Deliverable 2

PCA, CA and Clustering

Júlia Gasull i Claudia Sánchez

November 14, 2020

Contents

1 First setups

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

1.1 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/delivera

# Load Required Packages
options(contrasts=c("contr.treatment","contr.treatment"))
requiredPackages <- c("missMDA","chemometrics","mvoutlier","effects","FactoMineR","car", "factoextra","F
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]
if(length(missingPackages)) install.packages(missingPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

1.2 Load processed data from first deliverable

```
load(paste0(filepath,"/Taxi5000_del1.RData"))
```

1.3 Clean data

```
# remove some columns
names(df)
```

```
##  [1] "VendorID"              "lpep_pickup_datetime"  "Lpep_dropoff_datetime"
##  [4] "Store_and_fwd_flag"    "RateCodeID"            "Pickup_longitude"
##  [7] "Pickup_latitude"       "Dropoff_longitude"     "Dropoff_latitude"
## [10] "Passenger_count"       "Trip_distance"         "Fare_amount"
## [13] "Extra"                 "MTA_tax"               "Tip_amount"
## [16] "Tolls_amount"          "Ehail_fee"             "improvement_surcharge"
## [19] "Total_amount"          "Payment_type"          "Trip_type"
## [22] "hour"                  "period"                "tlenkm"
## [25] "traveltime"            "espeed"                "pickup"
## [28] "dropoff"               "Trip_distance_range"   "yearGt2015"
## [31] "CashTips"              "paidTolls"             "Sum_total_amount"
## [34] "TipIsGiven"            "passenger_groups"
```

```
df$lpep_pickup_datetime <- NULL
df$Lpep_dropoff_datetime <- NULL
df$Store_and_fwd_flag <- NULL
```

```r
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
```

---

2 Principal Component Analysis (PCA)

```r
names(df)
```

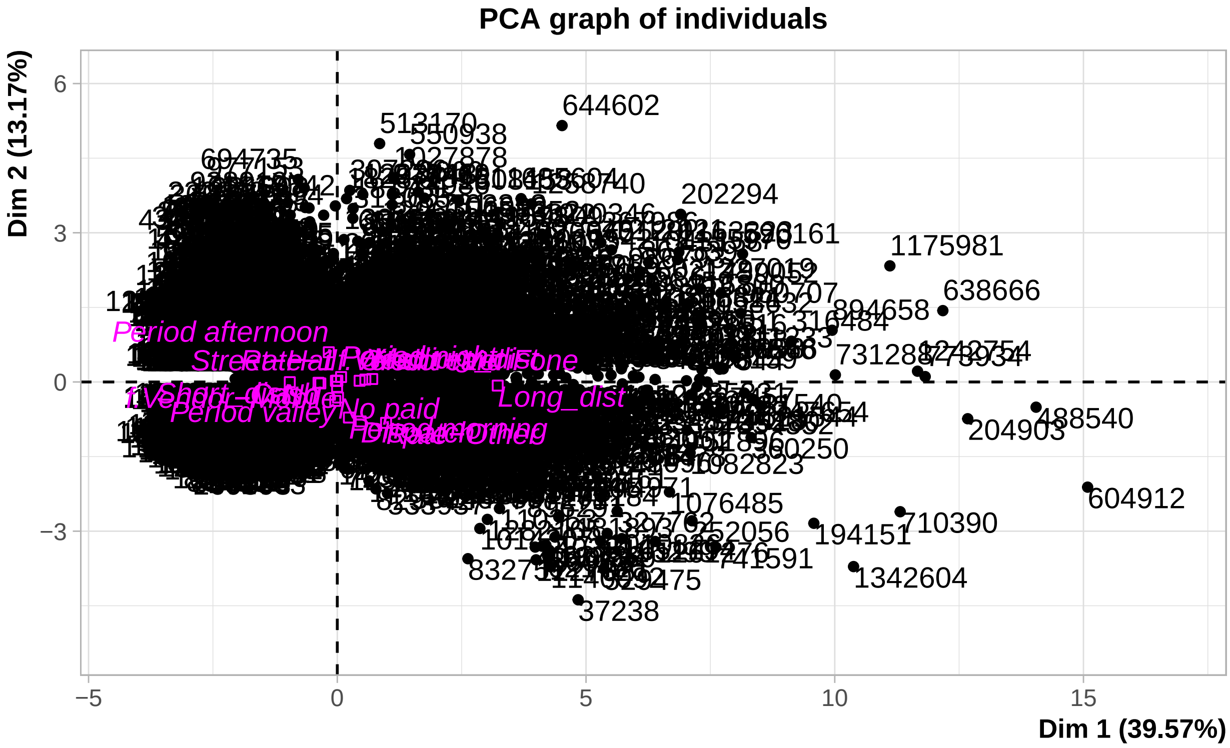
```
##  [1] "VendorID"              "RateCodeID"              "Pickup_longitude"
##  [4] "Pickup_latitude"       "Dropoff_longitude"       "Dropoff_latitude"
##  [7] "Passenger_count"       "Trip_distance"           "Fare_amount"
## [10] "Extra"                 "MTA_tax"                 "Tip_amount"
## [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"
```

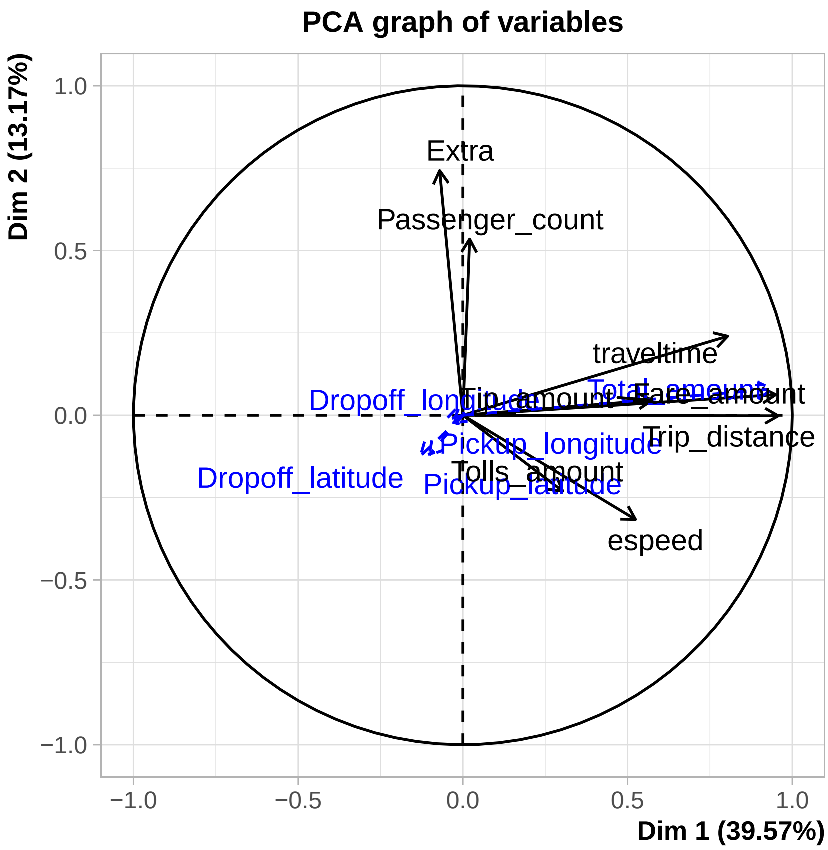
```r
vars_res<-names(df)[c(15,27)]
vars_quantitatives<-names(df)[c(3:10,12,20:22)]
vars_categorical<-names(df)[c(1,2,16:17,19,25,28)]
```

Note - web that we used in order to use factoextra
* http://www.sthda.com/english/wiki/wiki.php?id_contents=7851&fbclid=IwAR01E5XVvCrSKnpkCdAppbpvv7YMGvxSWaSSwb4SIgrXjrxoIpMIlNblYFY

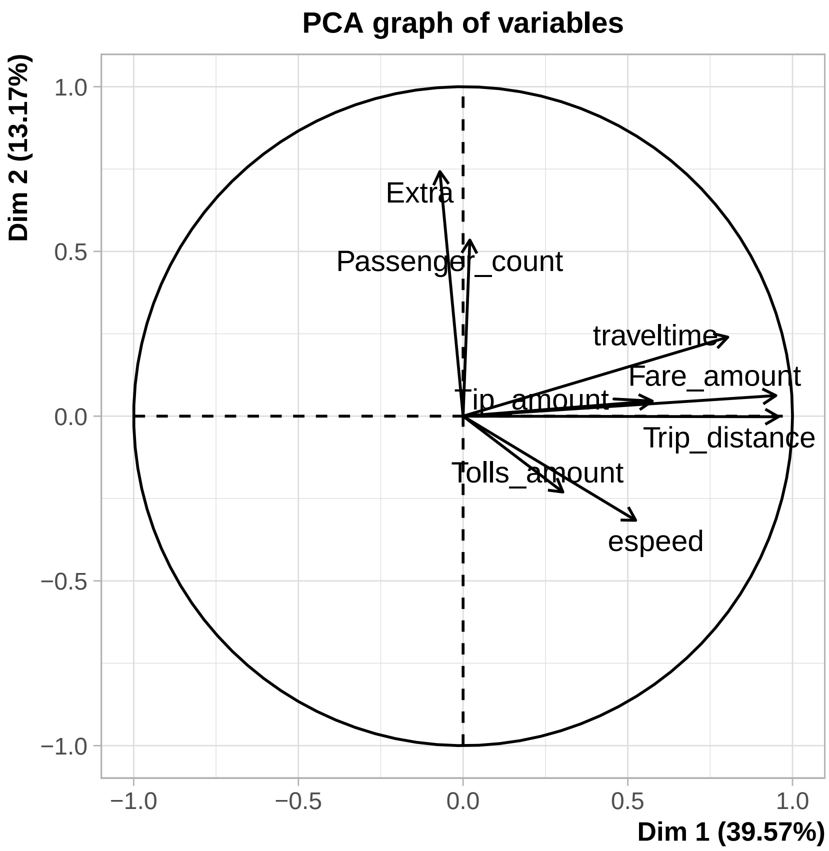
We have already seen profiling in the previous installment. So now, let's proceed to look at the main components.

```r
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)
)
```

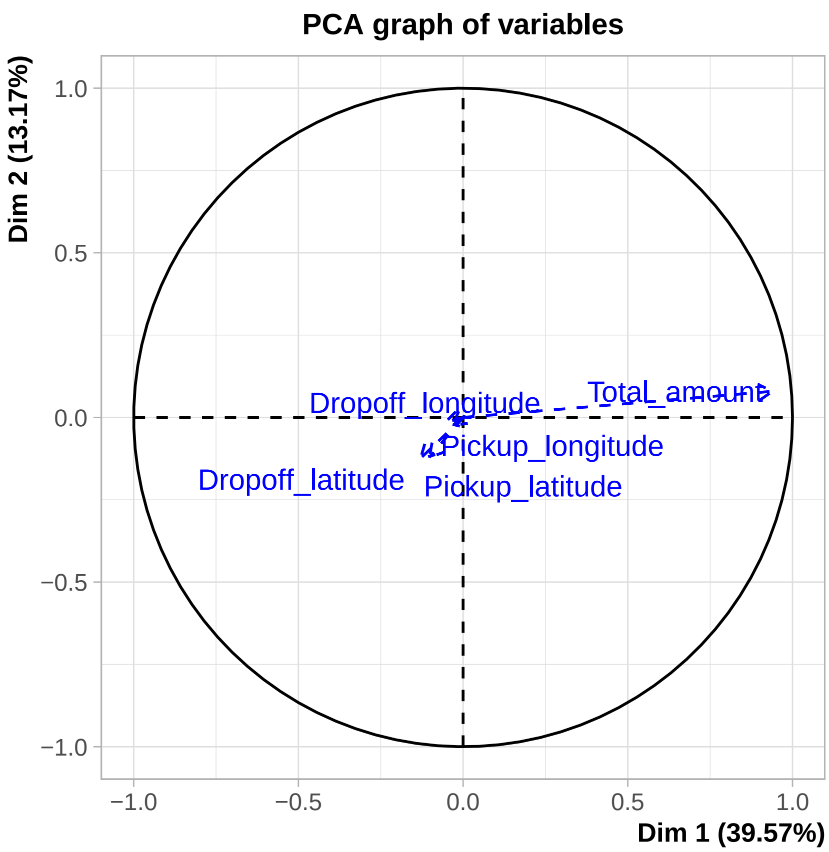




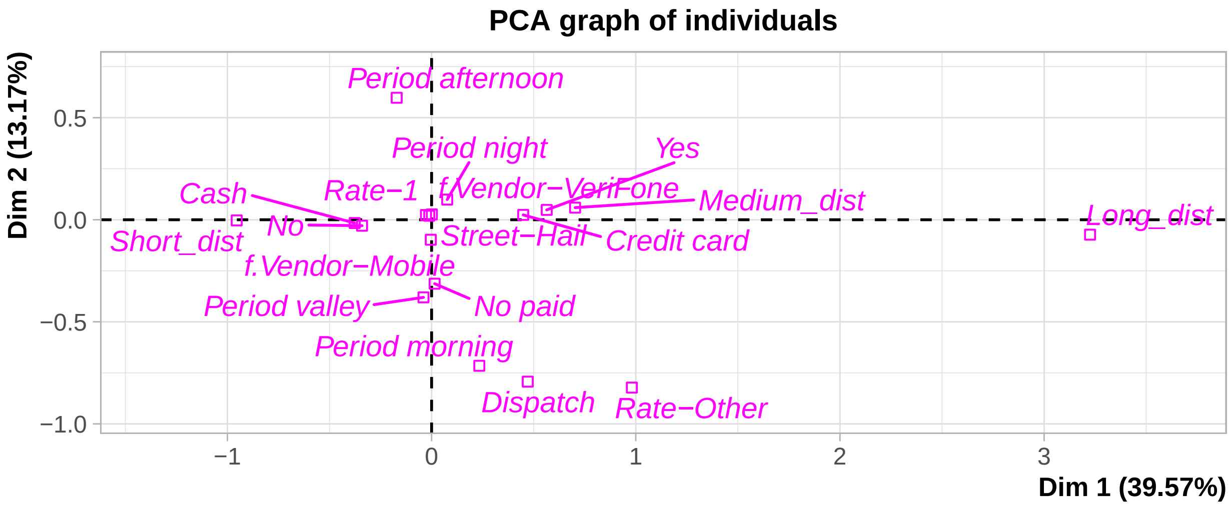

```r
plot.PCA(res.pca,choix=c("var"), invisible=c("quanti.sup"))
```




```
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:

```
# TO DO: explicar quins son multivariant outliers, la profe diu al video del 23/10 que aquests son uns p
```

2.1 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.

```
summary(res.pca, nb.dec=2,nbind=1, nbelements = 1000, ncp=5)
```

```
##
## Call:
## PCA(X = 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))
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8
## Variance                3.17   1.05   1.04   0.95   0.90   0.72   0.11   0.06
## % of var.              39.57  13.17  12.99  11.92  11.21   9.01   1.40   0.72
## Cumulative % of var.   39.57  52.74  65.73  77.66  88.87  97.88  99.28 100.00
##
## Individuals (the 1 first)
##                         Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 311                  |  1.48 | -1.24   0.01   0.70 |   0.05   0.00   0.00 |   0.00   0.00
##                       cos2    Dim.4    ctr   cos2    Dim.5    ctr   cos2
## 311                   0.00 |   0.66   0.01   0.20 |   0.00   0.00   0.00 |
##
## Variables
##                       Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
## Passenger_count     |  0.02   0.01   0.00 |   0.53  27.12   0.29 |   0.53  27.48   0.29 |
## Trip_distance       |  0.96  28.95   0.92 |   0.00   0.00   0.00 |  -0.01   0.01   0.00 |
## Fare_amount         |  0.95  28.49   0.90 |   0.06   0.37   0.00 |  -0.14   1.79   0.02 |
## Extra               | -0.07   0.16   0.00 |   0.74  52.33   0.55 |   0.14   1.84   0.02 |
## Tip_amount          |  0.57  10.41   0.33 |   0.05   0.20   0.00 |   0.06   0.30   0.00 |
## Tolls_amount        |  0.30   2.90   0.09 |  -0.23   5.03   0.05 |   0.53  27.38   0.28 |
```

```
## traveltime        |  0.80 20.40  0.65 |  0.24  5.46  0.06 | -0.41 15.85  0.16 |
## espeed            |  0.52  8.67  0.27 | -0.32  9.49  0.10 |  0.51 25.34  0.26 |
##                   Dim.4   ctr  cos2   Dim.5   ctr  cos2
## Passenger_count   -0.61 39.44  0.38 |  0.21  5.14  0.05 |
## Trip_distance     -0.07  0.58  0.01 | -0.15  2.53  0.02 |
## Fare_amount       -0.07  0.59  0.01 | -0.01  0.02  0.00 |
## Extra              0.56 33.04  0.32 | -0.31 10.45  0.09 |
## Tip_amount         0.27  7.66  0.07 |  0.16  2.93  0.03 |
## Tolls_amount       0.41 17.86  0.17 |  0.57 35.76  0.32 |
## traveltime        -0.07  0.57  0.01 |  0.21  5.11  0.05 |
## espeed            -0.05  0.27  0.00 | -0.58 38.06  0.34 |
##
## Supplementary continuous variables
##                     Dim.1  cos2   Dim.2  cos2   Dim.3  cos2   Dim.4  cos2
## Pickup_longitude  | -0.03  0.00 | -0.02  0.00 |  0.08  0.01 | -0.01  0.00 |
## Pickup_latitude   | -0.10  0.01 | -0.12  0.01 |  0.04  0.00 | -0.04  0.00 |
## Dropoff_longitude | -0.05  0.00 | -0.02  0.00 |  0.09  0.01 |  0.00  0.00 |
## Dropoff_latitude  | -0.13  0.02 | -0.12  0.02 |  0.04  0.00 | -0.03  0.00 |
## Total_amount      |  0.94  0.88 |  0.08  0.01 | -0.06  0.00 |  0.03  0.00 |
##                   Dim.5  cos2
## Pickup_longitude  -0.08  0.01 |
## Pickup_latitude   -0.01  0.00 |
## Dropoff_longitude -0.11  0.01 |
## Dropoff_latitude   0.00  0.00 |
## Total_amount       0.03  0.00 |
##
## Supplementary categories
##                     Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## f.Vendor-Mobile   |   0.16 |   0.00   0.00  -0.08 |  -0.10   0.36  -3.35 |
## f.Vendor-VeriFone |   0.04 |   0.00   0.00   0.08 |   0.03   0.36   3.35 |
## Rate-1            |   0.04 |  -0.03   0.43  -6.30 |   0.02   0.30   9.15 |
## Rate-Other        |   1.49 |   0.98   0.43   6.30 |  -0.82   0.30  -9.15 |
## Credit card       |   0.72 |   0.45   0.39  15.61 |   0.02   0.00   1.43 |
## Cash              |   0.60 |  -0.38   0.40 -15.60 |  -0.02   0.00  -1.16 |
## No paid           |   0.75 |   0.01   0.00   0.05 |  -0.31   0.17  -1.68 |
## Street-Hail       |   0.03 |  -0.01   0.14  -2.83 |   0.02   0.41   8.28 |
## Dispatch          |   1.24 |   0.47   0.14   2.83 |  -0.79   0.41  -8.28 |
## Period night      |   0.37 |   0.08   0.04   2.16 |   0.10   0.07   4.86 |
## Period morning    |   1.00 |   0.23   0.05   3.25 |  -0.72   0.51 -17.27 |
## Period valley     |   0.58 |  -0.04   0.00  -0.93 |  -0.38   0.43 -15.42 |
## Period afternoon  |   0.76 |  -0.17   0.05  -3.83 |   0.60   0.62  23.16 |
## Long_dist         |   3.25 |   3.22   0.98  50.51 |  -0.07   0.00  -1.98 |
## Medium_dist       |   0.74 |   0.70   0.90  13.98 |   0.06   0.01   2.05 |
## Short_dist        |   0.96 |  -0.95   0.99 -48.95 |   0.00   0.00  -0.30 |
## No                |   0.58 |  -0.34   0.34 -16.74 |  -0.03   0.00  -2.48 |
## Yes               |   0.97 |   0.56   0.34  16.74 |   0.05   0.00   2.48 |
##                    Dim.3   cos2 v.test    Dim.4   cos2 v.test    Dim.5   cos2
## f.Vendor-Mobile    -0.07   0.16  -2.24 |   0.10   0.41   3.72 |  -0.04   0.06
## f.Vendor-VeriFone   0.02   0.16   2.24 |  -0.03   0.41  -3.72 |   0.01   0.06
## Rate-1              0.00   0.00  -1.14 |   0.02   0.14   6.55 |   0.00   0.00
## Rate-Other          0.10   0.00   1.14 |  -0.56   0.14  -6.55 |   0.02   0.00
## Credit card         0.07   0.01   4.23 |   0.20   0.08  12.58 |   0.09   0.02
## Cash               -0.06   0.01  -4.08 |  -0.17   0.08 -12.46 |  -0.07   0.01
## No paid            -0.17   0.05  -0.91 |  -0.12   0.03  -0.69 |  -0.33   0.19
## Street-Hail         0.00   0.00   0.82 |   0.02   0.35   8.04 |   0.00   0.01
## Dispatch           -0.08   0.00  -0.82 |  -0.73   0.35  -8.04 |  -0.10   0.01
## Period night        0.23   0.37  11.32 |   0.07   0.04   3.70 |  -0.26   0.47
## Period morning     -0.26   0.07  -6.31 |  -0.41   0.17 -10.41 |   0.44   0.19
## Period valley      -0.20   0.12  -8.11 |  -0.30   0.26 -12.63 |   0.25   0.19
## Period afternoon    0.01   0.00   0.51 |   0.41   0.29  16.52 |  -0.11   0.02
## Long_dist           0.07   0.00   1.82 |  -0.18   0.00  -5.13 |  -0.32   0.01
## Medium_dist        -0.17   0.05  -6.02 |  -0.02   0.00  -0.83 |  -0.01   0.00
## Short_dist          0.04   0.00   3.81 |   0.05   0.00   4.47 |   0.07   0.01
## No                 -0.05   0.01  -4.57 |  -0.16   0.08 -14.54 |  -0.08   0.02
```

```
## Yes                  0.09    0.01    4.57 |    0.27    0.08   14.54 |    0.13    0.02
##                    v.test
## f.Vendor-Mobile     -1.46 |
## f.Vendor-VeriFone    1.46 |
## Rate-1              -0.20 |
## Rate-Other           0.20 |
## Credit card          5.95 |
## Cash                -5.64 |
## No paid             -1.89 |
## Street-Hail          1.08 |
## Dispatch            -1.08 |
## Period night       -13.71 |
## Period morning      11.41 |
## Period valley       11.11 |
## Period afternoon    -4.72 |
## Long_dist           -9.37 |
## Medium_dist         -0.35 |
## Short_dist           7.17 |
## No                  -7.42 |
## Yes                  7.42 |
```

2.1.1 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
```

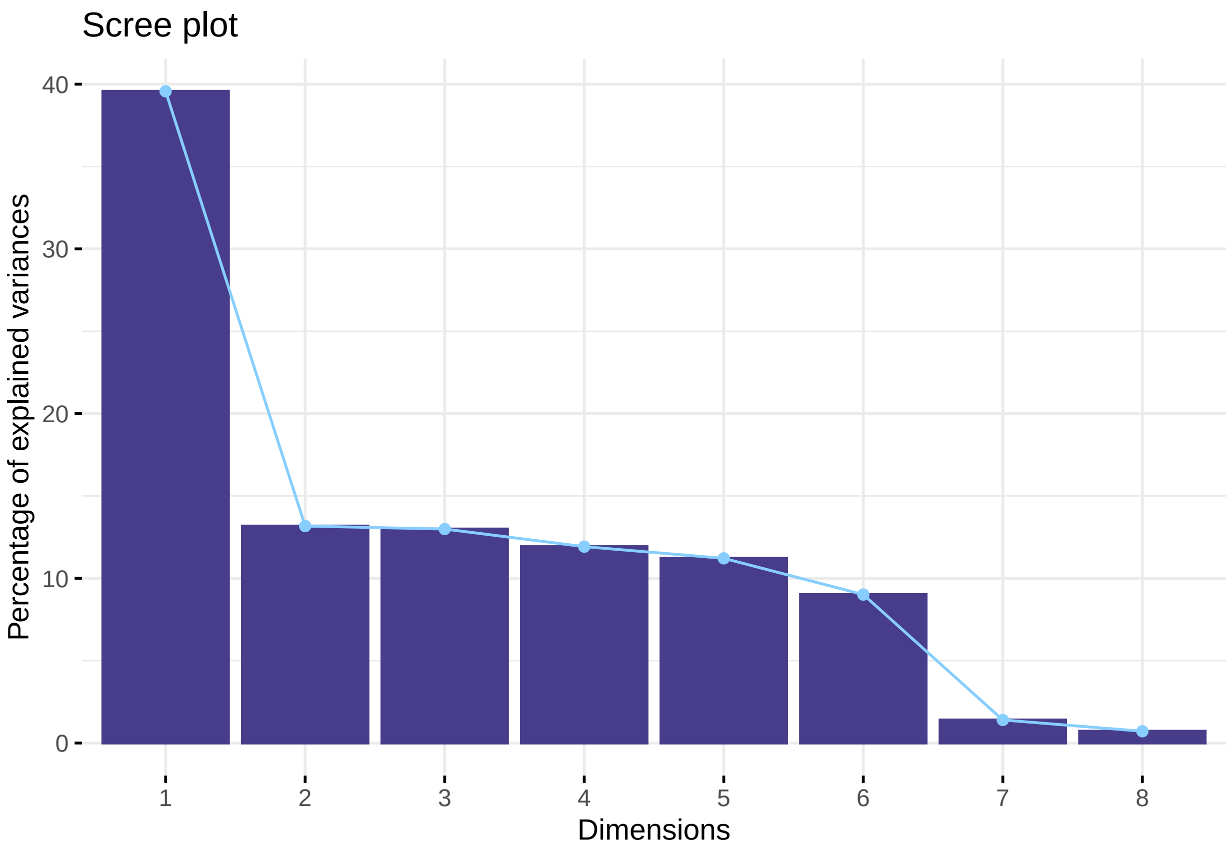
In this case, then, we will use up to dimension 3, and they will explain 65.73% of the total inertia.

2.1.2 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,
  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.

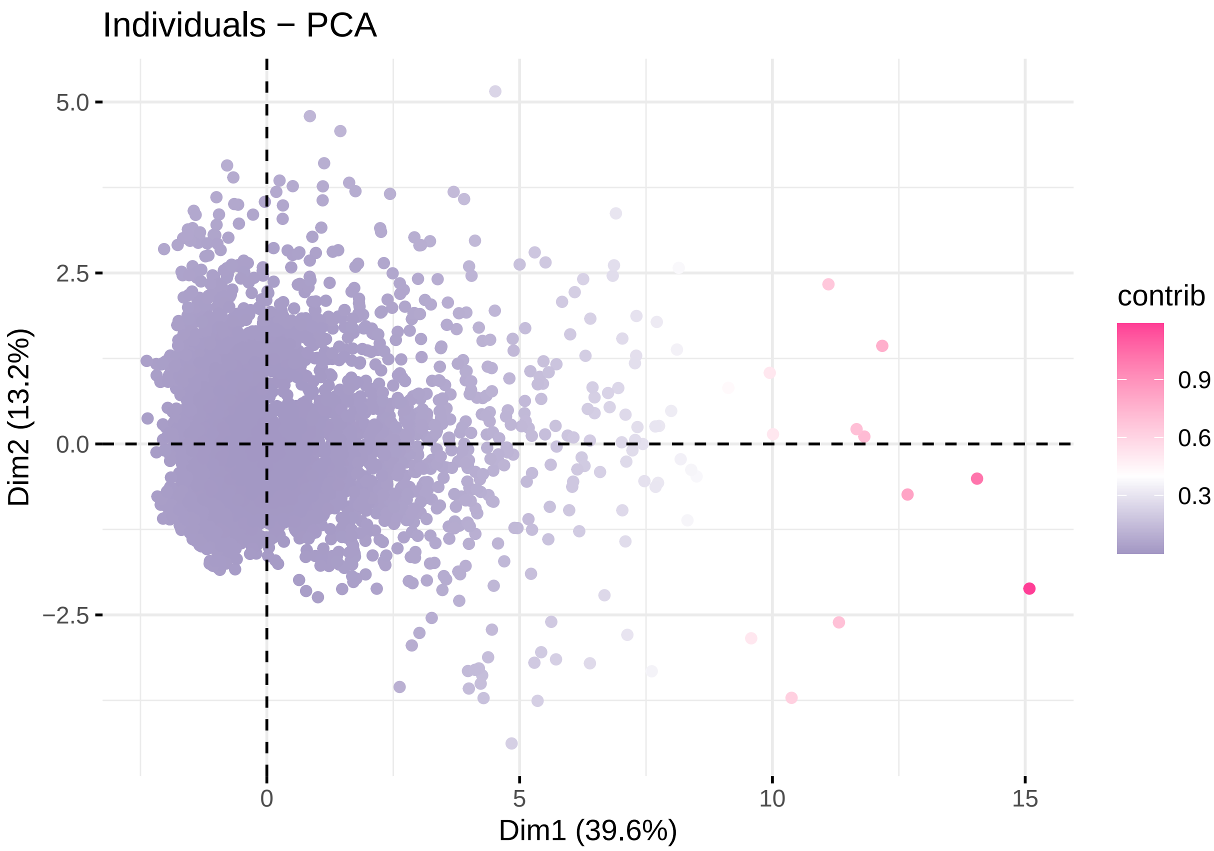
2.2 Individuals point of view

2.2.1 Contribution

```
head(res.pca$ind$contrib) # contribition of individuals to the princial components
```

```
##             Dim.1        Dim.2        Dim.3       Dim.4        Dim.5
## 311  0.010426834 6.030826e-05 3.705470e-07 0.009891871 4.706081e-10
## 749  0.155265882 4.964735e-03 7.015047e-03 0.007524551 6.357976e-03
## 907  0.003557855 1.759607e-05 1.207026e-04 0.002605736 2.737022e-03
## 1187 0.003978458 2.597782e-02 9.407763e-05 0.009387996 5.272289e-03
## 1200 0.004182317 3.839182e-06 4.542485e-04 0.010923895 8.799043e-04
## 1807 0.009131625 3.380623e-05 1.298368e-05 0.009512722 5.867972e-05
```

```
fviz_pca_ind(res.pca, col.ind="contrib", geom = "point") +
scale_color_gradient2(low="darkslateblue", mid="white",
                      high="violetred1", midpoint=0.40)
```

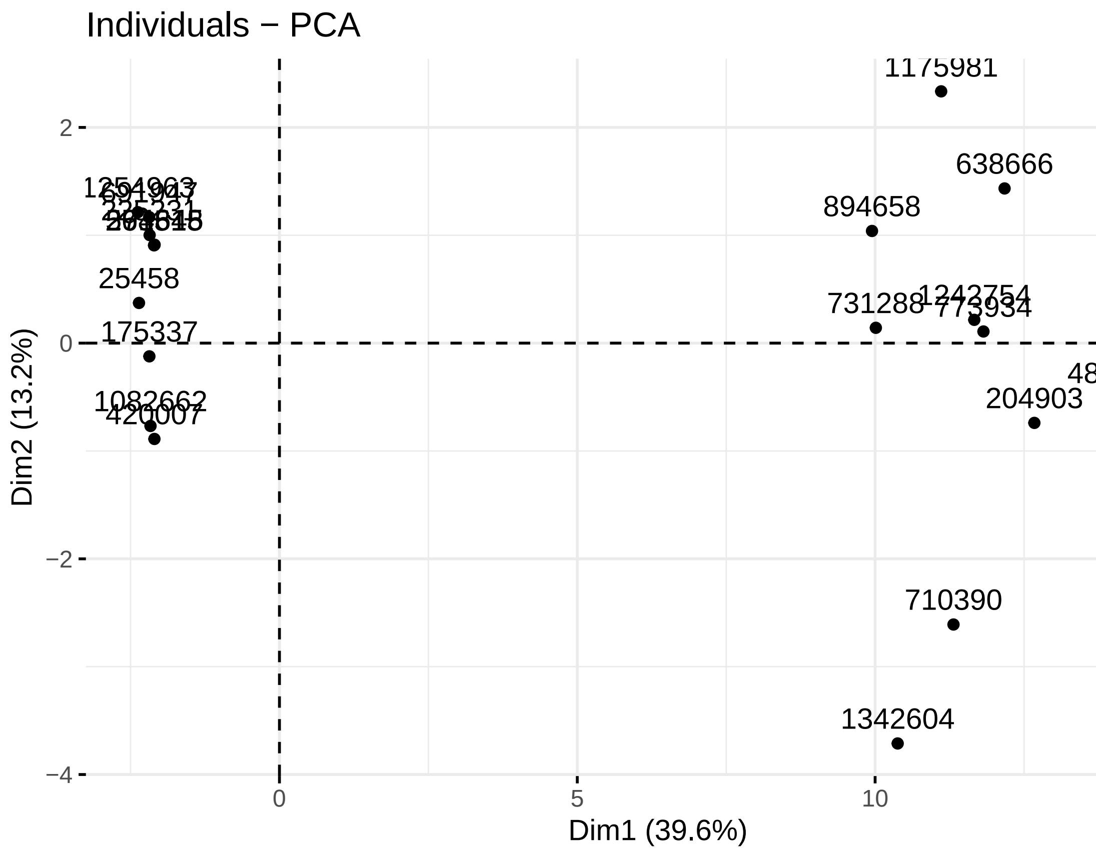


We can see that there are some individuals that are too contributive. So now, let's try to understand them better with extreme individuals.

2.2.2 Extreme individuals

```
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))
```



2.2.2.1 In dimension 1:

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
```

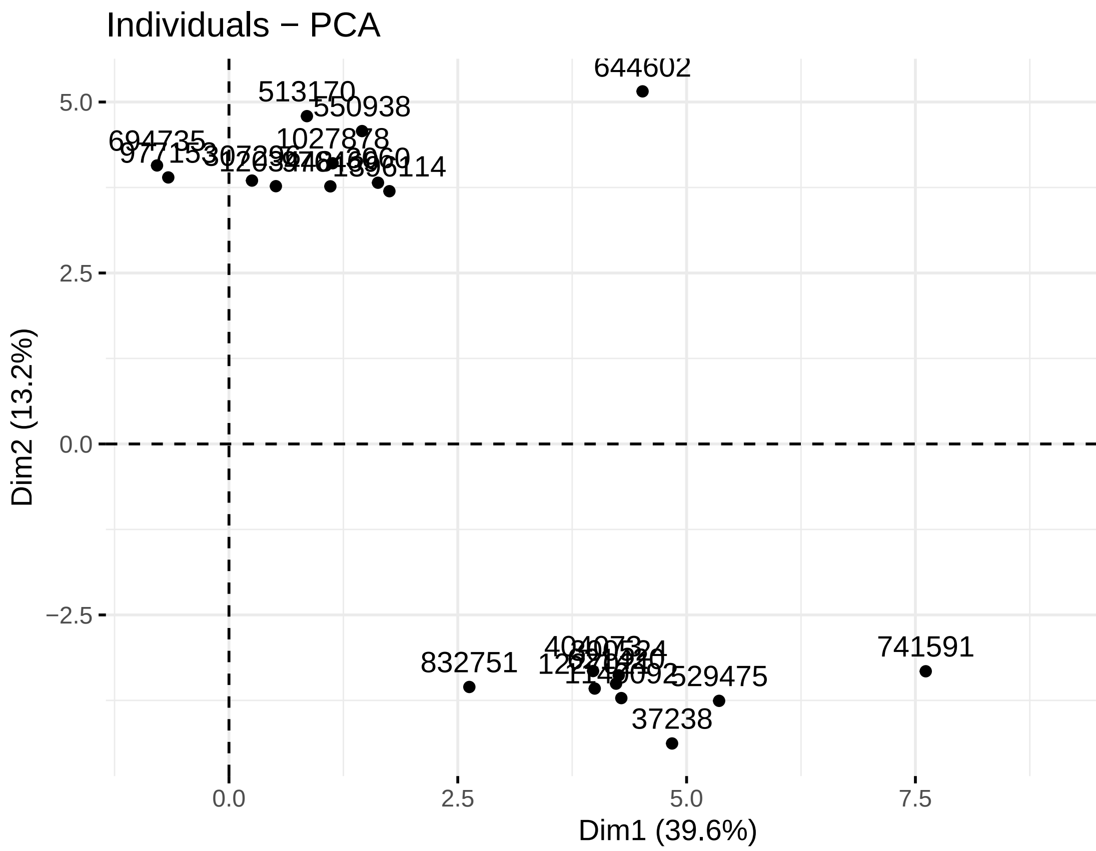
```r
df[which(row.names(df) %in% row.names(df)[rang[1:10]]),1:28]
```

```
##                  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
```

```r
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))
```



2.2.2.2 In dimension 2:

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
```

2.2.3 Detection of multivariant outliers and influent data.

```r
# no sé què posar aquí
```

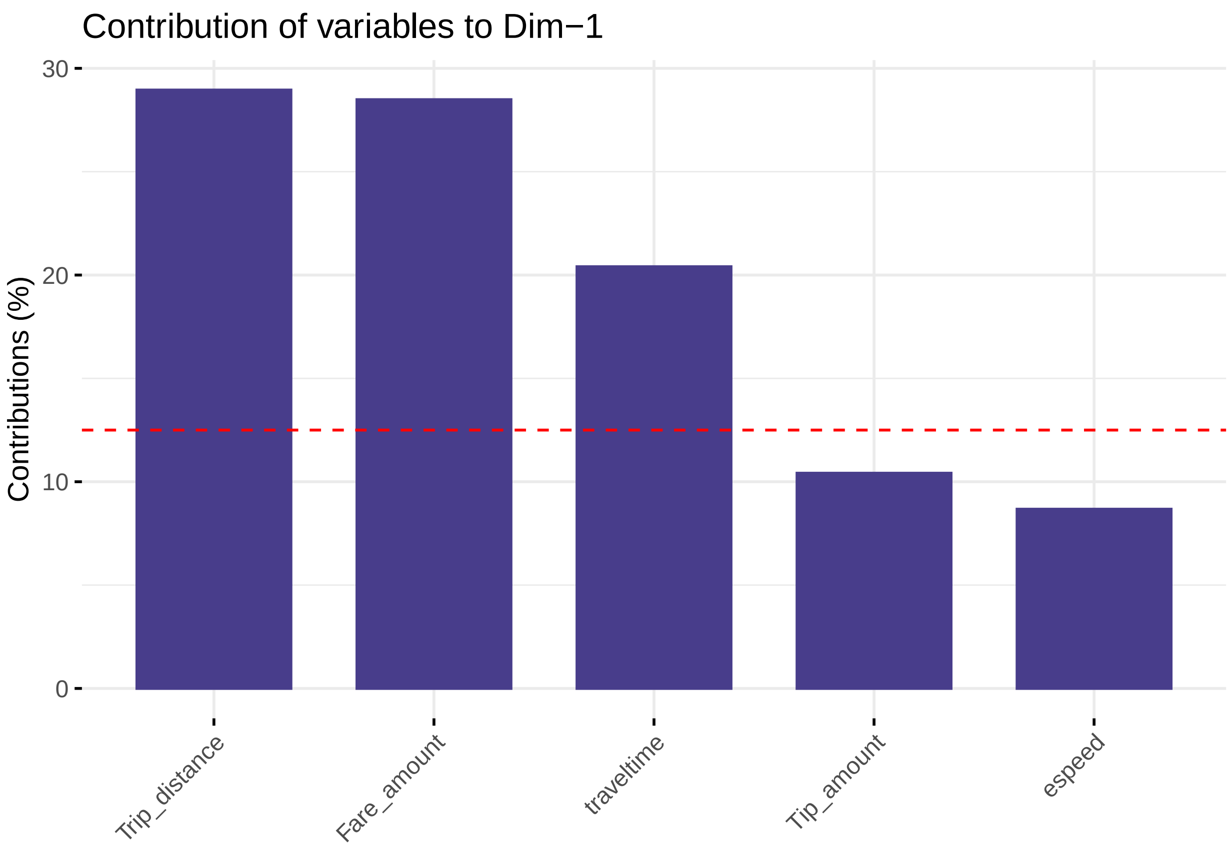
2.3 Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables

```r
res.des <- dimdesc(res.pca)
```

2.3.1 First dimension

```r
fviz_contrib(  # contributions of variables to PC1
  res.pca,
  fill = "darkslateblue",
  color = "darkslateblue",
  choice = "var",
```

```r
  axes = 1,
  top = 5)
```




```r
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 "
```

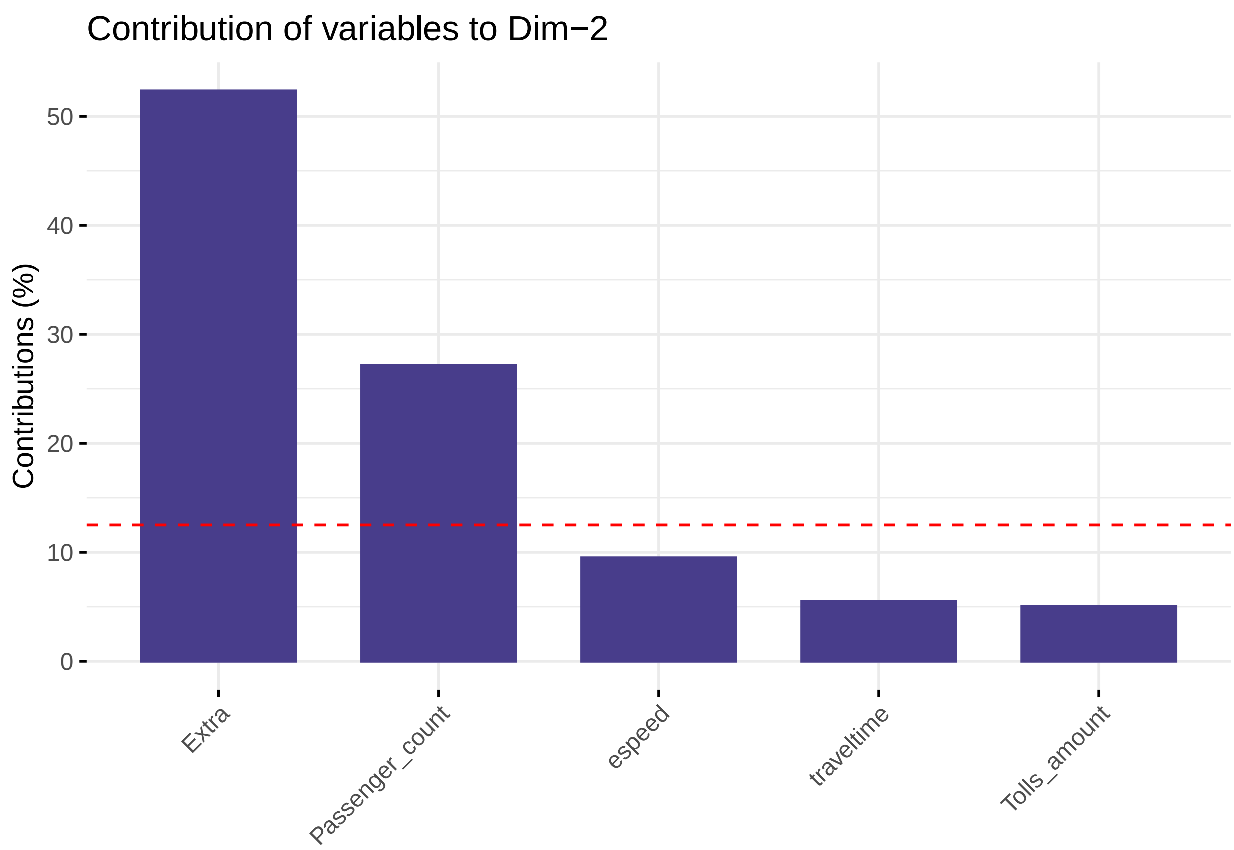
In the first dimension we see that for the **quantitative** variables the most positively related, from more to less, are: * Trip_distance (0.95) * Fare_amount (0.94) * Total_amount (0.93) * traveltime (0.80)

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.

2.3.2 Second dimension

```r
fviz_contrib(  # contributions of variables to PC1
  res.pca,
  fill = "darkslateblue",
  color = "darkslateblue",
  choice = "var",
  axes = 2,
  top = 5)
```




```r
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 "
```

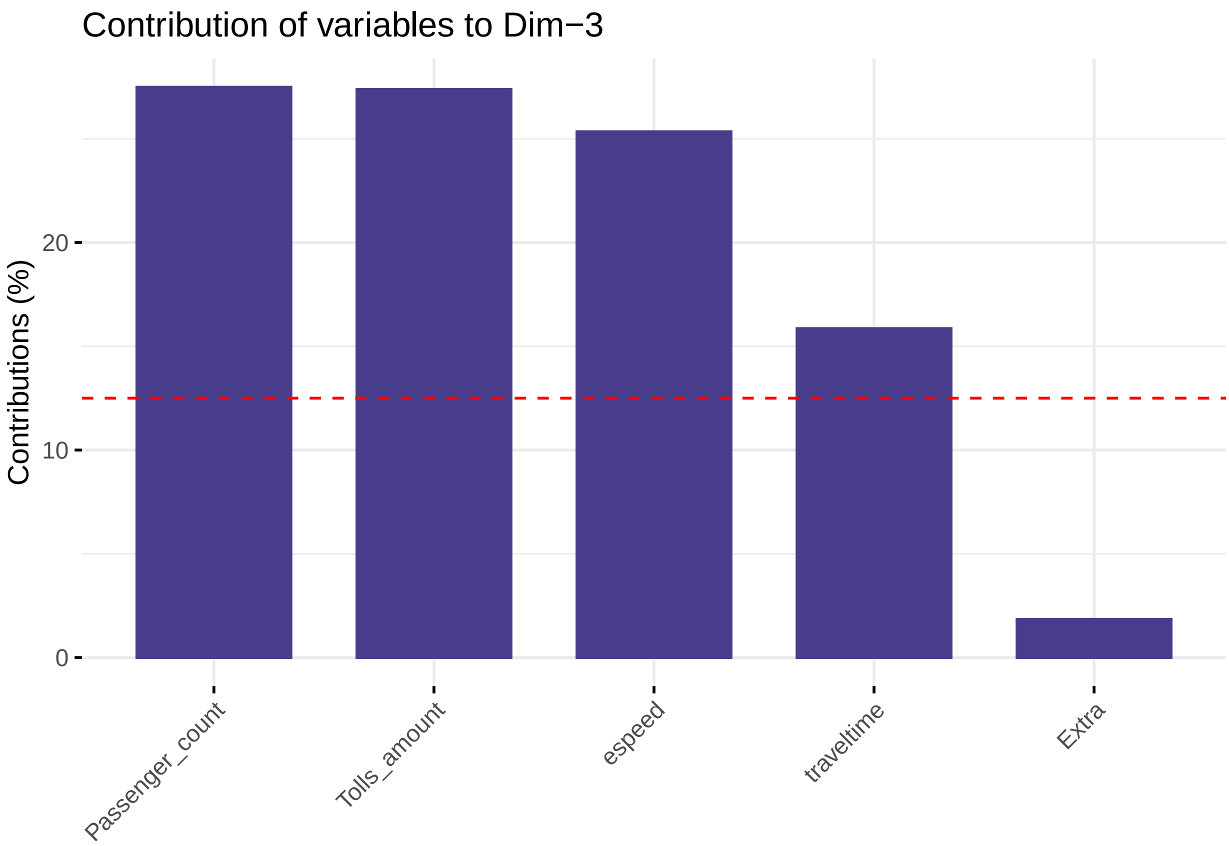
For the second dimension we see that or the **quantitative** variables Extra and Passenger_count are the most positively related ones with 0.74 and 0.53 respectively.

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.

2.3.3 Third dimension

```r
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.

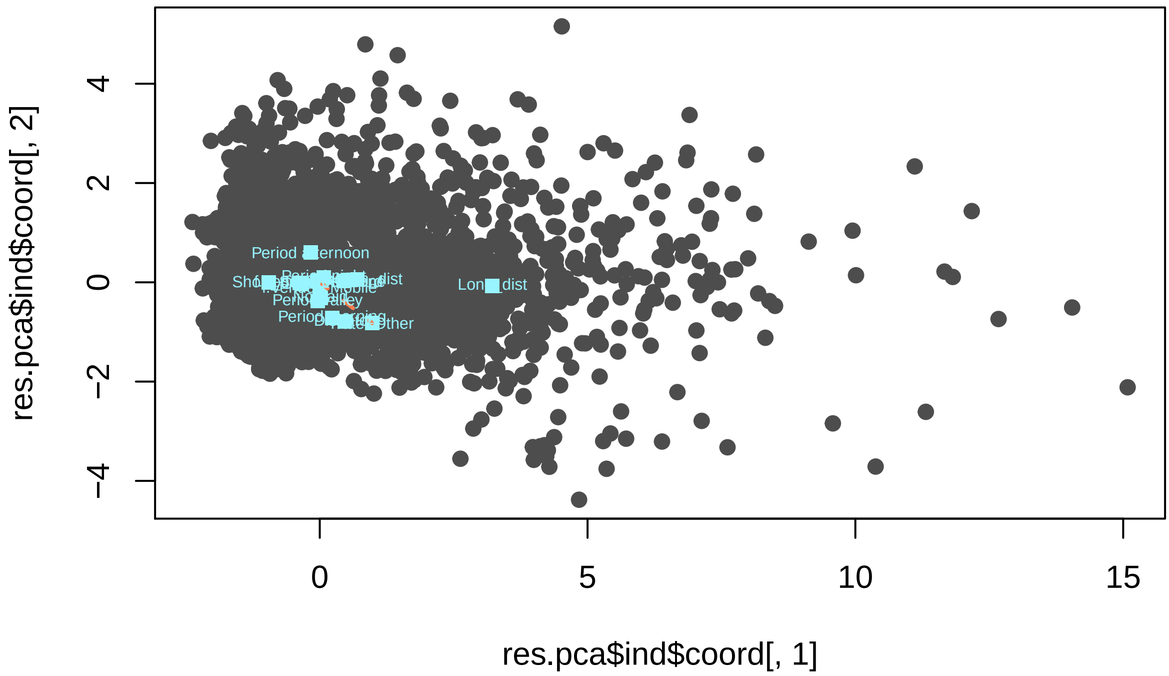
And we see that for this **category**, period afternoon mean is 0.28 units over the global mean and period valley mean, on the contrary, is -0.14 units under the global mean, hough it is not either a big relation.

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

2.4 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 associ
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 li
text(res.pca$quali.sup$coord[,1],res.pca$quali.sup$coord[,2],labels=names(res.pca$quali.sup$coord[,1]),c
```




```
res.pca$quali.sup$coord
```

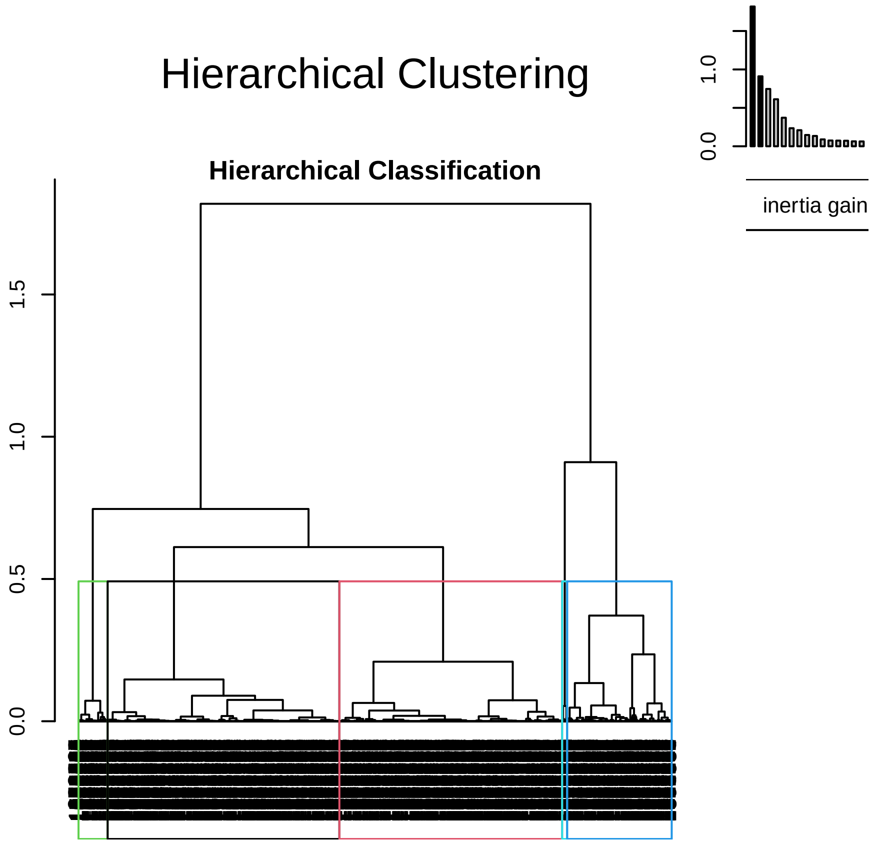
```
##                         Dim.1        Dim.2        Dim.3       Dim.4
## f.Vendor-Mobile   -0.004156948 -0.097911797 -0.065078791  0.10360028
## f.Vendor-VeriFone  0.001108140  0.026100871  0.017348401 -0.02761728
## Rate-1            -0.027703540  0.023224716 -0.002872324  0.01581731
## Rate-Other         0.980748959 -0.822191535  0.101684798 -0.55995764
## Credit card        0.448567849  0.023712582  0.069655549  0.19849333
## Cash              -0.376708753 -0.016140706 -0.056441297 -0.16514488
```

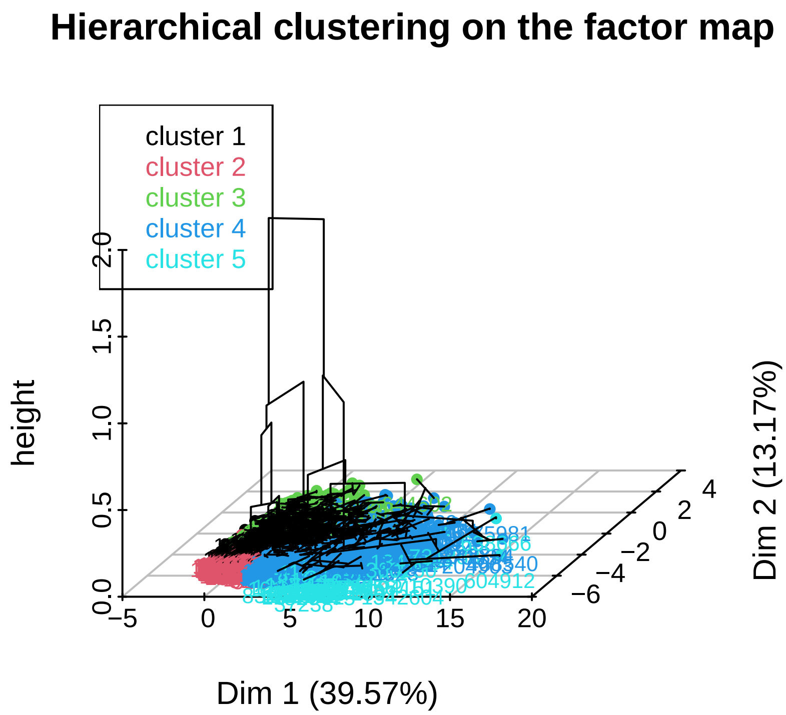
```
## No paid             0.014784804 -0.313274270 -0.168803729 -0.12250913
## Street-Hail         -0.011687857  0.019691230  0.001939789  0.01820524
## Dispatch             0.470749330 -0.793099463 -0.078128473 -0.73324858
## Period night         0.076291336  0.098881548  0.228743172  0.07154962
## Period morning       0.233587759 -0.715398722 -0.259847300 -0.41033341
## Period valley       -0.039783073 -0.380318373 -0.198651365 -0.29635439
## Period afternoon    -0.171118123  0.597611328  0.013182077  0.40570210
## Long_dist            3.224961311 -0.073035870  0.066607415 -0.17988023
## Medium_dist          0.702747017  0.059312533 -0.173420254 -0.02279226
## Short_dist          -0.954746915 -0.003335567  0.042630700  0.04781074
## No                  -0.340564204 -0.029130594 -0.053300310 -0.16235463
## Yes                  0.563759928  0.048221926  0.088231759  0.26875706
##                            Dim.5
## f.Vendor-Mobile     -0.0394669280
## f.Vendor-VeriFone    0.0105209098
## Rate-1              -0.0004798539
## Rate-Other           0.0169875844
## Credit card          0.0910111180
## Cash                -0.0724785949
## No paid             -0.3260083954
## Street-Hail          0.0023731798
## Dispatch            -0.0955840530
## Period night        -0.2573284053
## Period morning       0.4363196447
## Period valley        0.2527668547
## Period afternoon    -0.1123309948
## Long_dist           -0.3185982266
## Medium_dist         -0.0094686345
## Short_dist           0.0744293050
## No                  -0.0803119784
## Yes                  0.1329460780
```

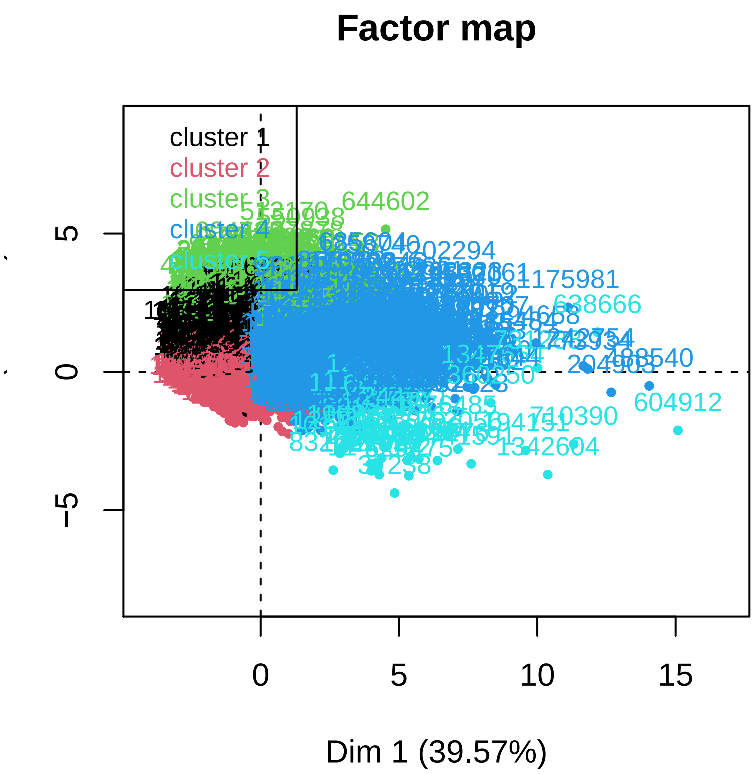
---

3 Hierarchical Clustering

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


Hierarchical Clustering
1.0
0.0
inertia gain
Hierarchical Classification
1.5
1.0
0.5
0.0


Hierarchical clustering on the factor map
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
height
2.0
1.5
1.0
0.5
0.0
−5
0
5
10
15
20
Dim 1 (39.57%)
4
2
0
−2
−4
−6
Dim 2 (13.17%)



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.

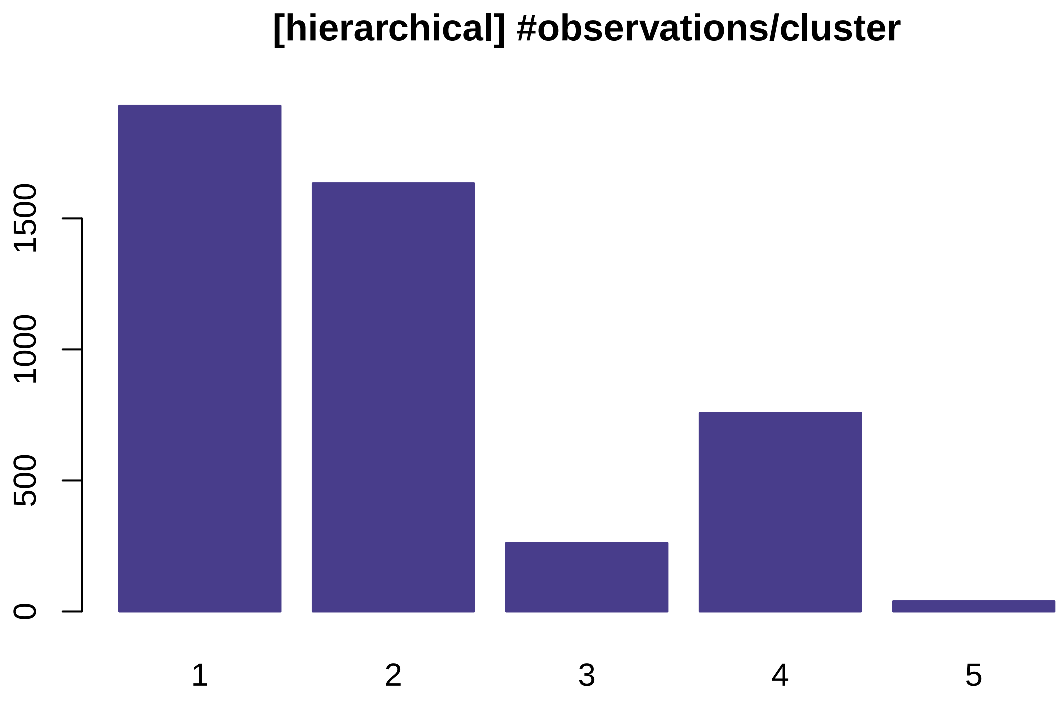
3.1 Description of clusters

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="[hierarchic
```



##Interpret the results of the classification

3.1.1 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 characterizes 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 variable that affects more to the clustering is **period** because is the one with the smallest p.value. The variables associated to the clusters are:

- period
- Trip_distance_range
- TipIsGiven
- Paymnet_type
- VendorID

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 Cluster 3 Cluster 4 Cluster 5** res.hcpc*desc.var*quanti.var # quantitative variables which characterizes the clusters res.hcpc*desc.var*quanti # description of each cluster by the quantitative variables

```
### The description of the clusters by the axes
It doens't help that much to identify the characteristics of each cluster.
!!! Segons ella, diu que no és important, que no creu que aporti res.

```r
dim(res.hcpc$data.clust)
```

```
## [1] 4623   21
```

```r
# catdes(res.hcpc$data.clust,21) this is to justify the content of the description
names(res.hcpc$desc.axes)
```

```
## [1] "quanti.var" "quanti"     "call"
```

```r
res.hcpc$desc.axes$quanti.var # ?
```

```
##            Eta2 P-value
## Dim.1 0.6542388       0
## Dim.2 0.5837298       0
## Dim.3 0.5228059       0
## Dim.4 0.6831260       0
## Dim.5 0.4740890       0
```

```r
res.hcpc$desc.axes$quanti     # principal dimensions that are the most associated with clusters
```

```
## $`1`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.4  36.894986       0.62608435  6.735165e-14      0.4891709  0.9766545
## Dim.2  25.601622       0.45664540 -1.654322e-14      0.6497623  1.0265664
## Dim.3   2.735761       0.04846122 -3.557521e-14      0.5280745  1.0195101
## Dim.5 -20.005331      -0.32921814 -2.666411e-14      0.4968160  0.9471384
## Dim.1 -22.826491      -0.70563827 -3.297048e-15      0.7491891  1.7791740
##             p.value
## Dim.4 5.560905e-298
## Dim.2 1.463416e-144
## Dim.3  6.223614e-03
## Dim.5  4.948907e-89
## Dim.1 2.502574e-115
##
## $`2`
##         v.test Mean in category  Overall mean sd in category Overall sd
## Dim.5  17.19818        0.3240544 -2.666411e-14      0.5137420  0.9471384
## Dim.3 -15.05803       -0.3054089 -3.557521e-14      0.5480496  1.0195101
## Dim.1 -17.86040       -0.6321665 -3.297048e-15      0.7787471  1.7791740
## Dim.4 -21.19665       -0.4118415  6.735165e-14      0.2881193  0.9766545
## Dim.2 -39.94197       -0.8157151 -1.654322e-14      0.3474022  1.0265664
##            p.value
## Dim.5 2.740050e-66
## Dim.3 3.057482e-51
## Dim.1 2.399338e-71
## Dim.4 1.025406e-99
## Dim.2 0.000000e+00
##
## $`3`
```
```

```
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.2  33.38936        2.0569445 -1.654322e-14      0.8628949  1.0265664
## Dim.3  30.55804        1.8695818 -3.557521e-14      0.7531158  1.0195101
## Dim.5  13.14483        0.7471295 -2.666411e-14      0.7485736  0.9471384
## Dim.1  -2.52769       -0.2698793 -3.297048e-15      1.2568926  1.7791740
## Dim.4 -36.81264       -2.1575722  6.735165e-14      0.7796728  0.9766545
##             p.value
## Dim.2 1.956593e-244
## Dim.3 4.421861e-205
## Dim.5  1.822110e-39
## Dim.1  1.148157e-02
## Dim.4 1.159038e-296
##
## $`4`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1  49.941195        2.9512265 -3.297048e-15      1.7274782  1.7791740
## Dim.4  -5.662788       -0.1836946  6.735165e-14      0.8750428  0.9766545
## Dim.3 -12.098664       -0.4096889 -3.557521e-14      1.1826753  1.0195101
## Dim.5 -13.238848       -0.4164749 -2.666411e-14      1.1865652  0.9471384
##            p.value
## Dim.1 0.000000e+00
## Dim.4 1.489331e-08
## Dim.3 1.073435e-33
## Dim.5 5.234580e-40
##
## $`5`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.5  38.33727         5.790414 -2.666411e-14       1.233806  0.9471384
## Dim.3  35.67823         5.800559 -3.557521e-14       1.189339  1.0195101
## Dim.4  27.84466         4.336685  6.735165e-14       1.242127  0.9766545
## Dim.1  20.65225         5.859503 -3.297048e-15       2.818065  1.7791740
## Dim.2 -15.32598        -2.508940 -1.654322e-14       1.178332  1.0265664
##             p.value
## Dim.5  0.000000e+00
## Dim.3 8.603348e-279
## Dim.4 1.250219e-170
## Dim.1  9.318300e-95
## Dim.2  5.127836e-53
```

3.1.2 The description of the clusters by the individuals

```r
names(res.hcpc$desc.ind)
```

```
## [1] "para" "dist"
```

```r
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
```

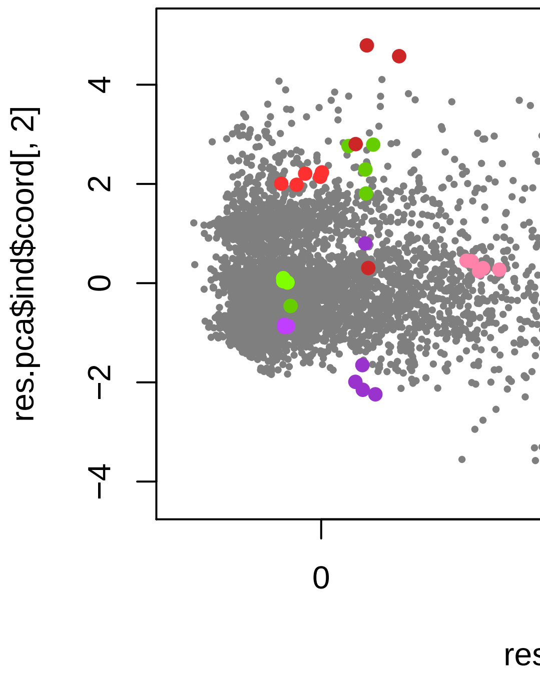
```
## 1261276  1016299   327762  1010826   529475
## 1.151077 1.224596 1.305726 1.472585 1.482492
```

```r
res.hcpc$desc.ind$dist  # ?
```

```
## 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
```

```r
# 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="chartreuse",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="darkorchid1",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="firebrick1",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="palevioletred1",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="royalblue1",cex=1,pch=16)
points(res.pca$ind$coord[dist5,1],res.pca$ind$coord[dist5,2],col="royalblue3",cex=1,pch=16)
```



3.1.2.1 Examine the values of individuals that characterize classes

3.1.3 Partition quality

```
((res.hcpc$call$t$within[1]-res.hcpc$call$t$within[5])/res.hcpc$call$t$within[1])*100
```

3.1.3.1 Gain in inertia (in %)

```
## [1] 57.49171
```

```
((res.hcpc$call$t$within[1]-res.hcpc$call$t$within[1:50])/res.hcpc$call$t$within[1])*100
```

3.1.3.2 Per assolir una representetivitat de clustering del 80% necessitem…

```
##  [1]  0.00000 25.58180 38.38958 48.88272 57.49171 62.71420 66.02096 68.96350
##  [9] 71.02825 72.91535 74.17668 75.22981 76.26582 77.27876 78.18306 79.06611
## [17] 79.84616 80.59951 81.27272 81.91954 82.45480 82.98288 83.46113 83.92761
## [25] 84.37742 84.80262 85.13118 85.45794 85.77559 86.06950 86.33585 86.59220
## [33] 86.84304 87.08620 87.31737 87.54760 87.75821 87.96757 88.17583 88.38194
## [41] 88.58074 88.76754 88.94710 89.11580 89.28410 89.44633 89.60389 89.76073
## [49] 89.90790 90.04816
```

…18 clusters.

```
names(res.hcpc$call$t)           # results for the hierarchical tree
```

3.1.3.3 Hierarchical tree

```
## [1] "res"        "tree"       "nb.clust"   "within"     "inert.gain"
## [6] "quot"
```

```
res.hcpc$call$t$nb.clust         # the suggested level to cut the tree
```

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

```
res.hcpc$call$t$within[1:5]      # within inertias
```

```
## [1] 7.109625 5.290855 4.380269 3.634247 3.022180
```

```r
res.hcpc$call$t$quot[1:5]          # ratio between within inertias
```

```
## [1] 0.8278944 0.8296858 0.8315835 0.8771419 0.9113131
```

```r
res.hcpc$call$t$inert.gain[1:5]   # inertia gain
```

```
## [1] 1.8187697 0.9105858 0.7460223 0.6120673 0.3712993
```

3.1.4 Save the results into dataframe

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

---

4 K-Means Classification

4.1 Description of clusters

```r
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
dim(ppcc)
```

```
## [1] 4623    3
```

4.1.1 Optimal number of clusters

```r
# library("factoextra")
# fviz_nbclust(ppcc, kmeans, method = "gap_stat")
# no funciona bé --> s'ha de repassar
```

According to the previous plot, the optimal number of clusters per k-means is ???.

4.1.2 Whatever

```r
# library("NbClust") # It takes a lot ....
# set.seed(123)
# res.nbclust <- NbClust(ppcc, distance = "euclidean",
#                   min.nc = 2, max.nc = 10,
#                   method = "complete", index ="all") # Time consuming
# # time consuming su madre, porto literal 10 min executant-lo i segueix igual
```

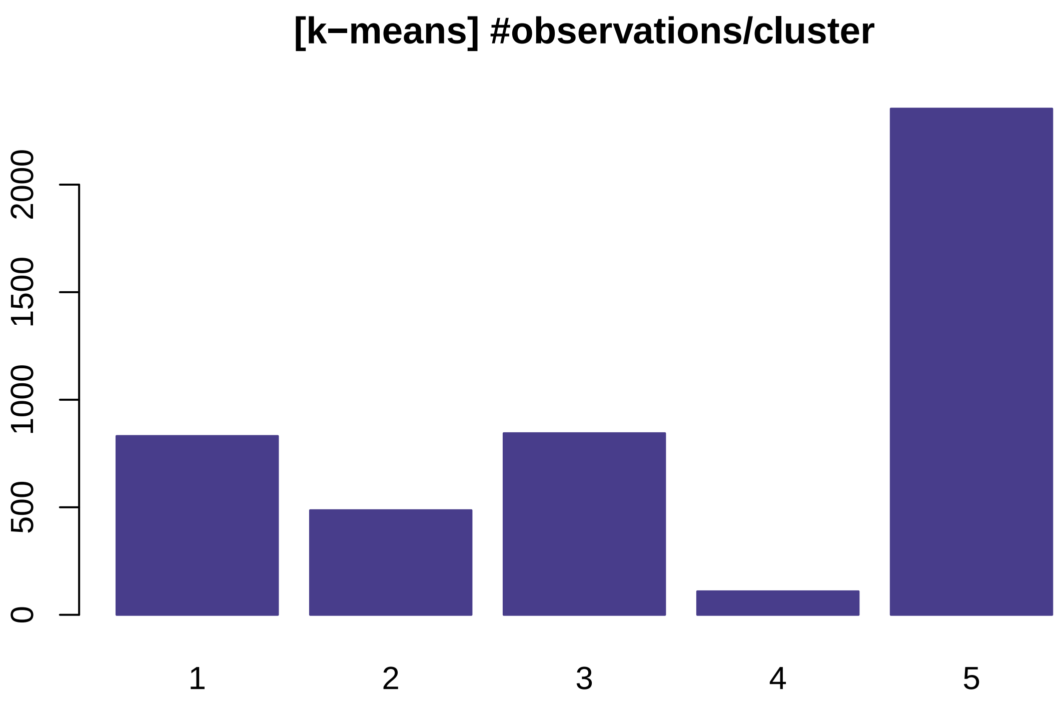
4.2 Classification

```r
dist<-dist(ppcc)  # coordenades són reals - Euclidea
kc<-kmeans(dist, centers=5, iter.max=30, trace=TRUE)
```

```
## 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
```

```
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"
)
```



4.2.1 Gain in inertia (in %)

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

```
## [1] 79.40953
```

4.2.2 Comparison of clusters

```
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
```

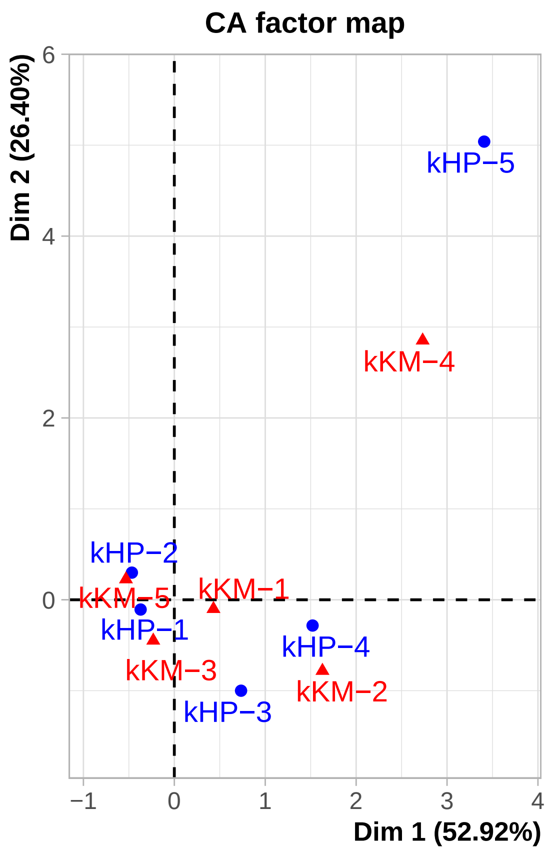
```
100*sum(diag(tt)/sum(tt))
```

```
## [1] 54.72637
```

---

5 CA analysis

```
res.ca <- CA(tt)
```


CA factor map
Dim 2 (26.40%)
Dim 1 (52.92%)
kHP−5
kKM−4
kHP−2
kKM−5
kKM−1
kHP−1
kHP−4
kKM−3
kKM−2
kHP−3
6
4
2
0
−1
0
1
2
3
4

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

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

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

7 MCA analysis for your data should contain:

7.1 Eigenvalues and dominant axes analysis. How many axes we have to consider for next Hierarchical Classification stage?

7.2 Individuals point of view: Are they any individuals “too contributive”? Are there any groups?

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

7.4 Interpreting the axes association to factor map.

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

8 Hierarchical Clustering (from MCA)

8.1 Description of clusters

8.2 Parangons and class-specific individuals.

8.3 Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA) focusing on Duration target.

8.4 Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA) focusing on the binary target.