Single -10.084471  
## hcpck=2 -20.648355  
## hcpck=1 -22.168450  
## Trip\_distance\_range=Short\_dist -23.542477  
##   
## $`3`  
## Cla/Mod Mod/Cla Global p.value  
## hcpck=1 35.9585492 82.2274882 41.7477828 2.590084e-157  
## period=Period afternoon 41.4758270 57.9383886 25.5029202 1.523397e-112  
## Trip\_distance\_range=Short\_dist 25.3364738 89.2180095 64.2872594 1.347784e-72  
## passenger\_groups=Group 50.1265823 23.4597156 8.5442353 3.342170e-52  
## dropoff=18 54.3408360 20.0236967 6.7272334 1.875281e-50  
## pickup=18 53.5031847 19.9052133 6.7921263 7.274603e-49  
## hcpck=3 54.1984733 16.8246445 5.6673156 7.896015e-42  
## dropoff=19 50.1607717 18.4834123 6.7272334 1.779927e-40  
## pickup=17 52.1912351 15.5213270 5.4293749 3.955139e-36  
## pickup=16 49.2957746 16.5876777 6.1431971 4.664129e-35  
## pickup=19 47.7272727 17.4170616 6.6623405 8.386326e-35  
## dropoff=17 48.6166008 14.5734597 5.4726368 6.125909e-30  
## dropoff=16 45.9074733 15.2843602 6.0783041 2.700838e-28  
## passenger\_groups=Couple 39.6501458 16.1137441 7.4194246 3.433183e-22  
## RateCodeID=Rate-1 18.7277580 99.7630332 97.2528661 2.499491e-09  
## MTA\_tax=Yes 18.6944938 99.7630332 97.4259139 1.160586e-08  
## improvement\_surcharge=Yes 18.6903441 99.7630332 97.4475449 1.405074e-08  
## Trip\_type=Street-Hail 18.6654844 99.7630332 97.5773307 4.407526e-08  
## paidTolls=No 18.4440559 100.0000000 98.9833441 7.285175e-05  
## TipIsGiven=No 19.5697432 66.8246445 62.3404716 2.794274e-03  
## Payment\_type=Cash 19.7837405 58.5308057 54.0125460 3.530940e-03  
## pickup=20 14.0522876 5.0947867 6.6190785 4.445448e-02  
## pickup=00 13.0630631 3.4360190 4.8020766 3.503728e-02  
## dropoff=22 13.0252101 3.6729858 5.1481722 2.744997e-02  
## pickup=05 6.0000000 0.3554502 1.0815488 1.485557e-02  
## dropoff=05 5.8823529 0.3554502 1.1031798 1.275707e-02  
## Payment\_type=Credit card 16.5553435 41.1137441 45.3385248 6.314097e-03  
## pickup=22 11.7886179 3.4360190 5.3212200 4.957197e-03  
## dropoff=01 10.1190476 2.0142180 3.6340039 3.321822e-03  
## TipIsGiven=Yes 16.0827111 33.1753555 37.6595284 2.794274e-03  
## pickup=01 9.2592593 1.7772512 3.5042180 1.300278e-03  
## pickup=23 10.0961538 2.4881517 4.4992429 9.689147e-04  
## hcpck=5 0.0000000 0.0000000 0.8436080 3.715552e-04  
## paidTolls=Yes 0.0000000 0.0000000 0.8652390 3.031450e-04  
## dropoff=06 0.0000000 0.0000000 0.9301319 1.645898e-04  
## dropoff=21 9.3984962 2.9620853 5.7538395 3.865807e-05  
## pickup=21 8.8607595 2.4881517 5.1265412 3.633993e-05  
## dropoff=23 8.0717489 2.1327014 4.8237075 1.214810e-05  
## pickup=06 0.0000000 0.0000000 1.2329656 9.463000e-06  
## period=Period valley 13.8888889 20.7345972 27.2550292 1.568281e-06  
## pickup=15 6.6371681 1.7772512 4.8886005 3.029225e-07  
## dropoff=08 4.5161290 0.8293839 3.3528012 2.967543e-07  
## Trip\_type=Dispatch 1.7857143 0.2369668 2.4226693 4.407526e-08  
## improvement\_surcharge=No 1.6949153 0.2369668 2.5524551 1.405074e-08  
## MTA\_tax=No 1.6806723 0.2369668 2.5740861 1.160586e-08  
## dropoff=07 0.9433962 0.1184834 2.2928834 1.052048e-08  
## pickup=08 3.6144578 0.7109005 3.5907419 8.563816e-09  
## pickup=07 1.6528926 0.2369668 2.6173480 7.914224e-09  
## dropoff=12 3.5714286 0.7109005 3.6340039 6.049325e-09  
## RateCodeID=Rate-Other 1.5748031 0.2369668 2.7471339 2.499491e-09  
## dropoff=10 3.7433155 0.8293839 4.0449924 1.342739e-09  
## pickup=11 2.9761905 0.5924171 3.6340039 9.095265e-10  
## dropoff=15 4.5454545 1.1848341 4.7588146 7.656623e-10  
## pickup=12 3.3333333 0.7109005 3.8935756 7.364870e-10  
## pickup=10 3.6649215 0.8293839 4.1315163 6.711789e-10  
## dropoff=13 2.2222222 0.4739336 3.8935756 1.197572e-11  
## pickup=13 1.7241379 0.3554502 3.7637897 3.526453e-12  
## dropoff=11 2.1164021 0.4739336 4.0882544 2.186471e-12  
## dropoff=14 2.7397260 0.7109005 4.7371836 6.370576e-13  
## dropoff=09 1.6216216 0.3554502 4.0017305 4.183562e-13  
## pickup=09 1.6216216 0.3554502 4.0017305 4.183562e-13  
## pickup=14 2.6315789 0.7109005 4.9318624 1.202828e-13  
## Trip\_distance\_range=Medium\_dist 9.1277890 10.6635071 21.3281419 6.109449e-19  
## period=Period night 9.9878197 19.4312796 35.5180619 3.460442e-29  
## period=Period morning 2.9520295 1.8957346 11.7239888 1.532512e-30  
## Trip\_distance\_range=Long\_dist 0.1503759 0.1184834 14.3845987 7.046119e-62  
## hcpck=4 0.0000000 0.0000000 16.3962795 5.906224e-74  
## passenger\_groups=Single 13.1274131 60.4265403 84.0363400 1.731005e-79  
## hcpck=2 0.4895961 0.9478673 35.3450141 1.859514e-164  
## v.test  
## hcpck=1 26.722331  
## period=Period afternoon 22.544416  
## Trip\_distance\_range=Short\_dist 18.020395  
## passenger\_groups=Group 15.203697  
## dropoff=18 14.937630  
## pickup=18 14.691808  
## hcpck=3 13.550251  
## dropoff=19 13.319628  
## pickup=17 12.550399  
## pickup=16 12.353493  
## pickup=19 12.306217  
## dropoff=17 11.366701  
## dropoff=16 11.031250  
## passenger\_groups=Couple 9.686738  
## RateCodeID=Rate-1 5.961489  
## MTA\_tax=Yes 5.705417  
## improvement\_surcharge=Yes 5.672769  
## Trip\_type=Street-Hail 5.473693  
## paidTolls=No 3.966775  
## TipIsGiven=No 2.989508  
## Payment\_type=Cash 2.917284  
## pickup=20 -2.009780  
## pickup=00 -2.107927  
## dropoff=22 -2.205059  
## pickup=05 -2.435881  
## dropoff=05 -2.490480  
## Payment\_type=Credit card -2.731008  
## pickup=22 -2.809802  
## dropoff=01 -2.936273  
## TipIsGiven=Yes -2.989508  
## pickup=01 -3.215918  
## pickup=23 -3.299401  
## hcpck=5 -3.559504  
## paidTolls=Yes -3.612598  
## dropoff=06 -3.767956  
## dropoff=21 -4.115357  
## pickup=21 -4.129597  
## dropoff=23 -4.374913  
## pickup=06 -4.429093  
## period=Period valley -4.802332  
## pickup=15 -5.121620  
## dropoff=08 -5.125497  
## Trip\_type=Dispatch -5.473693  
## improvement\_surcharge=No -5.672769  
## MTA\_tax=No -5.705417  
## dropoff=07 -5.722117  
## pickup=08 -5.756968  
## pickup=07 -5.770275  
## dropoff=12 -5.815388  
## RateCodeID=Rate-Other -5.961489  
## dropoff=10 -6.062198  
## pickup=11 -6.124528  
## dropoff=15 -6.151887  
## pickup=12 -6.158044  
## pickup=10 -6.172737  
## dropoff=13 -6.780504  
## pickup=13 -6.954967  
## dropoff=11 -7.022043  
## dropoff=14 -7.192307  
## dropoff=09 -7.249486  
## pickup=09 -7.249486  
## pickup=14 -7.416470  
## Trip\_distance\_range=Medium\_dist -8.890026  
## period=Period night -11.214524  
## period=Period morning -11.487053  
## Trip\_distance\_range=Long\_dist -16.599340  
## hcpck=4 -18.192608  
## passenger\_groups=Single -18.877974  
## hcpck=2 -27.330149  
##   
## $`4`  
## Cla/Mod Mod/Cla Global p.value  
## Trip\_distance\_range=Long\_dist 14.7368421 89.9082569 14.3845987 4.103928e-72  
## hcpck=5 100.0000000 35.7798165 0.8436080 1.655363e-67  
## paidTolls=Yes 95.0000000 34.8623853 0.8652390 8.094914e-63  
## hcpck=4 9.1029024 63.3027523 16.3962795 7.818532e-29  
## TipIsGiven=Yes 3.9058013 62.3853211 37.6595284 1.424189e-07  
## Payment\_type=Credit card 3.6259542 69.7247706 45.3385248 2.269663e-07  
## paidTolls=NA 57.1428571 3.6697248 0.1514168 9.830213e-06  
## dropoff=05 9.8039216 4.5871560 1.1031798 7.727506e-03  
## RateCodeID=Rate-Other 6.2992126 7.3394495 2.7471339 1.250518e-02  
## pickup=05 8.0000000 3.6697248 1.0815488 3.539482e-02  
## dropoff=02 0.0000000 0.0000000 2.7687649 4.516816e-02  
## pickup=21 0.4219409 0.9174312 5.1265412 2.418729e-02  
## dropoff=22 0.4201681 0.9174312 5.1481722 2.366511e-02  
## hcpck=3 0.3816794 0.9174312 5.6673156 1.395311e-02  
## RateCodeID=Rate-1 2.2464413 92.6605505 97.2528661 1.250518e-02  
## Trip\_distance\_range=Medium\_dist 0.8113590 7.3394495 21.3281419 7.343759e-05  
## Payment\_type=Cash 1.2815378 29.3577982 54.0125460 1.612428e-07  
## TipIsGiven=No 1.4226232 37.6146789 62.3404716 1.424189e-07  
## hcpck=2 0.0000000 0.0000000 35.3450141 1.114755e-21  
## hcpck=1 0.0000000 0.0000000 41.7477828 1.033106e-26  
## Trip\_distance\_range=Short\_dist 0.1009421 2.7522936 64.2872594 2.624922e-44  
## paidTolls=No 1.4641608 61.4678899 98.9833441 8.076067e-67  
## v.test  
## Trip\_distance\_range=Long\_dist 17.958688  
## hcpck=5 17.360065  
## paidTolls=Yes 16.728728  
## hcpck=4 11.142176  
## TipIsGiven=Yes 5.262100  
## Payment\_type=Credit card 5.175775  
## paidTolls=NA 4.420875  
## dropoff=05 2.663750  
## RateCodeID=Rate-Other 2.497558  
## pickup=05 2.103812  
## dropoff=02 -2.003085  
## pickup=21 -2.254141  
## dropoff=22 -2.262523  
## hcpck=3 -2.458468  
## RateCodeID=Rate-1 -2.497558  
## Trip\_distance\_range=Medium\_dist -3.964865  
## Payment\_type=Cash -5.239236  
## TipIsGiven=No -5.262100  
## hcpck=2 -9.565671  
## hcpck=1 -10.698615  
## Trip\_distance\_range=Short\_dist -13.962910  
## paidTolls=No -17.268832  
##   
## $`5`  
## Cla/Mod Mod/Cla Global p.value  
## Trip\_distance\_range=Short\_dist 70.794078 89.4177646 64.2872594 9.651284e-310  
## hcpck=2 83.414933 57.9260518 35.3450141 2.733939e-250  
## passenger\_groups=Single 57.503218 94.9426264 84.0363400 2.755721e-101  
## TipIsGiven=No 57.078418 69.9107522 62.3404716 2.371984e-27  
## Payment\_type=Cash 57.348819 60.8584785 54.0125460 1.744648e-21  
## paidTolls=No 51.420455 100.0000000 98.9833441 2.373794e-15  
## dropoff=14 69.863014 6.5023374 4.7371836 5.963055e-09  
## pickup=14 69.298246 6.7148321 4.9318624 8.357817e-09  
## period=Period night 56.516443 39.4390140 35.5180619 1.386893e-08  
## pickup=12 70.555556 5.3973651 3.8935756 5.154684e-08  
## dropoff=12 70.238095 5.0148746 3.6340039 2.395181e-07  
## period=Period morning 61.254613 14.1096473 11.7239888 2.614236e-07  
## period=Period valley 56.507937 30.2592435 27.2550292 2.961664e-06  
## dropoff=13 67.777778 5.1848704 3.8935756 3.191718e-06  
## dropoff=08 67.096774 4.4198895 3.3528012 3.613675e-05  
## pickup=20 61.764706 8.0322992 6.6190785 7.936518e-05  
## pickup=08 65.662651 4.6323842 3.5907419 9.795433e-05  
## pickup=15 63.274336 6.0773481 4.8886005 1.281302e-04  
## pickup=13 64.942529 4.8023799 3.7637897 1.477385e-04  
## dropoff=15 62.727273 5.8648534 4.7588146 3.094087e-04  
## dropoff=09 63.243243 4.9723757 4.0017305 5.848407e-04  
## pickup=07 66.115702 3.3999150 2.6173480 6.540118e-04  
## dropoff=07 66.981132 3.0174246 2.2928834 7.601315e-04  
## pickup=11 63.095238 4.5048874 3.6340039 1.238666e-03  
## pickup=09 62.162162 4.8873778 4.0017305 1.722289e-03  
## pickup=21 59.493671 5.9923502 5.1265412 6.539225e-03  
## dropoff=10 60.427807 4.8023799 4.0449924 7.744511e-03  
## dropoff=20 57.876712 7.1823204 6.3162449 1.370576e-02  
## improvement\_surcharge=No 61.864407 3.1024224 2.5524551 1.578392e-02  
## dropoff=22 58.403361 5.9073523 5.1481722 1.740070e-02  
## dropoff=21 57.518797 6.5023374 5.7538395 2.612277e-02  
## pickup=23 58.173077 5.1423714 4.4992429 3.182685e-02  
## MTA\_tax=No 60.504202 3.0599235 2.5740861 3.388949e-02  
## Trip\_type=Dispatch 60.714286 2.8899278 2.4226693 3.563260e-02  
## dropoff=23 57.399103 5.4398640 4.8237075 4.669475e-02  
## pickup=02 59.398496 3.3574161 2.8769197 4.691042e-02  
## Trip\_type=Street-Hail 50.653957 97.1100722 97.5773307 3.563260e-02  
## MTA\_tax=Yes 50.643872 96.9400765 97.4259139 3.388949e-02  
## improvement\_surcharge=Yes 50.610433 96.8975776 97.4475449 1.578392e-02  
## paidTolls=NA 0.000000 0.0000000 0.1514168 6.849611e-03  
## hcpck=5 0.000000 0.0000000 0.8436080 7.589373e-13  
## paidTolls=Yes 0.000000 0.0000000 0.8652390 3.693694e-13  
## dropoff=16 28.825623 3.4424139 6.0783041 1.123668e-14  
## passenger\_groups=Couple 28.862974 4.2073948 7.4194246 9.285954e-18  
## Payment\_type=Credit card 43.129771 38.4190395 45.3385248 5.816687e-22  
## pickup=16 22.887324 2.7624309 6.1431971 2.305954e-23  
## dropoff=19 24.115756 3.1874203 6.7272334 2.013643e-23  
## dropoff=17 20.553360 2.2099448 5.4726368 2.042389e-24  
## pickup=19 23.376623 3.0599235 6.6623405 1.809388e-24  
## pickup=17 20.318725 2.1674458 5.4293749 1.333510e-24  
## TipIsGiven=Yes 40.666284 30.0892478 37.6595284 2.371984e-27  
## pickup=18 21.974522 2.9324267 6.7921263 1.574639e-27  
## dropoff=18 20.257235 2.6774331 6.7272334 1.293445e-30  
## period=Period afternoon 32.315522 16.1920952 25.5029202 3.634586e-50  
## hcpck=3 0.000000 0.0000000 5.6673156 3.426844e-85  
## Trip\_distance\_range=Medium\_dist 23.326572 9.7747556 21.3281419 3.883089e-88  
## passenger\_groups=Group 5.063291 0.8499788 8.5442353 8.681102e-96  
## Trip\_distance\_range=Long\_dist 2.857143 0.8074798 14.3845987 6.703159e-192  
## hcpck=4 0.000000 0.0000000 16.3962795 1.365785e-268  
## v.test  
## Trip\_distance\_range=Short\_dist 37.621276  
## hcpck=2 33.790344  
## passenger\_groups=Single 21.366218  
## TipIsGiven=No 10.834134  
## Payment\_type=Cash 9.519231  
## paidTolls=No 7.920069  
## dropoff=14 5.817791  
## pickup=14 5.761078  
## period=Period night 5.674999  
## pickup=12 5.445891  
## dropoff=12 5.165718  
## period=Period morning 5.149326  
## period=Period valley 4.673461  
## dropoff=13 4.658080  
## dropoff=08 4.130886  
## pickup=20 3.946309  
## pickup=08 3.895604  
## pickup=15 3.830025  
## pickup=13 3.794840  
## dropoff=15 3.607293  
## dropoff=09 3.438549  
## pickup=07 3.408166  
## dropoff=07 3.366918  
## pickup=11 3.229824  
## pickup=09 3.134361  
## pickup=21 2.719442  
## dropoff=10 2.663010  
## dropoff=20 2.464884  
## improvement\_surcharge=No 2.413874  
## dropoff=22 2.378130  
## dropoff=21 2.224382  
## pickup=23 2.146579  
## MTA\_tax=No 2.121384  
## Trip\_type=Dispatch 2.101095  
## dropoff=23 1.989058  
## pickup=02 1.987108  
## Trip\_type=Street-Hail -2.101095  
## MTA\_tax=Yes -2.121384  
## improvement\_surcharge=Yes -2.413874  
## paidTolls=NA -2.704069  
## hcpck=5 -7.168374  
## paidTolls=Yes -7.266336  
## dropoff=16 -7.724417  
## passenger\_groups=Couple -8.582467  
## Payment\_type=Credit card -9.632724  
## pickup=16 -9.958902  
## dropoff=19 -9.972371  
## dropoff=17 -10.197120  
## pickup=19 -10.208881  
## pickup=17 -10.238453  
## TipIsGiven=Yes -10.834134  
## pickup=18 -10.871572  
## dropoff=18 -11.501699  
## period=Period afternoon -14.893461  
## hcpck=3 -19.559460  
## Trip\_distance\_range=Medium\_dist -19.902348  
## passenger\_groups=Group -20.766588  
## Trip\_distance\_range=Long\_dist -29.549307  
## hcpck=4 -35.014246  
##   
##   
## Link between the cluster variable and the quantitative variables  
## ================================================================  
## Eta2 P-value  
## Trip\_distance 0.682333867 0.000000e+00  
## Fare\_amount 0.700072899 0.000000e+00  
## Extra 0.346854642 0.000000e+00  
## Tolls\_amount 0.347118692 0.000000e+00  
## Total\_amount 0.688303660 0.000000e+00  
## tlenkm 0.672008922 0.000000e+00  
## traveltime 0.555354040 0.000000e+00  
## Tip\_amount 0.246746337 4.109487e-282  
## espeed 0.199180783 8.408988e-221  
## Passenger\_count 0.175757629 6.029127e-192  
## hour 0.032768593 2.980266e-32  
## Dropoff\_latitude 0.013838854 3.496069e-13  
## Pickup\_latitude 0.008063685 1.491934e-07  
## Dropoff\_longitude 0.006916752 1.860293e-06  
## Pickup\_longitude 0.005886284 1.753776e-05  
##   
## Description of each cluster by quantitative variables  
## =====================================================  
## $`1`  
## v.test Mean in category Overall mean sd in category  
## Fare\_amount 20.042407 16.8096151 11.61104706 3.74747126  
## traveltime 19.432544 18.6125953 12.48732425 5.92050797  
## Trip\_distance 17.591543 4.2630905 2.72449524 1.21907679  
## tlenkm 17.577688 6.8373493 4.34905091 2.00322478  
## Total\_amount 17.328080 19.3954753 13.92640493 3.88246513  
## espeed 13.362174 23.9908565 20.33575305 9.09332230  
## Tip\_amount 9.087511 1.5456197 1.02203842 1.76395923  
## hour -2.085754 12.9530686 13.39757733 6.92604649  
## Tolls\_amount -3.004488 0.0000000 0.04769564 0.00000000  
## Pickup\_latitude -4.220712 40.7394347 40.74676502 0.05593978  
## Pickup\_longitude -4.602008 -73.9409325 -73.93496823 0.04162304  
## Dropoff\_longitude -4.902556 -73.9414715 -73.93460830 0.04690465  
## Extra -5.246122 0.2918171 0.35226044 0.32723242  
## Dropoff\_latitude -5.362187 40.7357173 40.74500568 0.05465058  
## Passenger\_count -6.078029 1.1732852 1.37107208 0.52270578  
## Overall sd p.value  
## Fare\_amount 8.25496368 2.351166e-89  
## traveltime 10.03175633 4.095543e-84  
## Trip\_distance 2.78356770 2.859844e-69  
## tlenkm 4.50528246 3.651657e-69  
## Total\_amount 10.04487145 2.888095e-67  
## espeed 8.70570362 1.005835e-40  
## Tip\_amount 1.83366715 1.013320e-19  
## hour 6.78263699 3.700093e-02  
## Tolls\_amount 0.50523041 2.660280e-03  
## Pickup\_latitude 0.05527371 2.435315e-05  
## Pickup\_longitude 0.04124656 4.184366e-06  
## Dropoff\_longitude 0.04455396 9.459752e-07  
## Extra 0.36668354 1.553337e-07  
## Dropoff\_latitude 0.05512875 8.222053e-08  
## Passenger\_count 1.03565723 1.216689e-09  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category  
## Fare\_amount 31.618439 22.8122649 11.6110471 9.22680134  
## traveltime 31.356926 25.9868999 12.4873242 14.03897959  
## Trip\_distance 31.015924 6.4295631 2.7244952 3.06837162  
## tlenkm 30.335416 10.2142348 4.3490509 5.06506125  
## Total\_amount 29.183431 26.5066872 13.9264049 9.88546826  
## Tip\_amount 18.948178 2.5131070 1.0220384 2.90146068  
## Passenger\_count 18.548676 2.1954733 1.3710721 1.88366928  
## espeed 17.970433 27.0496099 20.3357531 13.80572702  
## Extra 2.000758 0.3837449 0.3522604 0.37865254  
## hour -1.966744 12.8251029 13.3975773 7.02701046  
## Dropoff\_latitude -2.235052 40.7397179 40.7450057 0.05038104  
## Pickup\_latitude -2.283116 40.7413493 40.7467650 0.05618931  
## Overall sd p.value  
## Fare\_amount 8.25496368 2.060133e-219  
## traveltime 10.03175633 7.828132e-216  
## Trip\_distance 2.78356770 3.288235e-211  
## tlenkm 4.50528246 3.912848e-202  
## Total\_amount 10.04487145 3.147246e-187  
## Tip\_amount 1.83366715 4.571378e-80  
## Passenger\_count 1.03565723 8.358632e-77  
## espeed 8.70570362 3.321124e-72  
## Extra 0.36668354 4.541845e-02  
## hour 6.78263699 4.921271e-02  
## Dropoff\_latitude 0.05512875 2.541391e-02  
## Pickup\_latitude 0.05527371 2.242356e-02  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category  
## Extra 38.691376 0.7938389 0.35226044 0.32313622  
## Passenger\_count 17.820318 1.9454976 1.37107208 1.46017142  
## hour 12.277044 15.9893365 13.39757733 5.49158167  
## Dropoff\_longitude 2.293129 -73.9314284 -73.93460830 0.04406971  
## Tolls\_amount -3.033102 0.0000000 0.04769564 0.00000000  
## Tip\_amount -7.445308 0.5971201 1.02203842 0.94944352  
## traveltime -12.103859 8.7080964 12.48732425 4.76347424  
## Total\_amount -12.119349 10.1373934 13.92640493 4.45694924  
## espeed -13.436913 16.6948791 20.33575305 5.45449827  
## tlenkm -14.668192 2.2922093 4.34905091 1.23050317  
## Fare\_amount -14.695506 7.8353081 11.61104706 2.95164582  
## Trip\_distance -14.966258 1.4278617 2.72449524 0.76358087  
## Overall sd p.value  
## Extra 0.36668354 0.000000e+00  
## Passenger\_count 1.03565723 4.915602e-71  
## hour 6.78263699 1.203142e-34  
## Dropoff\_longitude 0.04455396 2.184058e-02  
## Tolls\_amount 0.50523041 2.420542e-03  
## Tip\_amount 1.83366715 9.671838e-14  
## traveltime 10.03175633 1.007608e-33  
## Total\_amount 10.04487145 8.341900e-34  
## espeed 8.70570362 3.674477e-41  
## tlenkm 4.50528246 1.030547e-48  
## Fare\_amount 8.25496368 6.888191e-49  
## Trip\_distance 2.78356770 1.219954e-50  
##   
## $`4`  
## v.test Mean in category Overall mean sd in category  
## Tolls\_amount 40.05093 1.963074 0.04769564 2.63278950  
## Total\_amount 39.12185 51.124128 13.92640493 18.90835873  
## tlenkm 37.16394 20.197849 4.34905091 9.64419649  
## Trip\_distance 37.12125 12.505354 2.72449524 5.86941865  
## Fare\_amount 35.87332 39.642089 11.61104706 12.56020461  
## traveltime 27.17098 38.288226 12.48732425 14.95322699  
## Tip\_amount 22.93020 5.002018 1.02203842 4.90894443  
## espeed 15.01648 32.710163 20.33575305 13.86530272  
## Pickup\_latitude -2.51323 40.733616 40.74676502 0.06075561  
## Dropoff\_latitude -4.24057 40.722877 40.74500568 0.06697507  
## Overall sd p.value  
## Tolls\_amount 0.50523041 0.000000e+00  
## Total\_amount 10.04487145 0.000000e+00  
## tlenkm 4.50528246 2.610729e-302  
## Trip\_distance 2.78356770 1.275899e-301  
## Fare\_amount 8.25496368 7.964808e-282  
## traveltime 10.03175633 1.430929e-162  
## Tip\_amount 1.83366715 2.322683e-116  
## espeed 8.70570362 5.727137e-51  
## Pickup\_latitude 0.05527371 1.196314e-02  
## Dropoff\_latitude 0.05512875 2.229528e-05  
##   
## $`5`  
## v.test Mean in category Overall mean sd in category  
## Dropoff\_latitude 5.958416 40.7497513 40.74500568 0.05498413  
## Pickup\_latitude 4.803646 40.7506010 40.74676502 0.05450429  
## Dropoff\_longitude 3.210705 -73.9325416 -73.93460830 0.04115731  
## Pickup\_longitude 2.570962 -73.9334362 -73.93496823 0.03988268  
## hour -6.221468 12.7879303 13.39757733 6.90562873  
## Tolls\_amount -6.534347 0.0000000 0.04769564 0.00000000  
## espeed -15.463075 18.3909003 20.33575305 5.57665243  
## Tip\_amount -19.811493 0.4972011 1.02203842 0.83589332  
## Passenger\_count -20.838846 1.0592717 1.37107208 0.26946914  
## Extra -26.784004 0.2103697 0.35226044 0.24683890  
## tlenkm -32.057736 2.2624381 4.34905091 1.22971408  
## Trip\_distance -32.242607 1.4278567 2.72449524 0.74929076  
## traveltime -33.057788 7.6961963 12.48732425 4.04125063  
## Total\_amount -33.723082 9.0324649 13.92640493 3.54907115  
## Fare\_amount -34.325210 7.5173531 11.61104706 2.67938944  
## Overall sd p.value  
## Dropoff\_latitude 0.05512875 2.546951e-09  
## Pickup\_latitude 0.05527371 1.558026e-06  
## Dropoff\_longitude 0.04455396 1.324096e-03  
## Pickup\_longitude 0.04124656 1.014166e-02  
## hour 6.78263699 4.925237e-10  
## Tolls\_amount 0.50523041 6.388760e-11  
## espeed 8.70570362 6.158740e-54  
## Tip\_amount 1.83366715 2.369546e-87  
## Passenger\_count 1.03565723 1.924230e-96  
## Extra 0.36668354 4.963017e-158  
## tlenkm 4.50528246 1.712776e-225  
## Trip\_distance 2.78356770 4.466041e-228  
## traveltime 10.03175633 1.202241e-239  
## Total\_amount 10.04487145 2.653016e-249  
## Fare\_amount 8.25496368 3.301578e-258

We proceed to explain the data obtained.

### The description of the clusters by the variables

We start wit the description of the categorical variables that characterize the clusters, so in this output we do not have dimensions because it is the total association. We can see the intensity of the variables, in our case the variables that affect more to the clustering are **Trip\_distance\_range**, **paidTolls** and **hcpck** because are the one with the smallest p.value.

Next, we want to see for each cluster which are the categories that characterize them.

* Cluster 1
  + The first thing we can notice is that almost observation in the cluster is of the kind **paidTolls=No** (99.88%), we can also see that 87.61% of the observations in the cluster are **passenger\_groups=Single** and we have the 18.74% of the observations of this kind from the sample present in this cluster. We can see that 70.88% of the observations in the cluster are **Trip\_distance\_range=Medium\_dist** and we have the 59.74% of the observations of this kind from sample present in this cluster. We can also notice that 76.05% if the observations in the cluster are **VendorID=f.Vendor-VeriFone**. We can see that the cluster 4 from the hierarchical clustering (**hcpck=4**) is present in this cluster, we observe that 38.87% of the observations in the cluster are from that cluster 4 and we have the 42.61% of the observations from the sample present in this cluster.
* Cluster 2
  + We can see that 95.88% of the observations in the cluster are **improvement\_surcharge=Yes** and **Trip\_type=Street-Hail**. We can also see that 95.27% of the observations in the cluster are **MTA\_tax=Yes**, 94.86% of the observations in the cluster are **RateCodeID=Rate-1**. We can also see that we have the 70.37% of the observations in the cluster are **Trip\_distance\_range=Long\_dist** and we have 51.43% of the observations of this kind from the sample present in this cluster. We can see that the clusters 3 and 4 from the hierarchical clustering (**hacpck=3**, **hcpck=4**) are present in the cluster. We observe that 22.84% and 75.31% of the observations in the cluster are from those clusters respectively, and we have the 42.37% and 48.28% of the observations from the sample present in this cluster.
* Cluster 3
  + The first thing we can notice is that all observations in the cluster are **paidTolls=No**. Then, we see that we the 99.76% of the observations in the cluster are **RateCodeID=Rate-1**, **MTA\_tax=Yes**, **improvement\_surcharge=Yes** and **Trip\_type=Street-Hail**. We can also see that the majority of the observations in the cluster (89.22%) are **Trip\_distance\_range=Short\_dist** and we have 25.37% of the observations of this kind from the sample in this cluster. We can see that we have 54.34% of the observations of **dropoff=18**, 53.50% of **pickup=18**, 52.19% of **pickup=17**, 50.16% of **dropoff=19** and 50.13% of **passenger\_groups=Group** kinds from the sample in this cluster. We can notice that 54.20% of the observations of **hcpck=3** (cluster 3 from hierarchical clustering) and 35.96% observations of **hcpck=1** (cluster 1 from hierarchical clustering) kinds from the sample are present in this cluster.
* Cluster 4
  + The first thing we can notice is that the 100% of the observations from the sample that represent the cluster 5 from hierarchical clustering (**hcpck=5**) are present in this cluster, we can also see that the 95% of the observations from the sample that are of the kind **paidTolls=yes** are present in this cluster. We can see that 89.91% of the observations in the cluster are **Trip\_distance\_range=Long\_dist** and we have 14,74% of the observations of this kind from the sample present in this cluster. We can also notice that 69.72% of the observations in the cluster are **Payment\_type=Credit card**, 92.25% of the observations in the cluster are **RateCodeID=Rate-1**, 63.30% of the observations in the cluster are from the cluster 4 from the hierarchical clustering (**hcpck=4**), 62.39% of the observations in the cluster left some tip (**TipIsGiven=Yes**).
* Cluster 5
  + The first thing we can notice is that every observation in the cluster had not paid any toll (**paidTolls=No**) and we have 51.42% of the observations of this kind from the sample are present in this cluster. We have the 97.11% of the observations in the cluster are **Trip\_type=Street-Hail**, 96.94% are **MTA\_tax=Yes** and 96.90% are **improvement\_surcharge=Yes**, and we have around the 50% of the observations of these kinds from the sample present in this cluster. The majority of the observations in the cluster (94.94%) are **passenger\_groups=Single** and we have the 57.08% of the observations of this kind from the sample present in this cluster. We also see that 89.42% of the observations from the sample are **Trip\_distance\_range=Short\_dist** and we have 70.79% of the observations of this kind from the sample present in this cluster. From this cluster we can notice that is the one with biggest data representation from the sample, probably because it is a big cluster so we have a lot of data present here, that is why a lot of the categories present here are highly represented.

We now proceed to see the quantitative variables that characterizes the clusters. We can see in the output that all the variables that appear are slightly over represented in the clusters. We can notice that the greatest represented is the **Fare\_amount** with 0.70 units over the global mean, **Total\_amount** with 0.69 units over the mean and **Trip\_distance** with 0.68 units over the mean. The other variables are not remarkably over the mean.

We want to know now which variables are associated with the quantitative variables.

* Cluster 1
  + We can see that almost every variable is over the overall mean. We can see that **Total\_amount** and **traveltime** are around 6 units over the overall mean. **Fare\_amount** is around 5 units over the overall mean, **espeed** is around 3 units over the overall mean and **Trip\_distance** and **tlenkm** are around 2 units over the overall mean.
* Cluster 2
  + We can see almost every variable is over the overall mean. We can see that **Total\_amount** and **traveltime** are around 13 units over the overall mean, **Fare\_amount** is around 11 units over the overall mean, **espeed** is around 7 units over the overall mean, **tlenkm** is around 6 units over the overall mean and **Trip\_distance** is around 4 units over the overall mean. **Tip\_amount**, **Passenger\_count** and **hour** are around 1 unis under the overall mean.
* Cluster 3
  + We can see that **hour** is around 2 units over the overall mean and **Passenger\_count** is around 0.6 units over the overall mean, the rest of the variables in the cluster are under the mean. **traveltime**, **Fare\_amount** and **espeed** are around 4 units under the overall mean. **Total\_amount** is around 3 units under the overall mean, **tlenkm** is around 2 units under the overall mean and **Trip\_distance** is around 1 unit under the overall mean.
* Cluster 4
  + We can see that every variable except **Pickup\_latitude** and **Dropoff\_latitude** are over the mean. We can see that **Total\_amount** is around 38 units over the overall mean, **Fare\_amount** is around 28 units over the overall mean, **traveltime** is around 26 units over the overall mean, **tlenkm** is around 16 units over the overall mean, **espeed** is around 12 units over the overall mean, **Trip\_distance** is around 10 units over the overall mean and **Tip\_amount** is around 4 units over the overall mean.
* Cluster 5
  + We can see that almost every variable is under the overall mean. **traveltime** is around 5 units under the overall mean, **Fare\_amount** and **Total\_amount** are around 4 units under the overall mean, **tlenkm** and **espeed** are around 2 units under the overall mean, **hour** and **Trip\_distance** are around 1 unit under the overall mean.

### Comparison of clusters (confusion table)

We want to compare the hierarchical clustering, previously done, and the k-means clustering, so proceed to do the following.

table(df$hcpck,df$claKM)

##   
## 1 2 3 4 5  
## 1 239 7 694 0 990  
## 2 261 2 8 0 1363  
## 3 8 111 142 1 0  
## 4 323 366 0 69 0  
## 5 0 0 0 39 0

# we must do a relabel  
df$hcpck<-factor(df$hcpck,labels=c("kHP-1","kHP-2","kHP-3","kHP-4","kHP-5"))  
df$claKM<-factor(df$claKM,levels=c(3,5,2,1,4),labels=c("kKM-3","kKM-5","kKM-2","kKM-1","kKM-4"))  
tt<-table(df$hcpck,df$claKM); tt

##   
## kKM-3 kKM-5 kKM-2 kKM-1 kKM-4  
## kHP-1 694 990 7 239 0  
## kHP-2 8 1363 2 261 0  
## kHP-3 142 0 111 8 1  
## kHP-4 0 0 366 323 69  
## kHP-5 0 0 0 0 39

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

## [1] 54.72637

We have a concordance of the 54.73% so we can say that they are different, if we had a greater concordance, this would mean that they would be more similar.

# CA analysis

## Are there any row categories that can be combined/avoided to explain the discretization of the numeric target.

### CA analysis for your data should contain your factor version of the numeric target (previous) in K= 7 (maximum 10) levels and 2 factors.

The first thing we need to do is factor our numeric target variable, Total\_amount, and name it f.cost. We are going to set 6 different categories.

df$f.cost[df$Total\_amount<=8] = "[0,8]"  
df$f.cost[(df$Total\_amount>8) & (df$Total\_amount<=11)] = "(8,11]"  
df$f.cost[(df$Total\_amount>11) & (df$Total\_amount<=18)] = "(11,18]"  
df$f.cost[(df$Total\_amount>18) & (df$Total\_amount<= 30)] = "(18,30]"  
df$f.cost[(df$Total\_amount>30) & (df$Total\_amount<= 50)] = "(30,50]"  
df$f.cost[df$Total\_amount>50] = "(50,129)"  
df$f.cost<-factor(df$f.cost)  
table(df$f.cost)

##   
## (11,18] (18,30] (30,50] (50,129) (8,11] [0,8]   
## 1188 724 221 63 1151 1276

Once we have this factor, proceed to create a variable that associates the cost with the passenger groups, and we we a contingency table with 5 rows, one per kind of cost and 3 columns, one per each kind of group.

tt<-table(df[,c("f.cost","passenger\_groups")]);tt

## passenger\_groups  
## f.cost Couple Group Single  
## (11,18] 77 89 1022  
## (18,30] 58 72 594  
## (30,50] 20 20 181  
## (50,129) 5 7 51  
## (8,11] 81 104 966  
## [0,8] 102 103 1071

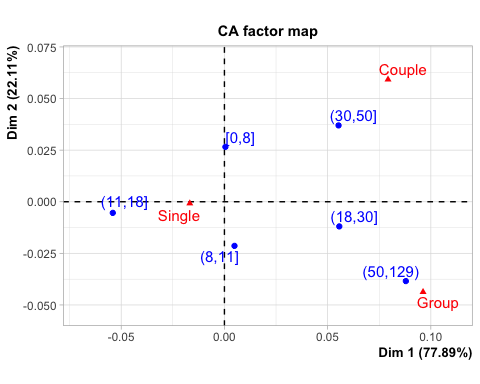
chisq.test(tt, simulate.p.value = TRUE) #to see if the rows and columns are independents. H0: Rows and columns are independent

##   
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##   
## data: tt  
## X-squared = 8.8677, df = NA, p-value = 0.5212

We get a p-value greater than 0.05 so we can assume the H0. ( 0.5217 < 0.05 = FALSE).

We are now going to take a look to the simple correspondences.

res.ca <- CA(tt)



Those observations far away from the gravity center will mean that represent less observations on the sample. If rows and columns are nearby, this will mean that there is a correspondence between them, which means that they occur simultaneously in the sample.

summary(res.ca)

##   
## Call:  
## CA(X = tt)   
##   
## The chi square of independence between the two variables is equal to 8.867721 (p-value = 0.5447017 ).  
##   
## Eigenvalues  
## Dim.1 Dim.2  
## Variance 0.001 0.000  
## % of var. 77.890 22.110  
## Cumulative % of var. 77.890 100.000  
##   
## Rows  
## Iner\*1000 Dim.1 ctr cos2 Dim.2 ctr cos2   
## (11,18] | 0.759 | -0.054 50.310 0.990 | -0.005 1.763 0.010 |  
## (18,30] | 0.507 | 0.056 32.461 0.956 | -0.012 5.273 0.044 |  
## (30,50] | 0.212 | 0.055 9.782 0.691 | 0.037 15.413 0.309 |  
## (50,129) | 0.125 | 0.088 7.047 0.839 | -0.038 4.746 0.161 |  
## (8,11] | 0.120 | 0.005 0.396 0.049 | -0.021 26.828 0.951 |  
## [0,8] | 0.195 | 0.000 0.004 0.000 | 0.027 45.976 1.000 |  
##   
## Columns  
## Iner\*1000 Dim.1 ctr cos2 Dim.2 ctr cos2   
## Couple | 0.726 | 0.079 31.197 0.642 | 0.059 61.383 0.358 |  
## Group | 0.955 | 0.096 52.961 0.829 | -0.044 38.494 0.171 |  
## Single | 0.237 | -0.017 15.841 0.998 | -0.001 0.122 0.002 |

We conclude that we can not reject the H0 for these pair of factors, and now we are going to see if we can see if there is independence between the cost and the travel time, so the first thing we are going to do is factor the travel time.

df$f.tt[df$traveltime<=5] = "[0,5]"  
df$f.tt[(df$traveltime>5) & (df$traveltime<=10)] = "(5,10]"  
df$f.tt[(df$traveltime>10) & (df$traveltime<=15)] = "(10,15]"  
df$f.tt[(df$traveltime>15) & (df$traveltime<= 20)] = "(15,20]"  
df$f.tt[(df$traveltime>20) & (df$traveltime<= 50)] = "(20,50]"  
df$f.tt<-factor(df$f.tt)  
table(df$f.tt)

##   
## (10,15] (15,20] (20,50] (5,10] [0,5]   
## 913 549 694 1511 894

Once we have this factor, proceed to create a variable that associates the cost with the traveltime.

tt<-table(df[,c("f.cost","f.tt")]);tt

## f.tt  
## f.cost (10,15] (15,20] (20,50] (5,10] [0,5]  
## (11,18] 613 314 88 156 8  
## (18,30] 106 205 388 3 15  
## (30,50] 1 23 175 2 4  
## (50,129) 1 1 35 0 7  
## (8,11] 189 3 4 864 85  
## [0,8] 3 3 4 486 775

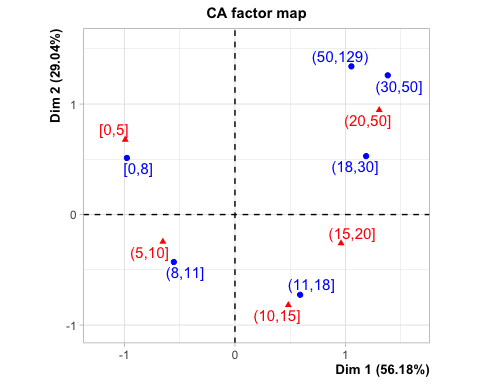
chisq.test(tt) #to see if the rows and columns are independents. H0: Rows and columns are independent

##   
## Pearson's Chi-squared test  
##   
## data: tt  
## X-squared = 6099.3, df = 20, p-value < 2.2e-16

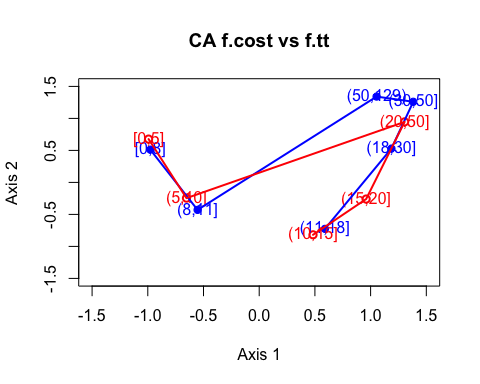
We get a p-value smaller than 0.05 so we can reject the H0. ((< 2.2e-16) < 0.05). So there is dependence between the traveltime and the cost, as we suspected.

We are now going to take a look to the simple correspondences.

res.ca <- CA(tt)



plot(res.ca$row$coord[,1],res.ca$row$coord[,2],pch=19,col="blue",xlim=c(-1.5,1.5),ylim=c(-1.5,1.5),xlab="Axis 1",ylab="Axis 2", main="CA f.cost vs f.tt")  
points(res.ca$col$coord[,1],res.ca$col$coord[,2],lwd=2,col="red")  
text(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="blue",labels=levels(df$f.cost))  
text(res.ca$col$coord[,1],res.ca$col$coord[,2],lwd=2,col="red",labels=levels(df$f.tt))  
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="blue")  
lines(res.ca$col$coord[,1],res.ca$col$coord[,2],lwd=2,col="red")



We can see in the plot, clearly that there are some categories that occur simultaneously in the sample, for instant the trips up to 5 minutes with the cost up to 8, the trips between 5-10 minutes and the costs between 8-11, the same happen with the trips between 10-15 minutes and the costs between 11-18. There is a clear relation between the f.cost and f.tt categories, even though we can not see a Guttman’s effect from manual the relation is there.

summary(res.ca)

##   
## Call:  
## CA(X = tt)   
##   
## The chi square of independence between the two variables is equal to 6099.333 (p-value = 0 ).  
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4  
## Variance 0.751 0.388 0.189 0.009  
## % of var. 56.176 29.038 14.129 0.656  
## Cumulative % of var. 56.176 85.215 99.344 100.000  
##   
## Rows  
## Iner\*1000 Dim.1 ctr cos2 Dim.2 ctr cos2   
## (11,18] | 266.105 | 0.590 11.967 0.338 | -0.726 35.079 0.512 |  
## (18,30] | 269.624 | 1.187 29.477 0.821 | 0.529 11.324 0.163 |  
## (30,50] | 175.119 | 1.383 11.441 0.491 | 1.260 18.373 0.407 |  
## (50,129) | 31.782 | 1.054 1.425 0.337 | 1.341 4.467 0.546 |  
## (8,11] | 221.698 | -0.553 10.223 0.346 | -0.429 11.924 0.209 |  
## [0,8] | 372.951 | -0.978 35.466 0.714 | 0.512 18.833 0.196 |  
## Dim.3 ctr cos2   
## (11,18] 0.391 20.884 0.148 |  
## (18,30] -0.063 0.333 0.002 |  
## (30,50] -0.582 8.062 0.087 |  
## (50,129) -0.419 0.895 0.053 |  
## (8,11] -0.627 52.158 0.445 |  
## [0,8] 0.346 17.668 0.090 |  
##   
## Columns  
## Iner\*1000 Dim.1 ctr cos2 Dim.2 ctr cos2   
## (10,15] | 200.286 | 0.483 6.218 0.233 | -0.819 34.577 0.670 |  
## (15,20] | 143.488 | 0.960 14.763 0.773 | -0.260 2.095 0.057 |  
## (20,50] | 415.261 | 1.305 34.509 0.624 | 0.946 35.059 0.328 |  
## (5,10] | 236.860 | -0.653 18.786 0.596 | -0.246 5.145 0.084 |  
## [0,5] | 341.385 | -0.993 25.724 0.566 | 0.677 23.123 0.263 |  
## Dim.3 ctr cos2   
## (10,15] 0.288 8.805 0.083 |  
## (15,20] 0.398 10.107 0.133 |  
## (20,50] -0.357 10.289 0.047 |  
## (5,10] -0.477 39.954 0.319 |  
## [0,5] 0.545 30.844 0.171 |

The first thing we can see from the summary is that we have a chi square statistic of 6099.333, great enough to reject the H0, which means the intensity of the relation is high. If we take a look at the variances from the different dimensions, we can see that all together sum more than 1.

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

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

## [1] 0.3343199

Following the kaiser kriteria and the value got in the output, we should retain dimensions with a variance greater than 0.3343199. In this case, the first dimension fulfills this because its variance is 0.751, but it is not enough to work with data so, we would choose 2 o 3 dimensions for this case.

# MCA analysis

The Multiple correspondence analysis (MCA) is an extension of the simple correspondence analysis for summarizing and visualizing a data table containing more than two categorical variables.

MCA is generally used to analyse a data set from survey. The goal is to identify:

* A group of individuals with similar profile in their answers to the questions
* The associations between variable categories

First, we load the libraries we’ll use:

library(FactoMineR)  
library(factoextra)

Now, we can start computing the MCA for our categorical variables:

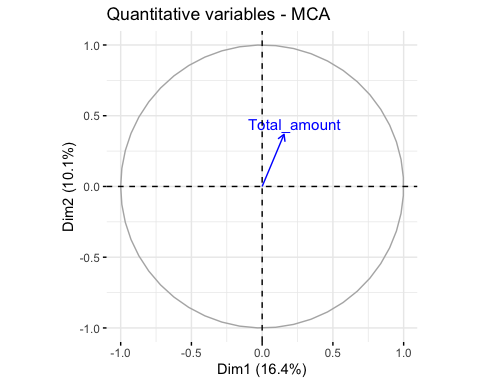
names(df[,c(1:2,15:17,19,25,27:28,31)])

## [1] "VendorID" "RateCodeID" "Total\_amount"   
## [4] "Payment\_type" "Trip\_type" "period"   
## [7] "Trip\_distance\_range" "TipIsGiven" "passenger\_groups"   
## [10] "f.cost"

res.mca <- MCA(  
 df[,c(1:2,15:17,19,25,27:28,31)],   
 quanti.sup=c(3),   
 quali.sup=c(8,10),  
 graph=FALSE  
)

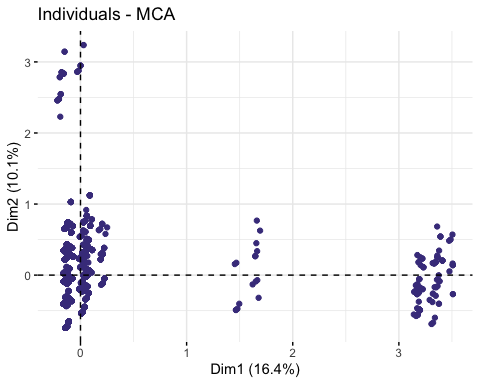
Let’s look at the supplementary quantitative variable Total\_amount. We can see that it is closer to the Dim2 than to the Dim1.

fviz\_mca\_var(res.mca, choice="quanti.sup", repel=TRUE, ggtheme=theme\_minimal())



Cloud of individuals:

fviz\_mca\_ind(  
 res.mca,  
 geom=c("point"),  
 col.ind="darkslateblue"  
)



## Eigenvalues and dominant axes analysis

**How many axes we have to consider for next Hierarchical Classification stage?**

We consider, according to the generalized Kaiser theorem, all those dimensions such that their eigenvalue is greater than the mean. We see that the average gives us 0.1428571. Therefore, we will take up to dimension 6, which represents the 62.07% of the sample.

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

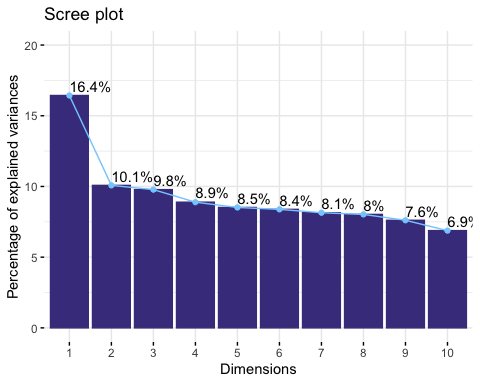
## [1] 0.1428571

head(get\_eigenvalue(res.mca), 10)

## eigenvalue variance.percent cumulative.variance.percent  
## Dim.1 0.2817102 16.433095 16.43310  
## Dim.2 0.1727341 10.076157 26.50925  
## Dim.3 0.1676074 9.777097 36.28635  
## Dim.4 0.1523716 8.888343 45.17469  
## Dim.5 0.1459733 8.515108 53.68980  
## Dim.6 0.1436861 8.381688 62.07149  
## Dim.7 0.1396003 8.143350 70.21484  
## Dim.8 0.1375543 8.024001 78.23884  
## Dim.9 0.1304320 7.608536 85.84738  
## Dim.10 0.1179063 6.877867 92.72524

We can also visualize the percentages of inertia explained by each MCA dimensions:

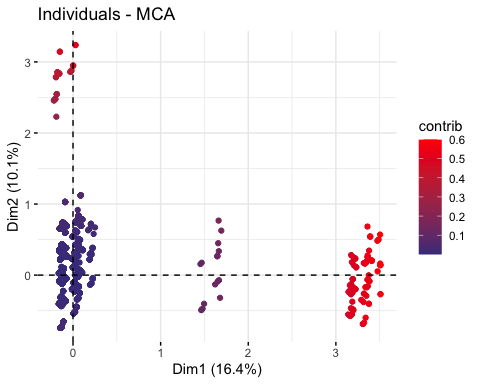
fviz\_screeplot(  
 res.mca,   
 addlabels=TRUE,   
 ylim=c(0,20),   
 barfill="darkslateblue",   
 barcolor="darkslateblue",  
 linecolor="skyblue1"  
)



## Individuals point of view

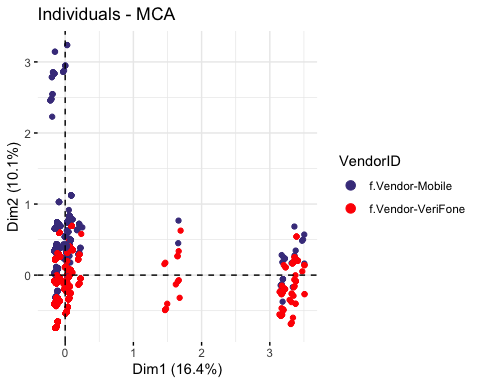
Are they any individuals “too contributive”?

fviz\_mca\_ind(  
 res.mca,   
 geom=c("point"),  
 col.ind="contrib",   
 gradient.cols=c("darkslateblue", "red")  
)

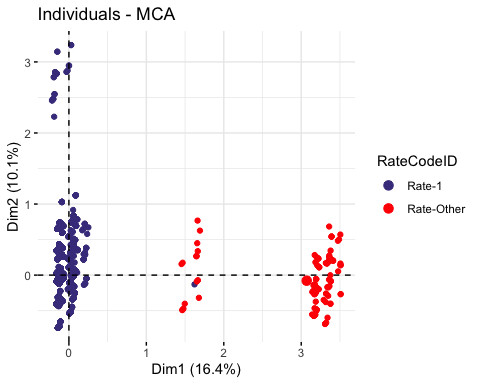


Are there any groups?

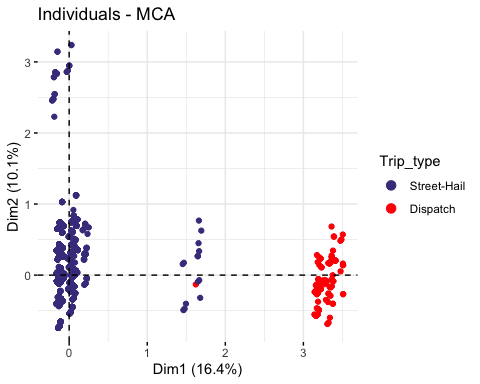
fviz\_mca\_ind(res.mca, label="none", habillage="VendorID", palette=c("darkslateblue", "red"))



fviz\_mca\_ind(res.mca, label="none", habillage="RateCodeID", palette=c("darkslateblue", "red"))



fviz\_mca\_ind(res.mca, label="none", habillage="Trip\_type", palette=c("darkslateblue", "red"))



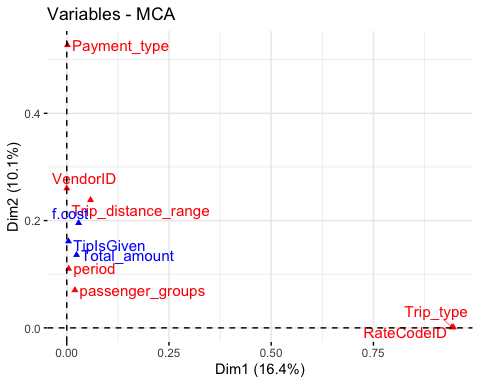
We can see that individuals are more grouped according to some variables than others. For example, the f.VendorID-Mobile is along the entire dimension 1 but also in the center of gravity. In contrast, the Rate-Other is only in the first dimension and does not touch the second at all.

## Interpreting map of categories: average profile versus extreme profiles (rare categories)

Before looking at the categories, let’s look at its variables:

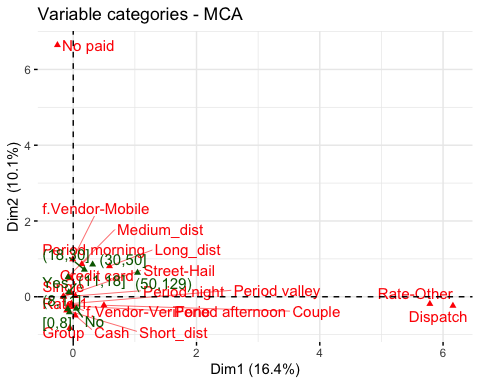
As we can see in the plot “Variables representation”, the correlation between the Payment\_type factor taking into account the eta2 and the second factorial axis is a value greater than 0.5. On the other hand, we can see that something similar happens with the Trip\_type factor and RateCodeID in dimension 1.

fviz\_mca\_var(res.mca, choice="mca.cor", repel=TRUE)



Now, let’s analyze the categories.

fviz\_mca\_var(res.mca, repel=TRUE)



As we can see, the “No paid” category (“Payment\_type” variable) is the one farthest from the center of the plot (in dimension 2). The farther from the center of gravity, the more rarely this feature value appears in the sample represented by the dimension. In addition, we see that in dimension 1 we also have two extremes, the “Rate-Other” category (“RateCodeID” variable) and the “Dispatch” category (“Trip\_type” variable). As we have said, this means that these categories are rarely represented in this dimension.

Regardering the center of mass, we can say that we find the categories most represented by the dimensions.

To give an example, let’s suppose we look at the first dimension. An observation that we could find with high probability would be the following:

* RateCodeID = Rate-1
* Trip\_type = Street-Hail

On the other hand, an observation that we could rarely find there would be…

* RateCodeID = Rate-Other
* Trip\_type = Street-Dispatch

We would follow the same logic for dimension 2 considering the Payment\_type variable.

## Interpreting the axes association to factor map

res.desc <- dimdesc(res.mca, axes = c(1,2))

### Description of dimension 1

res.desc[[1]]

## $quanti  
## correlation p.value  
## Total\_amount 0.1547222 3.65431e-26  
##   
## $quali  
## R2 p.value  
## RateCodeID 0.945537593 0.000000e+00  
## Trip\_type 0.942072409 0.000000e+00  
## Trip\_distance\_range 0.058205469 6.898258e-61  
## f.cost 0.028972784 1.405425e-27  
## passenger\_groups 0.019901125 6.814707e-21  
## TipIsGiven 0.004240936 9.364240e-06  
## period 0.004628593 8.564400e-05  
## Payment\_type 0.001608040 2.429314e-02  
##   
## $category  
## Estimate p.value  
## Trip\_type=Dispatch 1.67529735 0.000000e+00  
## RateCodeID=Rate-Other 1.57877258 0.000000e+00  
## Trip\_distance\_range=Long\_dist 0.24028354 4.637674e-62  
## passenger\_groups=Couple 0.19279452 5.856637e-22  
## f.cost=(50,129) 0.43727781 5.906344e-17  
## f.cost=(30,50] 0.05054341 1.602061e-06  
## TipIsGiven=No 0.03566808 9.364240e-06  
## period=Period morning 0.06536718 5.700992e-04  
## Payment\_type=Cash 0.06349408 1.434472e-02  
## Payment\_type=Credit card 0.02679756 2.616189e-02  
## f.cost=[0,8] -0.14970203 8.537458e-03  
## Trip\_distance\_range=Medium\_dist -0.11215628 6.996595e-03  
## f.cost=(11,18] -0.15476359 3.894367e-03  
## period=Period afternoon -0.05178612 1.144725e-03  
## f.cost=(8,11] -0.16266832 6.499724e-04  
## TipIsGiven=Yes -0.03566808 9.364240e-06  
## f.cost=(18,30] -0.02068728 1.202545e-07  
## passenger\_groups=Single -0.09190735 2.059738e-09  
## Trip\_distance\_range=Short\_dist -0.12812726 2.015102e-22  
## Trip\_type=Street-Hail -1.67529735 0.000000e+00  
## RateCodeID=Rate-1 -1.57877258 0.000000e+00  
##   
## attr(,"class")  
## [1] "condes" "list "

There is no info for the **quantitative** variables here.

In the first dimension we see that for the **qualitative** variables the most positively related, from more to less, are:

* RateCodeID (0.95)
* Trip\_type (0.94)

If we look at the **categories**, we see that the most related are,

* for Trip\_type:
  + Dispatch (1.68)
  + Long\_dist (0.24)
* and for RateCodeID:
  + Rate-Other (1.58)

### Description of dimension 2

res.desc[[2]]

## $quanti  
## correlation p.value  
## Total\_amount 0.3688482 5.757656e-149  
##   
## $quali  
## R2 p.value  
## Payment\_type 0.5272544813 0.000000e+00  
## VendorID 0.2602830667 6.879178e-305  
## Trip\_distance\_range 0.2384878813 4.714678e-274  
## f.cost 0.1956079989 4.287815e-215  
## TipIsGiven 0.1613968295 6.956769e-179  
## period 0.1103532182 9.429917e-117  
## passenger\_groups 0.0703669803 6.304633e-74  
## Trip\_type 0.0013941924 1.111798e-02  
## RateCodeID 0.0009990214 3.163284e-02  
##   
## $category  
## Estimate p.value  
## Payment\_type=No paid 1.84096016 0.000000e+00  
## VendorID=f.Vendor-Mobile 0.26007767 6.879178e-305  
## TipIsGiven=Yes 0.17229953 6.956769e-179  
## Trip\_distance\_range=Long\_dist 0.19939818 5.880829e-119  
## period=Period morning 0.30980763 1.193381e-106  
## f.cost=(18,30] 0.18702736 1.831882e-102  
## Trip\_distance\_range=Medium\_dist 0.08653538 8.235254e-84  
## passenger\_groups=Single 0.17356325 4.157410e-60  
## f.cost=(30,50] 0.24385380 3.076322e-39  
## f.cost=(50,129) 0.15326671 3.834075e-07  
## passenger\_groups=Couple 0.03719691 1.495679e-05  
## Trip\_type=Street-Hail 0.05046600 1.111798e-02  
## RateCodeID=Rate-1 0.04018420 3.163284e-02  
## RateCodeID=Rate-Other -0.04018420 3.163284e-02  
## Trip\_type=Dispatch -0.05046600 1.111798e-02  
## f.cost=(11,18] -0.06069884 1.647278e-06  
## period=Period valley -0.12396133 4.566127e-14  
## period=Period afternoon -0.14612741 8.436539e-21  
## f.cost=(8,11] -0.24322507 1.869439e-36  
## passenger\_groups=Group -0.21076016 2.204053e-67  
## f.cost=[0,8] -0.28022396 5.282753e-68  
## TipIsGiven=No -0.17229953 6.956769e-179  
## Payment\_type=Credit card -0.71587782 4.558246e-227  
## Trip\_distance\_range=Short\_dist -0.28593356 2.059524e-267  
## VendorID=f.Vendor-VeriFone -0.26007767 6.879178e-305  
## Payment\_type=Cash -1.12508234 0.000000e+00  
##   
## attr(,"class")  
## [1] "condes" "list "

There is no info for the **quantitative** variables here.

For the second dimension we see that for the **qualitative** variables the most positively related, from more to less, are:

* Payment\_type (0.53)
* VendorID (0.26)

We see that they are not very large numbers, however.

If we look at the **categories**, we see that the most related are,

* for Payment\_type:
  + No paid (1.84)
* and for VendorID:
  + f.Vendor-Mobile (0.26)

## Perform a MCA taking into account also supplementary variables (use all numeric variables) quantitative and/or categorical. How supplementary variables enhance the axis interpretation?

res.mca\_all <- MCA(  
 df[,c(1:32)],   
 quanti.sup=c(3:10, 12:13, 15, 18, 20:22),   
 quali.sup=c(27,31),  
 graph=FALSE  
)

### Description of dimensions

res.desc <- dimdesc(res.mca\_all, axes = c(1,2))

#### Description of dimension 1

res.desc[[1]]

## $quanti  
## correlation p.value  
## Fare\_amount 0.34704329 5.687723e-131  
## Trip\_distance 0.31264071 2.305988e-105  
## Total\_amount 0.28704716 2.116125e-88  
## tlenkm 0.28360598 2.991362e-86  
## traveltime 0.23128431 3.455149e-57  
## espeed 0.18449624 1.122581e-36  
## Tolls\_amount 0.11567250 3.040161e-15  
## Tip\_amount 0.10081884 6.393352e-12  
## Pickup\_latitude 0.09471249 1.100053e-10  
## Dropoff\_latitude 0.08750941 2.525109e-09  
## Pickup\_longitude 0.04599144 1.760667e-03  
## Passenger\_count -0.06437422 1.184978e-05  
## hour -0.20861841 1.253392e-46  
## Extra -0.46952211 3.175111e-252  
##   
## $quali  
## R2 p.value  
## RateCodeID 0.693923341 0.000000e+00  
## MTA\_tax 0.711903229 0.000000e+00  
## improvement\_surcharge 0.698232732 0.000000e+00  
## Trip\_type 0.708486163 0.000000e+00  
## hcpck 0.297939266 0.000000e+00  
## dropoff 0.209345234 3.392119e-214  
## pickup 0.207487287 6.821630e-212  
## period 0.164815275 5.012350e-180  
## claKM 0.163714821 1.972284e-177  
## Trip\_distance\_range 0.136491381 5.970680e-148  
## f.cost 0.102309739 1.704572e-105  
## f.tt 0.076192183 6.211428e-77  
## paidTolls 0.019509924 1.713157e-20  
## passenger\_groups 0.006558016 2.507248e-07  
##   
## $category  
## Estimate p.value  
## Trip\_type=Dispatch 1.43031511 0.000000e+00  
## improvement\_surcharge=improvement\_surcharge\_No 1.38427751 0.000000e+00  
## MTA\_tax=MTA\_tax\_No 1.39203218 0.000000e+00  
## RateCodeID=Rate-Other 1.33153381 0.000000e+00  
## Trip\_distance\_range=Long\_dist 0.32675153 8.100939e-136  
## hcpck=kHP-2 0.07977521 1.681574e-104  
## period=Period morning 0.37766782 8.601718e-102  
## hcpck=kHP-4 0.20181507 3.099380e-90  
## f.tt=(20,50] 0.18168927 6.096325e-53  
## dropoff=dropoff\_09 0.47527824 1.556093e-45  
## pickup=pickup\_09 0.43741728 3.021897e-39  
## claKM=kKM-2 0.17416148 2.127247e-38  
## f.cost=(18,30] 0.04742755 7.002029e-37  
## f.cost=(30,50] 0.21181762 3.678115e-30  
## pickup=pickup\_10 0.35502449 2.166357e-28  
## dropoff=dropoff\_10 0.35598916 5.081215e-28  
## pickup=pickup\_08 0.37525538 4.215535e-27  
## f.cost=(50,129) 0.51778721 1.154869e-26  
## claKM=kKM-4 0.40726332 3.051827e-26  
## period=Period valley 0.06316429 4.676156e-24  
## dropoff=dropoff\_08 0.31036705 1.118775e-18  
## claKM=kKM-1 0.02088140 1.810760e-16  
## dropoff=dropoff\_11 0.24202770 2.191530e-15  
## hcpck=kHP-5 0.51040471 4.740775e-15  
## dropoff=dropoff\_13 0.23740406 2.794296e-14  
## paidTolls=paidTolls\_Yes 0.01649022 1.300670e-13  
## pickup=pickup\_12 0.20658375 1.248113e-11  
## pickup=pickup\_13 0.20900204 1.839034e-11  
## f.tt=f.tt.NA 0.32116637 2.544896e-10  
## paidTolls=paidTolls.NA 0.58172801 2.637481e-09  
## pickup=pickup\_11 0.18243315 3.201149e-09  
## dropoff=dropoff\_12 0.17393741 1.042928e-08  
## dropoff=dropoff\_06 0.34833432 4.281223e-07  
## pickup=pickup\_06 0.29293154 5.357562e-07  
## dropoff=dropoff\_15 0.10947712 2.502414e-06  
## pickup=pickup\_14 0.08865893 3.225767e-05  
## dropoff=dropoff\_14 0.06535148 6.420665e-04  
## pickup=pickup\_07 0.10272201 9.978763e-04  
## pickup=pickup\_05 0.18403737 1.347096e-03  
## passenger\_groups=Couple 0.09533822 1.673249e-03  
## pickup=pickup\_15 0.05360616 1.924293e-03  
## dropoff=dropoff\_05 0.11200689 2.399701e-02  
## dropoff=dropoff\_07 0.04844411 4.477600e-02  
## Trip\_distance\_range=Medium\_dist -0.09239324 3.587226e-02  
## pickup=pickup\_03 -0.17632814 8.861076e-03  
## dropoff=dropoff\_16 -0.13845127 4.312258e-03  
## pickup=pickup\_16 -0.14870023 1.210472e-03  
## dropoff=dropoff\_22 -0.16127609 9.445790e-04  
## f.tt=(15,20] -0.02276355 5.656303e-04  
## pickup=pickup\_22 -0.17078247 2.323145e-04  
## f.tt=(10,15] -0.15233505 2.086366e-04  
## dropoff=dropoff\_03 -0.23113265 1.435539e-04  
## f.cost=[0,8] -0.23016247 1.876733e-05  
## f.cost=(11,18] -0.23321044 1.639065e-05  
## pickup=pickup\_21 -0.20005018 6.903869e-06  
## dropoff=dropoff\_23 -0.21249012 2.617862e-06  
## pickup=pickup\_00 -0.21451652 1.857404e-06  
## passenger\_groups=Group -0.11005910 1.742479e-06  
## pickup=pickup\_23 -0.22469398 9.767269e-07  
## dropoff=dropoff\_00 -0.22732617 2.822646e-07  
## dropoff=dropoff\_21 -0.22321151 3.701867e-08  
## period=Period night -0.12234903 1.052033e-08  
## hcpck=kHP-3 -0.34574730 5.171016e-11  
## dropoff=dropoff\_17 -0.27675451 1.836772e-12  
## pickup=pickup\_19 -0.27361333 9.619675e-15  
## dropoff=dropoff\_19 -0.28797827 1.382374e-16  
## pickup=pickup\_17 -0.31883145 6.076516e-17  
## dropoff=dropoff\_20 -0.30303289 1.825453e-17  
## pickup=pickup\_20 -0.30264483 2.466439e-18  
## paidTolls=paidTolls\_No -0.59821823 5.109733e-20  
## pickup=pickup\_18 -0.33381152 2.133837e-23  
## dropoff=dropoff\_18 -0.33632575 1.896016e-23  
## f.cost=(8,11] -0.31365948 7.123600e-25  
## f.tt=(5,10] -0.22721770 1.228615e-33  
## Trip\_distance\_range=Short\_dist -0.23435829 4.137407e-87  
## period=Period afternoon -0.31848308 1.175534e-87  
## claKM=kKM-3 -0.49342136 5.050918e-128  
## hcpck=kHP-1 -0.44624768 2.882408e-285  
## Trip\_type=Street-Hail -1.43031511 0.000000e+00  
## improvement\_surcharge=improvement\_surcharge\_Yes -1.38427751 0.000000e+00  
## MTA\_tax=MTA\_tax\_Yes -1.39203218 0.000000e+00  
## RateCodeID=Rate-1 -1.33153381 0.000000e+00  
##   
## attr(,"class")  
## [1] "condes" "list "

In this dimension, since we have taken into account all the variables, we now have information for the **quantitative** variables. We see that, more or less, the most positively related are:

* Fare\_amount (0.35)
* Trip\_distance (0.31)
* Total\_amount (0.29)

We also see that they do not contribute much given the numbers.

However, there is a little more inverse relationship with Extra, with a -0.47.

Regarding the **qualitative** variables, the new relationship is as follows:

* RateCodeID (0.69)
* MTA\_tax (0.71)
* improvement\_surcharge (0.70)
* Trip\_type (0.71)

If we look at the **categories**, we see that the most related are,

* for Trip\_type:
  + Dispatch (1.43) -> same as before but less related
* for improvement\_surcharge:
  + improvement\_surcharge\_No (1.38)
* for MTA\_tax:
  + MTA\_tax\_No (1.39)
* for Trip\_distance\_range:
  + Long\_dist (0.24)
* and for RateCodeID:
  + Rate-Other (1.33) -> same as before but less related

#### Description of dimension 2

res.desc[[2]]

## $quanti  
## correlation p.value  
## Extra 0.59540871 0.000000e+00  
## Passenger\_count 0.18753711 7.367467e-38  
## hour 0.14546401 2.768090e-23  
## Dropoff\_longitude 0.10780500 1.991105e-13  
## espeed 0.10518904 7.497280e-13  
## Pickup\_longitude 0.08329485 1.413350e-08  
## Total\_amount 0.04423863 2.624881e-03  
## Trip\_distance 0.04404583 2.740527e-03  
## Fare\_amount 0.03440690 1.931080e-02  
## tlenkm 0.03204240 2.936007e-02  
## traveltime -0.03531017 1.635340e-02  
## Tolls\_amount -0.05868397 6.539683e-05  
## Dropoff\_latitude -0.08128077 3.127258e-08  
## Pickup\_latitude -0.08469170 8.059026e-09  
##   
## $quali  
## R2 p.value  
## period 0.7193448269 0.000000e+00  
## pickup 0.7762688275 0.000000e+00  
## dropoff 0.7624477783 0.000000e+00  
## hcpck 0.4545819701 0.000000e+00  
## MTA\_tax 0.1619886885 1.358849e-179  
## Trip\_type 0.1582247481 4.316437e-175  
## improvement\_surcharge 0.1533670876 2.604975e-169  
## RateCodeID 0.1514542007 4.820984e-167  
## claKM 0.1244134404 1.691964e-131  
## passenger\_groups 0.0437705123 1.254658e-45  
## f.cost 0.0076558568 1.198591e-06  
## Trip\_distance\_range 0.0055181933 2.809998e-06  
## paidTolls 0.0044565106 3.304810e-05  
## f.tt 0.0041361451 1.808199e-03  
## VendorID 0.0009197986 3.920678e-02  
## Payment\_type 0.0012977242 4.980251e-02  
##   
## $category  
## Estimate p.value  
## hcpck=kHP-1 0.31938183 0.000000e+00  
## period=Period night 0.40222038 3.365631e-247  
## period=Period afternoon 0.45882535 5.397000e-213  
## MTA\_tax=MTA\_tax\_No 0.61577827 1.358849e-179  
## Trip\_type=Dispatch 0.62682523 4.316437e-175  
## improvement\_surcharge=improvement\_surcharge\_No 0.60163400 2.604975e-169  
## RateCodeID=Rate-Other 0.57687316 4.820984e-167  
## claKM=kKM-3 0.28367351 8.583686e-105  
## dropoff=dropoff\_19 0.38381622 1.832601e-46  
## pickup=pickup\_19 0.38522256 2.503341e-46  
## dropoff=dropoff\_18 0.38168754 6.015986e-46  
## pickup=pickup\_18 0.37954972 6.838557e-46  
## pickup=pickup\_20 0.37329421 3.371096e-43  
## dropoff=dropoff\_20 0.37801770 3.497189e-42  
## dropoff=dropoff\_22 0.38091527 1.100903e-34  
## pickup=pickup\_22 0.36184277 2.051986e-32  
## passenger\_groups=Group 0.13528200 1.069335e-31  
## dropoff=dropoff\_21 0.32784913 6.528817e-29  
## dropoff=dropoff\_01 0.40551849 1.201710e-27  
## pickup=pickup\_01 0.41106345 2.203563e-27  
## pickup=pickup\_17 0.32837866 2.379219e-27  
## hcpck=kHP-3 0.32908692 2.832286e-27  
## pickup=pickup\_21 0.33383417 1.161176e-26  
## pickup=pickup\_00 0.33610624 2.614122e-25  
## dropoff=dropoff\_00 0.32779268 3.212179e-24  
## pickup=pickup\_02 0.40676906 3.883490e-22  
## dropoff=dropoff\_02 0.41364408 4.972724e-22  
## dropoff=dropoff\_23 0.30192512 1.126132e-20  
## pickup=pickup\_23 0.30110187 3.234219e-19  
## dropoff=dropoff\_04 0.42108454 4.954886e-19  
## pickup=pickup\_04 0.40921819 2.566232e-15  
## pickup=pickup\_03 0.35653630 2.723112e-15  
## dropoff=dropoff\_03 0.33499061 5.956436e-14  
## dropoff=dropoff\_17 0.22947689 6.600147e-14  
## passenger\_groups=Couple 0.04411718 7.518697e-13  
## claKM=kKM-2 0.10747201 4.561136e-12  
## pickup=pickup\_05 0.35086823 4.782705e-07  
## Trip\_distance\_range=Long\_dist 0.06959039 4.957575e-07  
## dropoff=dropoff\_05 0.33403293 1.329113e-06  
## f.cost=(8,11] 0.02021781 4.875813e-04  
## f.tt=[0,5] 0.03435830 1.662342e-03  
## hcpck=kHP-4 0.05473367 1.732551e-02  
## paidTolls=paidTolls.NA 0.36969056 2.729367e-02  
## VendorID=f.Vendor-VeriFone 0.01802595 3.920678e-02  
## dropoff=dropoff\_07 -0.08624471 4.263781e-02  
## VendorID=f.Vendor-Mobile -0.01802595 3.920678e-02  
## Trip\_distance\_range=Short\_dist -0.03003534 2.057557e-02  
## Payment\_type=No paid -0.13844249 1.957223e-02  
## claKM=kKM-4 -0.15080413 1.234479e-02  
## pickup=pickup\_07 -0.10307362 1.055184e-02  
## paidTolls=paidTolls\_No -0.03224391 4.893648e-03  
## f.tt=(20,50] -0.06025223 3.961329e-03  
## claKM=kKM-1 -0.08122460 3.001909e-03  
## paidTolls=paidTolls\_Yes -0.33744664 6.994691e-05  
## hcpck=kHP-5 -0.30151068 3.827631e-05  
## f.cost=[0,8] -0.08481507 7.958326e-08  
## pickup=pickup\_16 -0.19161428 6.634026e-13  
## dropoff=dropoff\_16 -0.26024731 6.381674e-22  
## dropoff=dropoff\_08 -0.43184566 5.369589e-31  
## passenger\_groups=Single -0.17939918 2.073015e-45  
## pickup=pickup\_08 -0.53102690 2.383861e-49  
## pickup=pickup\_11 -0.54835024 3.477338e-53  
## dropoff=dropoff\_12 -0.54861363 3.112894e-53  
## dropoff=dropoff\_13 -0.53735642 6.910645e-55  
## pickup=pickup\_12 -0.53762203 6.088027e-55  
## pickup=pickup\_13 -0.55875049 3.267704e-57  
## dropoff=dropoff\_09 -0.54245620 1.605053e-57  
## pickup=pickup\_09 -0.55813145 7.767141e-61  
## dropoff=dropoff\_11 -0.55595768 9.959401e-62  
## dropoff=dropoff\_10 -0.59076056 7.773922e-69  
## pickup=pickup\_10 -0.59157063 1.412063e-70  
## claKM=kKM-5 -0.15911679 1.682905e-71  
## pickup=pickup\_15 -0.54732996 3.165865e-72  
## dropoff=dropoff\_15 -0.55708943 1.053768e-72  
## pickup=pickup\_14 -0.61682332 1.161024e-92  
## dropoff=dropoff\_14 -0.63592139 2.251034e-94  
## RateCodeID=Rate-1 -0.57687316 4.820984e-167  
## improvement\_surcharge=improvement\_surcharge\_Yes -0.60163400 2.604975e-169  
## Trip\_type=Street-Hail -0.62682523 4.316437e-175  
## MTA\_tax=MTA\_tax\_Yes -0.61577827 1.358849e-179  
## period=Period morning -0.47130282 8.452319e-206  
## hcpck=kHP-2 -0.40169174 0.000000e+00  
## period=Period valley -0.38974292 0.000000e+00  
##   
## attr(,"class")  
## [1] "condes" "list "

In this dimension, since we have taken into account all the variables, we now have information for the **quantitative** variables. We see that, more or less, the most positively related are:

* Extra (0.59540871)
* Passenger\_count (0.18753711)

For the second dimension we see that for the **qualitative** variables the most positively related, from more to less, are:

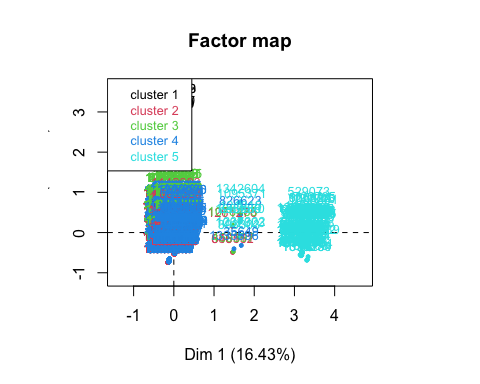
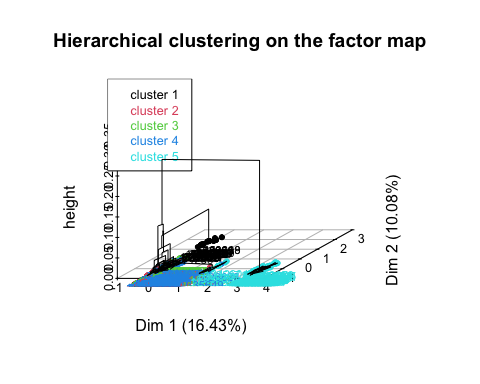
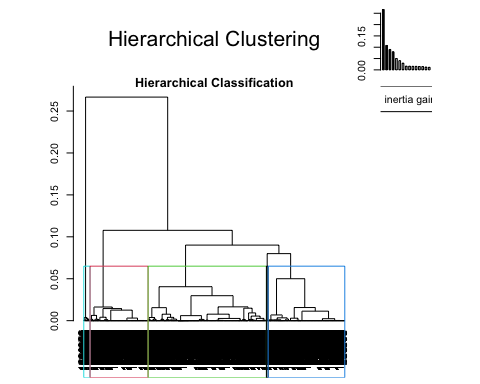
* period (0.72)
* pickup (0.78)
* dropoff (0.76)
* hcpck (0.45)
* MTA\_tax (0.16)
* …
* Payment\_type (0.0013) -> we see that it has lowed down in front of the other variables
* VendorID -> it does not even appear We see that they are not very large numbers, however.

If we look at the **categories**, we see that the most related are,

* for period:
  + Period night (0.40)
  + Period afternoon (0.46)
* …
* for Payment\_type:
  + No paid (1.84) -> now it’s inversed
* and for VendorID:
  + f.Vendor-Mobile -> it does not even appear

# Hierarchical Clustering (from MCA)

res.hcpcMCA <- HCPC(res.mca,nb.clust = 5, order = TRUE)



*Note*: If we chose the default number of cluster it would be 5, as we can guess from the inertia reduction plot, that follows the Elbow’s rule (number of black lines plus 1). In our case, after trying with bigger number of clusters, we decided that the default number of cluster was fine for our case and data.

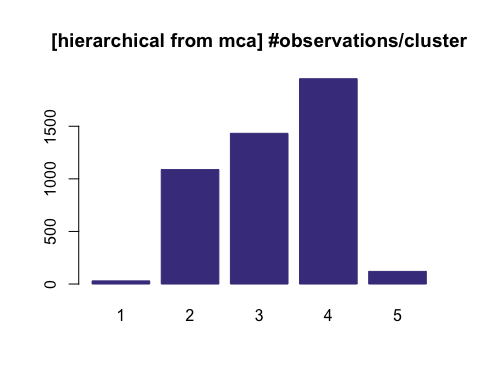
## Description of clusters

Number of observations in each cluster:

table(res.hcpcMCA$data.clust$clust)

##   
## 1 2 3 4 5   
## 30 1088 1433 1952 120

barplot(table(res.hcpcMCA$data.clust$clust), col="darkslateblue", border="darkslateblue", main="[hierarchical from mca] #observations/cluster")



## Interpret the results of the classification

### The description of the clusters by the variables

names(res.hcpcMCA$desc.var)

## [1] "test.chi2" "category" "quanti.var" "quanti" "call"

res.hcpcMCA$desc.var$test.chi2 # categorical variables which characterizes the clusters

## p.value df  
## RateCodeID 0.000000e+00 4  
## Payment\_type 0.000000e+00 8  
## Trip\_type 0.000000e+00 4  
## period 0.000000e+00 12  
## passenger\_groups 2.601045e-94 8  
## Trip\_distance\_range 6.685645e-92 8  
## f.cost 1.448630e-51 20  
## VendorID 2.325462e-27 4  
## TipIsGiven 2.455088e-11 4

We start wit the description of the categorical variables that characterize the clusters, so in this output we do not have dimensions because it is the total association. We can see the intensity of the variables, in our case the variables that affect more to the clustering are **RateCodeID**, **Payment\_type**, **Trip\_type** and **period** because are the one with the smallest p.value. The variables associated to the clusters are the ones that appear on the output.

Next, we want to see for each cluster which are the categories that characterize them. The clusters that contain more individuals are the first, the second and the fourth one. Clusters number 1 and 5 are the ones that have less individuals. We proceed to analyze them.

res.hcpcMCA$desc.var$category # description of each cluster by the categories

## $`1`  
## Cla/Mod Mod/Cla Global p.value  
## Payment\_type=No paid 100.0000000 100.00000 0.6489293 3.287724e-78  
## VendorID=f.Vendor-Mobile 3.0832477 100.00000 21.0469392 3.471103e-21  
## TipIsGiven=No 1.0409438 100.00000 62.3404716 6.580800e-07  
## period=Period morning 1.4760148 26.66667 11.7239888 2.482286e-02  
## passenger\_groups=Single 0.7464607 96.66667 84.0363400 4.121461e-02  
## TipIsGiven=Yes 0.0000000 0.00000 37.6595284 6.580800e-07  
## Payment\_type=Credit card 0.0000000 0.00000 45.3385248 1.248361e-08  
## Payment\_type=Cash 0.0000000 0.00000 54.0125460 6.774205e-11  
## VendorID=f.Vendor-VeriFone 0.0000000 0.00000 78.9530608 3.471103e-21  
## v.test  
## Payment\_type=No paid 18.721812  
## VendorID=f.Vendor-Mobile 9.447473  
## TipIsGiven=No 4.973343  
## period=Period morning 2.244148  
## passenger\_groups=Single 2.041364  
## TipIsGiven=Yes -4.973343  
## Payment\_type=Credit card -5.692987  
## Payment\_type=Cash -6.525573  
## VendorID=f.Vendor-VeriFone -9.447473  
##   
## $`2`  
## Cla/Mod Mod/Cla Global p.value  
## period=Period afternoon 88.379983 95.7720588 25.5029202 0.000000e+00  
## Trip\_distance\_range=Short\_dist 28.162853 76.9301471 64.2872594 2.073868e-24  
## Trip\_type=Street-Hail 24.118821 100.0000000 97.5773307 5.821121e-14  
## RateCodeID=Rate-1 24.132562 99.7242647 97.2528661 1.150890e-11  
## passenger\_groups=Couple 37.900875 11.9485294 7.4194246 5.792479e-10  
## f.cost=(11,18] 29.461279 32.1691176 25.6975990 3.920300e-08  
## f.cost=(8,11] 28.844483 30.5147059 24.8972529 1.397923e-06  
## VendorID=f.Vendor-Mobile 26.927030 24.0808824 21.0469392 5.477246e-03  
## VendorID=f.Vendor-VeriFone 22.630137 75.9191176 78.9530608 5.477246e-03  
## f.cost=(50,129) 9.523810 0.5514706 1.3627515 4.760384e-03  
## Payment\_type=No paid 0.000000 0.0000000 0.6489293 3.099747e-04  
## passenger\_groups=Single 22.265122 79.5036765 84.0363400 5.081032e-06  
## RateCodeID=Rate-Other 2.362205 0.2757353 2.7471339 1.150890e-11  
## f.cost=(18,30] 13.812155 9.1911765 15.6608263 1.988020e-12  
## Trip\_type=Dispatch 0.000000 0.0000000 2.4226693 5.821121e-14  
## f.cost=(30,50] 4.072398 0.8272059 4.7804456 5.272518e-16  
## period=Period morning 4.428044 2.2058824 11.7239888 1.528422e-37  
## Trip\_distance\_range=Long\_dist 1.654135 1.0110294 14.3845987 1.258712e-66  
## period=Period valley 1.746032 2.0220588 27.2550292 6.479660e-137  
## period=Period night 0.000000 0.0000000 35.5180619 1.204220e-246  
## v.test  
## period=Period afternoon Inf  
## Trip\_distance\_range=Short\_dist 10.195634  
## Trip\_type=Street-Hail 7.512044  
## RateCodeID=Rate-1 6.786246  
## passenger\_groups=Couple 6.195976  
## f.cost=(11,18] 5.494405  
## f.cost=(8,11] 4.825301  
## VendorID=f.Vendor-Mobile 2.777538  
## VendorID=f.Vendor-VeriFone -2.777538  
## f.cost=(50,129) -2.822816  
## Payment\_type=No paid -3.606818  
## passenger\_groups=Single -4.561414  
## RateCodeID=Rate-Other -6.786246  
## f.cost=(18,30] -7.035322  
## Trip\_type=Dispatch -7.512044  
## f.cost=(30,50] -8.105047  
## period=Period morning -12.805447  
## Trip\_distance\_range=Long\_dist -17.243201  
## period=Period valley -24.905542  
## period=Period night -33.541337  
##   
## $`3`  
## Cla/Mod Mod/Cla Global p.value  
## period=Period valley 77.222222 67.8995115 27.2550292 0.000000e+00  
## period=Period morning 84.870849 32.1004885 11.7239888 2.187992e-171  
## passenger\_groups=Single 36.885457 100.0000000 84.0363400 7.053895e-133  
## Trip\_type=Street-Hail 31.766792 100.0000000 97.5773307 4.847071e-19  
## RateCodeID=Rate-1 31.828292 99.8604327 97.2528661 2.899525e-18  
## f.cost=[0,8] 36.990596 32.9378925 27.6011248 7.127666e-08  
## Trip\_distance\_range=Short\_dist 33.411844 69.2951849 64.2872594 1.662139e-06  
## Payment\_type=Cash 33.520224 58.4089323 54.0125460 5.704677e-05  
## TipIsGiven=No 32.616239 65.5966504 62.3404716 2.137595e-03  
## TipIsGiven=Yes 28.317059 34.4033496 37.6595284 2.137595e-03  
## f.cost=(18,30] 26.104972 13.1891137 15.6608263 1.731548e-03  
## Payment\_type=Credit card 28.435115 41.5910677 45.3385248 5.948993e-04  
## f.cost=(30,50] 20.814480 3.2100488 4.7804456 5.532609e-04  
## f.cost=(50,129) 11.111111 0.4884857 1.3627515 2.255397e-04  
## Payment\_type=No paid 0.000000 0.0000000 0.6489293 1.404592e-05  
## Trip\_distance\_range=Long\_dist 17.894737 8.3042568 14.3845987 1.903360e-16  
## RateCodeID=Rate-Other 1.574803 0.1395673 2.7471339 2.899525e-18  
## Trip\_type=Dispatch 0.000000 0.0000000 2.4226693 4.847071e-19  
## passenger\_groups=Couple 0.000000 0.0000000 7.4194246 1.245354e-58  
## passenger\_groups=Group 0.000000 0.0000000 8.5442353 6.606223e-68  
## period=Period afternoon 0.000000 0.0000000 25.5029202 4.668360e-228  
## period=Period night 0.000000 0.0000000 35.5180619 0.000000e+00  
## v.test  
## period=Period valley Inf  
## period=Period morning 27.907100  
## passenger\_groups=Single 24.530099  
## Trip\_type=Street-Hail 8.915708  
## RateCodeID=Rate-1 8.715315  
## f.cost=[0,8] 5.387923  
## Trip\_distance\_range=Short\_dist 4.790684  
## Payment\_type=Cash 4.024705  
## TipIsGiven=No 3.070418  
## TipIsGiven=Yes -3.070418  
## f.cost=(18,30] -3.132787  
## Payment\_type=Credit card -3.433929  
## f.cost=(30,50] -3.453549  
## f.cost=(50,129) -3.688545  
## Payment\_type=No paid -4.343142  
## Trip\_distance\_range=Long\_dist -8.228018  
## RateCodeID=Rate-Other -8.715315  
## Trip\_type=Dispatch -8.915708  
## passenger\_groups=Couple -16.144309  
## passenger\_groups=Group -17.412726  
## period=Period afternoon -32.241234  
## period=Period night -Inf  
##   
## $`4`  
## Cla/Mod Mod/Cla Global p.value  
## period=Period night 96.711328 81.3524590 35.5180619 0.000000e+00  
## Trip\_distance\_range=Long\_dist 71.578947 24.3852459 14.3845987 1.695159e-61  
## passenger\_groups=Group 74.430380 15.0614754 8.5442353 6.686185e-42  
## Trip\_type=Street-Hail 43.272002 100.0000000 97.5773307 7.579366e-28  
## RateCodeID=Rate-1 43.349644 99.8463115 97.2528661 2.409545e-26  
## f.cost=(30,50] 71.493213 8.0942623 4.7804456 2.347589e-19  
## f.cost=(18,30] 56.215470 20.8504098 15.6608263 1.698775e-16  
## passenger\_groups=Couple 55.685131 9.7848361 7.4194246 1.982848e-07  
## TipIsGiven=Yes 46.984492 41.9057377 37.6595284 3.681425e-07  
## VendorID=f.Vendor-VeriFone 43.945205 82.1721311 78.9530608 3.937983e-06  
## Payment\_type=Credit card 45.753817 49.1290984 45.3385248 9.740537e-06  
## f.cost=(50,129) 61.904762 1.9979508 1.3627515 1.700462e-03  
## f.cost=(8,11] 39.530843 23.3094262 24.8972529 3.262945e-02  
## Payment\_type=Cash 39.767721 50.8709016 54.0125460 2.505066e-04  
## f.cost=[0,8] 36.912226 24.1290984 27.6011248 5.881095e-06  
## VendorID=f.Vendor-Mobile 35.765673 17.8278689 21.0469392 3.937983e-06  
## TipIsGiven=No 39.347675 58.0942623 62.3404716 3.681425e-07  
## Payment\_type=No paid 0.000000 0.0000000 0.6489293 6.644475e-08  
## f.cost=(11,18] 35.521886 21.6188525 25.6975990 4.928571e-08  
## RateCodeID=Rate-Other 2.362205 0.1536885 2.7471339 2.409545e-26  
## Trip\_type=Dispatch 0.000000 0.0000000 2.4226693 7.579366e-28  
## Trip\_distance\_range=Short\_dist 36.238223 55.1741803 64.2872594 2.788750e-28  
## passenger\_groups=Single 37.760618 75.1536885 84.0363400 1.056095e-44  
## period=Period morning 5.350554 1.4856557 11.7239888 2.335274e-94  
## period=Period valley 18.015873 11.6290984 27.2550292 2.460280e-99  
## period=Period afternoon 9.160305 5.5327869 25.5029202 1.780977e-179  
## v.test  
## period=Period night Inf  
## Trip\_distance\_range=Long\_dist 16.546560  
## passenger\_groups=Group 13.562453  
## Trip\_type=Street-Hail 10.938073  
## RateCodeID=Rate-1 10.619847  
## f.cost=(30,50] 8.995687  
## f.cost=(18,30] 8.241632  
## passenger\_groups=Couple 5.200938  
## TipIsGiven=Yes 5.084734  
## VendorID=f.Vendor-VeriFone 4.614629  
## Payment\_type=Credit card 4.422854  
## f.cost=(50,129) 3.138101  
## f.cost=(8,11] -2.136613  
## Payment\_type=Cash -3.661741  
## f.cost=[0,8] -4.530620  
## VendorID=f.Vendor-Mobile -4.614629  
## TipIsGiven=No -5.084734  
## Payment\_type=No paid -5.400529  
## f.cost=(11,18] -5.453868  
## RateCodeID=Rate-Other -10.619847  
## Trip\_type=Dispatch -10.938073  
## Trip\_distance\_range=Short\_dist -11.028370  
## passenger\_groups=Single -14.027639  
## period=Period morning -20.607817  
## period=Period valley -21.155413  
## period=Period afternoon -28.565936  
##   
## $`5`  
## Cla/Mod Mod/Cla Global p.value  
## RateCodeID=Rate-Other 93.70078740 99.1666667 2.747134 3.098738e-225  
## Trip\_type=Dispatch 100.00000000 93.3333333 2.422669 2.173170e-216  
## Trip\_distance\_range=Long\_dist 7.66917293 42.5000000 14.384599 3.518497e-14  
## f.cost=(50,129) 15.87301587 8.3333333 1.362751 4.263359e-06  
## TipIsGiven=No 3.33102012 80.0000000 62.340472 2.655335e-05  
## passenger\_groups=Couple 6.41399417 18.3333333 7.419425 7.020893e-05  
## passenger\_groups=Single 2.34234234 75.8333333 84.036340 1.837786e-02  
## TipIsGiven=Yes 1.37851809 20.0000000 37.659528 2.655335e-05  
## Trip\_distance\_range=Short\_dist 1.68236878 41.6666667 64.287259 3.637606e-07  
## Trip\_type=Street-Hail 0.17734427 6.6666667 97.577331 2.173170e-216  
## RateCodeID=Rate-1 0.02224199 0.8333333 97.252866 3.098738e-225  
## v.test  
## RateCodeID=Rate-Other 32.039255  
## Trip\_type=Dispatch 31.397728  
## Trip\_distance\_range=Long\_dist 7.577658  
## f.cost=(50,129) 4.598112  
## TipIsGiven=No 4.201175  
## passenger\_groups=Couple 3.975577  
## passenger\_groups=Single -2.357916  
## TipIsGiven=Yes -4.201175  
## Trip\_distance\_range=Short\_dist -5.087006  
## Trip\_type=Street-Hail -31.397728  
## RateCodeID=Rate-1 -32.039255

* Cluster 1
  + The first thing we can notice from this cluster is that all observations are of **Payment\_type=No paid**, even though this category only intervents in the sample 0.65% this cluster contains all the individuals of this payment type and all of the observations in the cluster are of **VendorID=f.Vendor-Mobile**, a category that intervents a 21.05% from the sample, but this cluster is that small that we only have a 3.08% of observations of this kind represented in the cluster. So, what is logical is that the other payment types represent a 0% in this cluster as well as the other vendor type. We can also see that all the observations in the did not left a tip, and again and because of the size of the cluster, even though the **TipIsGive=No** represents a 62.34% of the observations from sample, we only have a representation of the 1.04% of these individuals in this cluster. We can also notice that the majority of the trips are made by just one person (96.67%) and we have some morning trips (26.67%).
* Cluster 2
  + The first thing we can see from the cluster is that all of the observations present are of the category **Trip\_type=Street-Hail** and we have in this cluster a representation of the 24.12% of the observations of this category from sample. Something similar happens to the category **RateCodeID=Rate-1**. We can also see that we have the 88.38% of the observations from sample of the category **period=Period afternoon** represented in this cluster and they represent the 95.77% of the observations of the cluster. We can also notice that around the 80% of the observations in this cluster are single passengers and we have 22.27% of the observations of this category from the sample represented here.
* Cluster 3
  + The first thing we can notice is that every observation in the cluster is of the kind of **passenger\_groups=Single** and **Trip\_type=Street-Hail** and we have represented the 36.89% and 31.77%, respectively, of the observations from the sample of these categories. We can also see that almost every observation in the cluster (99.86%) is of **RateCodeID=Rate-1** and we have represented in this cluster the 31.83% of the observations with this category from the sample. We can see that we have the 84.87% of the **period=Period morning** observations of the sample represented in this cluster, and the 77.22% of the **period=Period valley** observations as well. The 67.90% of the observations of the cluster are **period=Period morning**. The 69.29% of the observations in the cluster are short distance trips and the 65.60% observations in the cluster did not left any tips.
* Cluster 4
  + The first thing we can see is that every observation in the cluster is of the kind **Trip\_type=Street-Hail** and we have the 43.27% of the observations from the sample of this kind are represented in this cluster. We can also notice that almost every observation in the cluster is of the kind **RateCodeID=Rate-1** and we have 43.35% of the observations of this kind from the sample represented here. We can see that the 96.71% of the **period=Period night** observations from the sample are represented in the cluster, and the 81.35% of the observations in the cluster are of this kind too. We can see that we have represented the 74.43% of **passenger\_groups=Group**, the 71.58% of **Trip\_distance=Long\_dist** and the 71.49% of **f.cost=(30,50]** observations of these kinds from the sample represented in this cluster.
* Cluster 5
  + The first thing we can notice from this cluster is that we have represented in this cluster all the observations of **Trip\_type=Dispatch** from the sample here and they represent the 93.33% of the observations of this kind in the cluster, so the rest are **Trip\_type=Street-Hail** and we only have a representation of 0.18% of the observations from the sample in this cluster. We can also see that the 80% of the observations in the cluster did not left any tip and the other 20% left some tips, we have a very small representation of observations from the sample of these two categories in this cluster. We can also see that almost every observation in the cluster (99.17%) is of **RateCodeID=Rate-Ohter** and we have the 93.70% of the observations from the sample of this category represented in this cluster. We can see that in this cluster we have represented the 15.87% of the observations from the sample of the category **f.cost=(50,129)**.

We now proceed to see the quantitative variables that characterizes the clusters.

res.hcpcMCA$desc.var$quanti.var # quantitative variables which characterizes the clusters

## Eta2 P-value  
## Total\_amount 0.03950465 3.518655e-39

We can see in the output that the variable that appears is slightly over represented in the clusters. We can notice that **Total\_amount** is over represented with 0.04 units over the global mean. So it is practically the same as the global mean.

We want to know now which variables are associated with the quantitative variables.

res.hcpcMCA$desc.var$quanti # description of each cluster by the quantitative variables

## $`1`  
## NULL  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Total\_amount -7.859152 11.83333 13.9264 7.170368 10.04487  
## p.value  
## Total\_amount 3.867431e-15  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Total\_amount -6.69081 12.45144 13.9264 7.604782 10.04487  
## p.value  
## Total\_amount 2.219385e-11  
##   
## $`4`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Total\_amount 11.26398 15.87319 13.9264 11.44962 10.04487  
## p.value  
## Total\_amount 1.976246e-29  
##   
## $`5`  
## v.test Mean in category Overall mean sd in category Overall sd  
## Total\_amount 5.641927 19.03283 13.9264 19.88545 10.04487  
## p.value  
## Total\_amount 1.681571e-08

We can notice that every cluster has remarked the Total\_amount variable except the first one, that does not have any variable to be described.

* Cluster 2
  + We can see that the **Total\_amount** is around 2 units under the overall mean.
* Cluster 3
  + We can see that the **Total\_amount** is around 1 unit under the overall mean.
* Cluster 4
  + We can see that the **Total\_amount** is around 2 units over the overall mean.
* Cluster 5
  + We can see that the **Total\_amount** is around 6 units over the overall mean.

### Partition quality

We are going to evaluate the partition quality.

#### Gain in inertia (in %)

# ( between sum of squares / total sum of squares ) \* 100  
((res.hcpcMCA$call$t$within[1]-res.hcpcMCA$call$t$within[5])/res.hcpcMCA$call$t$within[1])\*100

## [1] 59.14975

The quality of this reduction if of 59.15%.

In case we wanted to achieve an 80% of the clustering representativity we would need 13 clusters.

((res.hcpcMCA$call$t$within[1]-res.hcpcMCA$call$t$within[13])/res.hcpcMCA$call$t$within[1])\*100

## [1] 80.77602

## Parangons and class-specific individuals.

### The description of the clusters by the individuals

res.hcpcMCA$desc.ind$para # representative individuals of each cluster

## Cluster: 1  
## 632100 1421036 64149 154087 437922   
## 0.2538258 0.2538258 0.3519479 0.3519479 0.3519479   
## ------------------------------------------------------------   
## Cluster: 2  
## 48587 53670 55526 93463 96109   
## 0.2668603 0.2668603 0.2668603 0.2668603 0.2668603   
## ------------------------------------------------------------   
## Cluster: 3  
## 43055 85690 135038 135275 139019   
## 0.1708958 0.1708958 0.1708958 0.1708958 0.1708958   
## ------------------------------------------------------------   
## Cluster: 4  
## 1200 13382 14314 21607 22076   
## 0.222467 0.222467 0.222467 0.222467 0.222467   
## ------------------------------------------------------------   
## Cluster: 5  
## 485688 1399808 1399419 747830 27974   
## 0.2623554 0.2623554 0.2979732 0.3158258 0.4450544

What we obtain are the more representative individuals, paragons, for each cluster. We get the rownames of each paragon in every single cluster.

res.hcpcMCA$desc.ind$dist # individuals distant from each cluster

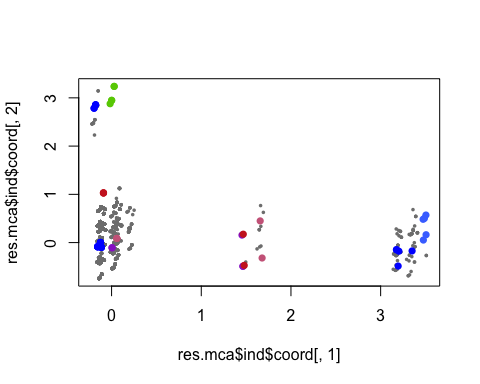
## Cluster: 1  
## 881540 209928 453619 24990 329000   
## 3.776488 3.763555 3.763555 3.753329 3.753329   
## ------------------------------------------------------------   
## Cluster: 2  
## 1261276 646551 856112 187123 226984   
## 1.936593 1.817659 1.817659 1.553835 1.553835   
## ------------------------------------------------------------   
## Cluster: 3  
## 459397 1076485 128467 163845 168358   
## 1.834493 1.735617 1.342113 1.342113 1.342113   
## ------------------------------------------------------------   
## Cluster: 4  
## 826623 35649 202294 245448 321262   
## 2.123598 2.034772 1.818039 1.818039 1.818039   
## ------------------------------------------------------------   
## Cluster: 5  
## 1083301 173366 720785 131915 810930   
## 3.739454 3.714631 3.708608 3.654759 3.652079

What we obtain are those individuals of each cluster that that far away in the same cluster from the rest of the individuals. We also obtain the rownames of each individual with the bigger distance respect the other ones in the cluster.

#### Examine the values of individuals that characterize classes

We get the grpahical representation for the individuals that characterize classes (para and dist).

# characteristic individuals  
para1<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[1]]))  
dist1<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[1]]))  
para2<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[2]]))  
dist2<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[2]]))  
para3<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[3]]))  
dist3<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[3]]))  
para4<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[4]]))  
dist4<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[4]]))  
para5<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[5]]))  
dist5<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[5]]))  
  
plot(res.mca$ind$coord[,1],res.mca$ind$coord[,2],col="grey50",cex=0.5,pch=16)  
points(res.mca$ind$coord[para1,1],res.mca$ind$coord[para1,2],col="blue",cex=1,pch=16)  
points(res.mca$ind$coord[dist1,1],res.mca$ind$coord[dist1,2],col="chartreuse3",cex=1,pch=16)  
points(res.mca$ind$coord[para2,1],res.mca$ind$coord[para2,2],col="blue",cex=1,pch=16)  
points(res.mca$ind$coord[dist2,1],res.mca$ind$coord[dist2,2],col="darkorchid3",cex=1,pch=16)  
points(res.mca$ind$coord[para3,1],res.mca$ind$coord[para3,2],col="blue",cex=1,pch=16)  
points(res.mca$ind$coord[dist3,1],res.mca$ind$coord[dist3,2],col="firebrick3",cex=1,pch=16)  
points(res.mca$ind$coord[para4,1],res.mca$ind$coord[para4,2],col="blue",cex=1,pch=16)  
points(res.mca$ind$coord[dist4,1],res.mca$ind$coord[dist4,2],col="palevioletred3",cex=1,pch=16)  
points(res.mca$ind$coord[para5,1],res.mca$ind$coord[para5,2],col="blue",cex=1,pch=16)  
points(res.mca$ind$coord[dist5,1],res.mca$ind$coord[dist5,2],col="royalblue1",cex=1,pch=16)



## Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA)focusing on…

df$hcpckMCA<-res.hcpcMCA$data.clust$clust  
  
# With Hierarchical Clustering (PCA)  
table(df$hcpck,df$hcpckMCA)

##   
## 1 2 3 4 5  
## kHP-1 12 719 140 1059 0  
## kHP-2 11 242 1107 191 83  
## kHP-3 0 71 0 189 2  
## kHP-4 7 53 176 489 33  
## kHP-5 0 3 10 24 2

df$hcpckMCA\_hcpck<-factor(  
 df$hcpckMCA,  
 levels=c(4,3,2,1,5),  
 labels=c("kHPmca-4","kHPmca-3","kHPmca-2","kHPmca-1","kHPmca-5")  
)  
tt1<-table(df$hcpck,df$hcpckMCA\_hcpck); tt1

##   
## kHPmca-4 kHPmca-3 kHPmca-2 kHPmca-1 kHPmca-5  
## kHP-1 1059 140 719 12 0  
## kHP-2 191 1107 242 11 83  
## kHP-3 189 0 71 0 2  
## kHP-4 489 176 53 7 33  
## kHP-5 24 10 3 0 2

100\*sum(diag(tt1)/sum(tt1))

## [1] 48.58317

We have a concordance of the 48.58% so we can say that they are different, if we had a greater concordance, this would mean that they would be more similar.

# With k-means (PCA)  
table(df$claKM, df$hcpckMCA)

##   
## 1 2 3 4 5  
## kKM-3 3 491 119 229 2  
## kKM-5 17 398 938 931 69  
## kKM-2 4 57 86 317 22  
## kKM-1 5 138 271 396 21  
## kKM-4 1 4 19 79 6

df$hcpckMCA\_claKM<-factor(  
 df$hcpckMCA,  
 levels=c(2,3,1,4,5),  
 labels=c("kHPmca-2","kHPmca-3","kHPmca-1","kHPmca-4","kHPmca-5")  
)  
tt2<-table(df$claKM,df$hcpckMCA\_claKM); tt

## f.tt  
## f.cost (10,15] (15,20] (20,50] (5,10] [0,5]  
## (11,18] 613 314 88 156 8  
## (18,30] 106 205 388 3 15  
## (30,50] 1 23 175 2 4  
## (50,129) 1 1 35 0 7  
## (8,11] 189 3 4 864 85  
## [0,8] 3 3 4 486 775

100\*sum(diag(tt2)/sum(tt2))

## [1] 39.69284

We have a concordance of the 39.69% so we can say that they are different, if we had a greater concordance, this would mean that they would be more similar.

### Quantitative target (Total\_amount)

* hcpc

# res.hcpc$desc.var$quanti.var # quantitative variables which characterizes the clusters  
# # Eta2 P-value  
# # Total\_amount 0.539522699 0.000000e+00

* kmeans

# res.cat <-catdes(df,30)  
# res.cat  
# # Link between the cluster variable and the quantitative variables  
# # ================================================================  
# # Eta2 P-value  
# # Total\_amount 0.688303660 0.000000e+00

* hcpc\_mca

# # res.hcpcMCA$desc.var$quanti.var # quantitative variables which characterizes the clusters  
# # Eta2 P-value  
# # Total\_amount 0.03950465 3.518655e-39

#### Comment

To compare the variable Total\_amount in the three different classifications, we will look at Eta2:

* The closer to 1 is eta2 for a variable, the better the variance between groups is explained by this variable.
* We can see that, in descending order, we have:
  + k-means (0.69)
  + hcpc (0.54)
  + hcpc\_mca (0.04)
* This means that in the last classification the variable to define the clusters is not taken into account so much.

### Binary target (TipIsGiven)

#### hcpc

# res.hcpc$desc.var$category # description of each cluster by the categories  
# # $`1`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 43.6502429 65.18134715 62.340472   
# # TipIsGiven=Yes 38.5985066 34.81865285 37.659528   
# #  
# # $`2`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 38.965996 68.727050 62.340472   
# # TipIsGiven=Yes 29.350948 31.272950 37.659528   
# #   
# # $`3`  
# # Cla/Mod Mod/Cla Global   
# # nothing to see here  
# #   
# # $`4`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=Yes 24.6984492 56.728232 37.659528   
# # TipIsGiven=No 11.3809854 43.271768 62.340472  
# #   
# # $`5`  
# # Cla/Mod Mod/Cla Global  
# # TipIsGiven=Yes 1.60827111 71.794872 37.659528   
# # TipIsGiven=No 0.38167939 28.205128 62.340472

#### kmeans

# res.cat <-catdes(df,30)  
# res.cat  
# #   
# # Description of each cluster by the categories  
# # =============================================  
# # $`1`  
# # Cla/Mod Mod/Cla Global  
# # TipIsGiven=Yes 23.721999 49.6991576 37.659528  
# # TipIsGiven=No 14.503817 50.3008424 62.340472  
# #   
# # $`2`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=Yes 15.6232051 55.9670782 37.659528   
# # TipIsGiven=No 7.4253990 44.0329218 62.340472  
# #   
# # $`3`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 19.5697432 66.8246445 62.3404716  
# # TipIsGiven=Yes 16.0827111 33.1753555 37.6595284   
# #   
# # $`4`  
# # Cla/Mod Mod/Cla Global  
# # TipIsGiven=Yes 3.9058013 62.3853211 37.6595284  
# # TipIsGiven=No 1.4226232 37.6146789 62.3404716  
# #   
# # $`5`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 57.078418 69.9107522 62.3404716  
# # TipIsGiven=Yes 40.666284 30.0892478 37.6595284

#### hcpc\_mca

# res.hcpcMCA$desc.var$category # description of each cluster by the categories  
# # $`1`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 1.0409438 100.00000 62.3404716  
# # TipIsGiven=Yes 0.0000000 0.00000 37.6595284  
# #   
# # $`2`  
# # Cla/Mod Mod/Cla Global   
# # nothing to see here  
# #   
# # $`3`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 32.616239 65.5966504 62.3404716   
# # TipIsGiven=Yes 28.317059 34.4033496 37.6595284   
# #   
# # $`4`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=Yes 46.984492 41.9057377 37.6595284   
# # TipIsGiven=No 39.347675 58.0942623 62.3404716   
# #   
# # $`5`  
# # Cla/Mod Mod/Cla Global   
# # TipIsGiven=No 3.33102012 80.0000000 62.340472   
# # TipIsGiven=Yes 1.37851809 20.0000000 37.659528

#### Comment

To compare the variable TipIsGiven in the three different classifications, we will look at Cla / Mod, Mod / Cla and Global:

* Cluster 1:
  + hcpc: TipIsGiven = No is overrepresented
  + kmeans: TipIsGiven = Yes is overrepresented
  + hcpc\_mca: TipIsGiven = No is overrepresented
* Cluster 2:
  + hcpc: TipIsGiven = No is overrepresented
  + kmeans: TipIsGiven = Yes is overrepresented
  + hcpc\_mca: There is no data in the cluster of this variable
* Cluster 3:
  + hcpc: No data in the cluster of this variable
  + kmeans: TipIsGiven = No is overrepresented
  + hcpc\_mca: TipIsGiven = No is overrepresented
* Cluster 4:
  + hcpc: TipIsGiven = Yes is overrepresented
  + kmeans: TipIsGiven = Yes is overrepresented
  + hcpc\_mca: TipIsGiven = Yes is overrepresented
* Cluster 5:
  + hcpc: TipIsGiven = Yes is overrepresented
  + kmeans: TipIsGiven = No is overrepresented
  + hcpc\_mca: TipIsGiven = No is overrepresented

### Final comment

We think that at first glance, we do not find the relationship between the different clusters of the different types of analysis. As we can see in the data, they are not distributed in the same way with respect to the two variables we had to analyze.

It makes sense to think this, since these variables have not been taken into account in the analyzes, as they had the role of supplementary variables, which means that they only served us as explanatory variables, and not to decide how to form clusters.