PCA analysis of NYCABS datase

Katerina Dimitrova, Jose Romero, Sergi Munoz

March 18, 2018

Previous work

Load requiered packages

PCA analysis

1. The Kaiser rule is to drop all components with eigenvalues under 1.0 According to the Elbow rule when the drop ceases and the curve makes an elbow toward less steep declinewe should drop all further components after the one starting the elbow.

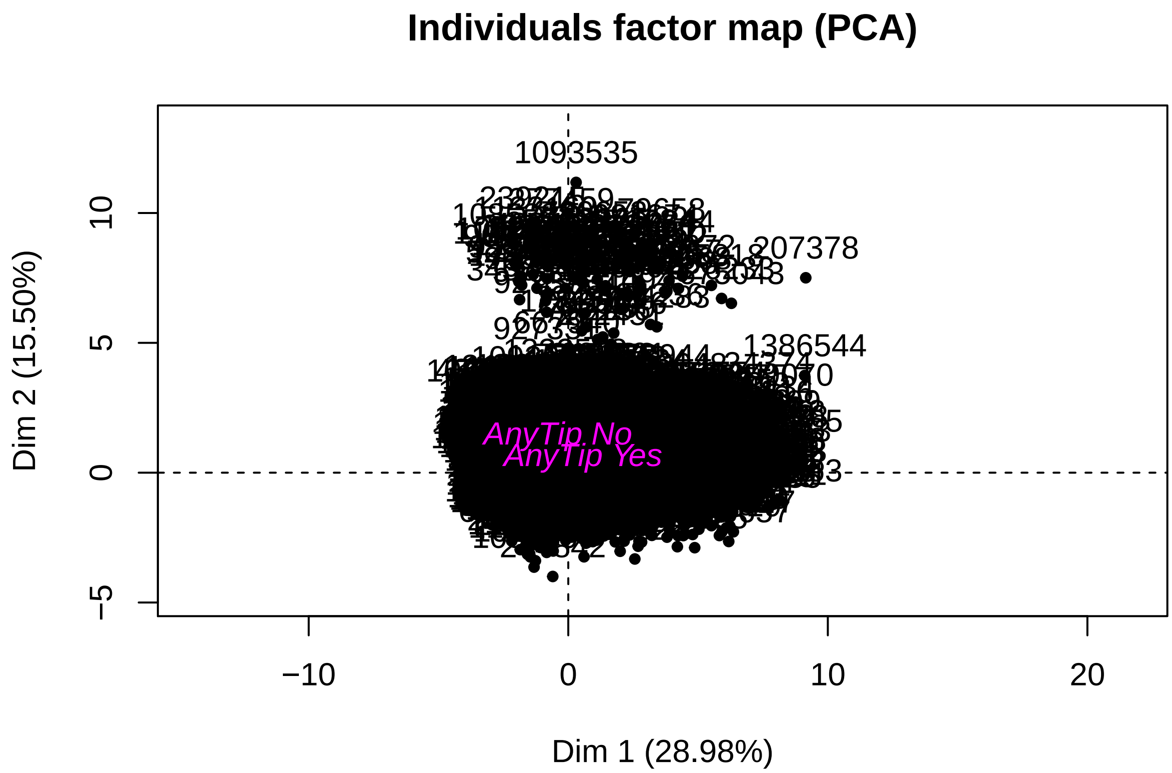
I. I. Eigenvalues and axes

For the PCA analysis we take all numerical variables as active, where TotalAmount and Anytip are suplementary.

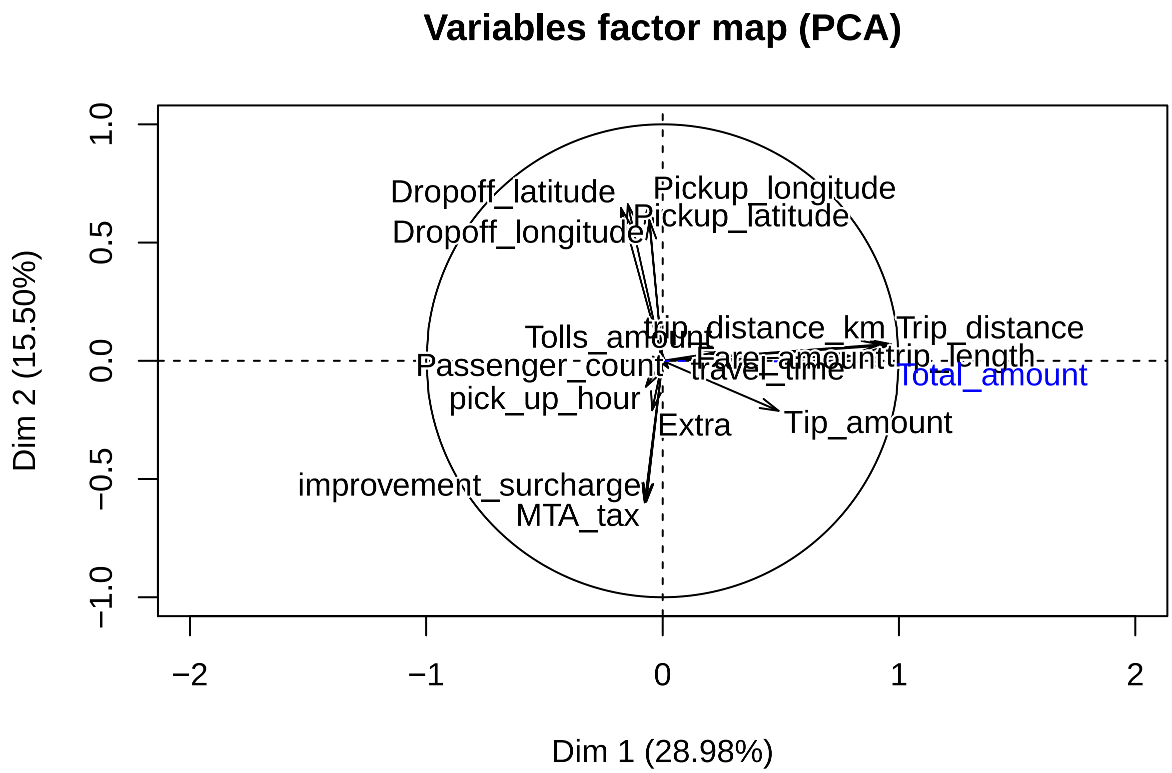
```
load("Taxi5000_raw_DataClean.RData")
library(FactoMineR)
names (df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"
```

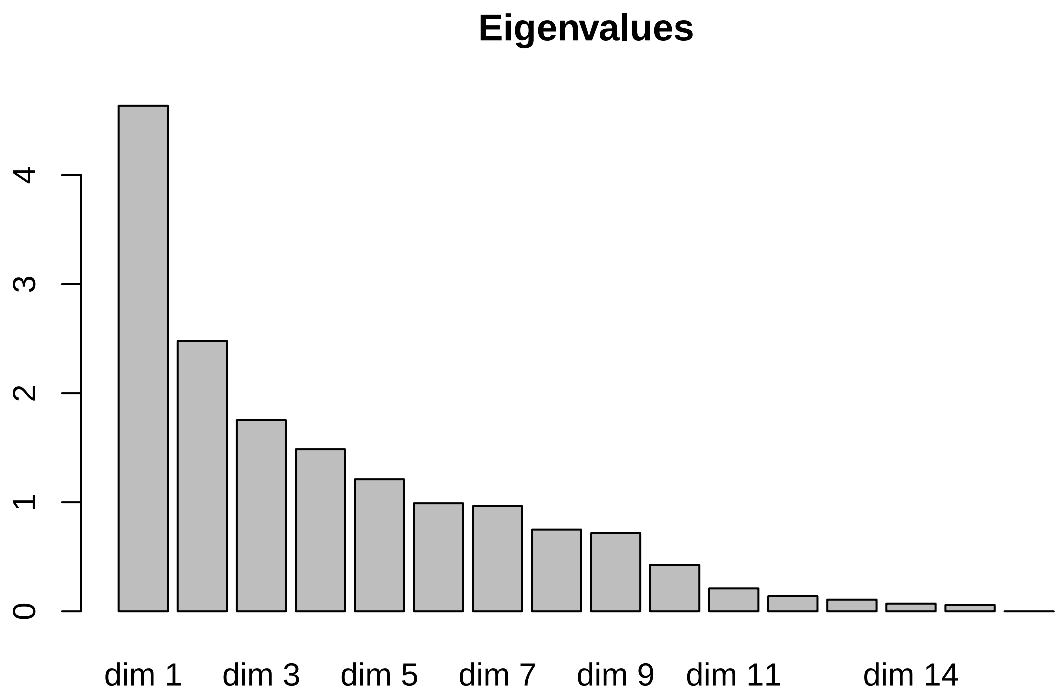
```
vars_con_pca<-c(6,7,8,9,10,11,12,13,14,15,16,17,18,22,23,24,25,26)

#From te plot we see that the variables "Trip_distance", "Trip_length", "Travel_time" and "Fare_amount"
res.pca<-PCA(df[,vars_con_pca], quanti.sup = 13, quali.sup = 14, ncp = 6 ) # TotalAmount and AnyTip
```



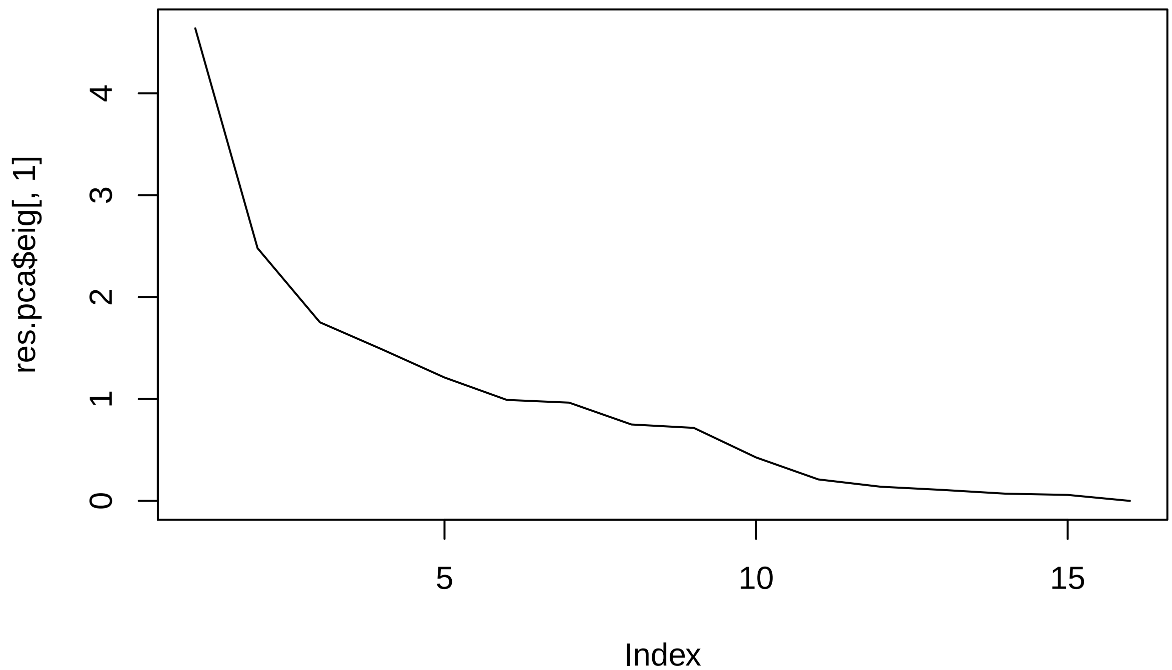
Individuals factor map (PCA)
Dim 2 (15.50%)
Dim 1 (28.98%)
AnyTip.No
AnyTip.Yes
1093535
207378
1386544




```
barplot(res.pca$eig[,1], main="Eigenvalues", names.arg = paste("dim", 1:nrow(res.pca$eig)))
```



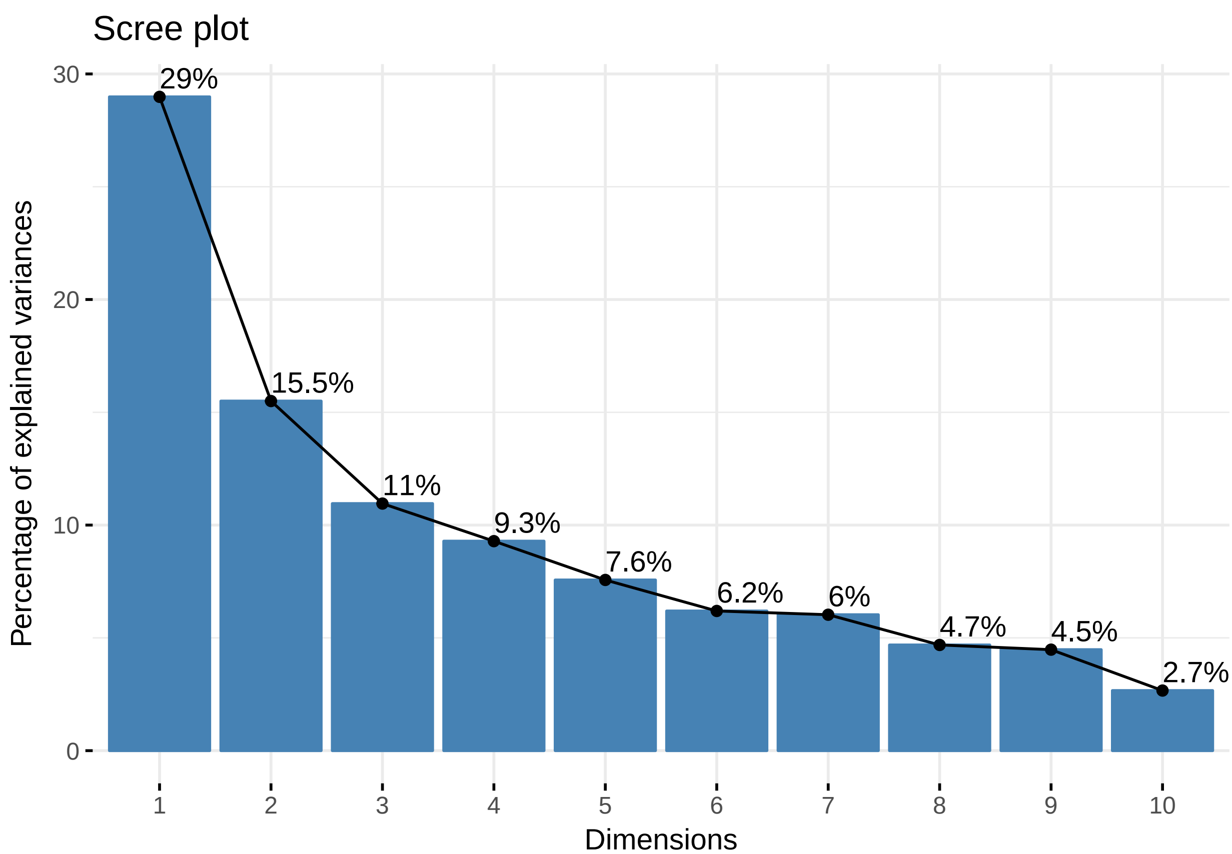

```r
# With the PCA transformation the PC1 covers 29% of the variance, PC2 - 15,5%, PCA3 - 11%, PCA4 - 9,3%
plot(res.pca$eig[,1], type = "l") # line chart
```



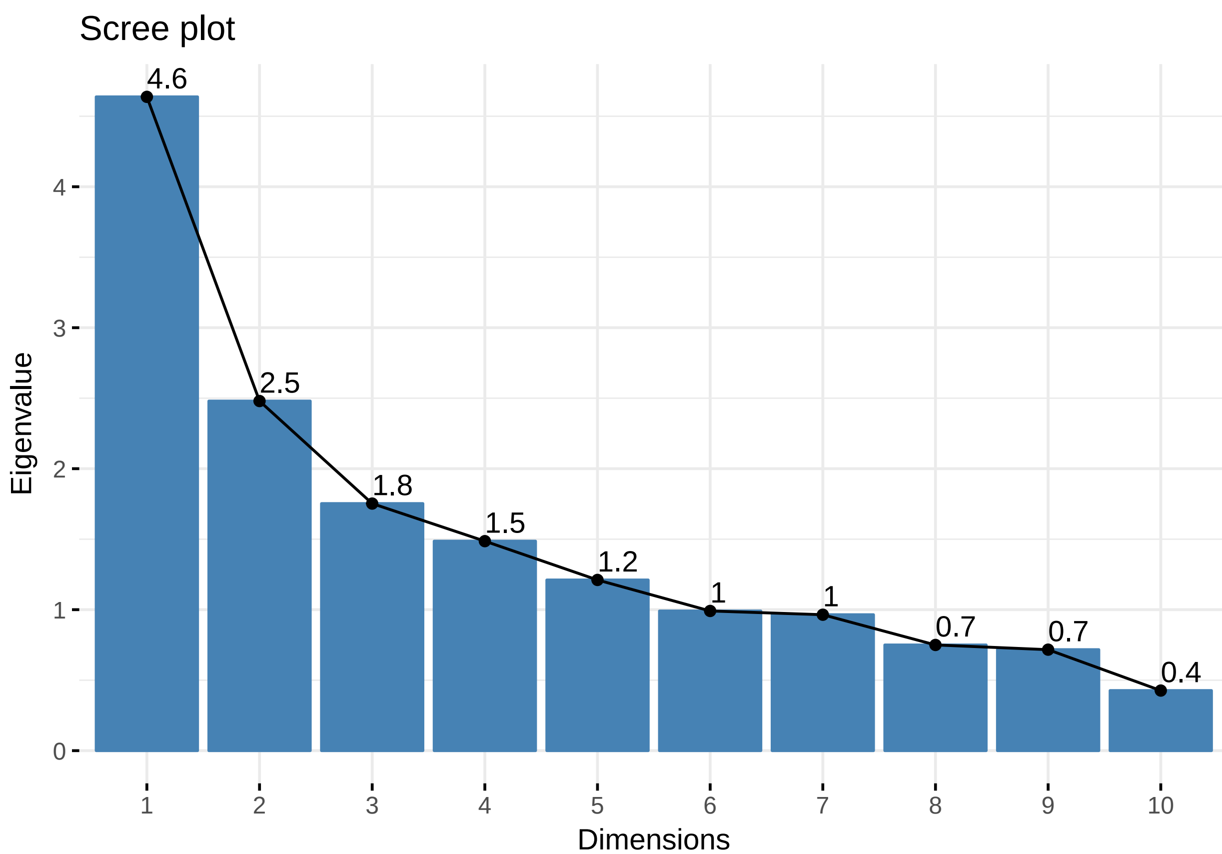

```
length <-length(which(res.pca$eig[,1]>=1));length
```

```
## [1] 5
```

```
kaiser <- res.pca$eig[1:length,1] #keep only EV >=1 ->first 7
#If we use the kaiser rule we have to keep all EV greater than 1, which results in saving the first 6 d
#facto extra
fviz_eig(res.pca, addlabels = TRUE)
```




```r
fviz_eig(res.pca, choice = "eigenvalue",addlabels = TRUE)
```




```
#According to the elbow rule we have to take the first 6 dimentions as the slope of the graphic shows.
elbow <- kaiser
```

II. Individuals point of view

Look at variables that are too contributive

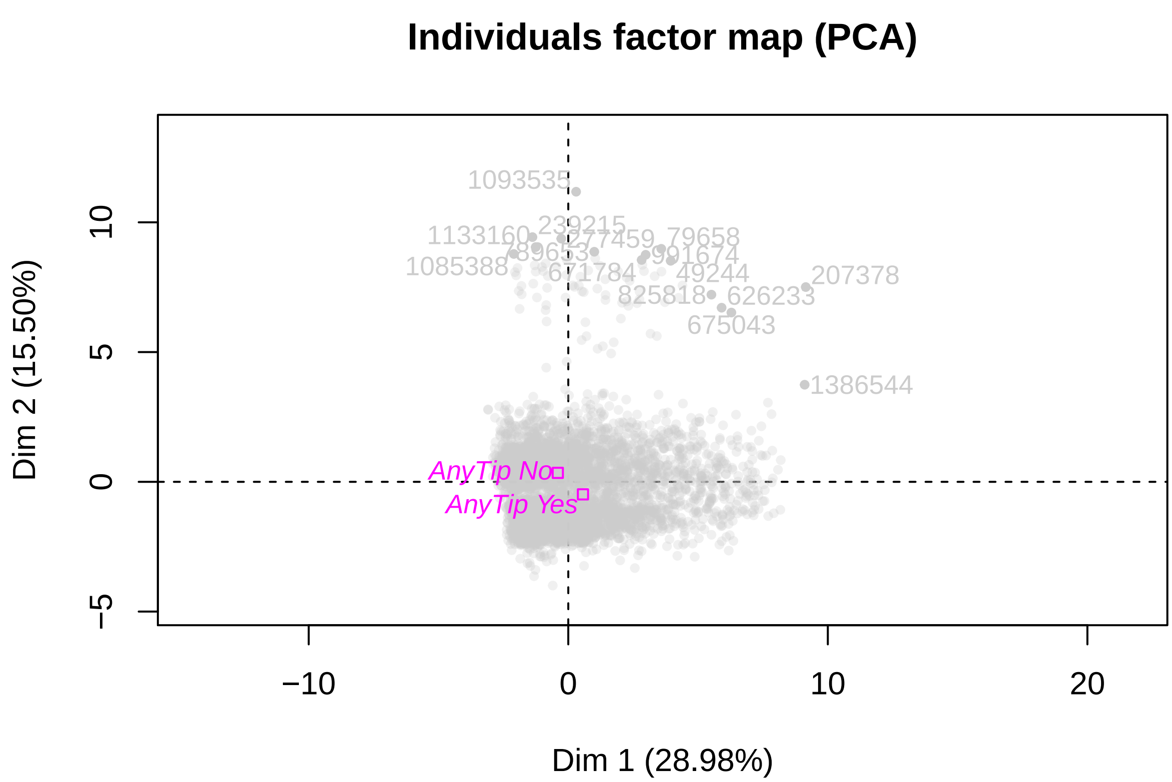
```
summary(res.pca, dig = 2, nbelements = 17, nbind=3, ncp=4)
```

```
##
## Call:
## PCA(X = df[, vars_con_pca], ncp = 6, quanti.sup = 13, quali.sup = 14)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               4.638   2.480   1.753   1.486   1.211   0.991
## % of var.             28.984  15.499  10.954   9.288   7.568   6.193
## Cumulative % of var.  28.984  44.483  55.437  64.725  72.293  78.485
##                        Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance               0.964   0.750   0.716   0.426   0.210   0.139
## % of var.              6.026   4.686   4.478   2.664   1.315   0.868
## Cumulative % of var.  84.512  89.198  93.675  96.339  97.654  98.522
##                       Dim.13  Dim.14  Dim.15  Dim.16
## Variance               0.107   0.071   0.058   0.000
```

```
## % of var.              0.670   0.442   0.365   0.000
## Cumulative % of var.  99.192  99.635 100.000 100.000
##
## Individuals (the 3 first)
##                          Dist    Dim.1    ctr   cos2    Dim.2    ctr
## 285                   |  3.346 |  1.366  0.008  0.167 | -0.018  0.000
## 307                   |  3.299 |  1.648  0.012  0.249 |  0.833  0.006
## 401                   |  2.613 |  0.939  0.004  0.129 | -0.259  0.001
##                        cos2    Dim.3    ctr   cos2    Dim.4    ctr   cos2
## 285                    0.000 |  0.066  0.000  0.000 |  1.838  0.047  0.302
## 307                    0.064 |  0.839  0.008  0.065 | -0.398  0.002  0.015
## 401                    0.010 |  0.007  0.000  0.000 |  0.383  0.002  0.021
##
## 285                   |
## 307                   |
## 401                   |
##
## Variables
##                         Dim.1    ctr   cos2    Dim.2    ctr   cos2
## Pickup_longitude      | -0.063  0.086  0.004 |  0.660 17.558  0.435 |
## Pickup_latitude       | -0.148  0.469  0.022 |  0.663 17.705  0.439 |
## Dropoff_longitude     | -0.055  0.066  0.003 |  0.593 14.203  0.352 |
## Dropoff_latitude      | -0.176  0.670  0.031 |  0.646 16.838  0.418 |
## Passenger_count       |  0.024  0.013  0.001 | -0.030  0.037  0.001 |
## Trip_distance         |  0.965 20.099  0.932 |  0.070  0.199  0.005 |
## Fare_amount           |  0.960 19.853  0.921 |  0.066  0.176  0.004 |
## Extra                 | -0.044  0.043  0.002 | -0.209  1.765  0.044 |
## MTA_tax               | -0.076  0.124  0.006 | -0.599 14.447  0.358 |
## Tip_amount            |  0.491  5.193  0.241 | -0.212  1.805  0.045 |
## Tolls_amount          |  0.234  1.176  0.055 |  0.036  0.052  0.001 |
## improvement_surcharge | -0.069  0.103  0.005 | -0.596 14.321  0.355 |
## trip_length           |  0.922 18.340  0.851 |  0.069  0.191  0.005 |
## trip_distance_km      |  0.965 20.099  0.932 |  0.070  0.199  0.005 |
## travel_time           |  0.793 13.557  0.629 |  0.020  0.016  0.000 |
## pick_up_hour          | -0.071  0.108  0.005 | -0.110  0.488  0.012 |
##                        Dim.3    ctr   cos2    Dim.4    ctr   cos2
## Pickup_longitude       0.306  5.349  0.094 |  0.572 21.996  0.327 |
## Pickup_latitude        0.418  9.953  0.174 | -0.513 17.683  0.263 |
## Dropoff_longitude      0.287  4.688  0.082 |  0.663 29.570  0.439 |
## Dropoff_latitude       0.415  9.836  0.172 | -0.527 18.706  0.278 |
## Passenger_count        0.026  0.039  0.001 |  0.112  0.845  0.013 |
## Trip_distance          0.078  0.350  0.006 |  0.007  0.003  0.000 |
## Fare_amount            0.042  0.102  0.002 |  0.003  0.001  0.000 |
## Extra                  0.129  0.951  0.017 |  0.325  7.128  0.106 |
## MTA_tax                0.763 33.210  0.582 |  0.022  0.034  0.001 |
## Tip_amount             0.013  0.010  0.000 | -0.145  1.407  0.021 |
## Tolls_amount           0.138  1.090  0.019 | -0.161  1.739  0.026 |
## improvement_surcharge  0.767 33.530  0.588 |  0.037  0.091  0.001 |
## trip_length            0.088  0.439  0.008 |  0.032  0.068  0.001 |
## trip_distance_km       0.078  0.350  0.006 |  0.007  0.003  0.000 |
## travel_time           -0.034  0.067  0.001 | -0.011  0.008  0.000 |
## pick_up_hour          -0.025  0.036  0.001 |  0.103  0.717  0.011 |
##
## Supplementary continuous variable
```

```
##                             Dim.1    cos2    Dim.2    cos2    Dim.3    cos2
## Total_amount              |  0.966   0.933 | -0.004   0.000 |  0.068   0.005 |
##                           Dim.4    cos2
## Total_amount             -0.029  0.001 |
##
## Supplementary categories
##                               Dist      Dim.1     cos2  v.test     Dim.2
## AnyTip No                 |   0.742 |  -0.404    0.296 -15.479 |   0.346
## AnyTip Yes                |   1.040 |   0.566    0.296  15.479 |  -0.484
##                             cos2   v.test     Dim.3     cos2  v.test     Dim.4
## AnyTip No                  0.217   18.114 |   0.031    0.002   1.951 |   0.168
## AnyTip Yes                 0.217 -18.114 |  -0.044    0.002  -1.951 |  -0.235
##                             cos2   v.test
## AnyTip No                  0.051   11.350 |
## AnyTip Yes                 0.051 -11.350 |
```

```
#The summary confirms the correlations between the variables that we already interpreted from the plots
#The plot show us that individuals that had to pay more tend to leave a tip.
plot.PCA(res.pca, choix=c("ind"),cex=0.8,col.ind="grey80",select="contrib15",axes=c(1,2))
```




```
#DIMENSION1
#Since the multivariant detection didnt manage to find outliers well enogh we are going to obtain them

#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,1])
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## -3.0923 -1.5874 -0.6932  0.0000  1.0294  9.1502
```

```r
iqrvar<-IQR(res.pca$ind$coord[,1])
quantil3<-quantile(res.pca$ind$coord[,1], .75);quantil3 #get 3rd quartile
```

```
##      75%
## 1.029432
```

```r
outliers<-which(res.pca$ind$coord[,1]>(iqrvar*3)+quantil3);length(outliers)
```

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

```r
df$f.outlierPCAd1<-0
df[outliers,"f.outlierPCAd1"]<-1
df$f.outlierPCAd1<-factor(df$f.outlierPCAd1,labels=c("NoOutDim1", "YesOutDim1"))
summary(df$f.outlierPCAd1)
```

```
##  NoOutDim1 YesOutDim1
##       4864          2
```

```r
names(df)
```

```
##  [1] "VendorID"                 "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"    "Store_and_fwd_flag"
##  [5] "RateCodeID"               "Pickup_longitude"
##  [7] "Pickup_latitude"          "Dropoff_longitude"
##  [9] "Dropoff_latitude"         "Passenger_count"
## [11] "Trip_distance"            "Fare_amount"
## [13] "Extra"                    "MTA_tax"
## [15] "Tip_amount"               "Tolls_amount"
## [17] "improvement_surcharge"    "Total_amount"
## [19] "Payment_type"             "Trip_type"
## [21] "mis_ind"                  "AnyTip"
## [23] "trip_length"              "trip_distance_km"
## [25] "travel_time"              "pick_up_hour"
## [27] "pick_up_period"           "espeed"
## [29] "f.passenger"              "f.distance"
## [31] "f.pickup_longitude"       "f.pickup_latitude"
## [33] "f.dropoff_longitude"      "f.dropoff_latitude"
## [35] "f.fare_amount"            "f.extra"
## [37] "f.MTA_tax"                "f.Improvement_surcharge"
## [39] "f.tip_amount"             "f.toll"
## [41] "f.total"                  "f.outlierPCAd1"
```

```r
#catdes(,names(df)[c(22)])

#DIMENSION2
#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9975 -1.1969  0.2116  0.0000  0.8072 11.1786
```

```r
iqrvar<-IQR(res.pca$ind$coord[,2])
quantil3<-quantile(res.pca$ind$coord[,2], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.8072157
```

```r
outliers2<-which(res.pca$ind$coord[,2]>(iqrvar*3)+quantil3);length(outliers2)
```

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

```r
df$f.outlierPCAd2<-0
df[outliers2,"f.outlierPCAd2"]<-1
df$f.outlierPCAd2<-factor(df$f.outlierPCAd2,labels=c("NoOutDim2", "YesOutDim2"))
summary(df$f.outlierPCAd2)
```

```
##  NoOutDim2 YesOutDim2
##       4806         60
```

```r
names(df)
```

```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.outlierPCAd1"
## [43] "f.outlierPCAd2"
```

```r
#DIMENSION3
#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,3])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.9758  -0.5999   0.3556   0.0000   0.6596   3.0814
```

```r
iqrvar<-IQR(res.pca$ind$coord[,3])
quantil3<-quantile(res.pca$ind$coord[,3], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.6595931
```

```r
outliers3<-which(res.pca$ind$coord[,3]>(iqrvar*3)+quantil3);length(outliers3)
```

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

```r
df$f.outlierPCAd3<-0
df$f.outlierPCAd3<-factor(df$f.outlierPCAd3,labels=c("NoOutDim3"))
summary(df$f.outlierPCAd3)
```

```
## NoOutDim3
##      4866
```

```r
names(df)
```

```
##  [1] "VendorID"                 "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"    "Store_and_fwd_flag"
##  [5] "RateCodeID"               "Pickup_longitude"
##  [7] "Pickup_latitude"          "Dropoff_longitude"
##  [9] "Dropoff_latitude"         "Passenger_count"
## [11] "Trip_distance"            "Fare_amount"
## [13] "Extra"                    "MTA_tax"
## [15] "Tip_amount"               "Tolls_amount"
## [17] "improvement_surcharge"    "Total_amount"
## [19] "Payment_type"             "Trip_type"
## [21] "mis_ind"                  "AnyTip"
## [23] "trip_length"              "trip_distance_km"
## [25] "travel_time"              "pick_up_hour"
## [27] "pick_up_period"           "espeed"
## [29] "f.passenger"              "f.distance"
## [31] "f.pickup_longitude"       "f.pickup_latitude"
## [33] "f.dropoff_longitude"      "f.dropoff_latitude"
## [35] "f.fare_amount"            "f.extra"
## [37] "f.MTA_tax"                "f.Improvement_surcharge"
## [39] "f.tip_amount"             "f.toll"
## [41] "f.total"                  "f.outlierPCAd1"
## [43] "f.outlierPCAd2"           "f.outlierPCAd3"
```

```r
#DIMENSION4
#characteristic of extreme otliers in dim1
summary(res.pca$ind$coord[,4])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -5.07857 -0.95074 -0.09769  0.00000  0.69112  4.57227
```

```r
iqrvar<-IQR(res.pca$ind$coord[,4])
quantil3<-quantile(res.pca$ind$coord[,4], .75);quantil3 #get 3rd quartile
```

```
##       75%
## 0.6911202
```

```r
outliers4<-which(res.pca$ind$coord[,4]>(iqrvar*3)+quantil3);length(outliers4)
```

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

```r
df$f.outlierPCAd4<-0
df$f.outlierPCAd4<-factor(df$f.outlierPCAd4,labels=c("NoOutDim4"))
summary(df$f.outlierPCAd4)
```

```
## NoOutDim4
##      4866
```

```r
names(df)
```

```
##  [1] "VendorID"                 "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"    "Store_and_fwd_flag"
##  [5] "RateCodeID"               "Pickup_longitude"
##  [7] "Pickup_latitude"          "Dropoff_longitude"
##  [9] "Dropoff_latitude"         "Passenger_count"
```

```
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [15] "Tip_amount"              "Tolls_amount"
## [17] "improvement_surcharge"   "Total_amount"
## [19] "Payment_type"            "Trip_type"
## [21] "mis_ind"                 "AnyTip"
## [23] "trip_length"             "trip_distance_km"
## [25] "travel_time"             "pick_up_hour"
## [27] "pick_up_period"          "espeed"
## [29] "f.passenger"             "f.distance"
## [31] "f.pickup_longitude"      "f.pickup_latitude"
## [33] "f.dropoff_longitude"     "f.dropoff_latitude"
## [35] "f.fare_amount"           "f.extra"
## [37] "f.MTA_tax"               "f.Improvement_surcharge"
## [39] "f.tip_amount"            "f.toll"
## [41] "f.total"                 "f.outlierPCAd1"
## [43] "f.outlierPCAd2"          "f.outlierPCAd3"
## [45] "f.outlierPCAd4"
```

```
#Finally we obtained 62 extreme outliers.
llvout<- c(outliers, outliers2);length(llvout)
```

```
## [1] 62
```

```
catdes(df, 42)
```

```
## $test.chi2
##                              p.value df
## f.outlierPCAd3          0.000000e+00  1
## f.outlierPCAd4          0.000000e+00  1
## Trip_type               7.361875e-28  1
## f.Improvement_surcharge 3.287829e-27  1
## RateCodeID              6.763510e-27  1
## f.MTA_tax               6.763510e-27  1
## f.outlierPCAd2          4.083972e-10  1
##
## $category
## $category$NoOutDim1
##                                          Cla/Mod   Mod/Cla      Global
## Trip_type=Street-hail                  100.00000 98.396382 98.35593917
## f.Improvement_surcharge=(0.1,0.8]      100.00000 98.355263 98.31483765
## f.MTA_tax=(0.4,0.5]                    100.00000 98.334704 98.29428689
## RateCodeID=Standard rate               100.00000 98.334704 98.29428689
## f.outlierPCAd2=NoOutDim2                99.97919 98.787007 98.76695438
## f.outlierPCAd2=YesOutDim2               98.33333  1.212993  1.23304562
## Lpep_dropoff_datetime=2016-01-31 02:00:28   0.00000  0.000000  0.02055076
## Lpep_dropoff_datetime=2016-01-05 09:29:16   0.00000  0.000000  0.02055076
## lpep_pickup_datetime=2016-01-30 22:25:55    0.00000  0.000000  0.02055076
## lpep_pickup_datetime=2016-01-05 08:34:06    0.00000  0.000000  0.02055076
## f.MTA_tax=(-0.1,0.4]                    97.59036  1.665296  1.70571311
## RateCodeID=Special rate                 97.59036  1.665296  1.70571311
## f.Improvement_surcharge=(-0.1,0.1]      97.56098  1.644737  1.68516235
## Trip_type=Dispatch                      97.50000  1.603618  1.64406083
##                                             p.value    v.test
## Trip_type=Street-hail                  0.0002669698  3.645406
```

```
## f.Improvement_surcharge=(0.1,0.8]        0.0002805717  3.632607
## f.MTA_tax=(0.4,0.5]                      0.0002874994  3.626311
## RateCodeID=Standard rate                 0.0002874994  3.626311
## f.outlierPCAd2=NoOutDim2                 0.0246609125  2.246673
## f.outlierPCAd2=YesOutDim2                0.0246609125 -2.246673
## Lpep_dropoff_datetime=2016-01-31 02:00:28 0.0004110152 -3.532908
## Lpep_dropoff_datetime=2016-01-05 09:29:16 0.0004110152 -3.532908
## lpep_pickup_datetime=2016-01-30 22:25:55  0.0004110152 -3.532908
## lpep_pickup_datetime=2016-01-05 08:34:06  0.0004110152 -3.532908
## f.MTA_tax=(-0.1,0.4]                     0.0002874994 -3.626311
## RateCodeID=Special rate                  0.0002874994 -3.626311
## f.Improvement_surcharge=(-0.1,0.1]       0.0002805717 -3.632607
## Trip_type=Dispatch                       0.0002669698 -3.645406
##
## $category$YesOutDim1
##                                              Cla/Mod Mod/Cla      Global
## Trip_type=Dispatch                          2.50000000     100  1.64406083
## f.Improvement_surcharge=(-0.1,0.1]          2.43902439     100  1.68516235
## f.MTA_tax=(-0.1,0.4]                        2.40963855     100  1.70571311
## RateCodeID=Special rate                     2.40963855     100  1.70571311
## Lpep_dropoff_datetime=2016-01-31 02:00:28 100.00000000      50  0.02055076
## Lpep_dropoff_datetime=2016-01-05 09:29:16 100.00000000      50  0.02055076
## lpep_pickup_datetime=2016-01-30 22:25:55  100.00000000      50  0.02055076
## lpep_pickup_datetime=2016-01-05 08:34:06  100.00000000      50  0.02055076
## f.outlierPCAd2=YesOutDim2                   1.66666667      50  1.23304562
## f.outlierPCAd2=NoOutDim2                    0.02080732      50 98.76695438
## f.MTA_tax=(0.4,0.5]                         0.00000000       0 98.29428689
## RateCodeID=Standard rate                    0.00000000       0 98.29428689
## f.Improvement_surcharge=(0.1,0.8]           0.00000000       0 98.31483765
## Trip_type=Street-hail                       0.00000000       0 98.35593917
##                                              p.value    v.test
## Trip_type=Dispatch                       0.0002669698  3.645406
## f.Improvement_surcharge=(-0.1,0.1]       0.0002805717  3.632607
## f.MTA_tax=(-0.1,0.4]                     0.0002874994  3.626311
## RateCodeID=Special rate                  0.0002874994  3.626311
## Lpep_dropoff_datetime=2016-01-31 02:00:28 0.0004110152  3.532908
## Lpep_dropoff_datetime=2016-01-05 09:29:16 0.0004110152  3.532908
## lpep_pickup_datetime=2016-01-30 22:25:55  0.0004110152  3.532908
## lpep_pickup_datetime=2016-01-05 08:34:06  0.0004110152  3.532908
## f.outlierPCAd2=YesOutDim2                0.0246609125  2.246673
## f.outlierPCAd2=NoOutDim2                 0.0246609125 -2.246673
## f.MTA_tax=(0.4,0.5]                      0.0002874994 -3.626311
## RateCodeID=Standard rate                 0.0002874994 -3.626311
## f.Improvement_surcharge=(0.1,0.8]        0.0002805717 -3.632607
## Trip_type=Street-hail                    0.0002669698 -3.645406
##
##
## $quanti.var
##                              Eta2      P-value
## travel_time           0.0654142628 1.567072e-73
## MTA_tax               0.0236951094 3.462099e-27
## improvement_surcharge 0.0232809064 9.797386e-27
## mis_ind               0.0072774270 2.520328e-09
## trip_length           0.0022486676 9.366881e-04
```

```
## trip_distance_km      0.0015711379 5.685930e-03
## Trip_distance         0.0015711379 5.685930e-03
## Dropoff_latitude      0.0013653916 9.942778e-03
## Fare_amount           0.0011964249 1.582427e-02
## Total_amount          0.0010996816 2.070749e-02
## espeed                0.0007912635 4.975093e-02
##
## $quanti
## $quanti$NoOutDim1
##                           v.test Mean in category Overall mean
## MTA_tax                10.736699        0.4916735    0.4914714
## improvement_surcharge  10.642444        0.2951624    0.2950411
## Dropoff_latitude        2.577330       40.7447051   40.7446630
## espeed                  1.962013       21.3229435   21.3177354
## Total_amount           -2.312996       13.4884005   13.4937485
## Fare_amount            -2.412593       11.1498725   11.1547431
## trip_distance_km       -2.764704        4.0616233    4.0643323
## Trip_distance          -2.764704        2.5237757    2.5254590
## trip_length            -3.307532        4.0010371    4.0037179
## mis_ind                -5.950183        2.4917763    2.4928072
## travel_time           -17.839293       12.0918446   12.1423035
##                       sd in category Overall sd      p.value
## MTA_tax                   0.06398367 0.06474215 6.844810e-27
## improvement_surcharge     0.03875921 0.03921036 1.891034e-26
## Dropoff_latitude          0.05621814 0.05624604 9.956682e-03
## espeed                    9.12782900 9.13058219 4.976092e-02
## Total_amount              7.94415294 7.95310822 2.072286e-02
## Fare_amount               6.93716919 6.94416418 1.583948e-02
## trip_distance_km          3.36666751 3.37034414 5.697454e-03
## Trip_distance             2.09195021 2.09423476 5.697454e-03
## trip_length               2.78445515 2.78797993 9.412196e-04
## mis_ind                   0.59390944 0.59595986 2.678422e-09
## travel_time               9.26784462 9.72933787 3.500950e-71
##
## $quanti$YesOutDim1
##                           v.test Mean in category Overall mean
## travel_time            17.839293       134.858333   12.1423035
## mis_ind                 5.950183         5.000000    2.4928072
## trip_length             3.307532        10.523515    4.0037179
## trip_distance_km        2.764704        10.652478    4.0643323
## Trip_distance           2.764704         6.619143    2.5254590
## Fare_amount             2.412593        23.000000   11.1547431
## Total_amount            2.312996        26.500000   13.4937485
## espeed                 -1.962013         8.651697   21.3177354
## Dropoff_latitude       -2.577330        40.642168   40.7446630
## improvement_surcharge -10.642444         0.000000    0.2950411
## MTA_tax               -10.736699         0.000000    0.4914714
##                       sd in category Overall sd      p.value
## travel_time              79.69166667 9.72933787 3.500950e-71
## mis_ind                   0.00000000 0.59595986 2.678422e-09
## trip_length               3.60776586 2.78797993 9.412196e-04
## trip_distance_km          5.30821789 3.37034414 5.697454e-03
## Trip_distance             3.29837368 2.09423476 5.697454e-03
## Fare_amount              12.00000000 6.94416418 1.583948e-02
```

```
## Total_amount             15.50000000 7.95310822 2.072286e-02
## espeed                    6.71767213 9.13058219 4.976092e-02
## Dropoff_latitude          0.01688957 0.05624604 9.956682e-03
## improvement_surcharge     0.00000000 0.03921036 1.891034e-26
## MTA_tax                   0.00000000 0.06474215 6.844810e-27
##
##
## attr(,"class")
## [1] "catdes" "list "
```

III Interpret axis

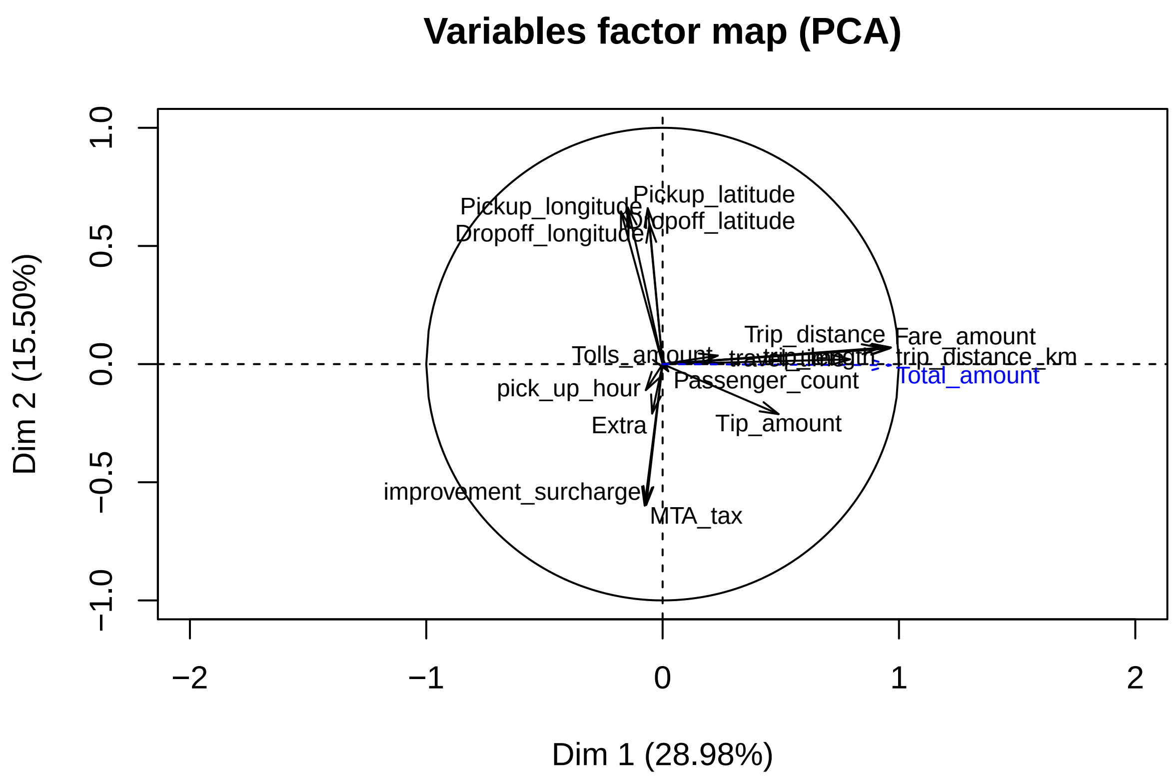
```r
# Interential criteria

dimdesc (res.pca, axes=1:4)
```

```
## $Dim.1
## $Dim.1$quanti
##                       correlation       p.value
## Total_amount           0.96578021  0.000000e+00
## trip_distance_km       0.96545892  0.000000e+00
## Trip_distance          0.96545892  0.000000e+00
## Fare_amount            0.95952463  0.000000e+00
## trip_length            0.92223605  0.000000e+00
## travel_time            0.79291771  0.000000e+00
## Tip_amount             0.49073894 2.286608e-293
## Tolls_amount           0.23354094  2.825897e-61
## Extra                 -0.04441593  1.941467e-03
## Dropoff_longitude     -0.05536809  1.114226e-04
## Pickup_longitude      -0.06305806  1.072702e-05
## improvement_surcharge -0.06923616  1.336498e-06
## pick_up_hour          -0.07089733  7.401367e-07
## MTA_tax               -0.07585972  1.170775e-07
## Pickup_latitude       -0.14751746  4.440143e-25
## Dropoff_latitude      -0.17625772  2.993390e-35
##
## $Dim.1$quali
##                R2      p.value
## AnyTip 0.04924904 2.338313e-55
##
## $Dim.1$category
##              Estimate      p.value
## AnyTip Yes  0.4847013 2.338313e-55
## AnyTip No  -0.4847013 2.338313e-55
##
##
## $Dim.2
## $Dim.2$quanti
##                       correlation      p.value
## Pickup_latitude        0.66260373 0.000000e+00
## Pickup_longitude       0.65985850 0.000000e+00
## Dropoff_latitude       0.64617666 0.000000e+00
## Dropoff_longitude      0.59347965 0.000000e+00
```

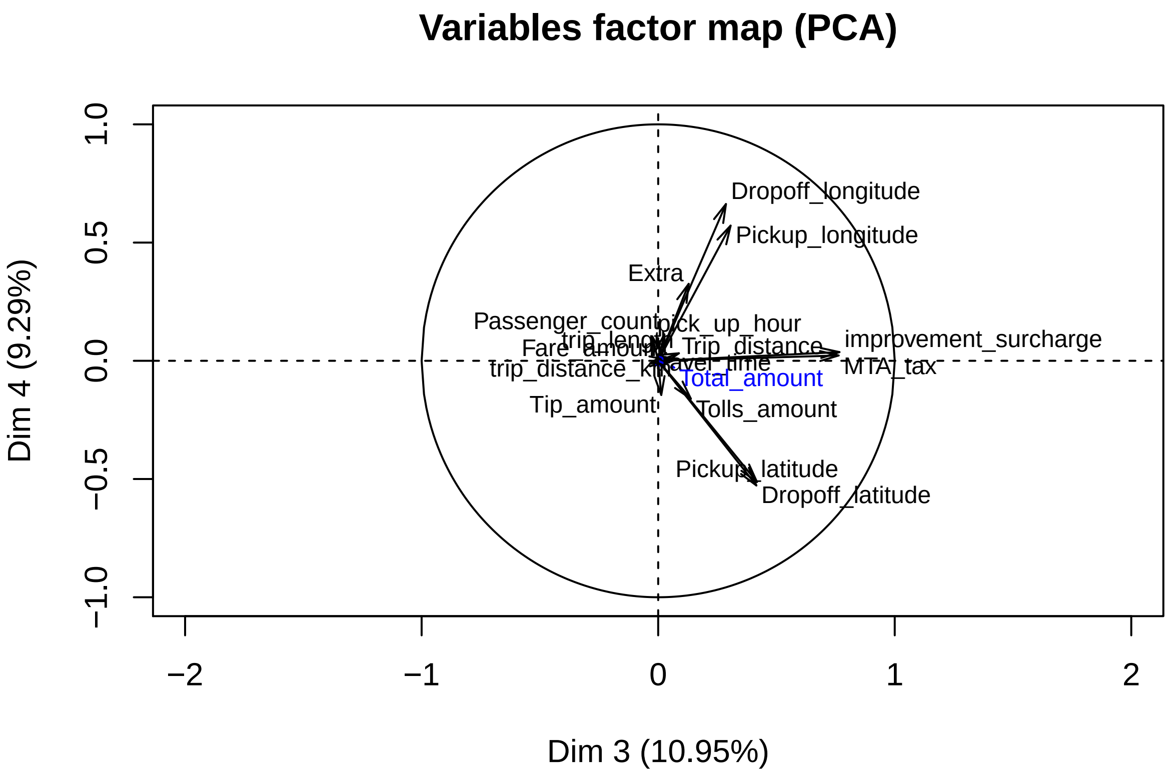
```
## trip_distance_km       0.07016348 9.624966e-07
## Trip_distance          0.07016348 9.624966e-07
## trip_length            0.06889982 1.503966e-06
## Fare_amount            0.06612191 3.906140e-06
## Tolls_amount           0.03596743 1.210264e-02
## Passenger_count       -0.03024353 3.489026e-02
## pick_up_hour          -0.11001618 1.406494e-14
## Extra                 -0.20918526 2.994770e-49
## Tip_amount            -0.21154095 2.378320e-50
## improvement_surcharge -0.59592529 0.000000e+00
## MTA_tax               -0.59855164 0.000000e+00
##
## $Dim.2$quali
##                R2      p.value
## AnyTip 0.06744473 7.785848e-76
##
## $Dim.2$category
##              Estimate      p.value
## AnyTip No   0.4147775 7.785848e-76
## AnyTip Yes -0.4147775 7.785848e-76
##
##
## $Dim.3
## $Dim.3$quanti
##                       correlation       p.value
## improvement_surcharge  0.76657280  0.000000e+00
## MTA_tax                0.76291440  0.000000e+00
## Pickup_latitude        0.41765439 9.463043e-205
## Dropoff_latitude       0.41519442 3.949101e-202
## Pickup_longitude       0.30618865 3.953837e-106
## Dropoff_longitude      0.28663331  1.124479e-92
## Tolls_amount           0.13824187  3.412117e-22
## Extra                  0.12912389  1.526769e-19
## trip_length            0.08771141  8.859842e-10
## trip_distance_km       0.07829221  4.541174e-08
## Trip_distance          0.07829221  4.541174e-08
## Total_amount           0.06766432  2.309738e-06
## Fare_amount            0.04231563  3.153515e-03
## travel_time           -0.03421414  1.699796e-02
##
##
## $Dim.4
## $Dim.4$quanti
##                       correlation       p.value
## Dropoff_longitude      0.66291516  0.000000e+00
## Pickup_longitude       0.57174226  0.000000e+00
## Extra                  0.32547576 1.885454e-120
## Passenger_count        0.11204182  4.564971e-15
## pick_up_hour           0.10324266  5.224927e-13
## improvement_surcharge  0.03681150  1.022695e-02
## trip_length            0.03176618  2.669876e-02
## Total_amount          -0.02884409  4.422305e-02
## Tip_amount            -0.14460125  3.760819e-24
## Tolls_amount          -0.16075354  1.570319e-29
```

```
## Pickup_latitude       -0.51263879  0.000000e+00
## Dropoff_latitude      -0.52725817  0.000000e+00
##
## $Dim.4$quali
##                R2      p.value
## AnyTip 0.02647713 3.174086e-30
##
## $Dim.4$category
##               Estimate      p.value
## AnyTip No   0.2011856 3.174086e-30
## AnyTip Yes -0.2011856 3.174086e-30
```

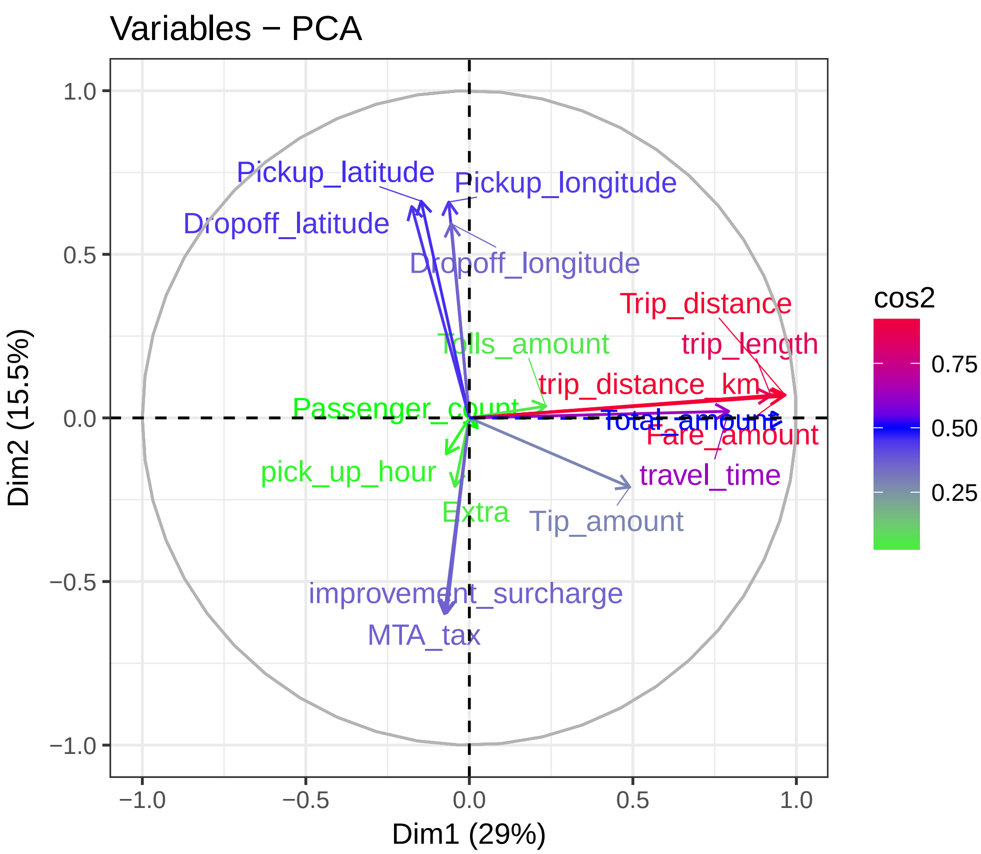
```
#The first Dimention is best described by the quantative variables Total_amount, trip_distance and Fare_
plot(res.pca,choix="var", cex = 0.75)
```




```
plot(res.pca,choix="var", cex = 0.75, axes = (3:4))# 3rd and 4th PCA
```




```r
#modern factoextra

fviz_pca_var(res.pca,col.var="cos2", repel=TRUE)+scale_color_gradient2(low="green", mid="blue", high="r
```

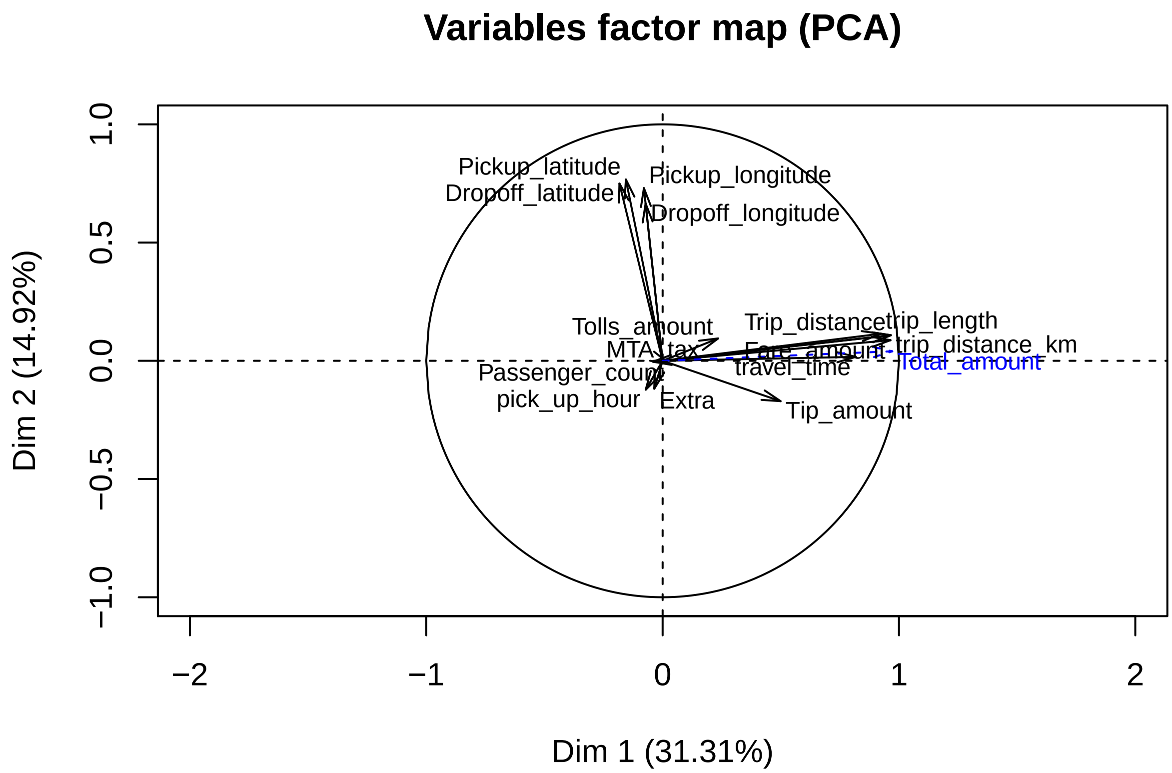


IV PCA execution with supplementary individuals

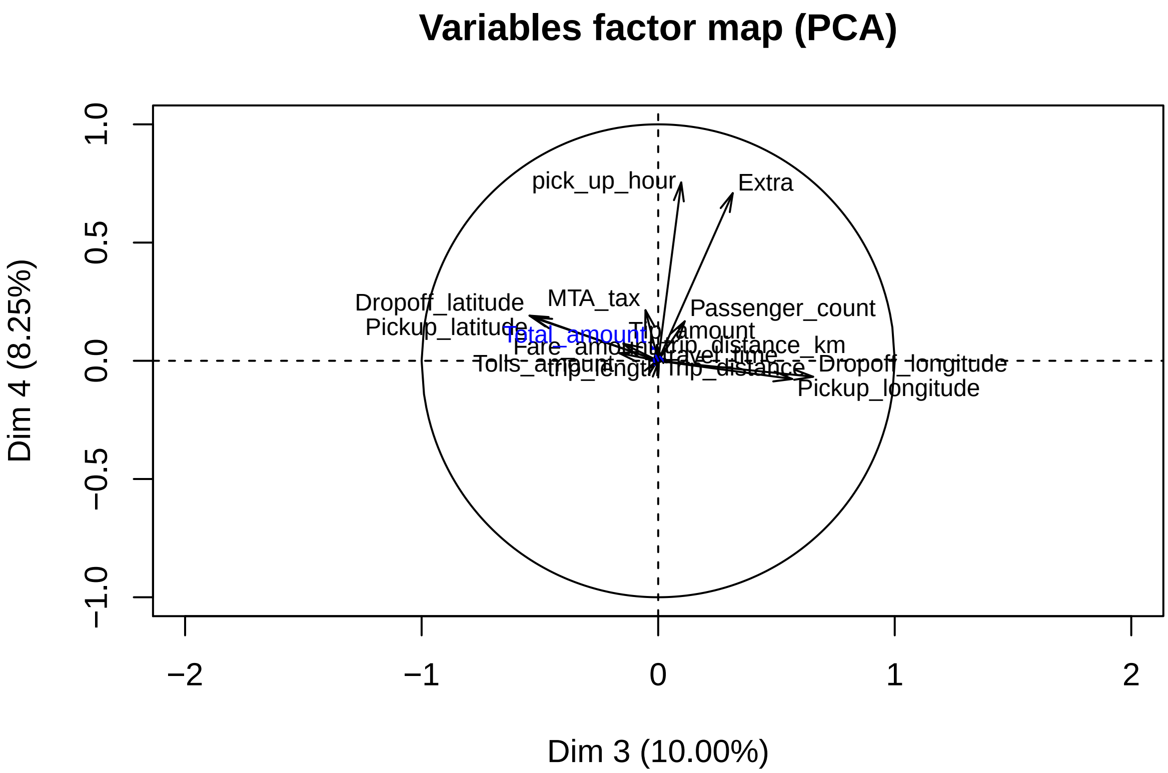
```
vec_out <- llvout
vars_con_pca <- names(df)[c(6:16,18,23:26)]
# We do a PCA analysis using the factorial variables Fare amount, total and the pickup perio in order t

res.pca<-PCA(df[,c(vars_con_pca, "f.fare_amount", "f.total", "pick_up_period", "f.passenger", "f.pickup
```

```
## Warning in PCA(df[, c(vars_con_pca, "f.fare_amount", "f.total",
## "pick_up_period", : Missing values are imputed by the mean of the variable:
## you should use the imputePCA function of the missMDA package
```

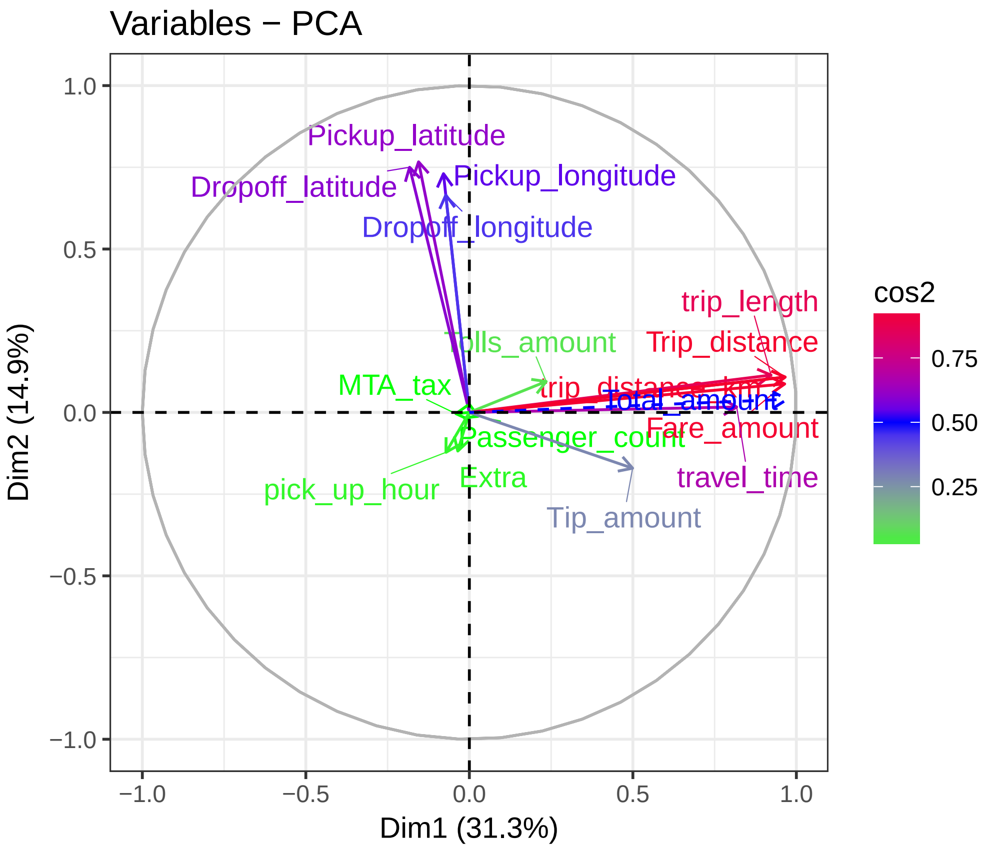
```
plot(res.pca,choix="var", cex = 0.75)
```



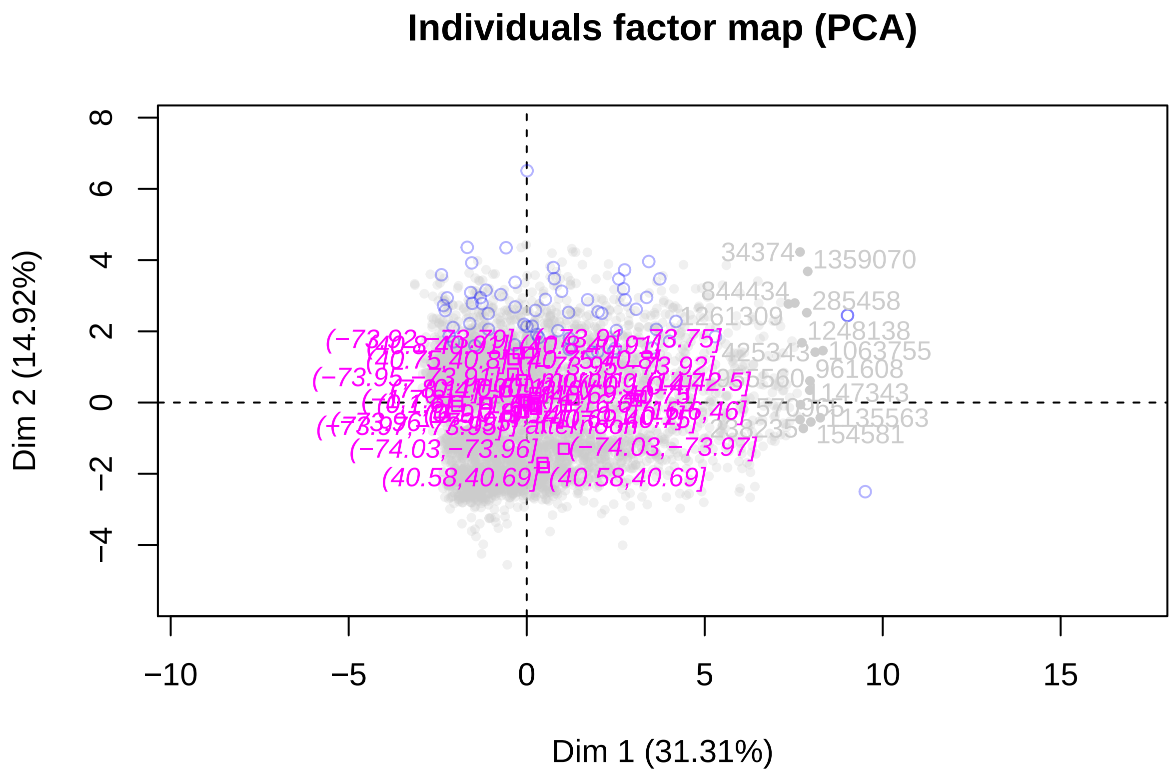

```r
plot(res.pca,choix="var", cex = 0.75, axes = (3:4))# 3rd and 4th PCA
```




```
fviz_pca_var(res.pca,col.var="cos2", repel=TRUE)+scale_color_gradient2(low="green", mid="blue", high="r
```



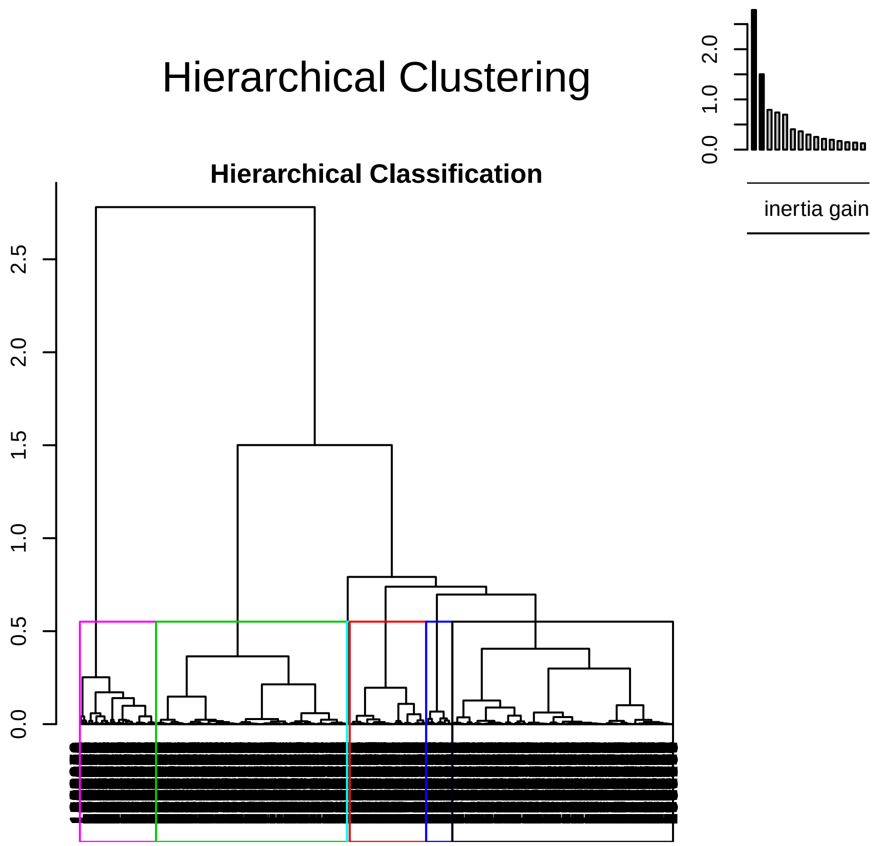

```r
#We can see that trips in the afternoon tend to be longer and thus also more expensive than the ones du
plot.PCA(res.pca, choix=c("ind"),cex=0.8,col.ind="grey80",select="contrib15",axes=c(1,2))
```



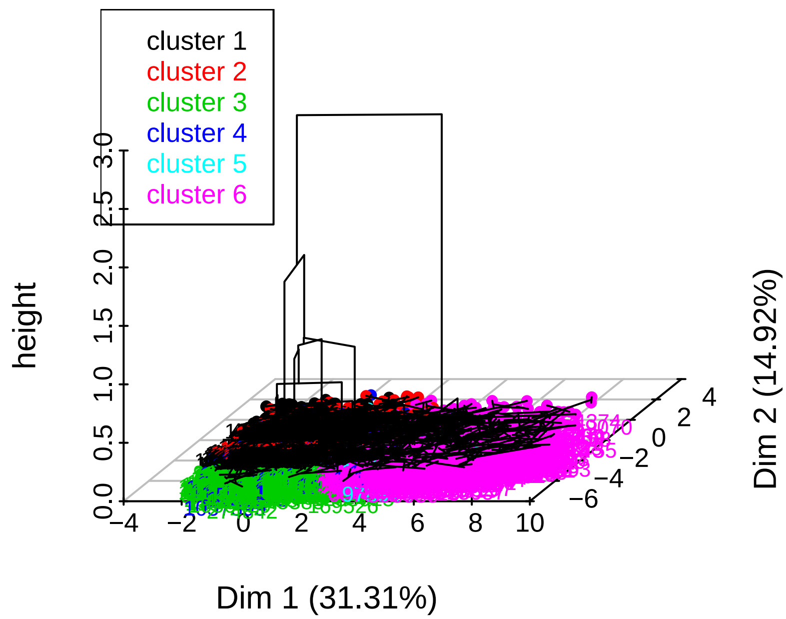
Hierarchical clustering

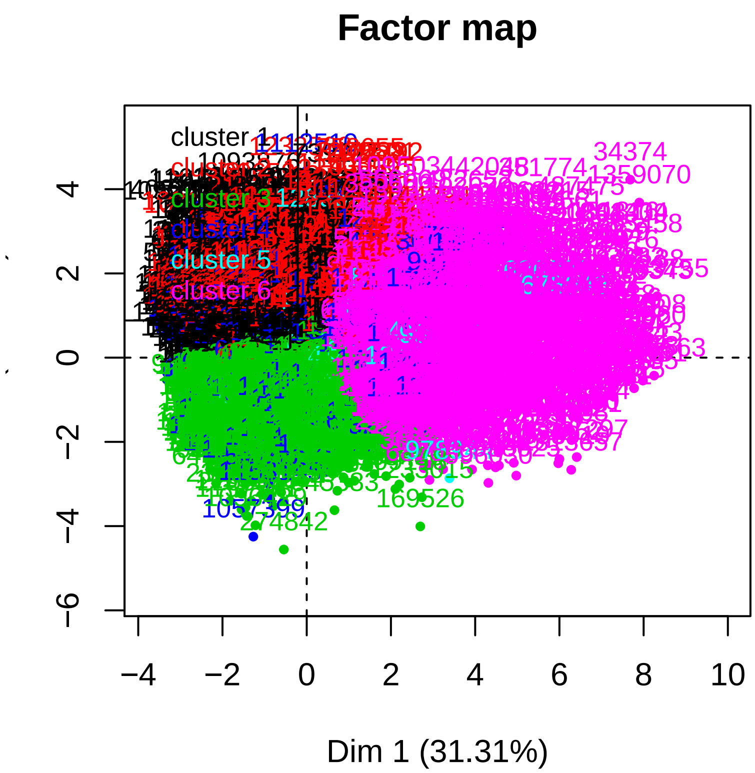
We generate 6 clusters using the hierarchical method, taking the projection obtained by the PCA as a source dataset. The resulting table is showing the distribution taken between the different clusters (excluding the multivariant outliers)

```
library(FactoMineR)
res.hcpc <-HCPC(res.pca, nb.clust = 6, order=TRUE)
```


Hierarchical Clustering
Hierarchical Classification
inertia gain
0.0
0.5
1.0
1.5
2.0
2.5
0.0
1.0
2.0

Hierarchical clustering on the factor map






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

```
##
##    1    2    3    4    5    6
## 1522  756 1455  241   22  809
```

Variable description

Below is listed the categorical description for each cluster and our explanation as it follows.

```r
#Block A descripcion per variables
res.hcpc$desc.var
```

```
## $test.chi2
##                                p.value df
## f.fare_amount             0.000000e+00 15
## f.total                   0.000000e+00 15
## f.pickup_longitude        0.000000e+00 20
## f.pickup_latitude         0.000000e+00 20
## f.dropoff_longitude       0.000000e+00 20
## f.dropoff_latitude        0.000000e+00 20
## f.Improvement_surcharge   0.000000e+00  5
## f.MTA_tax                 0.000000e+00  5
## f.passenger              5.811913e-308  5
## pick_up_period            1.388397e-34 15
##
```

```
## $category
## $category$`1`
##                                      Cla/Mod      Mod/Cla     Global
## f.dropoff_latitude=(40.8,40.91]     77.976190  60.24967148 24.4745057
## f.pickup_latitude=(40.8,40.91]      77.976190  60.24967148 24.4745057
## f.pickup_longitude=(-73.95,-73.92]  60.945274  48.29172142 25.0988554
## f.dropoff_longitude=(-73.95,-73.91] 56.761269  44.67805519 24.9323621
## f.total=(-1,7.8]                    47.996795  39.35611038 25.9729448
## f.dropoff_latitude=(40.75,40.8]     47.555924  37.71353482 25.1196670
## f.pickup_latitude=(40.75,40.8]      47.555924  37.71353482 25.1196670
## f.fare_amount=(0.1,6]               45.600000  37.45072273 26.0145682
## f.dropoff_longitude=(-73.97,-73.95] 43.858203  34.95400788 25.2445369
## f.passenger=(0,1]                   34.480216  92.18134034 84.6826223
## f.fare_amount=(6,9]                 42.028986  34.29697766 25.8480749
## f.pickup_longitude=(-73.96,-73.95]  40.939044  32.65440210 25.2653486
## f.total=(7.8,11]                    39.536878  30.28909330 24.2663892
## pick_up_period=morning              40.540541  25.62417871 20.0208117
## pick_up_period=valley               36.740741  32.58869908 28.0957336
## f.MTA_tax=(0.4,0.5]                 31.821033 100.00000000 99.5421436
## f.Improvement_surcharge=(0.1,0.8]   31.793478  99.93429698 99.5629553
## f.Improvement_surcharge=(-0.1,0.1]   4.761905   0.06570302  0.4370447
## pick_up_period=afternoon            28.904429  32.58869908 35.7127992
## f.MTA_tax=(-0.1,0.4]                 0.000000   0.00000000  0.4578564
## pick_up_period=night                18.018018   9.19842313 16.1706556
## f.passenger=(1,6]                   16.168478   7.81865966 15.3173777
## f.pickup_longitude=(-73.92,-73.79]  17.707442  13.60052562 24.3288241
## f.dropoff_longitude=(-73.91,-73.75] 15.293118  11.82654402 24.4953174
## f.dropoff_longitude=(-74.03,-73.97] 10.690789   8.54139290 25.3069719
## f.pickup_longitude=(-74.03,-73.96]   6.831276   5.45335085 25.2861602
## f.fare_amount=(14,42.5]              5.643739   4.20499343 23.6004162
## f.total=(16.6,46]                    5.643154   4.46780552 25.0780437
## f.dropoff_latitude=(40.69,40.75]     2.559868   2.03679369 25.2029136
## f.pickup_latitude=(40.69,40.75]      2.559868   2.03679369 25.2029136
## f.dropoff_latitude=(40.58,40.69]     0.000000   0.00000000 25.1821020
## f.pickup_latitude=(40.58,40.69]      0.000000   0.00000000 25.1821020
##                                          p.value     v.test
## f.dropoff_latitude=(40.8,40.91]      0.000000e+00        Inf
## f.pickup_latitude=(40.8,40.91]       0.000000e+00        Inf
## f.pickup_longitude=(-73.95,-73.92]  9.653530e-134  24.610910
## f.dropoff_longitude=(-73.95,-73.91]  6.012435e-98  21.004130
## f.total=(-1,7.8]                     3.010774e-45  14.116377
## f.dropoff_latitude=(40.75,40.8]      4.293051e-41  13.425392
## f.pickup_latitude=(40.75,40.8]       4.293051e-41  13.425392
## f.fare_amount=(0.1,6]                1.371224e-33  12.078546
## f.dropoff_longitude=(-73.97,-73.95]  3.489669e-25  10.367378
## f.passenger=(0,1]                    5.503716e-25  10.323739
## f.fare_amount=(6,9]                  2.926343e-19   8.971451
## f.pickup_longitude=(-73.96,-73.95]   2.523487e-15   7.912462
## f.total=(7.8,11]                     5.908761e-11   6.546028
## pick_up_period=morning               7.753913e-11   6.505299
## pick_up_period=valley                2.854485e-06   4.681022
## f.MTA_tax=(0.4,0.5]                  2.243829e-04   3.689854
## f.Improvement_surcharge=(0.1,0.8]    3.881550e-03   2.887631
## f.Improvement_surcharge=(-0.1,0.1]   3.881550e-03  -2.887631
```

```
## pick_up_period=afternoon             2.027034e-03  -3.086243
## f.MTA_tax=(-0.1,0.4]                 2.243829e-04  -3.689854
## pick_up_period=night                 1.286503e-20  -9.309323
## f.passenger=(1,6]                    5.503716e-25 -10.323739
## f.pickup_longitude=(-73.92,-73.79]   1.803454e-34 -12.244246
## f.dropoff_longitude=(-73.91,-73.75]  4.794654e-48 -14.563490
## f.dropoff_longitude=(-74.03,-73.97]  1.589431e-84 -19.481064
## f.pickup_longitude=(-74.03,-73.96]  1.447159e-123 -23.641396
## f.fare_amount=(14,42.5]             7.194958e-127 -23.960426
## f.total=(16.6,46]                   1.123791e-136 -24.883459
## f.dropoff_latitude=(40.69,40.75]    5.515651e-183 -28.847062
## f.pickup_latitude=(40.69,40.75]     5.515651e-183 -28.847062
## f.dropoff_latitude=(40.58,40.69]    6.991451e-240 -33.074168
## f.pickup_latitude=(40.58,40.69]     6.991451e-240 -33.074168
##
## $category$`2`
##                                        Cla/Mod     Mod/Cla     Global
## f.dropoff_longitude=(-73.91,-73.75] 60.66270178  94.4444444 24.4953174
## f.pickup_longitude=(-73.92,-73.79]  61.50556031  95.1058201 24.3288241
## f.dropoff_latitude=(40.75,40.8]     30.24026512  48.2804233 25.1196670
## f.pickup_latitude=(40.75,40.8]      30.24026512  48.2804233 25.1196670
## f.dropoff_latitude=(40.69,40.75]    28.98431049  46.4285714 25.2029136
## f.pickup_latitude=(40.69,40.75]     28.98431049  46.4285714 25.2029136
## f.fare_amount=(9,14]                21.20441052  33.0687831 24.5369407
## pick_up_period=night                21.49292149  22.0899471 16.1706556
## f.total=(11,16.6]                   19.81450253  31.0846561 24.6826223
## f.passenger=(0,1]                   16.63799459  89.5502646 84.6826223
## f.total=(7.8,11]                    19.29674099  29.7619048 24.2663892
## f.fare_amount=(6,9]                 18.03542673  29.6296296 25.8480749
## f.MTA_tax=(0.4,0.5]                 15.80597951 100.0000000 99.5421436
## f.Improvement_surcharge=(0.1,0.8]   15.80267559 100.0000000 99.5629553
## f.Improvement_surcharge=(-0.1,0.1]   0.00000000   0.0000000  0.4370447
## f.MTA_tax=(-0.1,0.4]                 0.00000000   0.0000000  0.4578564
## pick_up_period=morning              11.85031185  15.0793651 20.0208117
## f.passenger=(1,6]                   10.73369565  10.4497354 15.3173777
## f.fare_amount=(14,42.5]              8.55379189  12.8306878 23.6004162
## f.total=(16.6,46]                    7.71784232  12.3015873 25.0780437
## f.dropoff_longitude=(-73.95,-73.91]  3.50584307   5.5555556 24.9323621
## f.pickup_longitude=(-73.95,-73.92]   2.98507463   4.7619048 25.0988554
## f.dropoff_latitude=(40.8,40.91]      1.87074830   2.9100529 24.4745057
## f.pickup_latitude=(40.8,40.91]       1.87074830   2.9100529 24.4745057
## f.dropoff_latitude=(40.58,40.69]     1.48760331   2.3809524 25.1821020
## f.pickup_latitude=(40.58,40.69]      1.48760331   2.3809524 25.1821020
## f.pickup_longitude=(-74.03,-73.96]   0.08230453   0.1322751 25.2861602
## f.dropoff_longitude=(-73.97,-73.95]  0.00000000   0.0000000 25.2445369
## f.pickup_longitude=(-73.96,-73.95]   0.00000000   0.0000000 25.2653486
## f.dropoff_longitude=(-74.03,-73.97]  0.00000000   0.0000000 25.3069719
##                                          p.value     v.test
## f.dropoff_longitude=(-73.91,-73.75]  0.000000e+00        Inf
## f.pickup_longitude=(-73.92,-73.79]   0.000000e+00        Inf
## f.dropoff_latitude=(40.75,40.8]      8.586101e-52  15.141778
## f.pickup_latitude=(40.75,40.8]       8.586101e-52  15.141778
## f.dropoff_latitude=(40.69,40.75]     5.908326e-44  13.904980
## f.pickup_latitude=(40.69,40.75]      5.908326e-44  13.904980
```

```
## f.fare_amount=(9,14]                7.459414e-09   5.780240
## pick_up_period=night                3.267506e-06   4.653246
## f.total=(11,16.6]                   1.322624e-05   4.356328
## f.passenger=(0,1]                   2.659953e-05   4.200781
## f.total=(7.8,11]                    1.633699e-04   3.769813
## f.fare_amount=(6,9]                 1.051032e-02   2.558572
## f.MTA_tax=(0.4,0.5]                 2.293429e-02   2.274527
## f.Improvement_surcharge=(0.1,0.8]   2.723875e-02   2.208079
## f.Improvement_surcharge=(-0.1,0.1]  2.723875e-02  -2.208079
## f.MTA_tax=(-0.1,0.4]                2.293429e-02  -2.274527
## pick_up_period=morning              1.491105e-04  -3.792546
## f.passenger=(1,6]                   2.659953e-05  -4.200781
## f.fare_amount=(14,42.5]             1.119713e-15  -8.012970
## f.total=(16.6,46]                   6.665484e-21  -9.378915
## f.dropoff_longitude=(-73.95,-73.91]  1.459621e-51 -15.106845
## f.pickup_longitude=(-73.95,-73.92]   1.979850e-57 -15.972713
## f.dropoff_latitude=(40.8,40.91]      7.014169e-69 -17.540633
## f.pickup_latitude=(40.8,40.91]       7.014169e-69 -17.540633
## f.dropoff_latitude=(40.58,40.69]     1.122948e-76 -18.532797
## f.pickup_latitude=(40.58,40.69]      1.122948e-76 -18.532797
## f.pickup_longitude=(-74.03,-73.96]  6.756121e-104 -21.645124
## f.dropoff_longitude=(-73.97,-73.95] 3.324852e-106 -21.888753
## f.pickup_longitude=(-73.96,-73.95]  2.625078e-106 -21.899524
## f.dropoff_longitude=(-74.03,-73.97] 1.636007e-106 -21.921061
##
## $category$`3`
##                                        Cla/Mod      Mod/Cla     Global
## f.dropoff_latitude=(40.58,40.69]    74.8760331  62.26804124 25.1821020
## f.pickup_latitude=(40.58,40.69]     74.8760331  62.26804124 25.1821020
## f.pickup_longitude=(-74.03,-73.96]  68.5596708  57.25085911 25.2861602
## f.dropoff_longitude=(-74.03,-73.97] 52.7960526  44.12371134 25.3069719
## f.dropoff_latitude=(40.69,40.75]    44.8389761  37.31958763 25.2029136
## f.pickup_latitude=(40.69,40.75]     44.8389761  37.31958763 25.2029136
## f.dropoff_longitude=(-73.97,-73.95] 40.7254740  33.95189003 25.2445369
## f.total=(11,16.6]                   40.5564924  33.05841924 24.6826223
## f.fare_amount=(9,14]                40.2035623  32.57731959 24.5369407
## f.pickup_longitude=(-73.96,-73.95]  37.3970346  31.20274914 25.2653486
## pick_up_period=afternoon            35.3146853  41.64948454 35.7127992
## f.passenger=(0,1]                   31.8260015  89.00343643 84.6826223
## f.total=(7.8,11]                    35.1629503  28.17869416 24.2663892
## f.MTA_tax=(0.4,0.5]                 30.4202383 100.00000000 99.5421436
## f.fare_amount=(6,9]                 34.2995169  29.27835052 25.8480749
## f.fare_amount=(0.1,6]               33.8400000  29.07216495 26.0145682
## f.Improvement_surcharge=(0.1,0.8]   30.3929766  99.93127148 99.5629553
## pick_up_period=night                33.8481338  18.07560137 16.1706556
## f.Improvement_surcharge=(-0.1,0.1]   4.7619048   0.06872852  0.4370447
## pick_up_period=valley               27.2592593  25.29209622 28.0957336
## f.MTA_tax=(-0.1,0.4]                 0.0000000   0.00000000  0.4578564
## f.dropoff_longitude=(-73.95,-73.91] 24.6243740  20.27491409 24.9323621
## f.passenger=(1,6]                   21.7391304  10.99656357 15.3173777
## pick_up_period=morning              22.6611227  14.98281787 20.0208117
## f.total=(16.6,46]                   15.7676349  13.05841924 25.0780437
## f.fare_amount=(14,42.5]             11.6402116   9.07216495 23.6004162
## f.pickup_longitude=(-73.95,-73.92]  12.2719735  10.17182131 25.0988554
```

```
## f.dropoff_longitude=(-73.91,-73.75]  2.0390824   1.64948454 24.4953174
## f.pickup_longitude=(-73.92,-73.79]   1.6253208   1.30584192 24.3288241
## f.dropoff_latitude=(40.75,40.8]      0.4142502   0.34364261 25.1196670
## f.pickup_latitude=(40.75,40.8]       0.4142502   0.34364261 25.1196670
## f.dropoff_latitude=(40.8,40.91]      0.0000000   0.00000000 24.4745057
## f.pickup_latitude=(40.8,40.91]       0.0000000   0.00000000 24.4745057
##                                          p.value     v.test
## f.dropoff_latitude=(40.58,40.69]    2.901539e-318  38.138933
## f.pickup_latitude=(40.58,40.69]     2.901539e-318  38.138933
## f.pickup_longitude=(-74.03,-73.96]  1.758277e-234  32.696379
## f.dropoff_longitude=(-74.03,-73.97]  1.640153e-82  19.242225
## f.dropoff_latitude=(40.69,40.75]     1.021758e-35  12.475024
## f.pickup_latitude=(40.69,40.75]      1.021758e-35  12.475024
## f.dropoff_longitude=(-73.97,-73.95]  2.251270e-19   9.000288
## f.total=(11,16.6]                    2.734320e-18   8.721958
## f.fare_amount=(9,14]                 4.757062e-17   8.392551
## f.pickup_longitude=(-73.96,-73.95]   7.122257e-10   6.163348
## pick_up_period=afternoon             1.848224e-08   5.625639
## f.passenger=(0,1]                    1.962410e-08   5.615283
## f.total=(7.8,11]                     3.622051e-05   4.130354
## f.MTA_tax=(0.4,0.5]                  3.504531e-04   3.574832
## f.fare_amount=(6,9]                  3.810595e-04   3.552865
## f.fare_amount=(0.1,6]                1.561249e-03   3.163051
## f.Improvement_surcharge=(0.1,0.8]    5.628349e-03   2.768682
## pick_up_period=night                 1.900267e-02   2.345479
## f.Improvement_surcharge=(-0.1,0.1]   5.628349e-03  -2.768682
## pick_up_period=valley                4.192053e-03  -2.863336
## f.MTA_tax=(-0.1,0.4]                 3.504531e-04  -3.574832
## f.dropoff_longitude=(-73.95,-73.91]  6.403247e-07  -4.978640
## f.passenger=(1,6]                    1.962410e-08  -5.615283
## pick_up_period=morning               4.474485e-09  -5.865623
## f.total=(16.6,46]                    6.958904e-40 -13.217444
## f.fare_amount=(14,42.5]              2.444321e-62 -16.662771
## f.pickup_longitude=(-73.95,-73.92]   1.766270e-62 -16.682189
## f.dropoff_longitude=(-73.91,-73.75] 8.835619e-174 -28.103667
## f.pickup_longitude=(-73.92,-73.79]  8.888158e-180 -28.590226
## f.dropoff_latitude=(40.75,40.8]     1.619397e-213 -31.186592
## f.pickup_latitude=(40.75,40.8]      1.619397e-213 -31.186592
## f.dropoff_latitude=(40.8,40.91]     4.580988e-219 -31.593179
## f.pickup_latitude=(40.8,40.91]      4.580988e-219 -31.593179
##
## $category$`4`
##                                       Cla/Mod   Mod/Cla   Global
## f.passenger=(1,6]                   32.744565 100.00000 15.31738
## f.dropoff_latitude=(40.75,40.8]      7.373654  36.92946 25.11967
## f.pickup_latitude=(40.75,40.8]       7.373654  36.92946 25.11967
## f.pickup_longitude=(-73.92,-73.79]   6.501283  31.53527 24.32882
## pick_up_period=afternoon             6.118881  43.56846 35.71280
## pick_up_period=night                 6.821107  21.99170 16.17066
## f.dropoff_longitude=(-74.03,-73.97]  3.947368  19.91701 25.30697
## f.dropoff_latitude=(40.8,40.91]      3.486395  17.01245 24.47451
## f.pickup_latitude=(40.8,40.91]       3.486395  17.01245 24.47451
## f.dropoff_latitude=(40.58,40.69]     3.388430  17.01245 25.18210
## f.pickup_latitude=(40.58,40.69]      3.388430  17.01245 25.18210
```

```
## f.fare_amount=(14,42.5]             3.262787  15.35270 23.60042
## f.total=(16.6,46]                   3.236515  16.18257 25.07804
## pick_up_period=morning              2.806653  11.20332 20.02081
## f.passenger=(0,1]                   0.000000   0.00000 84.68262
##                                          p.value     v.test
## f.passenger=(1,6]                  9.642058e-214  31.203197
## f.dropoff_latitude=(40.75,40.8]     3.094430e-05   4.166403
## f.pickup_latitude=(40.75,40.8]      3.094430e-05   4.166403
## f.pickup_longitude=(-73.92,-73.79]  9.169751e-03   2.605660
## pick_up_period=afternoon            9.964341e-03   2.577064
## pick_up_period=night                1.521547e-02   2.427209
## f.dropoff_longitude=(-74.03,-73.97] 4.488910e-02  -2.005692
## f.dropoff_latitude=(40.8,40.91]     4.387282e-03  -2.848884
## f.pickup_latitude=(40.8,40.91]      4.387282e-03  -2.848884
## f.dropoff_latitude=(40.58,40.69]    1.948584e-03  -3.097959
## f.pickup_latitude=(40.58,40.69]     1.948584e-03  -3.097959
## f.fare_amount=(14,42.5]             1.315500e-03  -3.212577
## f.total=(16.6,46]                   6.878920e-04  -3.394360
## pick_up_period=morning              2.061717e-04  -3.711332
## f.passenger=(0,1]                  9.642058e-214 -31.203197
##
## $category$`5`
##                                        Cla/Mod    Mod/Cla     Global
## f.MTA_tax=(-0.1,0.4]              100.00000000 100.000000  0.4578564
## f.Improvement_surcharge=(-0.1,0.1]  90.47619048  86.363636  0.4370447
## f.fare_amount=(14,42.5]             1.23456790  63.636364 23.6004162
## f.total=(16.6,46]                   1.16182573  63.636364 25.0780437
## f.pickup_longitude=(-73.92,-73.79]  0.85543199  45.454545 24.3288241
## f.dropoff_latitude=(40.58,40.69]    0.82644628  45.454545 25.1821020
## f.pickup_latitude=(40.58,40.69]     0.82644628  45.454545 25.1821020
## f.pickup_longitude=(-74.03,-73.96]  0.08230453   4.545455 25.2861602
## f.fare_amount=(0.1,6]               0.00000000   0.000000 26.0145682
## f.Improvement_surcharge=(0.1,0.8]   0.06270903  13.636364 99.5629553
## f.MTA_tax=(0.4,0.5]                 0.00000000   0.000000 99.5421436
##                                        p.value     v.test
## f.MTA_tax=(-0.1,0.4]              1.186921e-60  16.428952
## f.Improvement_surcharge=(-0.1,0.1] 4.546168e-48  14.567126
## f.fare_amount=(14,42.5]           8.151930e-05   3.939889
## f.total=(16.6,46]                 1.682944e-04   3.762394
## f.pickup_longitude=(-73.92,-73.79] 3.224635e-02   2.141343
## f.dropoff_latitude=(40.58,40.69]  4.104187e-02   2.043107
## f.pickup_latitude=(40.58,40.69]   4.104187e-02   2.043107
## f.pickup_longitude=(-74.03,-73.96] 1.531000e-02  -2.424962
## f.fare_amount=(0.1,6]             1.299719e-03  -3.216042
## f.Improvement_surcharge=(0.1,0.8] 4.546168e-48 -14.567126
## f.MTA_tax=(0.4,0.5]               1.186921e-60 -16.428952
##
## $category$`6`
##                                        Cla/Mod     Mod/Cla     Global
## f.total=(16.6,46]                  66.47302905  99.0111248 25.0780437
## f.fare_amount=(14,42.5]            69.66490300  97.6514215 23.6004162
## f.dropoff_longitude=(-74.03,-73.97] 32.23684211  48.4548826 25.3069719
## pick_up_period=morning             21.72557173  25.8343634 20.0208117
## f.pickup_longitude=(-74.03,-73.96] 20.00000000  30.0370828 25.2861602
```

```
## f.dropoff_latitude=(40.58,40.69]     19.42148760  29.0482077 25.1821020
## f.pickup_latitude=(40.58,40.69]      19.42148760  29.0482077 25.1821020
## f.MTA_tax=(0.4,0.5]                  16.91407067 100.0000000 99.5421436
## f.Improvement_surcharge=(0.1,0.8]    16.91053512 100.0000000 99.5629553
## f.Improvement_surcharge=(-0.1,0.1]    0.00000000   0.0000000  0.4370447
## f.MTA_tax=(-0.1,0.4]                  0.00000000   0.0000000  0.4578564
## f.dropoff_latitude=(40.75,40.8]      14.16735708  21.1372064 25.1196670
## f.pickup_latitude=(40.75,40.8]       14.16735708  21.1372064 25.1196670
## pick_up_period=afternoon             13.63636364  28.9245983 35.7127992
## f.pickup_longitude=(-73.92,-73.79]   11.80496151  17.0580964 24.3288241
## f.dropoff_longitude=(-73.97,-73.95]  10.30502885  15.4511743 25.2445369
## f.dropoff_longitude=(-73.95,-73.91]   8.76460768  12.9789864 24.9323621
## f.fare_amount=(9,14]                  1.35708227   1.9777503 24.5369407
## f.total=(11,16.6]                     0.50590219   0.7416564 24.6826223
## f.total=(7.8,11]                      0.08576329   0.1236094 24.2663892
## f.fare_amount=(0.1,6]                 0.16000000   0.2472188 26.0145682
## f.fare_amount=(6,9]                   0.08051530   0.1236094 25.8480749
## f.total=(-1,7.8]                      0.08012821   0.1236094 25.9729448
##                                           p.value     v.test
## f.total=(16.6,46]                     0.000000e+00        Inf
## f.fare_amount=(14,42.5]               0.000000e+00        Inf
## f.dropoff_longitude=(-74.03,-73.97]   6.525688e-56  15.753234
## pick_up_period=morning                9.940138e-06   4.418472
## f.pickup_longitude=(-74.03,-73.96]    7.802749e-04   3.359699
## f.dropoff_latitude=(40.58,40.69]      6.043008e-03   2.745439
## f.pickup_latitude=(40.58,40.69]       6.043008e-03   2.745439
## f.MTA_tax=(0.4,0.5]                   1.715004e-02   2.383475
## f.Improvement_surcharge=(0.1,0.8]     2.064045e-02   2.314498
## f.Improvement_surcharge=(-0.1,0.1]    2.064045e-02  -2.314498
## f.MTA_tax=(-0.1,0.4]                  1.715004e-02  -2.383475
## f.dropoff_latitude=(40.75,40.8]       3.734544e-03  -2.899755
## f.pickup_latitude=(40.75,40.8]        3.734544e-03  -2.899755
## pick_up_period=afternoon              7.807610e-06  -4.470393
## f.pickup_longitude=(-73.92,-73.79]    5.206341e-08  -5.444116
## f.dropoff_longitude=(-73.97,-73.95]   2.390878e-13  -7.324894
## f.dropoff_longitude=(-73.95,-73.91]   8.723545e-20  -9.103787
## f.fare_amount=(9,14]                  1.190223e-83 -19.377712
## f.total=(11,16.6]                     3.820508e-99 -21.134645
## f.total=(7.8,11]                     9.049492e-107 -21.948001
## f.fare_amount=(0.1,6]                9.385963e-114 -22.667458
## f.fare_amount=(6,9]                  3.962574e-115 -22.806387
## f.total=(-1,7.8]                     8.498683e-116 -22.873665
##
##
## $quanti.var
##                          Eta2      P-value
## Pickup_longitude  0.595855793  0.000000e+00
## Pickup_latitude   0.611624674  0.000000e+00
## Dropoff_longitude 0.488354469  0.000000e+00
## Dropoff_latitude  0.590348251  0.000000e+00
## Passenger_count   0.763317675  0.000000e+00
## Trip_distance     0.636966471  0.000000e+00
## Fare_amount       0.631803706  0.000000e+00
## MTA_tax           1.000000000  0.000000e+00
```

```
## Total_amount      0.645064694  0.000000e+00
## trip_length       0.541374081  0.000000e+00
## trip_distance_km  0.636966471  0.000000e+00
## travel_time       0.432624782  0.000000e+00
## Tip_amount        0.186933260 1.597351e-212
## Tolls_amount      0.047464928  1.957846e-48
## Extra             0.015686740  6.033039e-15
## pick_up_hour      0.006025938  2.303082e-05
##
## $quanti
## $quanti$`1`
##                       v.test Mean in category Overall mean sd in category
## Pickup_latitude    47.598688      40.80159258  40.74593355     0.02844638
## Dropoff_latitude   47.278397      40.79971455  40.74390608     0.02975843
## MTA_tax             3.200623       0.50000000   0.49771072     0.00000000
## pick_up_hour        2.017448      13.77660972  13.48553590     5.95394321
## Dropoff_longitude  -3.712485     -73.94024312 -73.93655617     0.02167351
## Tolls_amount       -4.463365       0.01455979   0.07963788     0.28363577
## Extra              -6.412345       0.30486202   0.35411030     0.37369486
## Passenger_count   -12.108148       1.09001314   1.35150884     0.32490593
## Tip_amount        -12.837756       0.64638633   1.13662227     1.11570091
## travel_time       -19.134311       8.30912816  12.05645354     4.72184646
## trip_length       -21.503319       2.72623912   3.99248724     1.62720935
## Fare_amount       -22.444090       7.81865966  11.11978772     3.16906969
## Trip_distance     -22.823553       1.50456882   2.51292892     0.88814687
## trip_distance_km  -22.823553       2.42136881   4.04416709     1.42933384
## Total_amount      -23.124081       9.58427070  13.48665557     3.57102687
##                   Overall sd       p.value
## Pickup_latitude   0.05518411  0.000000e+00
## Dropoff_latitude  0.05570713  0.000000e+00
## MTA_tax           0.03375500  1.371308e-03
## pick_up_hour      6.80885517  4.364874e-02
## Dropoff_longitude 0.04686801  2.052341e-04
## Tolls_amount      0.68809078  8.068235e-06
## Extra             0.36244956  1.432984e-10
## Passenger_count   1.01920186  9.562854e-34
## Tip_amount        1.80214366  1.007526e-37
## travel_time       9.24234052  1.307976e-81
## trip_length       2.77898821 1.449431e-102
## Fare_amount       6.94118718 1.461647e-111
## Trip_distance     2.08499866 2.676475e-115
## trip_distance_km  3.35548009 2.676475e-115
## Total_amount      7.96414229 2.651072e-118
##
## $quanti$`2`
##                       v.test Mean in category Overall mean sd in category
## Pickup_longitude   49.023893      -73.8702054 -73.93692250     0.03053626
## Dropoff_longitude  45.779901      -73.8649150 -73.93655617     0.03464036
## Extra               3.147025        0.3921958   0.35411030     0.35160997
## MTA_tax             2.031186        0.5000000   0.49771072     0.00000000
## Dropoff_latitude   -2.079665       40.7400378  40.74390608     0.03123097
## Pickup_latitude    -2.539641       40.7412541  40.74593355     0.02618321
## Tolls_amount       -3.466270        0.0000000   0.07963788     0.00000000
## trip_length        -4.378894        3.5861724   3.99248724     2.14405977
```

```
## Passenger_count    -6.442221        1.1322751   1.35150884     0.42222090
## Trip_distance      -6.747609        2.0431785   2.51292892     1.27163247
## trip_distance_km   -6.747609        3.2881770   4.04416709     2.04649408
## travel_time        -7.055161        9.8792431  12.05645354     5.09360578
## Fare_amount        -7.599649        9.3584656  11.11978772     3.87603132
## Total_amount       -9.228985       11.0324868  13.48665557     4.13939940
## Tip_amount        -10.881924        0.4818254   1.13662227     1.03184768
##                   Overall sd      p.value
## Pickup_longitude  0.04075847 0.000000e+00
## Dropoff_longitude 0.04686801 0.000000e+00
## Extra             0.36244956 1.649408e-03
## MTA_tax           0.03375500 4.223615e-02
## Dropoff_latitude  0.05570713 3.755625e-02
## Pickup_latitude   0.05518411 1.109665e-02
## Tolls_amount      0.68809078 5.277321e-04
## trip_length       2.77898821 1.192830e-05
## Passenger_count   1.01920186 1.177374e-10
## Trip_distance     2.08499866 1.503008e-11
## trip_distance_km  3.35548009 1.503008e-11
## travel_time       9.24234052 1.724006e-12
## Fare_amount       6.94118718 2.969348e-14
## Total_amount      7.96414229 2.732091e-20
## Tip_amount        1.80214366 1.405639e-27
##
## $quanti$`3`
##                       v.test Mean in category Overall mean sd in category
## Extra               4.960478       0.39347079   0.35411030     0.35024584
## MTA_tax             3.097931       0.50000000   0.49771072     0.00000000
## pick_up_hour        2.838138      13.90859107  13.48553590     7.20799683
## Tolls_amount       -4.781175       0.00761512   0.07963788     0.20525539
## Passenger_count    -9.593327       1.13745704   1.35150884     0.41658783
## travel_time       -10.448783       9.94229731  12.05645354     5.37245981
## Total_amount      -12.953976      11.22809622  13.48665557     4.19896081
## Fare_amount       -14.496925       8.91686598  11.11978772     3.66431531
## trip_length       -14.646511       3.10142039   3.99248724     1.75093055
## Trip_distance     -14.854197       1.83490603   2.51292892     1.05465751
## trip_distance_km  -14.854197       2.95299500   4.04416709     1.69730673
## Dropoff_longitude -27.476019     -73.96474777 -73.93655617     0.02390941
## Pickup_longitude  -33.990571     -73.96725204 -73.93692250     0.01966000
## Dropoff_latitude  -42.986658      40.69148163  40.74390608     0.02523324
## Pickup_latitude   -44.479810      40.69219742  40.74593355     0.02195261
##                   Overall sd       p.value
## Extra             0.36244956  7.032004e-07
## MTA_tax           0.03375500  1.948768e-03
## pick_up_hour      6.80885517  4.537757e-03
## Tolls_amount      0.68809078  1.742740e-06
## Passenger_count   1.01920186  8.528912e-22
## travel_time       9.24234052  1.484176e-25
## Total_amount      7.96414229  2.230931e-38
## Fare_amount       6.94118718  1.266991e-47
## trip_length       2.77898821  1.418135e-48
## Trip_distance     2.08499866  6.534644e-50
## trip_distance_km  3.35548009  6.534644e-50
## Dropoff_longitude 0.04686801 3.396996e-166
```

```
## Pickup_longitude  0.04075847 3.070519e-253
## Dropoff_latitude  0.05570713  0.000000e+00
## Pickup_latitude   0.05518411  0.000000e+00
##
## $quanti$`4`
##                      v.test Mean in category Overall mean sd in category
## Passenger_count  60.386964        5.2157676    1.3515088      0.6271191
## Extra             2.855277        0.4190871    0.3541103      0.3449229
## trip_length      -2.724836        3.5170536    3.9924872      2.2793668
## Trip_distance    -3.209743        2.0927456    2.5129289      1.4345710
## trip_distance_km -3.209743        3.3679476    4.0441671      2.3087183
## travel_time      -3.502259       10.0241293   12.0564535      5.8153887
## Total_amount     -3.600101       11.6864730   13.4866556      5.3797204
## Fare_amount      -3.769101        9.4771784   11.1197877      4.4696415
##                  Overall sd      p.value
## Passenger_count   1.0192019 0.0000000000
## Extra             0.3624496 0.0042999365
## trip_length       2.7789882 0.0064333407
## Trip_distance     2.0849987 0.0013285377
## trip_distance_km  3.3554801 0.0013285377
## travel_time       9.2423405 0.0004613314
## Total_amount      7.9641423 0.0003180931
## Fare_amount       6.9411872 0.0001638369
##
## $quanti$`5`
##                  v.test Mean in category Overall mean sd in category
## Fare_amount    5.568728       19.3427273   11.1197877      9.6130201
## Total_amount   4.050705       20.3495455   13.4866556      9.9789244
## travel_time    2.216411       16.4142775   12.0564535     10.9292303
## Extra         -3.118768        0.1136364    0.3541103      0.4245805
## MTA_tax      -69.310894        0.0000000    0.4977107      0.0000000
##              Overall sd      p.value
## Fare_amount   6.9411872 2.566060e-08
## Total_amount  7.9641423 5.106344e-05
## travel_time   9.2423405 2.666334e-02
## Extra         0.3624496 1.816088e-03
## MTA_tax       0.0337550 0.000000e+00
##
## $quanti$`6`
##                       v.test Mean in category Overall mean sd in category
## Total_amount      55.004828       27.5334363  13.48665557     7.11825283
## trip_distance_km  54.835178        9.9441521   4.04416709     3.05279683
## Trip_distance     54.835178        6.1790096   2.51292892     1.89692001
## Fare_amount       54.293271       23.2039555  11.11978772     6.31006097
## trip_length       50.382267        8.4820225   3.99248724     2.11630881
## travel_time       45.126987       25.4302677  12.05645354    12.50733914
## Tip_amount        28.621246        2.7905439   1.13662227     2.86474236
## Tolls_amount      15.035141        0.4113721   0.07963788     1.52942437
## MTA_tax            2.115067        0.5000000   0.49771072     0.00000000
## Pickup_latitude   -2.075765       40.7422605  40.74593355     0.05650315
## Passenger_count   -2.169922        1.2805933   1.35150884     0.79810136
## Extra             -2.284046        0.3275649   0.35411030     0.35888046
## Dropoff_latitude  -3.986456       40.7367852  40.74390608     0.05490046
## pick_up_hour      -4.341319       12.5377009  13.48553590     6.76765193
```

```
## Pickup_longitude  -5.657573      -73.9443166 -73.93692250     0.03693034
## Dropoff_longitude -7.103771      -73.9472320 -73.93655617     0.05503831
##                   Overall sd       p.value
## Total_amount      7.96414229  0.000000e+00
## trip_distance_km  3.35548009  0.000000e+00
## Trip_distance     2.08499866  0.000000e+00
## Fare_amount       6.94118718  0.000000e+00
## trip_length       2.77898821  0.000000e+00
## travel_time       9.24234052  0.000000e+00
## Tip_amount        1.80214366 3.655747e-180
## Tolls_amount      0.68809078  4.321258e-51
## MTA_tax           0.03375500  3.442422e-02
## Pickup_latitude   0.05518411  3.791566e-02
## Passenger_count   1.01920186  3.001276e-02
## Extra             0.36244956  2.236883e-02
## Dropoff_latitude  0.05570713  6.706750e-05
## pick_up_hour      6.80885517  1.416300e-05
## Pickup_longitude  0.04075847  1.535284e-08
## Dropoff_longitude 0.04686801  1.213984e-12
##
##
## attr(,"class")
## [1] "catdes" "list "
```

So, first we can assume that all the null hipothesis of independence for the qualitative variables taken can be denied by the chisquare test. It means that all of them have been used somehow to calculate the clustering distances and their splittings, as expected (because we took them from the axis interpretation analysis so we knew they were significative for PCA projections).

Diving inside each category, we can determine the following characterization: ### Category 1: It is defined by individuals contained between this coordinates rangs: - pickup_latitude=(40.8,40.91] - pickup_longitude=(-73.95,-73.92]

- dropoff_latitude=(40.8,40.91]
- dropoff_longitude=(-73.95,-73.91]

It also have a significative representation (almost 48 in Cla/Mod) of rows which total amount are contained in (-1,7.8] rang.

Category 2:

Very similar to previous case, defined also by coordinates values, but this times the rangs are the nexts: - f.pickup_latitude=(40.75,40.8] - f.pickup_longitude=(-73.92,-73.79]

- f.pickup_latitude=(40.75,40.8]
- f.dropoff_longitude=(-73.91,-73.75]

Category 3

As category 1 and 2, it seems to be characterized depending on the coordinates points where the client has been picked up and dropped off. This time, this rangs are: - f.pickup_latitude=(40.58,40.69] - f.pickup_longitude=(-74.03,-73.96] - f.dropoff_longitude=(-74.03,-73.97] - f.dropoff_latitude=(40.58,40.69] However, we can appreciate that in any of the different clusters the rangs that defines the cluster itself are being overlapped (which makes totally sense). Furthermore, it also have a significative representation (almost 41%) of the rows which, this time, its total amount rang is: (11,16.6].

Category 4

This category is determined, with a huge difference between its 2 first v.test values, for passenger variable. Concretely, 100% of their individuals are included in (1,6] rang for f.passenger.

Category 5

This category is gathering all the rows with MTA_tax = 0 and also the 90% of the individuals which its improvement surcharge is 0 are in this category. So MTA_tax and Improvement_surcharge explains the behaviour of category 5 rows, and 63% of individuals of this category it has its total_amount value compressed between (16.6,46].

Category 6

Definitely, category 6 is defined by those rows which their total_amount is in the rang: (16.6,46] and their fare_amount between (14,42.5] (so , the most expensive one). We can appreciate how 100% of this individuals have: - MTA_tax = 0.5 - Improvement surcharge != 0

$quanti section

We can observe how all the peculiarities pointed at the cluster analysis are proved by the quantitative variables output. - For quanti 1, 2 and 3 the most significative quantitative variables are latitudes and longitudes - For quanti 4 we have passenger_count at the top - quanti 5 has a huge negative value for MTA_tax (which make sense with the description above) - And in quanti 6 Total_amount, trip_distance and Fare_amount are distinguished as more correlated.

Axes description

At this point, this output does not help anymore to detail our clusters. But we can also see how the interpretation of axis made in a past section corresponds to the characterization of the clusters, being the variables that explain the most each specific dimension the ones that are also distinguished for each cluster who has a higher v.test value for the dimension in question.

```r
#Block B descripcion per eixos
res.hcpc$desc.axes
```

```
## $quanti.var
##             Eta2     P-value
## Dim.1 0.67109834 0.00000e+00
## Dim.2 0.63739845 0.00000e+00
## Dim.3 0.52264765 0.00000e+00
## Dim.5 0.74730335 0.00000e+00
## Dim.6 0.68776489 0.00000e+00
## Dim.4 0.09331422 2.09914e-99
##
## $quanti
## $quanti$`1`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.2  27.443757       0.87005073 -5.062992e-13      0.7201106   1.496147
## Dim.4   8.479036       0.19987477  6.084197e-13      1.0456575   1.112461
## Dim.5  -3.241831      -0.06905784  4.682821e-13      0.2931861   1.005301
## Dim.6 -10.112275      -0.21334016  5.340819e-14      0.3734743   0.995628
```

```
## Dim.1 -26.553752      -1.21944878 -5.185978e-14      0.9965003   2.167260
## Dim.3 -37.000861      -0.96043352 -2.104400e-12      0.6572895   1.224980
##             p.value
## Dim.2 8.247983e-166
## Dim.4  2.270676e-17
## Dim.5  1.187646e-03
## Dim.6  4.873905e-24
## Dim.1 2.324363e-155
## Dim.3 1.109185e-299
##
## $quanti$`2`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.3 41.460145        1.6957865 -2.104400e-12      0.9695120   1.224980
## Dim.2 27.468540        1.3722128 -5.062992e-13      0.7081247   1.496147
## Dim.5  8.505206        0.2854910  4.682821e-13      0.2870454   1.005301
## Dim.4 -7.452525       -0.2768215  6.084197e-13      1.0713248   1.112461
## Dim.6 -7.870793       -0.2616538  5.340819e-14      0.4483228   0.995628
## Dim.1 -9.164301       -0.6631655 -5.185978e-14      1.2653732   2.167260
##             p.value
## Dim.3  0.000000e+00
## Dim.2 4.172984e-166
## Dim.5  1.812738e-17
## Dim.4  9.157017e-14
## Dim.6  3.523995e-15
## Dim.1  4.986901e-20
##
## $quanti$`3`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.3   7.038462        0.1887540 -2.104400e-12      0.6177491   1.224980
## Dim.5   6.740043        0.1483365  4.682821e-13      0.2850575   1.005301
## Dim.4  -6.995461       -0.1703690  6.084197e-13      1.0578699   1.112461
## Dim.6  -8.181636       -0.1783309  5.340819e-14      0.4418606   0.995628
## Dim.1 -10.116389       -0.4799831 -5.185978e-14      1.1555060   2.167260
## Dim.2 -52.249889       -1.7113907 -5.062992e-13      0.6579794   1.496147
##            p.value
## Dim.3 1.943727e-12
## Dim.5 1.583393e-11
## Dim.4 2.643875e-12
## Dim.6 2.800150e-16
## Dim.1 4.673400e-24
## Dim.2 0.000000e+00
##
## $quanti$`4`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.6  57.325846        3.5835246  5.340819e-14      0.6592727   0.995628
## Dim.4  10.014128        0.6994567  6.084197e-13      1.0751834   1.112461
## Dim.3   6.097473        0.4689662 -2.104400e-12      1.0864753   1.224980
## Dim.1  -3.228332       -0.4392907 -5.185978e-14      1.5069990   2.167260
## Dim.5 -10.334111       -0.6522768  4.682821e-13      0.4893148   1.005301
##            p.value
## Dim.6 0.000000e+00
## Dim.4 1.321221e-23
## Dim.3 1.077584e-09
## Dim.1 1.245146e-03
```

```
## Dim.5 4.939735e-25
##
## $quanti$`5`
##           v.test Mean in category  Overall mean sd in category Overall sd
## Dim.3   3.719915        0.9693915 -2.104400e-12      1.4406635   1.224980
## Dim.1   2.851316        1.3146003 -5.185978e-14      2.0845845   2.167260
## Dim.6  -2.838523       -0.6012109  5.340819e-14      1.0513642   0.995628
## Dim.4 -14.839165       -3.5118156  6.084197e-13      0.9565864   1.112461
## Dim.5 -58.231768      -12.4535564  4.682821e-13      0.3938538   1.005301
##            p.value
## Dim.3 1.992900e-04
## Dim.1 4.353871e-03
## Dim.6 4.532286e-03
## Dim.4 8.176680e-50
## Dim.5 0.000000e+00
##
## $quanti$`6`
##          v.test Mean in category  Overall mean sd in category Overall sd
## Dim.1 55.721029       3.87228363 -5.185978e-14      1.6171043   2.167260
## Dim.5  4.011825       0.12932279  4.682821e-13      1.0278268   1.005301
## Dim.2  3.136536       0.15047408 -5.062992e-13      1.4093541   1.496147
## Dim.4  2.136146       0.07619963  6.084197e-13      1.0737796   1.112461
## Dim.6 -2.649006      -0.08457014  5.340819e-14      0.9366946   0.995628
## Dim.3 -7.213345      -0.28333651 -2.104400e-12      1.2119561   1.224980
##            p.value
## Dim.1 0.000000e+00
## Dim.5 6.025103e-05
## Dim.2 1.709563e-03
## Dim.4 3.266749e-02
## Dim.6 8.072889e-03
## Dim.3 5.459386e-13
##
##
## attr(,"class")
## [1] "catdes" "list "
```

Invidual analysis

Again, this command can help us now to confirm the conclusions made until now. As an example, we will look a paragon of C6, and how its total_amount is served in some middle-point of the last rang (16.6,46] for this variable (total_amount = 26.3), and also look at how a distinguished C6 row has one of the possible highest values for total_amount (= 44.8).

If we keep tracking for the rest of the clusters, we can assume that the conclusions made below are concordant.

```
#Block C individus
res.hcpc$desc.ind
```

```
## $para
## Cluster: 1
##    419422    253799   1435375     87900     92598
## 0.3344669 0.3389733 0.4344772 0.4478629 0.4596034
## --------------------------------------------------------
## Cluster: 2
```

```
##    746656     90225   1362378   1372589   1363325
## 0.5465155 0.5990569 0.6160073 0.6186373 0.6641605
## ---------------------------------------------------------
## Cluster: 3
##    473230   1369263   1076497    474605    748186
## 0.5275761 0.5798164 0.6081080 0.6235652 0.6706031
## ---------------------------------------------------------
## Cluster: 4
##    473235    745377   1361206   1370940    415886
## 0.6884401 0.7522253 1.0351214 1.0543070 1.1090335
## ---------------------------------------------------------
## Cluster: 5
##   272451   725855   782156   829507   885046
## 1.361694 1.376902 1.737938 1.804818 1.842608
## ---------------------------------------------------------
## Cluster: 6
##    678042     53452     50328   1406084     11713
## 0.9302321 1.0911229 1.1229448 1.1666184 1.1683024
##
## $dist
## Cluster: 1
##   572868   915921  1178619   955214   529632
## 5.229943 5.148077 5.109972 5.064625 4.344363
## ---------------------------------------------------------
## Cluster: 2
##   532659  1404537   657301     9207   576477
## 6.142807 5.957720 5.839240 5.818861 5.809140
## ---------------------------------------------------------
## Cluster: 3
##   274842   645383  1157271   229968   573367
## 5.790128 5.294909 5.290883 5.217173 5.210914
## ---------------------------------------------------------
## Cluster: 4
##  1137082   329313  1112510   749823  1123289
## 7.439163 7.017419 6.044010 5.771125 5.661158
## ---------------------------------------------------------
## Cluster: 5
##   675043   978944   984283  1283504   424236
## 13.83254 13.78378 13.61879 13.53750 13.51863
## ---------------------------------------------------------
## Cluster: 6
##    285458    868718   1135563    425343    154581
## 12.172935 10.987981  9.975215  9.874141  9.794942
```

```
#parangon C6
df["678042",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 678042 VeriFone Inc.  2016-01-15 13:24:58   2016-01-15 14:00:17
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 678042      Store_and_fwd Standard rate        -73.90332        40.74579
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 678042         -73.98235          40.7681               1          4.95
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 678042          25     0     0.5          3            0
```

```
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 678042                   0.3         28.8  Credit card Street-hail       3
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 678042 AnyTip Yes      8.1186         7.966253    35.31667           13
##        pick_up_period   espeed f.passenger  f.distance f.pickup_longitude
## 678042         valley 13.79281       (0,1] (3.31,11.1]    (-73.92,-73.79]
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 678042     (40.75,40.8]      (-74.03,-73.97]       (40.75,40.8]
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 678042     (14,42.5] (-0.1,0.5] (0.4,0.5]               (0.1,0.8]
##        f.tip_amount f.toll    f.total f.outlierPCAd1 f.outlierPCAd2
## 678042       (1,22] (-1,1] (16.6,46]      NoOutDim1      NoOutDim2
##        f.outlierPCAd3 f.outlierPCAd4
## 678042      NoOutDim3      NoOutDim4
```

```
#distinguished C6
df["285458",]
```

```
##             VendorID lpep_pickup_datetime Lpep_dropoff_datetime
## 285458 VeriFone Inc.  2016-01-07 01:26:54   2016-01-07 01:41:13
##        Store_and_fwd_flag    RateCodeID Pickup_longitude Pickup_latitude
## 285458      Store_and_fwd Standard rate        -73.94707        40.81071
##        Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
## 285458         -74.01742         40.85104               1         10.47
##        Fare_amount Extra MTA_tax Tip_amount Tolls_amount
## 285458          29   0.5     0.5          4         10.5
##        improvement_surcharge Total_amount Payment_type   Trip_type mis_ind
## 285458                   0.3         44.8  Credit card Street-hail       2
##            AnyTip trip_length trip_distance_km travel_time pick_up_hour
## 285458 AnyTip Yes    12.83019         16.84983    14.31667            1
##        pick_up_period  espeed f.passenger  f.distance f.pickup_longitude
## 285458          night 53.7703       (0,1] (3.31,11.1]    (-73.96,-73.95]
##        f.pickup_latitude f.dropoff_longitude f.dropoff_latitude
## 285458      (40.8,40.91]     (-74.03,-73.97]       (40.8,40.91]
##        f.fare_amount    f.extra f.MTA_tax f.Improvement_surcharge
## 285458     (14,42.5] (-0.1,0.5] (0.4,0.5]               (0.1,0.8]
##        f.tip_amount f.toll    f.total f.outlierPCAd1 f.outlierPCAd2
## 285458       (1,22] (1,50] (16.6,46]      NoOutDim1      NoOutDim2
##        f.outlierPCAd3 f.outlierPCAd4
## 285458      NoOutDim3      NoOutDim4
```

Assigning clusters groups

Now we assign to each row the cluster group decided by HPC method and we also consider as group 7 the outliers (which they haven't been taking into consideration until now).

```
#Donar-li una classe (the last one) a tots els outliers multidimensionals (sup.)
df$claHP<-7
df[row.names(res.hcpc$data.clust),"claHP"]<-res.hcpc$data.clust$clust
table(df$claHP)
```

```
##
##    1    2    3    4    5    6    7
## 1522  756 1455  241   22  809   61
```

K-Means Classification

We execute kmeans command defining 6 clusters in order to get the same number of groups as in the hierarchical process.

```
ppcc<-res.pca$ind$coord[,1:6]
dim(ppcc)
```

```
## [1] 4805    6
```

```
kc<-kmeans(ppcc,6,iter.max = 30, trace=T)
```

```
## KMNS(*, k=6): iter=  1, indx=3
##  QTRAN(): istep=4805, icoun=6
##  QTRAN(): istep=9610, icoun=104
##  QTRAN(): istep=14415, icoun=38
##  QTRAN(): istep=19220, icoun=411
##  QTRAN(): istep=24025, icoun=1625
##  QTRAN(): istep=28830, icoun=1020
##  QTRAN(): istep=33635, icoun=1950
##  QTRAN(): istep=38440, icoun=1950
## KMNS(*, k=6): iter=  2, indx=12
##  QTRAN(): istep=4805, icoun=43
##  QTRAN(): istep=9610, icoun=108
##  QTRAN(): istep=14415, icoun=41
##  QTRAN(): istep=19220, icoun=103
##  QTRAN(): istep=24025, icoun=35
##  QTRAN(): istep=28830, icoun=174
##  QTRAN(): istep=33635, icoun=38
##  QTRAN(): istep=38440, icoun=145
##  QTRAN(): istep=43245, icoun=104
##  QTRAN(): istep=48050, icoun=613
##  QTRAN(): istep=52855, icoun=333
##  QTRAN(): istep=57660, icoun=254
##  QTRAN(): istep=62465, icoun=41
##  QTRAN(): istep=67270, icoun=434
##  QTRAN(): istep=72075, icoun=59
##  QTRAN(): istep=76880, icoun=112
##  QTRAN(): istep=81685, icoun=364
##  QTRAN(): istep=86490, icoun=987
## KMNS(*, k=6): iter=  3, indx=3
##  QTRAN(): istep=4805, icoun=16
##  QTRAN(): istep=9610, icoun=41
##  QTRAN(): istep=14415, icoun=1
##  QTRAN(): istep=19220, icoun=138
##  QTRAN(): istep=24025, icoun=1488
##  QTRAN(): istep=28830, icoun=4182
##  QTRAN(): istep=33635, icoun=232
##  QTRAN(): istep=38440, icoun=980
##  QTRAN(): istep=43245, icoun=2375
##  QTRAN(): istep=48050, icoun=295
##  QTRAN(): istep=52855, icoun=3164
## KMNS(*, k=6): iter=  4, indx=12
##  QTRAN(): istep=4805, icoun=39
##  QTRAN(): istep=9610, icoun=123
```

```
##  QTRAN(): istep=14415, icoun=20
##  QTRAN(): istep=19220, icoun=1393
##  QTRAN(): istep=24025, icoun=116
##  QTRAN(): istep=28830, icoun=1393
##  QTRAN(): istep=33635, icoun=2399
##  QTRAN(): istep=38440, icoun=773
##  QTRAN(): istep=43245, icoun=232
##  QTRAN(): istep=48050, icoun=804
##  QTRAN(): istep=52855, icoun=2149
##  QTRAN(): istep=57660, icoun=1003
##  QTRAN(): istep=62465, icoun=773
## KMNS(*, k=6): iter=  5, indx=59
##  QTRAN(): istep=4805, icoun=292
##  QTRAN(): istep=9610, icoun=231
##  QTRAN(): istep=14415, icoun=2056
##  QTRAN(): istep=19220, icoun=85
## KMNS(*, k=6): iter=  6, indx=4805
```

```
table(kc$cluster)
```

```
##
##    1    2    3    4    5    6
##  508 1375  891  721  739  571
```

Assigning clusters groups

As we also did before in HPC, we assign the clusters in a way that group 7 is taken by the outliers.

```
df$claKM<-7
df[names(kc$cluster),"claKM"]<-kc$cluster
kc$betweenss/kc$totss
```

```
## [1] 0.5405259
```

```
table(df$claKM)
```

```
##
##    1    2    3    4    5    6    7
##  508 1375  891  721  739  571   61
```

Characterization of Kmeans clustering

As we didn't manage to execute catdes command, we've tried to be a bit creative and search for internet. At last, we found interesting "$centers" and we realized we could try to give it a try in order to get some notion about wether the clustering done by Kmeans was similar or not to HCPC.

```
kc$centers
```

```
##        Dim.1       Dim.2       Dim.3        Dim.4       Dim.5       Dim.6
## 1  1.4530232  1.05429532 -0.55385928 -0.003266492 -0.69657514 -0.17513429
## 2 -1.5167506  0.83512406 -0.91899485  0.243338520 -0.10379837 -0.01379821
## 3 -0.3681682 -1.55673754 -0.04676337 -1.035916295  0.24983708  0.14927969
## 4 -0.4327736 -1.80937599  0.49572014  0.995491293 -0.03659532 -0.17221895
## 5 -0.8283187  1.34566680  1.78442819 -0.225716979  0.24294974  0.05198588
## 6  4.5526995  0.02327121 -0.15667611  0.068521816  0.21159975  0.10627820
```

```r
#catdes(df,47)
#veure si s'han posat d'acord o no
table(df$claHP,df$claKM)
```

```
##
##         1    2    3    4    5    6    7
##    1  193 1301   17   11    0    0    0
##    2   63    4    6    4  679    0    0
##    3    8    0  804  643    0    0    0
##    4   22   70   39   38   59   13    0
##    5   22    0    0    0    0    0    0
##    6  200    0   25   25    1  558    0
##    7    0    0    0    0    0    0   61
```

The output it's a bit messy, but we can certainly appreciate how cluster1 it's centred by Dimension 1 (which is the axis that increases with total_amount prices) and, at least, also check how Dimension 2 and cluster4 is very correlated. These two clusters, as we can observe in the next section, are precisely associated with cluster6 and cluster4 (in this order) by the labels done in HCPC. So, even though we haven't been able to interpret the categorical description, we can predict that both methods (kmeans and hierarchical) are actually generating a very similar groups of individuals.

Re-labeling

After all, we generate a new label for Kmeans groups so the cluster numbers are referring to the same group and avoid further confusions.

To finalize, we check the diagonal summatory of the number of invidiuals from the contingency table generated, to have an idea of the bias taken by kmeans respect to HCPC.

```r
df$claHP<-factor(df$claHP,labels=paste("kHP-",1:7))
df$claKM<-factor(df$claKM,levels=c(6,4,1,5,2,3,7),labels=c("kKM-6","kKM-4","kKM-1","kKM-5","kKM-2","kKM
tt<-table(df$claHP,df$claKM)
tt
```

```
##
##           kKM-6 kKM-4 kKM-1 kKM-5 kKM-2 kKM-3 kKM-7
##    kHP- 1     0    11   193     0  1301    17     0
##    kHP- 2     0     4    63   679     4     6     0
##    kHP- 3     0   643     8     0     0   804     0
##    kHP- 4    13    38    22    59    70    39     0
##    kHP- 5     0     0    22     0     0     0     0
##    kHP- 6   558    25   200     1     0    25     0
##    kHP- 7     0     0     0     0     0     0    61
```

```r
sum(diag(tt)/sum(tt))
```

```
## [1] 0.03226469
```