

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 activ, where TotalAmount nd Anytip are supementary.

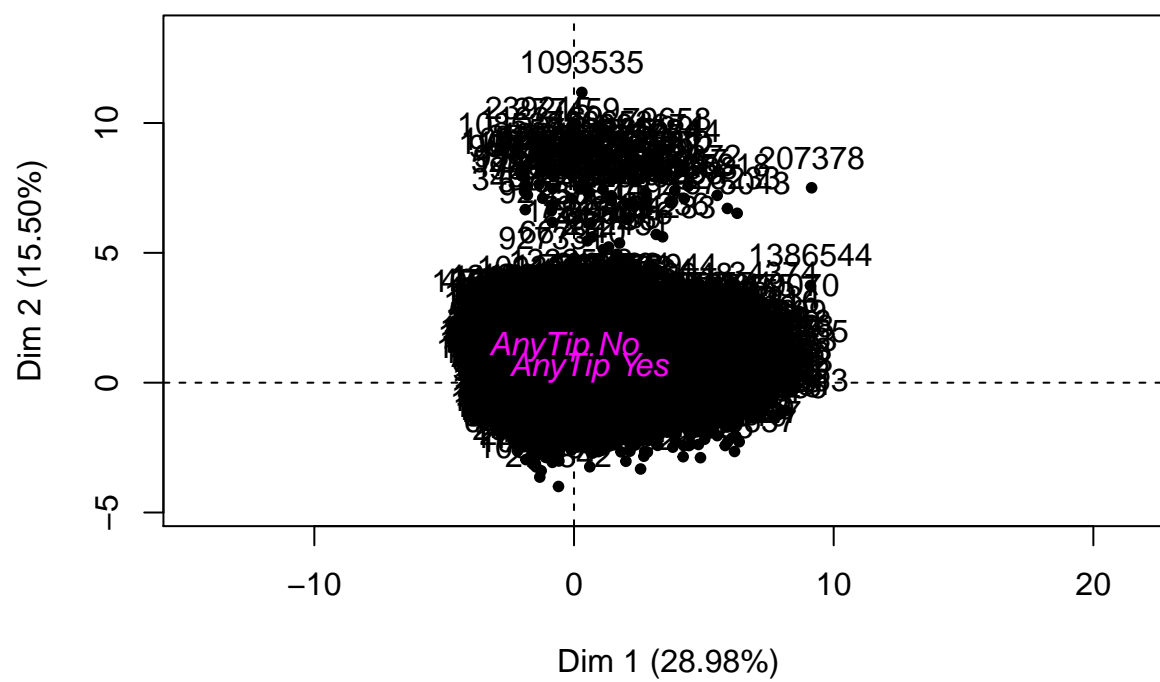
```
load("Taxi5000_raw_DataClean.RData")
library(FactoMineR)
names (df)
```

```
## [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"
```

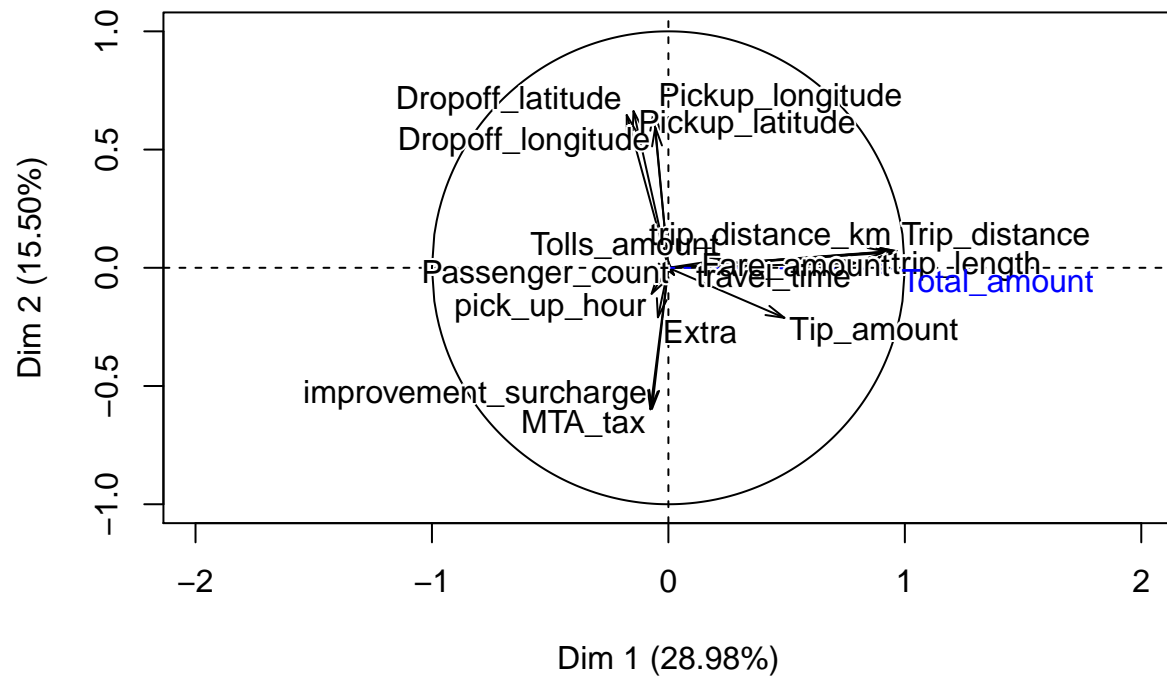
```
vars_con_pca<-c(6,7,8,9,10,11,12,13,14,15,16,17,18,22,23,24,25,26)
```

```
#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)

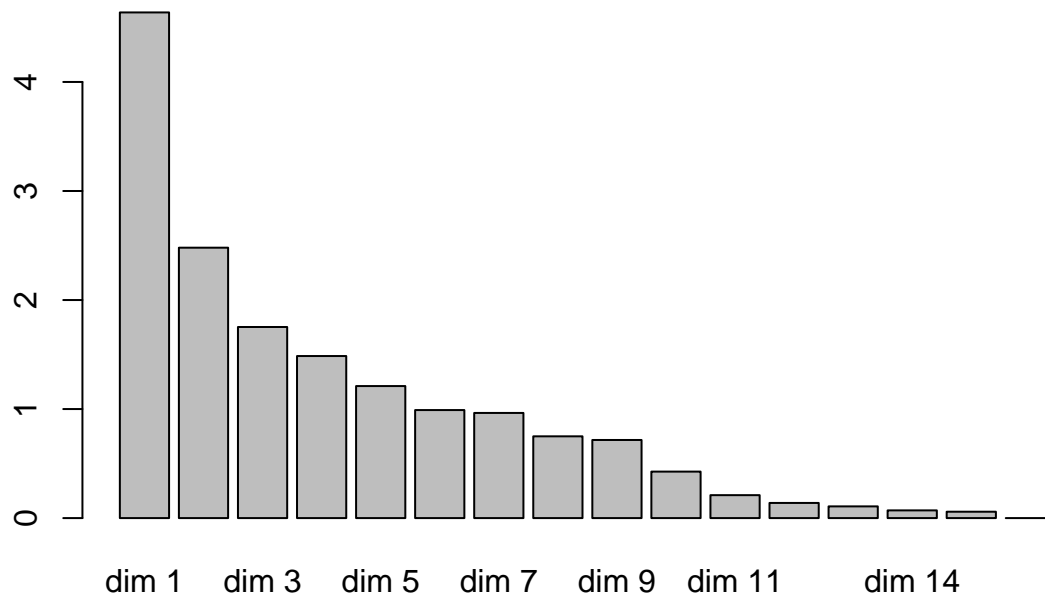


Variables factor map (PCA)

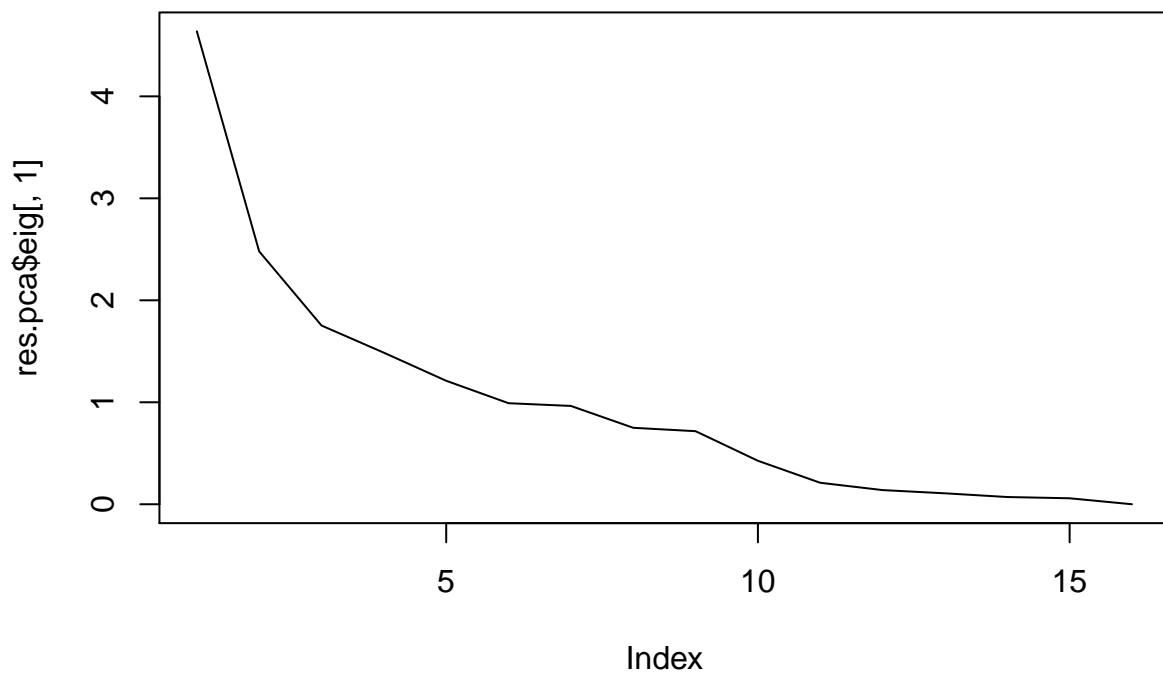


```
barplot(res.pca$eig[,1], main="Eigenvalues", names.arg = paste("dim", 1:nrow(res.pca$eig)))
```

Eigenvalues



```
# 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
```



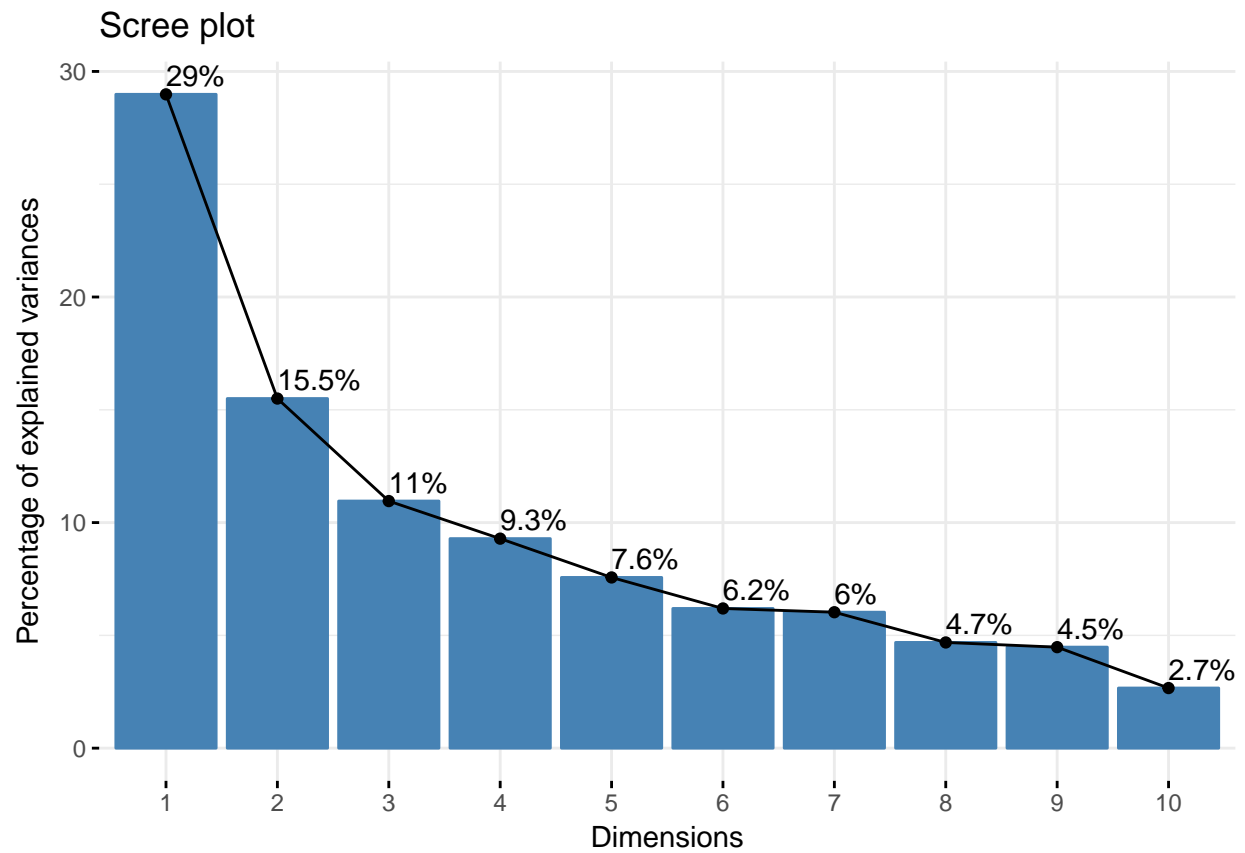
```
length <-length(which(res.pca$eig[,1]>=1));length
```

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

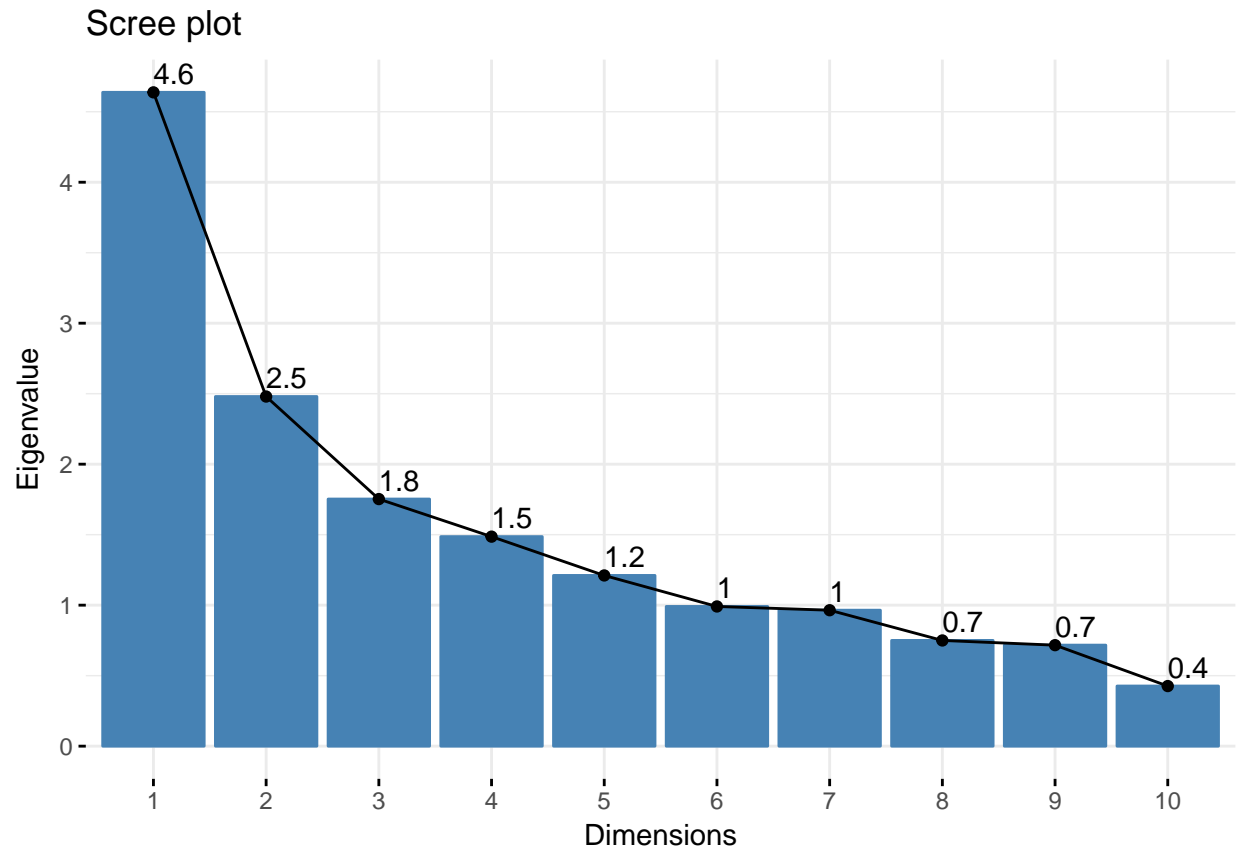
```
kaiser <- res.pca$eig[1:length,1] #keep only EV >=1 ->first 7
```

```
#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)
```



```
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

```

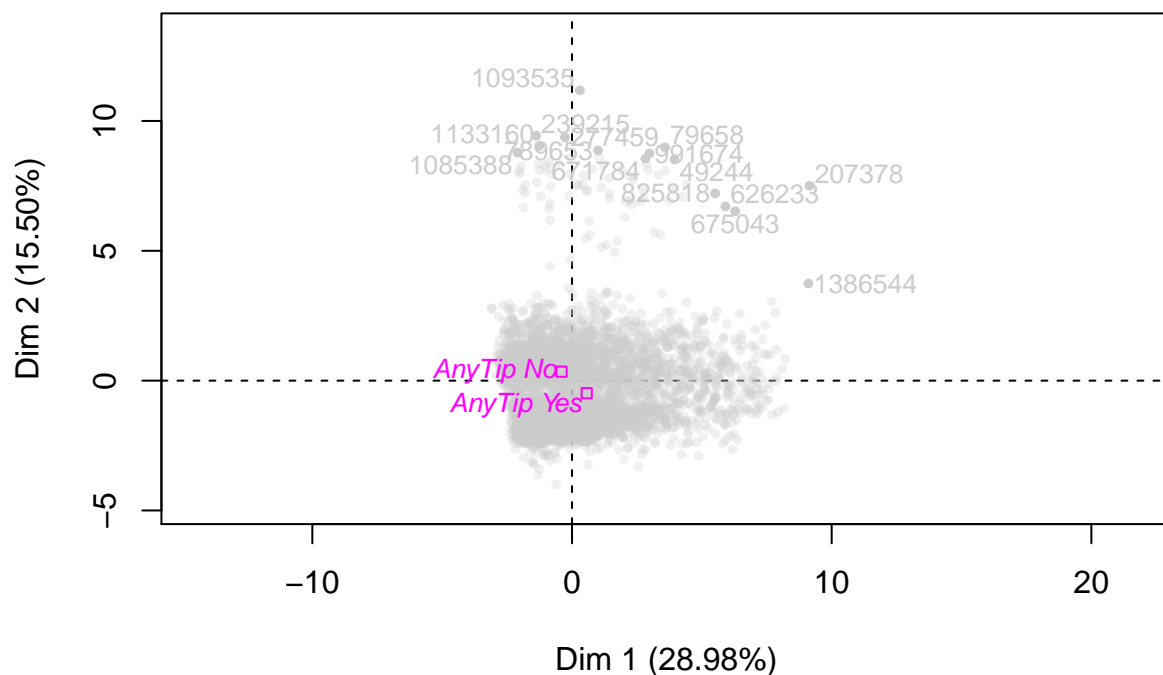


```
##          Dim.1  cos2   Dim.2  cos2   Dim.3  cos2
## Total_amount | 0.966 0.933 | -0.004 0.000 | 0.068 0.005 |
##          Dim.4  cos2
## Total_amount -0.029 0.001 |
##
## Supplementary categories
##          Dist   Dim.1   cos2  v.test   Dim.2
## AnyTip No      | 0.742 | -0.404 0.296 -15.479 | 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))
```

Individuals factor map (PCA)



```
#DIMENSION1
```

#Since the multivariant detection didnt manage to find outliers well enough 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
iqrvar<-IQR(res.pca$ind$coord[,1])
quantil3<-quantile(res.pca$ind$coord[,1], .75);quantil3 #get 3rd quartile

##      75%
## 1.029432

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

## [1] 2

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

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"

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

#DIMENSION2
#characteristic of extreme outliers 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

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

##      75%
## 0.8072157

```

```

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

## [1] 60

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

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"

#DIMENSION3
#characteristic of extreme outliers 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

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

## 75%
## 0.6595931

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

## [1] 0

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

```

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

```
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"
```

```
#DIMENSION4
```

```
#characteristic of extreme outliers 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
```

```
iqrvar<-IQR(res.pca$ind$coord[,4])
```

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

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

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

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

```
df$f.outlierPCAd4<-0
```

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

```
summary(df$f.outlierPCAd4)
```

```
## NoOutDim4
```

```
##      4866
```

```
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 have 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

```
# 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
```



```

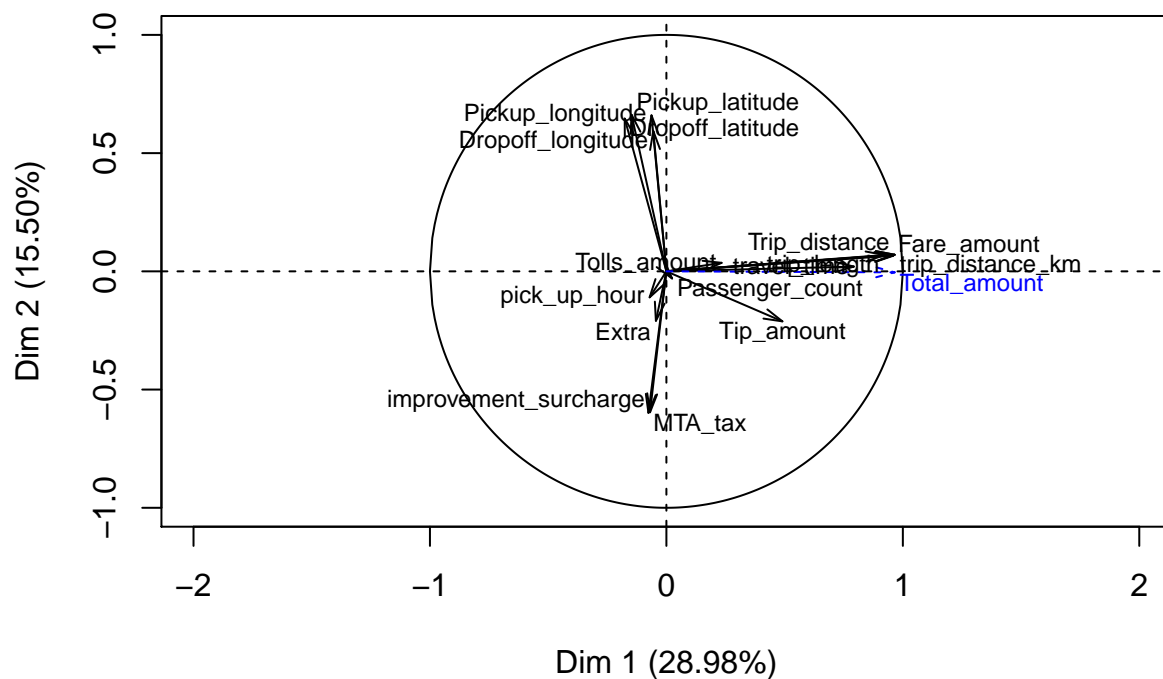
## trip_distance_km      0.07016348 9.624966e-07
## Trip_distance         0.07016348 9.624966e-07
## trip_length           0.06889982 1.503966e-06
## Fare_amount           0.06612191 3.906140e-06
## Tolls_amount          0.03596743 1.210264e-02
## Passenger_count       -0.03024353 3.489026e-02
## pick_up_hour          -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
```

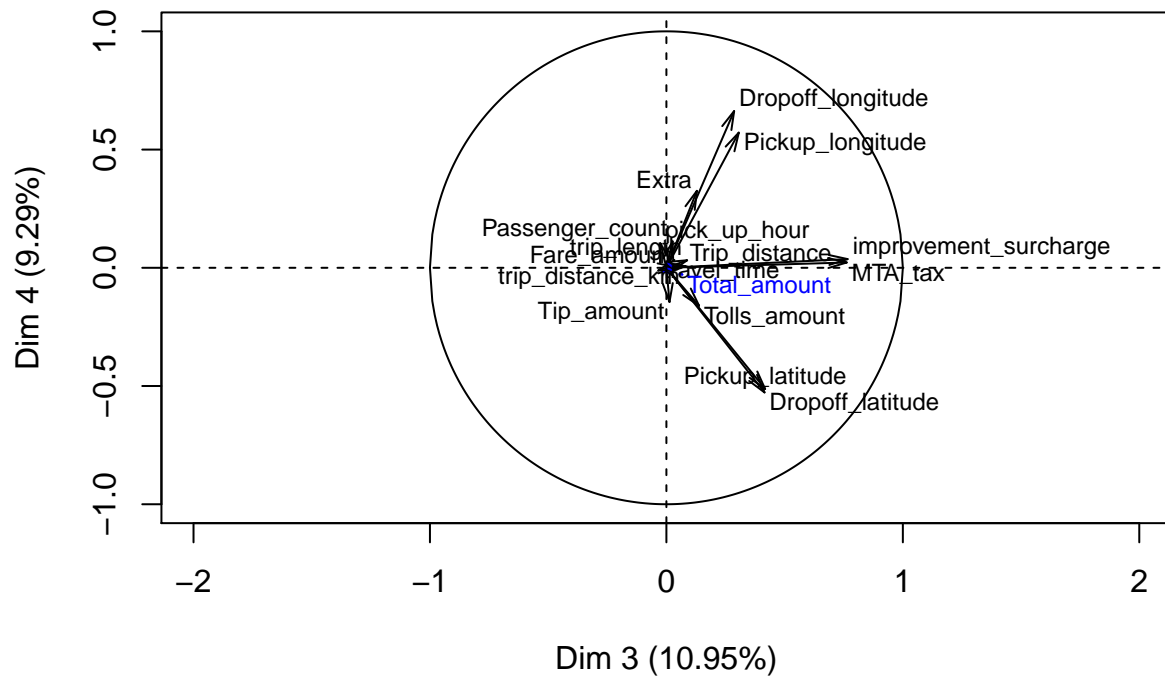
#The first Dimention is best described by the quantative variables Total_amount, trip_distance and Fare.
`plot(res.pca,choix="var", cex = 0.75)`

Variables factor map (PCA)



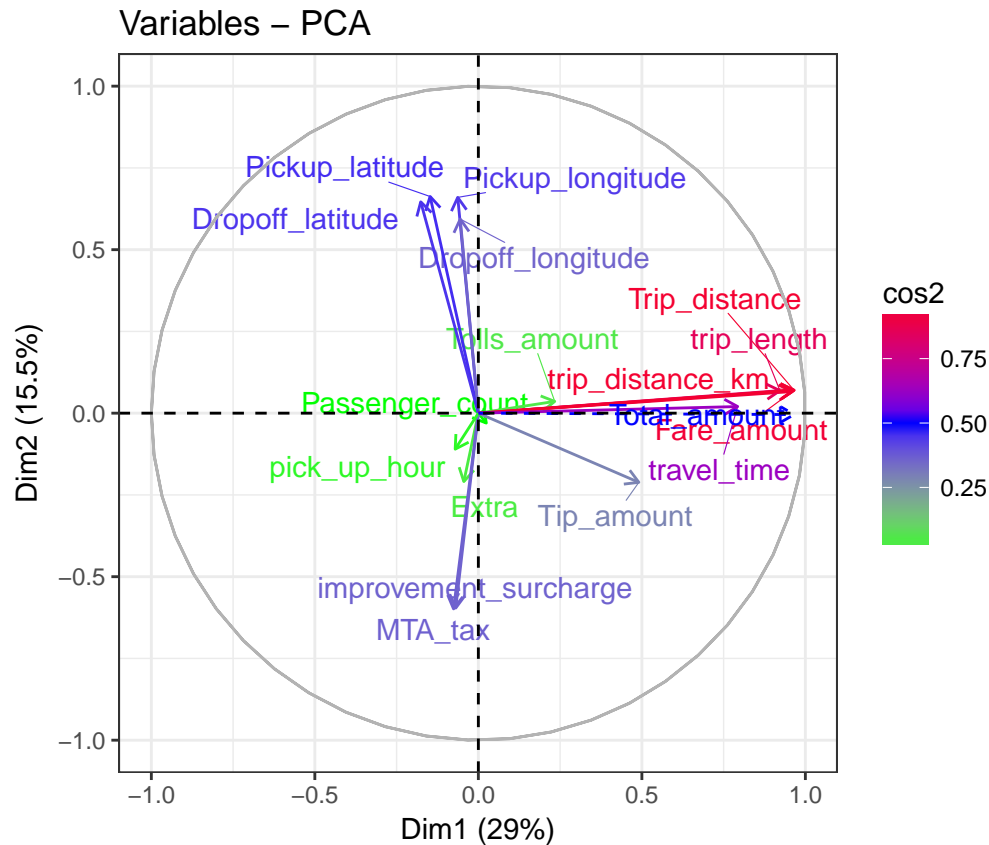
```
plot(res.pca,choix="var", cex = 0.75, axes = (3:4))# 3rd and 4th PCA
```

Variables factor map (PCA)



```
#modern factoeextra
```

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



IV PCA execution with supplementary individuals

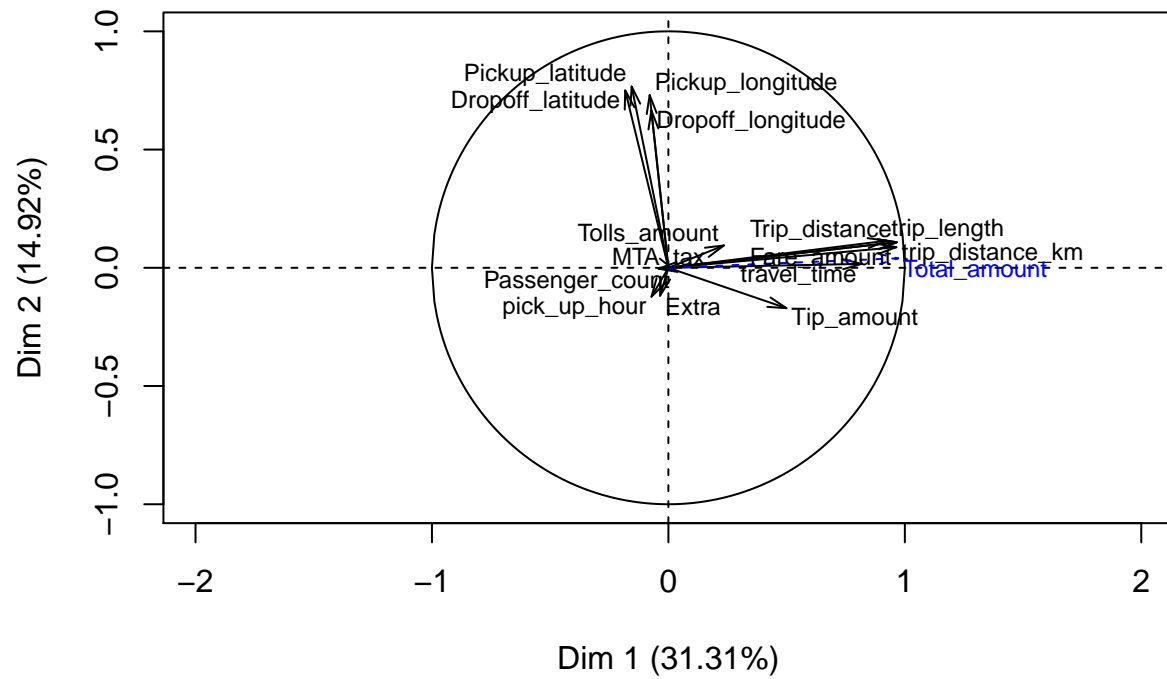
```
vec_out <- llvout
vars_con_pca <- names(df)[c(6:16,18,23:26)]
# We do a PCA analysis using the factorial variables Fare amount, total and the pickup perio in order t

res.pca<-PCA(df[,c(vars_con_pca, "f.fare_amount", "f.total", "pick_up_period")], ncp=6, quanti.sup=which

## 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

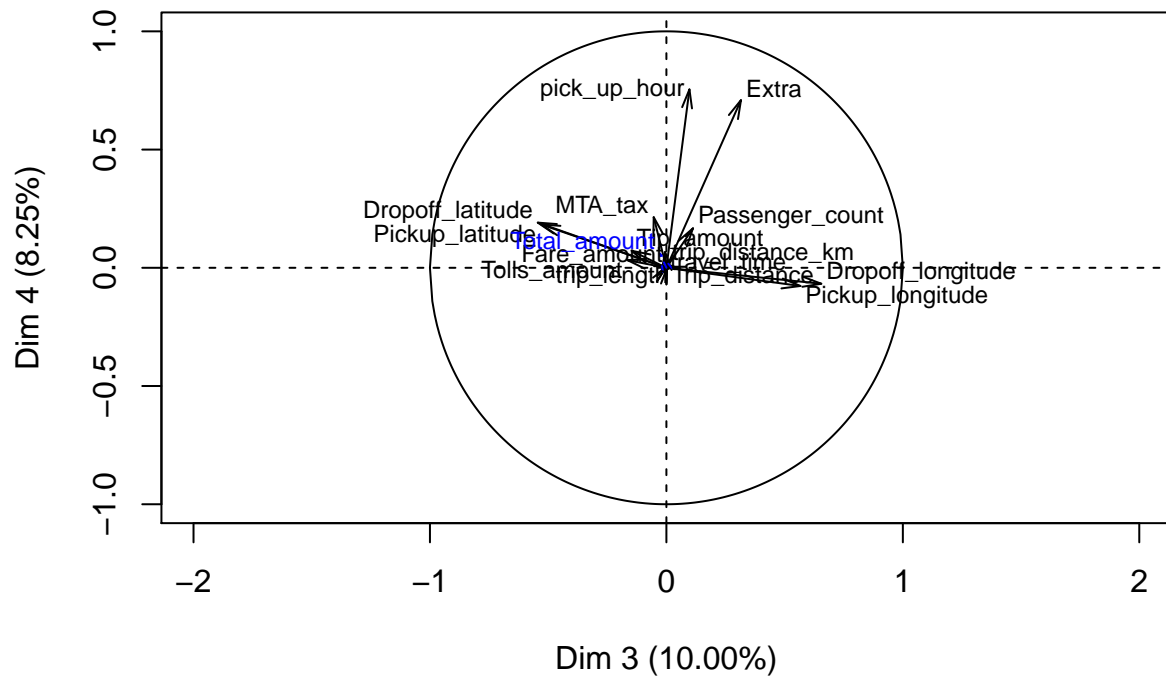
plot(res.pca,choix="var", cex = 0.75)
```

Variables factor map (PCA)

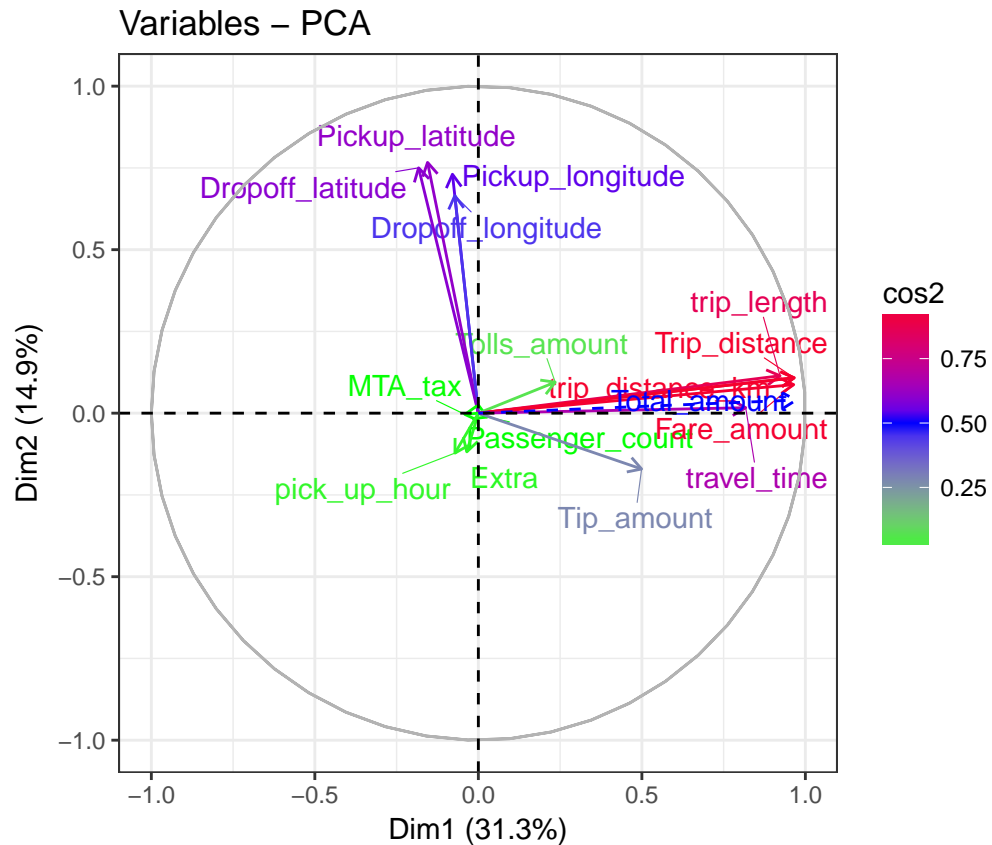


```
plot(res.pca,choix="var", cex = 0.75, axes = (3:4))# 3rd and 4th PCA
```

Variables factor map (PCA)



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



Hierarchical clustering

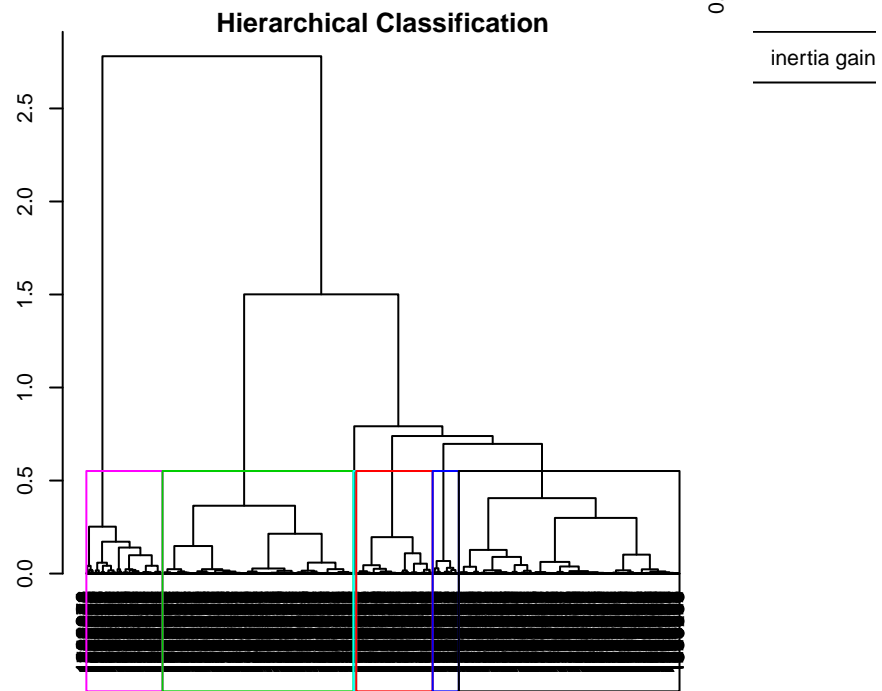
Generem 8 clusters amb el mètode Hierarchichal a partir de les projeccions obtingudes amb el PCA.

```
library(FactoMineR)
library(ggplot2)
library(ggdendro)
```

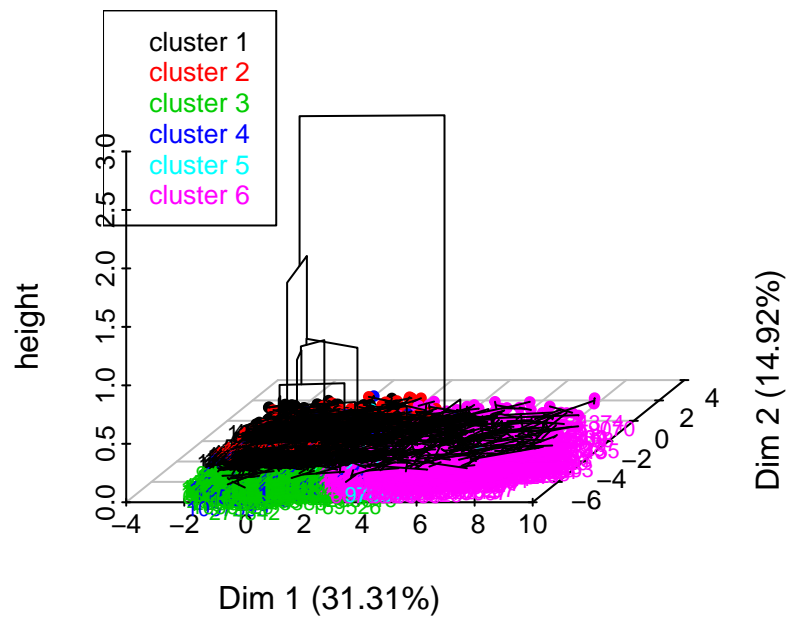
```
## Warning: package 'ggdendro' was built under R version 3.4.4
```

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

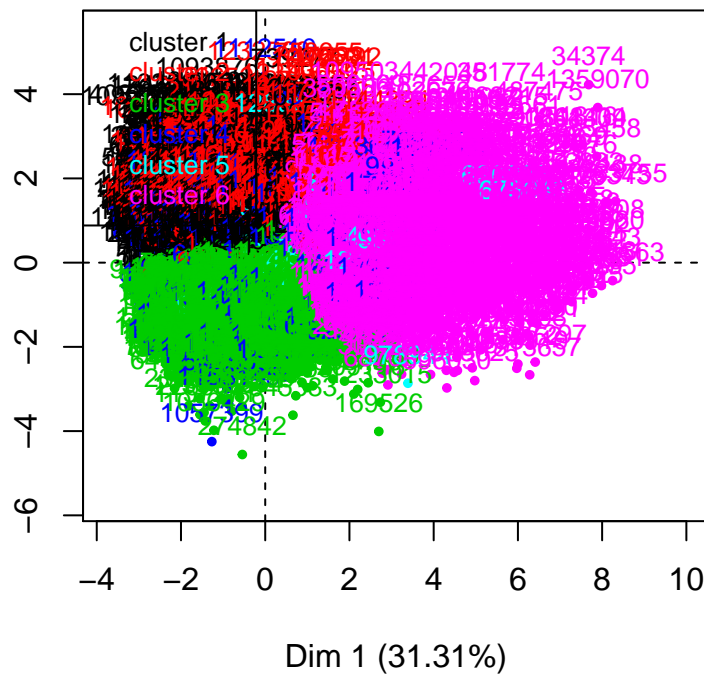
Hierarchical Clustering



Hierarchical clustering on the factor map



Factor map



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

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

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

```
## $test.chi2
##                p.value df
## f.fare_amount  0.000000e+00 15
## f.total        0.000000e+00 15
## pick_up_period 1.388397e-34 15
##
## $category
## $category$`1`
##                Cla/Mod  Mod/Cla  Global      p.value
## f.total=(-1,7.8]      47.996795 39.356110 25.97294  3.010774e-45
## f.fare_amount=(0.1,6] 45.600000 37.450723 26.01457  1.371224e-33
## f.fare_amount=(6,9]   42.028986 34.296978 25.84807  2.926343e-19
## f.total=(7.8,11]      39.536878 30.289093 24.26639  5.908761e-11
## pick_up_period=morning 40.540541 25.624179 20.02081  7.753913e-11
## pick_up_period=valley  36.740741 32.588699 28.09573  2.854485e-06
## pick_up_period=afternoon 28.904429 32.588699 35.71280  2.027034e-03
## pick_up_period=night   18.018018  9.198423 16.17066  1.286503e-20
## f.fare_amount=(14,42.5] 5.643739  4.204993 23.60042  7.194958e-127
```

```

## f.total=(16.6,46]          5.643154  4.467806 25.07804 1.123791e-136
##                               v.test
## f.total=(-1,7.8]          14.116377
## f.fare_amount=(0.1,6]     12.078546
## f.fare_amount=(6,9]       8.971451
## f.total=(7.8,11]          6.546028
## pick_up_period=morning    6.505299
## pick_up_period=valley     4.681022
## pick_up_period=afternoon  -3.086243
## pick_up_period=night      -9.309323
## f.fare_amount=(14,42.5]   -23.960426
## f.total=(16.6,46]         -24.883459
##
## $category$`2`
##                               Cla/Mod  Mod/Cla  Global      p.value    v.test
## f.fare_amount=(9,14]      21.204411 33.06878 24.53694 7.459414e-09 5.780240
## pick_up_period=night      21.492921 22.08995 16.17066 3.267506e-06 4.653246
## f.total=(11,16.6]         19.814503 31.08466 24.68262 1.322624e-05 4.356328
## f.total=(7.8,11]          19.296741 29.76190 24.26639 1.633699e-04 3.769813
## f.fare_amount=(6,9]       18.035427 29.62963 25.84807 1.051032e-02 2.558572
## pick_up_period=morning    11.850312 15.07937 20.02081 1.491105e-04 -3.792546
## f.fare_amount=(14,42.5]   8.553792 12.83069 23.60042 1.119713e-15 -8.012970
## f.total=(16.6,46]         7.717842 12.30159 25.07804 6.665484e-21 -9.378915
##
## $category$`3`
##                               Cla/Mod  Mod/Cla  Global      p.value
## f.total=(11,16.6]         40.55649 33.058419 24.68262 2.734320e-18
## f.fare_amount=(9,14]      40.20356 32.577320 24.53694 4.757062e-17
## pick_up_period=afternoon  35.31469 41.649485 35.71280 1.848224e-08
## f.total=(7.8,11]          35.16295 28.178694 24.26639 3.622051e-05
## f.fare_amount=(6,9]       34.29952 29.278351 25.84807 3.810595e-04
## f.fare_amount=(0.1,6]     33.84000 29.072165 26.01457 1.561249e-03
## pick_up_period=night      33.84813 18.075601 16.17066 1.900267e-02
## pick_up_period=valley     27.25926 25.292096 28.09573 4.192053e-03
## pick_up_period=morning    22.66112 14.982818 20.02081 4.474485e-09
## f.total=(16.6,46]         15.76763 13.058419 25.07804 6.958904e-40
## f.fare_amount=(14,42.5]   11.64021 9.072165 23.60042 2.444321e-62
##                               v.test
## f.total=(11,16.6]         8.721958
## f.fare_amount=(9,14]      8.392551
## pick_up_period=afternoon  5.625639
## f.total=(7.8,11]          4.130354
## f.fare_amount=(6,9]       3.552865
## f.fare_amount=(0.1,6]     3.163051
## pick_up_period=night      2.345479
## pick_up_period=valley     -2.863336
## pick_up_period=morning    -5.865623
## f.total=(16.6,46]         -13.217444
## f.fare_amount=(14,42.5]   -16.662771
##
## $category$`4`
##                               Cla/Mod  Mod/Cla  Global      p.value    v.test
## pick_up_period=afternoon  6.118881 43.56846 35.71280 0.0099643414 2.577064
## pick_up_period=night      6.821107 21.99170 16.17066 0.0152154712 2.427209

```

```

## f.fare_amount=(14,42.5] 3.262787 15.35270 23.60042 0.0013155002 -3.212577
## f.total=(16.6,46] 3.236515 16.18257 25.07804 0.0006878920 -3.394360
## pick_up_period=morning 2.806653 11.20332 20.02081 0.0002061717 -3.711332
##
## $category$`5`
## Cla/Mod Mod/Cla Global p.value v.test
## f.fare_amount=(14,42.5] 1.234568 63.63636 23.60042 0.0000815193 3.939889
## f.total=(16.6,46] 1.161826 63.63636 25.07804 0.0001682944 3.762394
## f.fare_amount=(0.1,6] 0.000000 0.000000 26.01457 0.0012997192 -3.216042
##
## $category$`6`
## Cla/Mod Mod/Cla Global p.value
## f.total=(16.6,46] 66.47302905 99.0111248 25.07804 0.000000e+00
## f.fare_amount=(14,42.5] 69.66490300 97.6514215 23.60042 0.000000e+00
## pick_up_period=morning 21.72557173 25.8343634 20.02081 9.940138e-06
## pick_up_period=afternoon 13.63636364 28.9245983 35.71280 7.807610e-06
## f.fare_amount=(9,14] 1.35708227 1.9777503 24.53694 1.190223e-83
## f.total=(11,16.6] 0.50590219 0.7416564 24.68262 3.820508e-99
## f.total=(7.8,11] 0.08576329 0.1236094 24.26639 9.049492e-107
## f.fare_amount=(0.1,6] 0.16000000 0.2472188 26.01457 9.385963e-114
## f.fare_amount=(6,9] 0.08051530 0.1236094 25.84807 3.962574e-115
## f.total=(-1,7.8] 0.08012821 0.1236094 25.97294 8.498683e-116
## v.test
## f.total=(16.6,46] Inf
## f.fare_amount=(14,42.5] Inf
## pick_up_period=morning 4.418472
## pick_up_period=afternoon -4.470393
## f.fare_amount=(9,14] -19.377712
## f.total=(11,16.6] -21.134645
## f.total=(7.8,11] -21.948001
## f.fare_amount=(0.1,6] -22.667458
## f.fare_amount=(6,9] -22.806387
## f.total=(-1,7.8] -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      0.50000000 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      0.00000000 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 0.00000000000
## 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 "

#Block B descripció per eixos
#She doesn't recommend that (not very usefull by her opinion)
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 "
```

```
#Block C individus
```

```
#parangons. per cadascun dels clusters tenim els seus parangons i els seus més espeicifis (dist) -> que
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
```

```
#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
```

```
#No ens hem de preocupar de la classe outliers, només caracteritzar els clusters del metode per defecte
```

K-Means Classification

```
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=11
## QTRAN(): istep=4805, icoun=3
## QTRAN(): istep=9610, icoun=1
## QTRAN(): istep=14415, icoun=8
## QTRAN(): istep=19220, icoun=126
## QTRAN(): istep=24025, icoun=375
## KMNS(*, k=6): iter= 2, indx=16
## QTRAN(): istep=4805, icoun=1
## QTRAN(): istep=9610, icoun=209
## QTRAN(): istep=14415, icoun=584
## QTRAN(): istep=19220, icoun=3572
## KMNS(*, k=6): iter= 3, indx=23
## QTRAN(): istep=4805, icoun=59
## QTRAN(): istep=9610, icoun=267
## QTRAN(): istep=14415, icoun=3409
## KMNS(*, k=6): iter= 4, indx=36
## QTRAN(): istep=4805, icoun=4786
## KMNS(*, k=6): iter= 5, indx=4805
```

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

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

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

```
## [1] 0.5923259
```

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

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

```
#caracteristació claKM  
#catdes(df, 44)  
#veure si s'han posat d'acord o no  
table(df$claHP,df$claKM)
```

```
##  
##      1      2      3      4      5      6      7  
## 1      0      0      0      0 1522      0      0  
## 2      0      0 756      0      0      0      0  
## 3      0      0      0 1455      0      0      0  
## 4      0 241      0      0      0      0      0  
## 5      0      0      0      0      0 22      0  
## 6 809      0      0      0      0      0      0  
## 7      0      0      0      0      0      0 61
```

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

```
##  
##      kKM-2 kKM-3 kKM-1 kKM-5 kKM-4 kKM-6 kKM-7  
## kHP- 1      0      0      0 1522      0      0      0  
## kHP- 2      0 756      0      0      0      0      0  
## kHP- 3      0      0      0      0 1455      0      0  
## kHP- 4 241      0      0      0      0      0      0  
## kHP- 5      0      0      0      0      0 22      0  
## kHP- 6      0      0 809      0      0      0      0  
## kHP- 7      0      0      0      0      0      0 61
```

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

```
## [1] 0.1678997
```