CA and MCA of NYCABS dataset

Katerina Dimitrova, Jose Romero, Sergi Munoz

March 18, 2018

Load Required Packages: to be increased over the course

```
rm(list=ls())
requiredPackages <- c("effects","FactoMineR","car", "factoextra","RColorBrewer","ggplot2","missMDA","mv
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]

if(length(missingPackages)) install.packages(missingPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
## [1] TRUE
##
## [[8]]
## [1] TRUE
##
## [[9]]
## [1] TRUE
##
## [[10]]
## [1] TRUE
##
## [[11]]
## [1] TRUE
##
## [[12]]
## [1] TRUE
##
## [[13]]
```

```
## [1] TRUE
```

```
# Useful function and Data
load("Taxi5000_raw_DataCleanv3.RData")
calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q2=s.x[3],        q3=s.x[5], ma
```

CA: f.cost (discretization of Total_amount) vs f.hour and f.tt and period ...

```
par(mfrow=c(1,1))
llvout<-which(df$mvout==TRUE);length(llvout)
```

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

```
#llout<-unique(c(llout,llvout))
tt<-table(df[,c("f.total","pick_up_period")])
tt
```

```
##            pick_up_period
## f.total     night morning valley afternoon
##   (-1,7.8]    203     275    363       412
##   (7.8,11]    177     218    331       461
##   (11,16.6]   185     204    340       475
##   (16.6,46]   222     280    333       387
```

```
prop.table(tt,1)
```

```
##            pick_up_period
## f.total         night   morning    valley afternoon
##   (-1,7.8]  0.1620112 0.2194733 0.2897047 0.3288109
##   (7.8,11]  0.1491154 0.1836563 0.2788543 0.3883741
##   (11,16.6] 0.1536545 0.1694352 0.2823920 0.3945183
##   (16.6,46] 0.1816694 0.2291326 0.2725041 0.3166939
```

```
prop.table(table(df$pick_up_period))
```

```
##
##     night   morning    valley afternoon
## 0.1617345 0.2007809 0.2809289 0.3565557
```

```
prop.table(tt,2)
```

```
##            pick_up_period
## f.total         night   morning    valley afternoon
##   (-1,7.8]  0.2579416 0.2814739 0.2655450 0.2374640
##   (7.8,11]  0.2249047 0.2231320 0.2421361 0.2657061
##   (11,16.6] 0.2350699 0.2088025 0.2487198 0.2737752
##   (16.6,46] 0.2820839 0.2865916 0.2435991 0.2230548
```

```
prop.table(table(df$f.total))
```

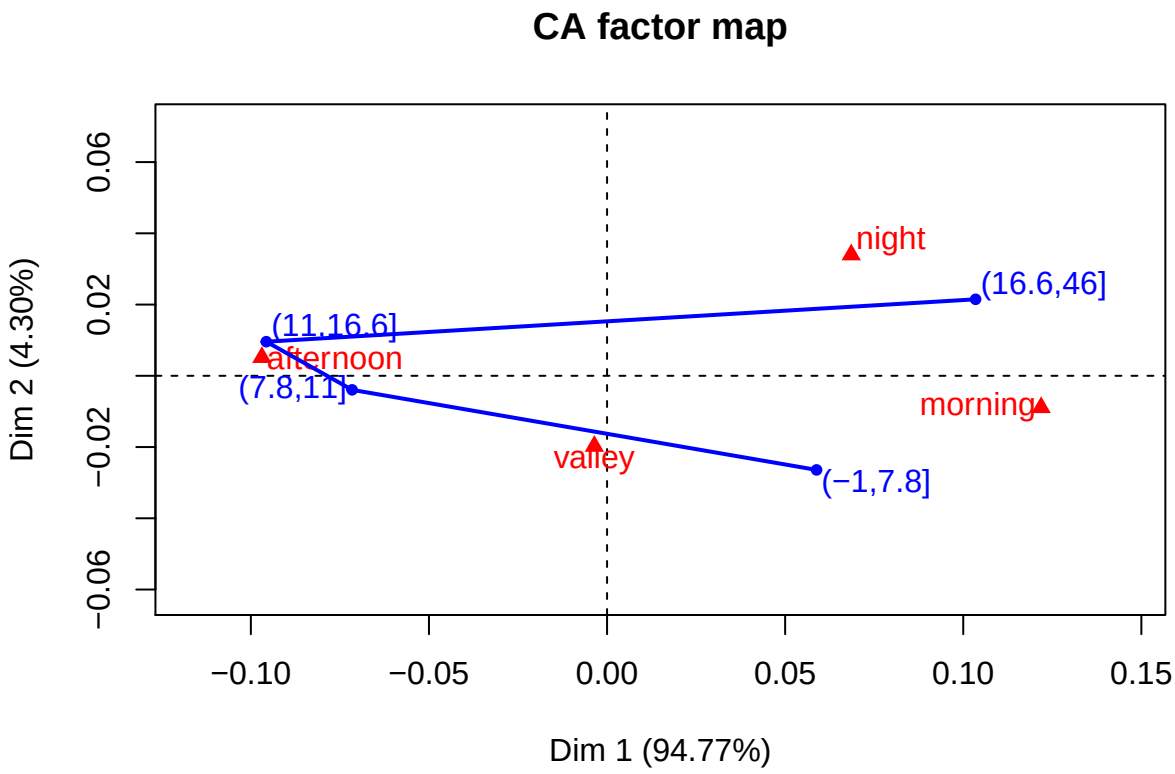
```
##
##  (-1,7.8]  (7.8,11] (11,16.6] (16.6,46]
```

```
## 0.2575010 0.2439375 0.2474312 0.2511303
```

```
# From the chi-square test we see that the p.value is less than 0.05 which meand that we can reject the
chisq.test(tt)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tt
## X-squared = 36.424, df = 9, p-value = 3.334e-05
```

```
res.ca<-CA(tt)
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="blue")
```




```
names(res.ca)
```

```
## [1] "eig"  "call" "row"  "col"  "svd"
```

```
res.ca$row$contrib[,1:2]
```

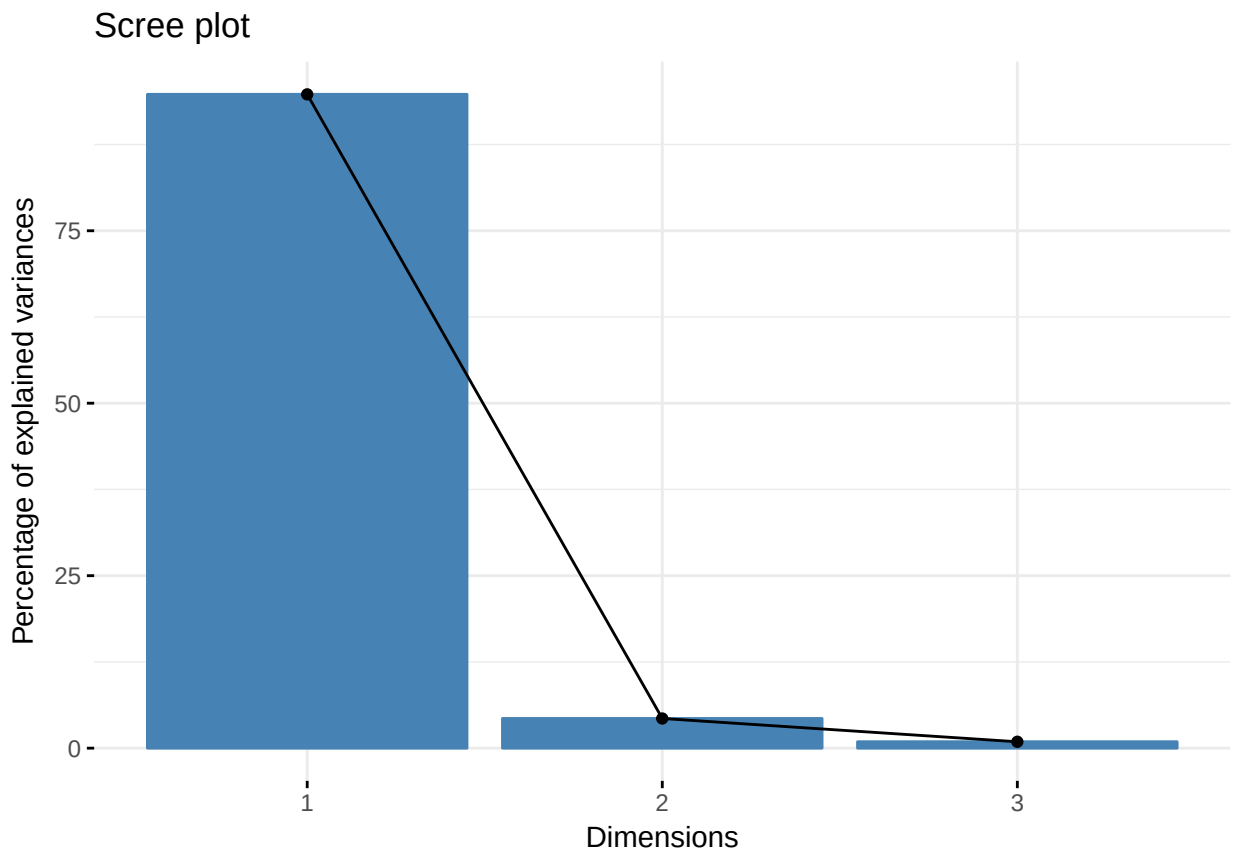
```
##              Dim 1     Dim 2
## (-1,7.8]  12.55998 55.799559
## (7.8,11]  17.61430  1.177242
## (11,16.6] 31.92607  7.050975
## (16.6,46] 37.89965 35.972224
```

```
res.ca$row$coord[,1:2]
```

```
##                 Dim 1        Dim 2
## (-1,7.8]   0.05882227 -0.026415570
```

```
## (7.8,11]  -0.07156986 -0.003942100
## (11,16.6] -0.09567143  0.009579249
## (16.6,46]  0.10346770  0.021476731
```

```
fviz_eig(res.ca)
```

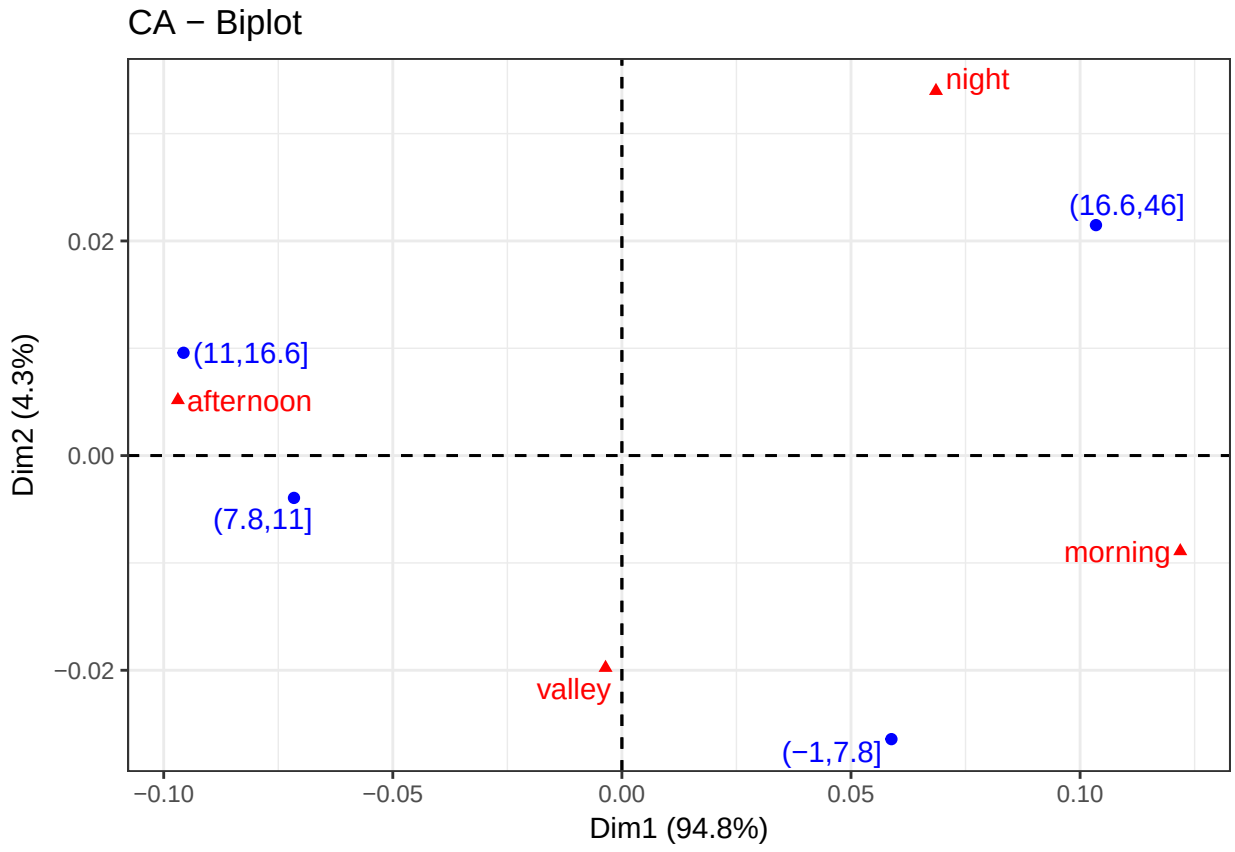



```
#From the variance and the eigenvalues we conclude that the first Dim explains a huge part of the data
summary(res.ca,dig=2)
```

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 36.42355 (p-value =  3.333947e-0
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.007   0.000   0.000
## % of var.             94.768   4.302   0.930
## Cumulative % of var.  94.768  99.070 100.000
##
## Rows
##              Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## (-1,7.8]  |     1.075 |  0.059 12.560  0.829 | -0.026 55.800  0.167 |
## (7.8,11]  |     1.293 | -0.072 17.614  0.966 | -0.004  1.177  0.003 |
## (11,16.6] |     2.313 | -0.096 31.926  0.979 |  0.010  7.051  0.010 |
## (16.6,46] |     2.805 |  0.103 37.900  0.958 |  0.021 35.972  0.041 |
```

```
##            Dim.3    ctr   cos2
## (-1,7.8]   0.004  5.890  0.004 |
## (7.8,11]  -0.013 56.815  0.031 |
## (11,16.6]  0.010 36.280  0.011 |
## (16.6,46] -0.002  1.015  0.000 |
##
## Columns
##              Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## night     |     0.957 |  0.069 10.713  0.794 |  0.034 57.995  0.195 |
## morning   |     3.021 |  0.122 42.034  0.987 | -0.009  4.937  0.005 |
## valley    |     0.140 | -0.004  0.050  0.026 | -0.020 34.103  0.786 |
## afternoon |     3.368 | -0.097 47.202  0.994 |  0.005  2.965  0.003 |
##            Dim.3    ctr   cos2
## night      0.008 15.119  0.011 |
## morning   -0.011 32.950  0.008 |
## valley     0.010 37.754  0.188 |
## afternoon -0.005 14.177  0.003 |
```

```
#We see that the the column points and the row points are far away from each other and thus not similar
fviz_ca_biplot(res.ca,repel=TRUE)+theme_bw()
```




```
tt<-table(df[,c("f.total","f.ttime")])
tt
```

```
##            f.ttime
## f.total     (-1,6] (6,9.78] (9.78,15.7] (15.7,215]
##   (-1,7.8]    1011      234           5          3
```

```
##   (7.8,11]     206      741         232          8
##   (11,16.6]      2      224         751        227
##   (16.6,46]      4       10         232        976
```

```
prop.table(tt,1)
```

```
##            f.ttime
## f.total          (-1,6]    (6,9.78] (9.78,15.7]  (15.7,215]
##   (-1,7.8]  0.806863528 0.186751796 0.003990423 0.002394254
##   (7.8,11]  0.173546757 0.624262848 0.195450716 0.006739680
##   (11,16.6] 0.001661130 0.186046512 0.623754153 0.188538206
##   (16.6,46] 0.003273322 0.008183306 0.189852700 0.798690671
```

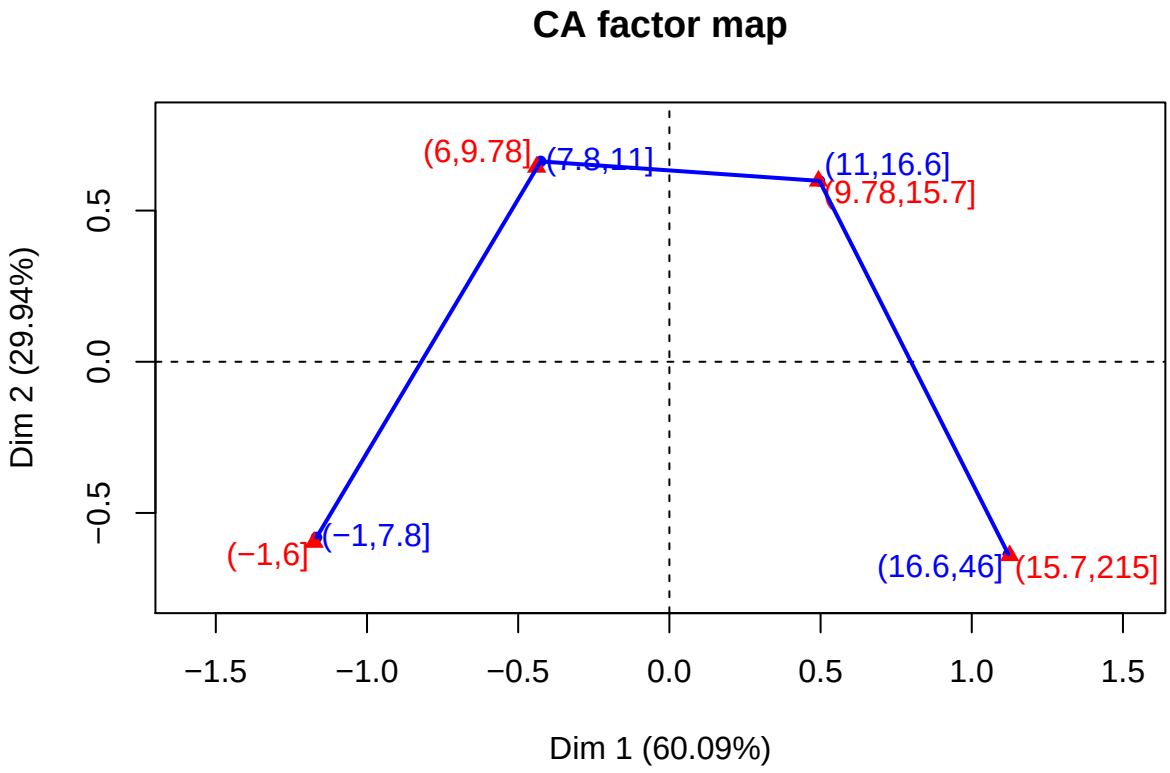
```
prop.table(table(df$f.ttime))
```

```
##
##      (-1,6]    (6,9.78] (9.78,15.7]  (15.7,215]
##   0.2513358   0.2484587   0.2507193   0.2494862
```

```
#Same as the Ca before this 2 facotrs are dependent because of the really small p.value and the big X-s
chisq.test(tt)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tt
## X-squared = 6249, df = 9, p-value < 2.2e-16
```

```
#From the plot we see that the longer the trips is the more it costs which was also expected.
res.ca<-CA(tt)
lines(res.ca$row$coord[,1],res.ca$row$coord[,2],lwd=2,col="blue")
```




```
#We need to consider the first 2 dimentions which together explain 90% of the data.
summary(res.ca)
```

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 6248.958 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.772   0.384   0.128
## % of var.             60.089  29.938   9.973
## Cumulative % of var.  60.089  90.027 100.000
##
## Rows
##               Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## (-1,7.8]    |   445.665 |  -1.167  45.457   0.787 |  -0.581  22.627
## (7.8,11]    |   205.119 |  -0.426   5.730   0.216 |   0.663  27.852
## (11,16.6]   |   206.262 |   0.496   7.900   0.296 |   0.598  23.024
## (16.6,46]   |   427.162 |   1.121  40.913   0.739 |  -0.637  26.497
##                cos2     Dim.3     ctr    cos2
## (-1,7.8]      0.195 |  -0.175   6.166   0.018 |
## (7.8,11]      0.522 |   0.470  42.025   0.262 |
## (11,16.6]     0.429 |  -0.479  44.333   0.275 |
## (16.6,46]     0.238 |   0.195   7.477   0.022 |
```

```
##
## Columns
##               Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## (-1,6]      |   444.908 |  -1.175  44.958   0.780 |  -0.597  23.277
## (6,9.78]    |   204.837 |  -0.439   6.204   0.234 |   0.644  26.778
## (9.78,15.7] |   206.535 |   0.493   7.895   0.295 |   0.598  23.307
## (15.7,215]  |   427.928 |   1.125  40.943   0.738 |  -0.641  26.639
##                cos2     Dim.3     ctr    cos2
## (-1,6]        0.201 |  -0.184   6.632   0.019 |
## (6,9.78]      0.503 |   0.466  42.173   0.264 |
## (9.78,15.7]   0.434 |  -0.473  43.726   0.271 |
## (15.7,215]    0.239 |   0.196   7.469   0.022 |
```

```
res.ca$row$contrib[,1:2]
```

```
##               Dim 1    Dim 2
## (-1,7.8]  45.456861 22.62710
## (7.8,11]   5.729620 27.85203
## (11,16.6]  7.900036 23.02413
## (16.6,46] 40.913483 26.49674
```

```
res.ca$row$coord[,1:2]
```

```
##                Dim 1      Dim 2
## (-1,7.8]  -1.1671449 -0.5812360
## (7.8,11]  -0.4257338  0.6625470
## (11,16.6]  0.4963658  0.5981251
## (16.6,46]  1.1212390 -0.6369044
```

```
res.ca$row$cos2[,1:2]
```

```
##               Dim 1     Dim 2
## (-1,7.8]  0.7870818 0.1951980
## (7.8,11]  0.2155503 0.5220423
## (11,16.6] 0.2955551 0.4291594
## (16.6,46] 0.7391000 0.2384818
```

```
#The 2st two eigenectors give us 90% of the variance so they are sufficient.
res.ca$eig
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.7716655              60.088810                           60.08881
## dim 2  0.3844636              29.937788                           90.02660
## dim 3  0.1280793               9.973402                          100.00000
```

```
sum(res.ca$eig[1:2,1])/sum(res.ca$eig[1:3,1])
```

```
## [1] 0.900266
```

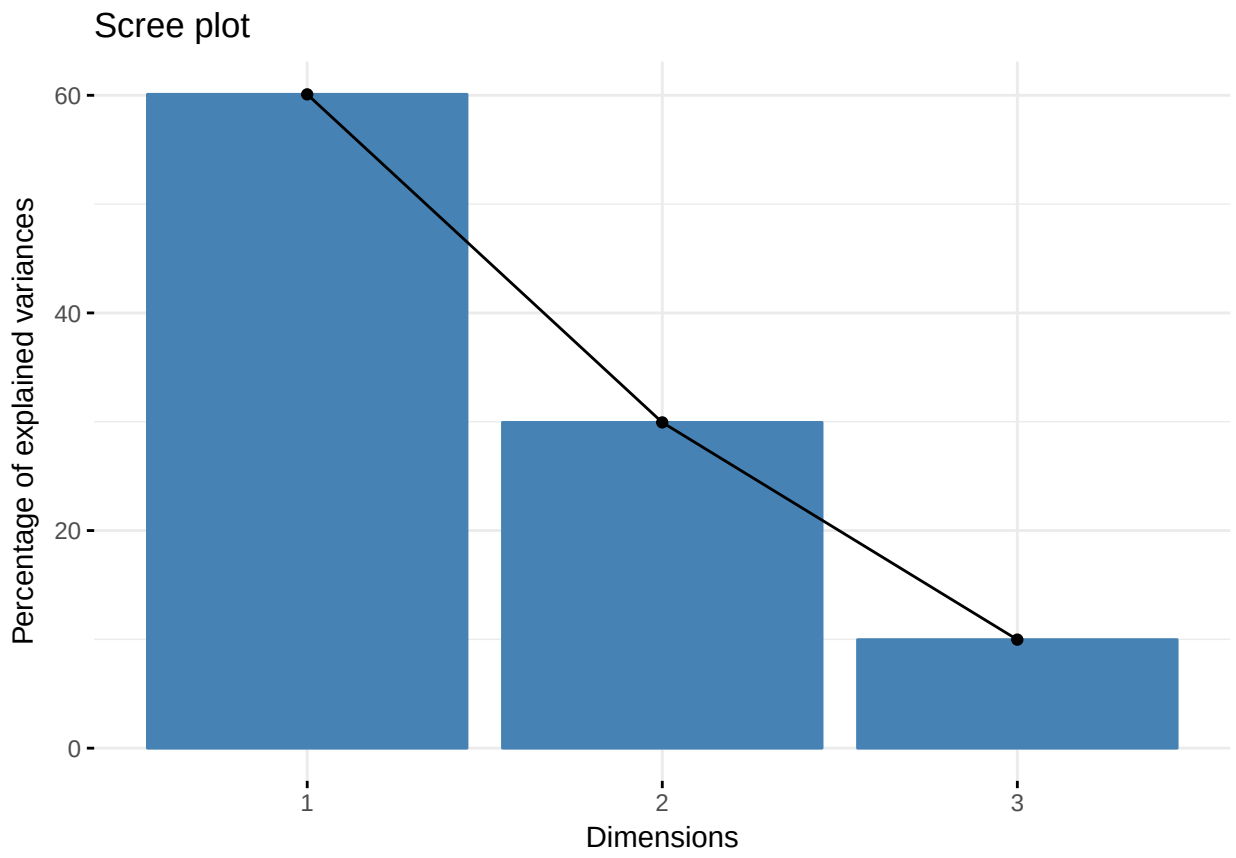
```
sum(res.ca$eig[1:3,1])
```

```
## [1] 1.284208
```

```
res.ca$call$marge.col
```

```
##      (-1,6]    (6,9.78] (9.78,15.7]  (15.7,215]
##   0.2513358   0.2484587   0.2507193   0.2494862
```

```
fviz_eig(res.ca)
```

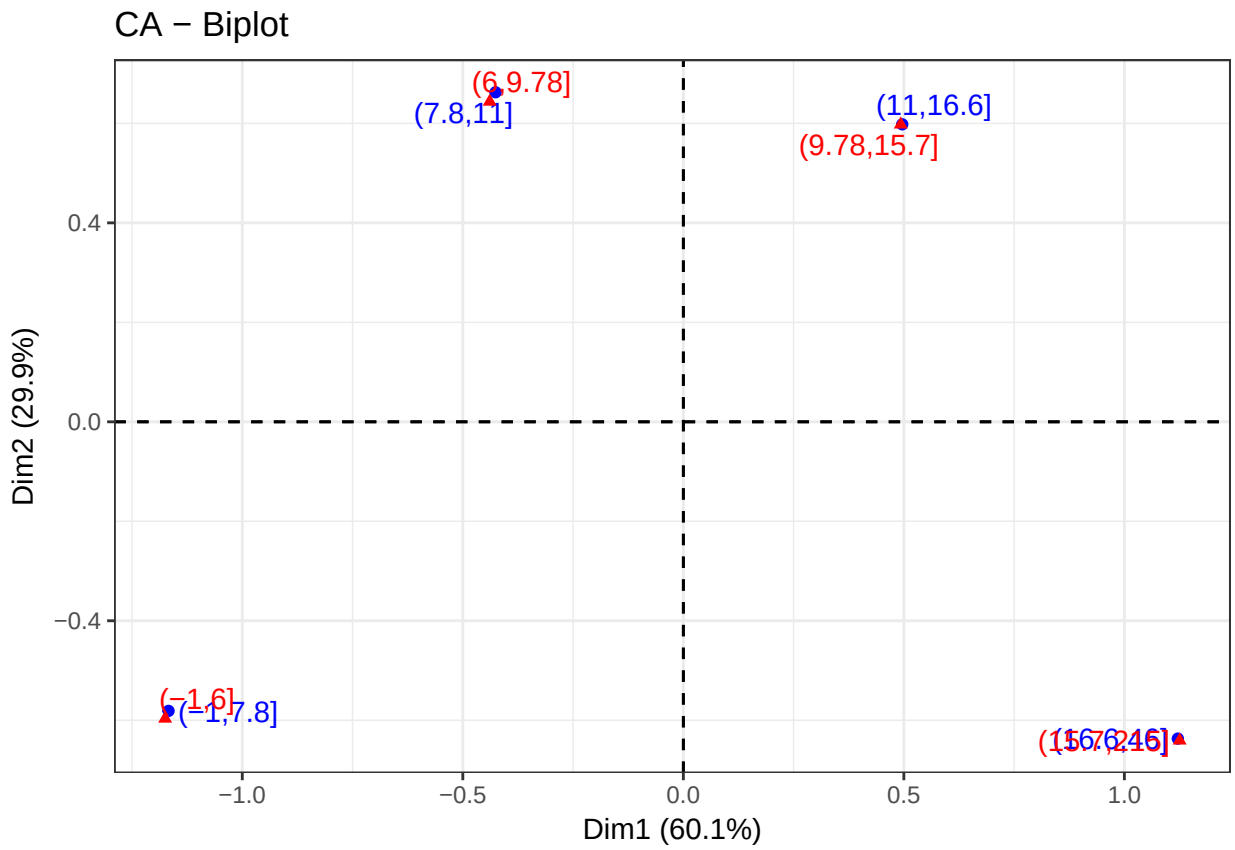



```
summary(res.ca,dig=2)
```

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 6248.958 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.772   0.384   0.128
## % of var.             60.089  29.938   9.973
## Cumulative % of var.  60.089  90.027 100.000
##
## Rows
##                Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## (-1,7.8]    |   445.665 |  -1.167  45.457   0.787 |  -0.581  22.627
## (7.8,11]    |   205.119 |  -0.426   5.730   0.216 |   0.663  27.852
## (11,16.6]   |   206.262 |   0.496   7.900   0.296 |   0.598  23.024
## (16.6,46]   |   427.162 |   1.121  40.913   0.739 |  -0.637  26.497
##                 cos2     Dim.3     ctr    cos2
## (-1,7.8]       0.195 |  -0.175   6.166   0.018 |
## (7.8,11]       0.522 |   0.470  42.025   0.262 |
## (11,16.6]      0.429 |  -0.479  44.333   0.275 |
## (16.6,46]      0.238 |   0.195   7.477   0.022 |
##
```

```
## Columns
##                Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## (-1,6]       |   444.908 |  -1.175  44.958   0.780 |  -0.597  23.277
## (6,9.78]     |   204.837 |  -0.439   6.204   0.234 |   0.644  26.778
## (9.78,15.7] |   206.535 |   0.493   7.895   0.295 |   0.598  23.307
## (15.7,215]   |   427.928 |   1.125  40.943   0.738 |  -0.641  26.639
##                 cos2     Dim.3     ctr    cos2
## (-1,6]         0.201 |  -0.184   6.632   0.019 |
## (6,9.78]       0.503 |   0.466  42.173   0.264 |
## (9.78,15.7]    0.434 |  -0.473  43.726   0.271 |
## (15.7,215]     0.239 |   0.196   7.469   0.022 |
```

```
fviz_ca_biplot(res.ca,repel=TRUE)+theme_bw()
```




```
#tt<-table(df[,c("AnyTip","f.ttime")])  # No tinc sentit fer-ho
#tt
#res.ca<-CA(tt)
#fviz_ca_biplot(res.ca,repel=TRUE)+theme_bw()
```

MCA using multivariant

```
par(mfrow=c(1,1))
names(df)
```

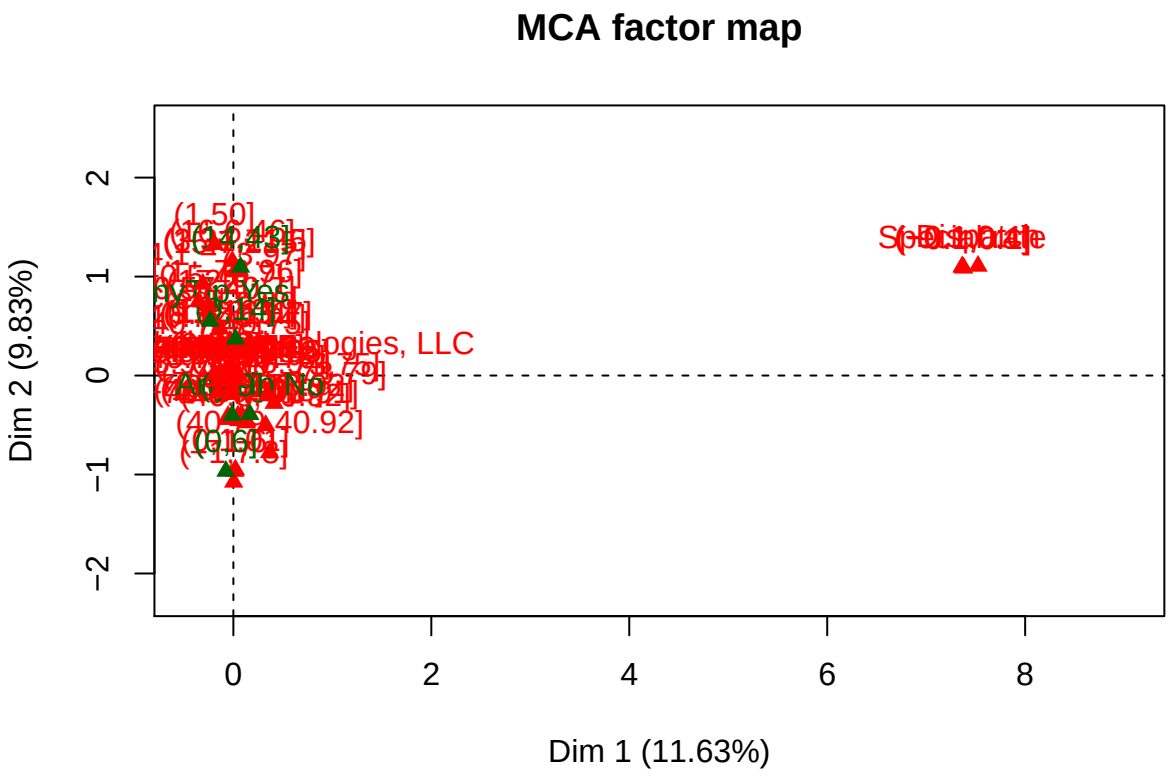
```
##  [1] "VendorID"                "lpep_pickup_datetime"
```

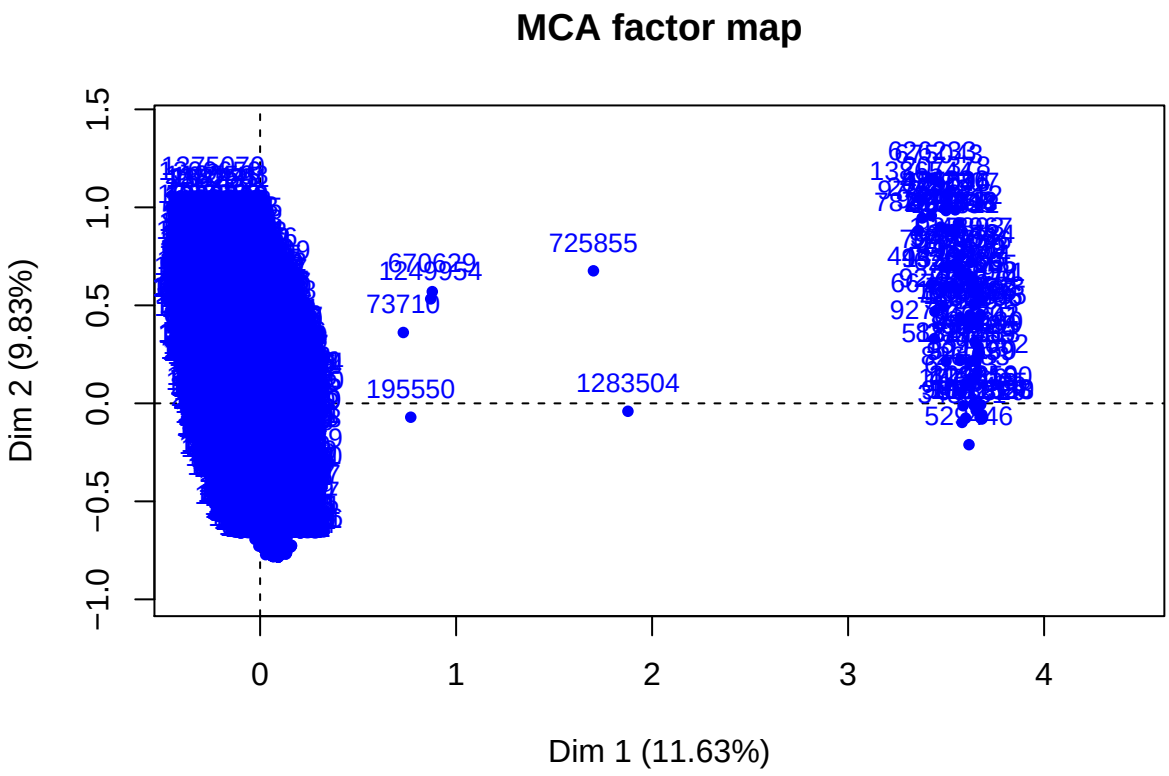
```
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"           "Fare_amount"
## [13] "Extra"                   "MTA_tax"
## [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.ttime"
```

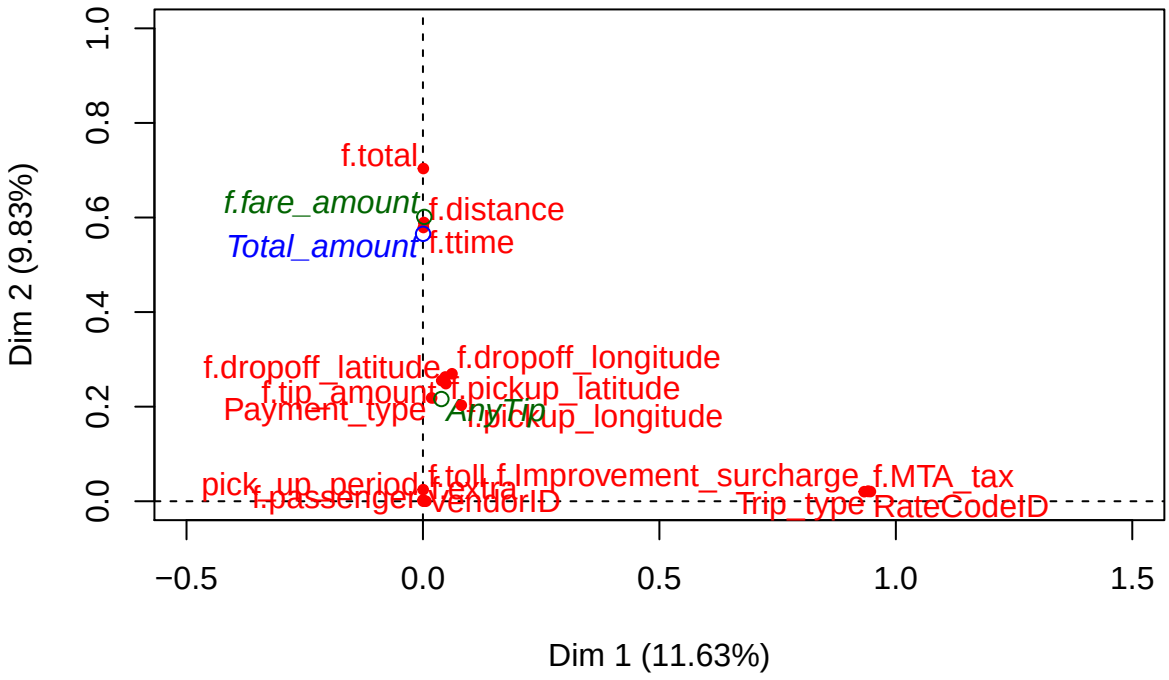
```
#explain we don't use this variable cause it is too bad balanced
summary(df$Store_and_fwd_flag)
```

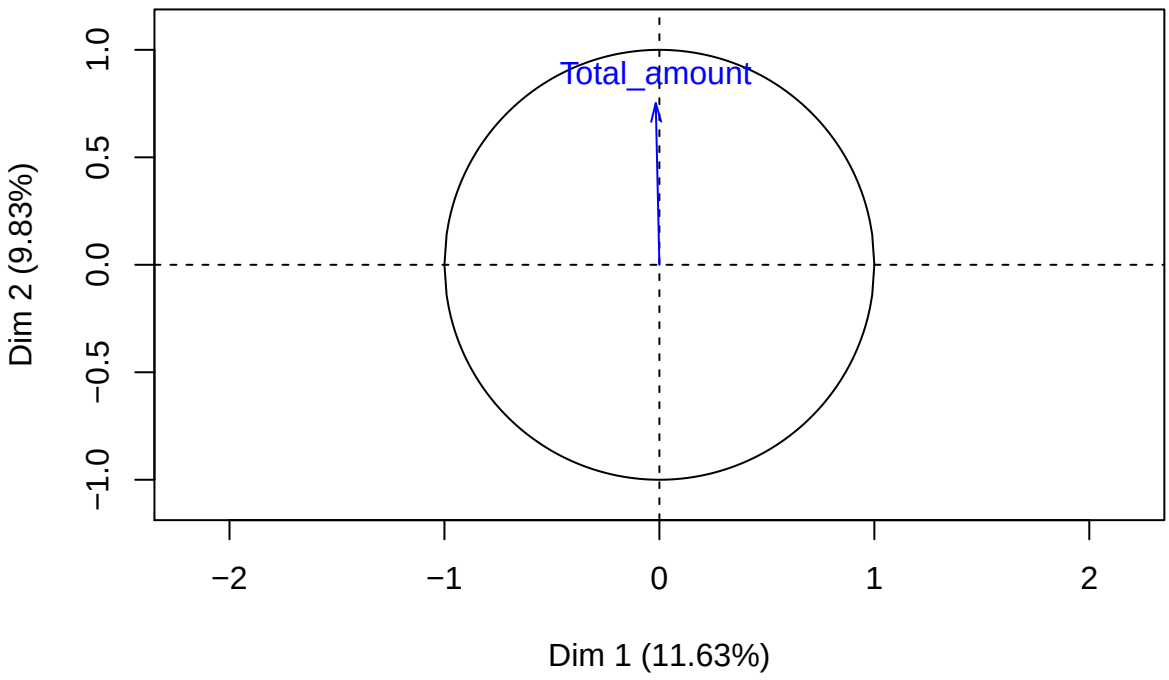
```
##     Store_and_fwd Not_Store_and_fwd
##              4848                18
```

```
#add espeed factor as active (we need to generate it)
vars_dis <- names(df)[c(1,19,20,22,29:40,42)]
vars_cat<-names(df)[c(1, 5,19:20,22,27, 29:42)]#10, 21 not working
#add f.totalamount as quali.sup
res.mca<-MCA(df[,c(vars_cat,"Total_amount")],quali.sup=c(5,13),quanti.sup=21)
```

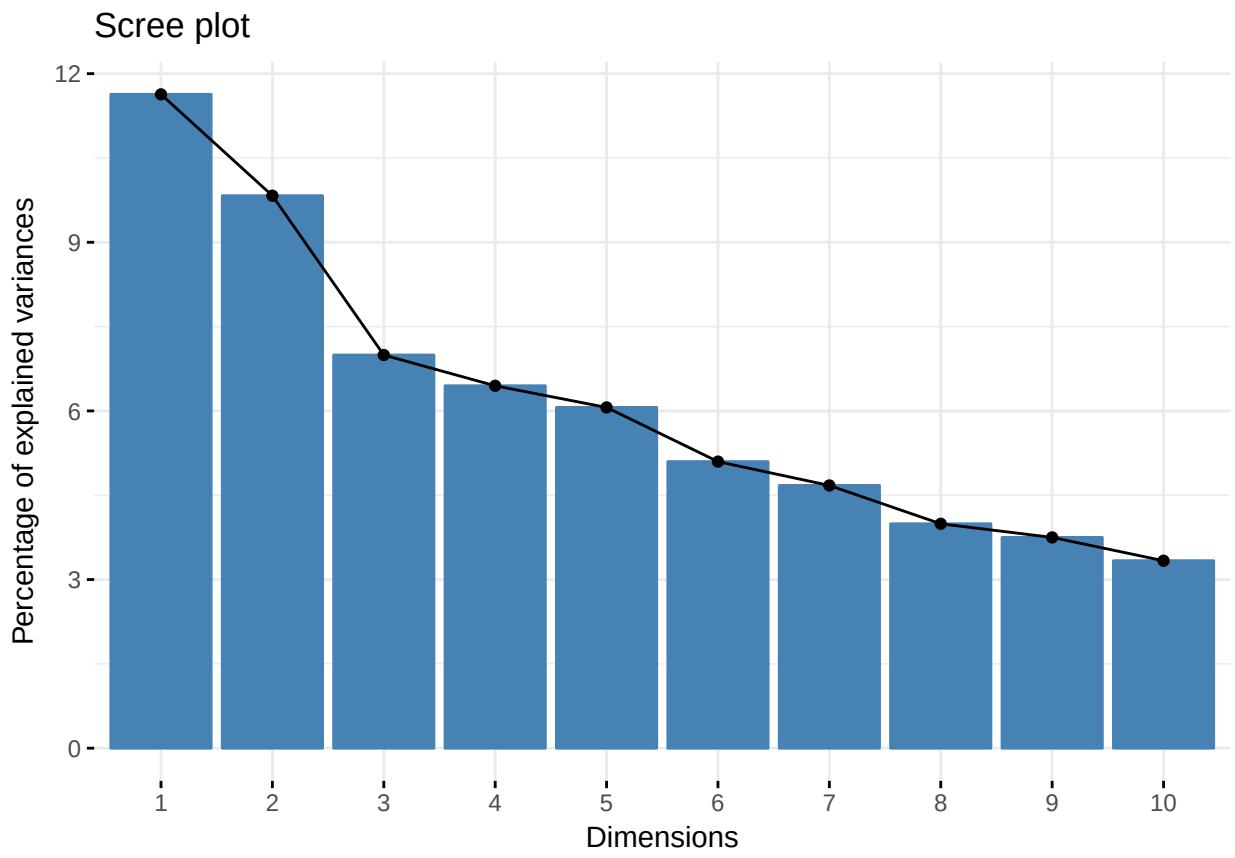

MCA factor map
Dim 2 (9.83%)
Dim 1 (11.63%)
2
1
0
−1
−2
0
2
4
6
8


MCA factor map
Dim 2 (9.83%)
Dim 1 (11.63%)
725855
670629
1249954
73710
195550
1283504
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4

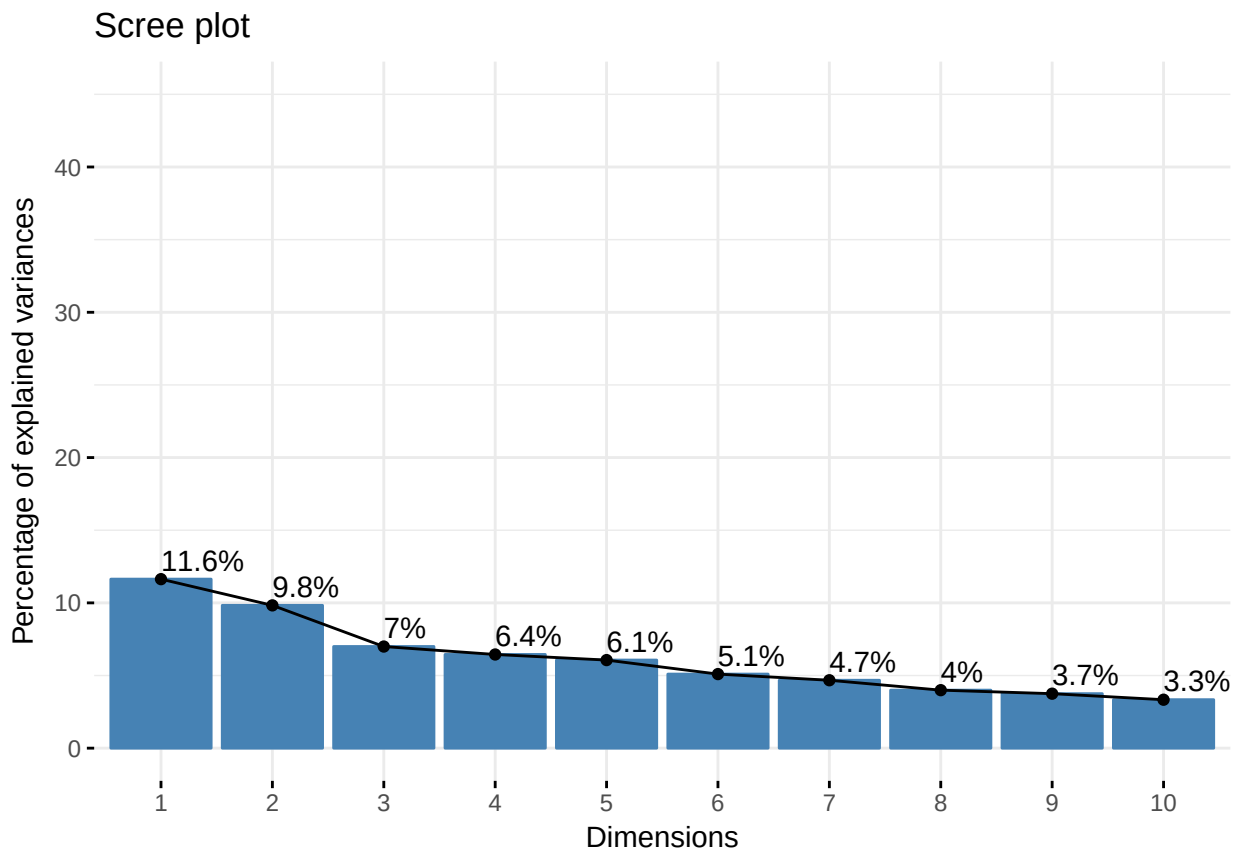

f.total
f.fare_amount
f.distance
Total_amount
f.ttime
f.dropoff_latitude
f.dropoff_longitude
f.tip_amount
f.pickup_latitude
Payment_type
Any_Tip
f.pickup_longitude
pick_up_period
f.toll
f.Improvement_surcharge
f.MTA_tax
f.passenger
f.extra
VendorID
Trip_type
RateCodeID
Dim 2 (9.83%)
Dim 1 (11.63%)
0.0
0.2
0.4
0.6
0.8
1.0
−0.5
0.0
0.5
1.0
1.5




```
# We see that all of the eigenvectors are really small, so we will use a similiar rule like in the PCA
fviz_eig(res.mca)
```




```
fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0, 45))
```




```
get_eigenvalue(res.mca)
```

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.2261765402       11.6319364                    11.63194
## Dim.2  0.1911102515        9.8285272                    21.46046
## Dim.3  0.1359984308        6.9942050                    28.45467
## Dim.4  0.1253575716        6.4469608                    34.90163
## Dim.5  0.1178666778        6.0617149                    40.96334
## Dim.6  0.0991214132        5.0976727                    46.06102
## Dim.7  0.0908755725        4.6736009                    50.73462
## Dim.8  0.0776108087        3.9914130                    54.72603
## Dim.9  0.0729017237        3.7492315                    58.47526
## Dim.10 0.0648386658        3.3345600                    61.80982
## Dim.11 0.0619088374        3.1838831                    64.99371
## Dim.12 0.0576966185        2.9672547                    67.96096
## Dim.13 0.0552112413        2.8394353                    70.80040
## Dim.14 0.0540182528        2.7780816                    73.57848
## Dim.15 0.0525391877        2.7020154                    76.28049
## Dim.16 0.0491398601        2.5271928                    78.80769
## Dim.17 0.0460239064        2.3669438                    81.17463
## Dim.18 0.0457852141        2.3546682                    83.52930
## Dim.19 0.0405587859        2.0858804                    85.61518
## Dim.20 0.0392198668        2.0170217                    87.63220
## Dim.21 0.0370060530        1.9031684                    89.53537
## Dim.22 0.0339188719        1.7443991                    91.27977
## Dim.23 0.0313546178        1.6125232                    92.89229
```

```
## Dim.24 0.0278196973        1.4307273                    94.32302
## Dim.25 0.0250900563        1.2903458                    95.61336
## Dim.26 0.0194981903        1.0027641                    96.61613
## Dim.27 0.0172792123        0.8886452                    97.50477
## Dim.28 0.0141305567        0.7267143                    98.23149
## Dim.29 0.0114623094        0.5894902                    98.82098
## Dim.30 0.0082052451        0.4219840                    99.24296
## Dim.31 0.0071780552        0.3691571                    99.61212
## Dim.32 0.0041939532        0.2156890                    99.82781
## Dim.33 0.0021405803        0.1100870                    99.93789
## Dim.34 0.0006779947        0.0348683                    99.97276
## Dim.35 0.0005296239        0.0272378                   100.00000
```

```
names(res.mca)
```

```
## [1] "eig"        "call"       "ind"        "var"        "svd"
## [6] "quali.sup"  "quanti.sup"
```

```
summary(res.mca,nbelements=50)
```

```
##
## Call:
## MCA(X = df[, c(vars_cat, "Total_amount")], quanti.sup = 21, quali.sup = c(5,
##      13))
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               0.226   0.191   0.136   0.125   0.118   0.099
## % of var.             11.632   9.829   6.994   6.447   6.062   5.098
## Cumulative % of var.  11.632  21.460  28.455  34.902  40.963  46.061
##                        Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance               0.091   0.078   0.073   0.065   0.062   0.058
## % of var.              4.674   3.991   3.749   3.335   3.184   2.967
## Cumulative % of var.  50.735  54.726  58.475  61.810  64.994  67.961
##                       Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18
## Variance               0.055   0.054   0.053   0.049   0.046   0.046
## % of var.              2.839   2.778   2.702   2.527   2.367   2.355
## Cumulative % of var.  70.800  73.578  76.280  78.808  81.175  83.529
##                       Dim.19  Dim.20  Dim.21  Dim.22  Dim.23  Dim.24
## Variance               0.041   0.039   0.037   0.034   0.031   0.028
## % of var.              2.086   2.017   1.903   1.744   1.613   1.431
## Cumulative % of var.  85.615  87.632  89.535  91.280  92.892  94.323
##                       Dim.25  Dim.26  Dim.27  Dim.28  Dim.29  Dim.30
## Variance               0.025   0.019   0.017   0.014   0.011   0.008
## % of var.              1.290   1.003   0.889   0.727   0.589   0.422
## Cumulative % of var.  95.613  96.616  97.505  98.231  98.821  99.243
##                       Dim.31  Dim.32  Dim.33  Dim.34  Dim.35
## Variance               0.007   0.004   0.002   0.001   0.001
## % of var.              0.369   0.216   0.110   0.035   0.027
## Cumulative % of var.  99.612  99.828  99.938  99.973 100.000
##
## Individuals (the 50 first)
##                                    Dim.1    ctr   cos2    Dim.2    ctr
## 285                              | -0.017  0.000  0.000 |  0.280  0.008
```

```
## 307                          |  0.029  0.000  0.001 |  0.074  0.001
## 401                          | -0.036  0.000  0.001 |  0.099  0.001
## 593                          |  0.080  0.001  0.003 | -0.617  0.041
## 636                          | -0.207  0.004  0.026 |  0.636  0.043
## 886                          | -0.228  0.005  0.031 |  0.276  0.008
## 904                          | -0.016  0.000  0.000 | -0.118  0.002
## 978                          | -0.030  0.000  0.000 |  0.266  0.008
## 1135                         | -0.093  0.001  0.006 | -0.362  0.014
## 1282                         | -0.055  0.000  0.002 | -0.348  0.013
## 1409                         |  0.056  0.000  0.002 | -0.449  0.022
## 1475                         | -0.103  0.001  0.006 |  0.216  0.005
## 1495                         | -0.203  0.004  0.026 |  0.783  0.066
## 1905                         |  0.091  0.001  0.005 | -0.131  0.002
## 2126                         |  0.064  0.000  0.002 | -0.063  0.000
## 2151                         |  0.084  0.001  0.004 | -0.555  0.033
## 2201                         | -0.185  0.003  0.021 |  0.738  0.059
## 2271                         | -0.009  0.000  0.000 | -0.162  0.003
## 2747                         |  0.074  0.000  0.003 | -0.569  0.035
## 3065                         | -0.181  0.003  0.020 |  0.774  0.064
## 3089                         | -0.085  0.001  0.005 |  0.470  0.024
## 3130                         |  0.045  0.000  0.001 | -0.649  0.045
## 3221                         |  0.056  0.000  0.002 | -0.512  0.028
## 3420                         |  0.035  0.000  0.001 | -0.693  0.052
## 3679                         | -0.031  0.000  0.001 |  0.073  0.001
## 4310                         | -0.020  0.000  0.000 |  0.030  0.000
## 4754                         | -0.063  0.000  0.002 | -0.253  0.007
## 5241                         | -0.219  0.004  0.027 |  0.604  0.039
## 5277                         | -0.011  0.000  0.000 | -0.431  0.020
## 5649                         | -0.195  0.003  0.018 |  0.322  0.011
## 6353                         |  0.045  0.000  0.001 | -0.207  0.005
## 6364                         | -0.163  0.002  0.013 | -0.040  0.000
## 6755                         | -0.082  0.001  0.004 | -0.244  0.006
## 6869                         | -0.016  0.000  0.000 | -0.330  0.012
## 7079                         |  0.059  0.000  0.002 | -0.455  0.022
## 7211                         |  0.038  0.000  0.001 | -0.360  0.014
## 7342                         | -0.128  0.001  0.010 | -0.178  0.003
## 7802                         | -0.143  0.002  0.013 | -0.129  0.002
## 8138                         | -0.125  0.001  0.009 |  0.258  0.007
## 8443                         | -0.084  0.001  0.005 |  0.271  0.008
## 8619                         |  0.017  0.000  0.000 |  0.106  0.001
## 8891                         |  0.038  0.000  0.001 | -0.192  0.004
## 8960                         | -0.096  0.001  0.006 | -0.266  0.008
## 9207                         |  0.038  0.000  0.001 |  0.152  0.002
## 9503                         |  0.048  0.000  0.001 |  0.147  0.002
## 9747                         |  0.000  0.000  0.000 |  0.281  0.009
## 9765                         |  3.643  1.206  0.911 |  0.575  0.036
## 9984                         | -0.215  0.004  0.024 |  0.617  0.041
## 10034                        | -0.137  0.002  0.011 |  0.679  0.050
## 10199                        |  0.066  0.000  0.003 | -0.499  0.027
##                                cos2    Dim.3    ctr   cos2
## 285                           0.050 | -0.112  0.002  0.008 |
## 307                           0.003 | -0.612  0.057  0.240 |
## 401                           0.006 | -0.302  0.014  0.057 |
## 593                           0.202 | -0.469  0.033  0.116 |
```

```
## 636                                0.243 |  0.266  0.011  0.043 |
## 886                                0.046 |  0.606  0.055  0.221 |
## 904                                0.009 | -0.516  0.040  0.164 |
## 978                                0.037 | -0.482  0.035  0.120 |
## 1135                               0.085 |  0.198  0.006  0.026 |
## 1282                               0.074 | -0.117  0.002  0.008 |
## 1409                               0.128 | -0.149  0.003  0.014 |
## 1475                               0.028 |  0.265  0.011  0.042 |
## 1495                               0.382 |  0.248  0.009  0.038 |
## 1905                               0.011 | -0.453  0.031  0.130 |
## 2126                               0.002 | -0.677  0.069  0.247 |
## 2151                               0.193 |  0.047  0.000  0.001 |
## 2201                               0.341 |  0.023  0.000  0.000 |
## 2271                               0.016 | -0.556  0.047  0.185 |
## 2747                               0.204 | -0.495  0.037  0.154 |
## 3065                               0.362 | -0.081  0.001  0.004 |
## 3089                               0.138 | -0.348  0.018  0.075 |
## 3130                               0.266 | -0.292  0.013  0.054 |
## 3221                               0.162 |  0.077  0.001  0.004 |
## 3420                               0.305 |  0.294  0.013  0.055 |
## 3679                               0.003 | -0.364  0.020  0.068 |
## 4310                               0.001 |  0.042  0.000  0.001 |
## 4754                               0.039 |  0.414  0.026  0.104 |
## 5241                               0.203 |  0.404  0.025  0.091 |
## 5277                               0.106 |  0.313  0.015  0.056 |
## 5649                               0.049 |  0.048  0.000  0.001 |
## 6353                               0.021 | -0.018  0.000  0.000 |
## 6364                               0.001 |  0.617  0.057  0.184 |
## 6755                               0.038 | -0.083  0.001  0.004 |
## 6869                               0.065 |  0.189  0.005  0.021 |
## 7079                               0.134 | -0.516  0.040  0.172 |
## 7211                               0.082 | -0.446  0.030  0.126 |
## 7342                               0.020 |  0.439  0.029  0.121 |
## 7802                               0.011 |  0.530  0.042  0.182 |
## 8138                               0.040 | -0.174  0.005  0.018 |
## 8443                               0.047 |  0.172  0.004  0.019 |
## 8619                               0.007 | -0.545  0.045  0.195 |
## 8891                               0.024 | -0.479  0.035  0.148 |
## 8960                               0.045 | -0.021  0.000  0.000 |
## 9207                               0.014 | -0.437  0.029  0.119 |
## 9503                               0.014 | -0.735  0.082  0.339 |
## 9747                               0.014 | -0.778  0.091  0.109 |
## 9765                               0.023 |  0.139  0.003  0.001 |
## 9984                               0.200 |  0.348  0.018  0.064 |
## 10034                              0.279 | -0.326  0.016  0.064 |
## 10199                              0.156 | -0.556  0.047  0.194 |
##
## Categories (the 50 first)
##                                      Dim.1     ctr    cos2  v.test
## Creative Mobile Technologies, LLC |  -0.048   0.012   0.001  -1.755 |
## VeriFone Inc.                     |   0.013   0.003   0.001   1.755 |
## Standard rate                     |  -0.128   0.394   0.941 -67.668 |
## Special rate                      |   7.365  22.724   0.941  67.668 |
## Credit card                       |  -0.138   0.228   0.018  -9.414 |
```

```
## Cash                                |   0.134   0.223   0.018   9.435 |
## Other                               |  -0.024   0.000   0.000  -0.139 |
## Street-hail                         |  -0.126   0.382   0.946 -67.847 |
## Dispatch                            |   7.524  22.859   0.946  67.847 |
## night                               |   0.046   0.009   0.000   1.421 |
## morning                             |   0.045   0.010   0.001   1.587 |
## valley                              |   0.014   0.001   0.000   0.622 |
## afternoon                           |  -0.058   0.029   0.002  -3.003 |
## (0,1]                               |   0.004   0.000   0.000   0.649 |
## (1,6]                               |  -0.022   0.002   0.000  -0.649 |
## (0,1.01]                            |   0.021   0.003   0.000   0.848 |
## (1.01,1.8]                          |  -0.057   0.020   0.001  -2.305 |
## (1.8,3.31]                          |  -0.024   0.004   0.000  -0.975 |
## (3.31,11.5]                         |   0.060   0.022   0.001   2.431 |
## (-74.1,-73.96]                      |  -0.350   0.755   0.041 -14.127 |
## (-73.96,-73.95]                     |  -0.158   0.145   0.008  -6.135 |
## (-73.95,-73.92]                     |   0.082   0.043   0.002   3.389 |
## (-73.92,-73.79]                     |   0.411   1.050   0.057  16.683 |
## (40.5,40.7]                         |  -0.259   0.412   0.022 -10.442 |
## (40.7,40.74]                        |  -0.125   0.090   0.005  -4.820 |
## (40.74,40.8]                        |   0.054   0.020   0.001   2.322 |
## (40.8,40.92]                        |   0.326   0.636   0.034  12.901 |
## (-74.1,-73.97]                      |  -0.309   0.538   0.028 -11.755 |
## (-73.97,-73.94]                     |  -0.155   0.168   0.010  -6.807 |
## (-73.94,-73.91]                     |   0.130   0.103   0.006   5.209 |
## (-73.91,-73.75]                     |   0.345   0.699   0.037  13.495 |
## (40.57,40.7]                        |  -0.234   0.354   0.020  -9.753 |
## (40.7,40.75]                        |  -0.081   0.044   0.002  -3.473 |
## (40.75,40.79]                       |  -0.018   0.002   0.000  -0.665 |
## (40.79,40.92]                       |   0.366   0.789   0.042  14.336 |
## (-0.1,0.5]                          |   0.034   0.024   0.006   5.485 |
## (0.5,2]                             |  -0.183   0.128   0.006  -5.485 |
## (-0.1,0.4]                          |   7.364  22.723   0.941  67.666 |
## (0.4,0.5]                           |  -0.128   0.394   0.941 -67.666 |
## (-0.1,0.1]                          |   7.379  22.536   0.933  67.380 |
## (0.1,0.8]                           |  -0.126   0.386   0.933 -67.380 |
## (-0.1,1]                            |   0.154   0.365   0.040  13.893 |
## (1,22]                              |  -0.258   0.610   0.040 -13.893 |
## (-1,1]                              |   0.003   0.000   0.001   1.567 |
## (1,50]                              |  -0.190   0.012   0.001  -1.567 |
## (-1,7.8]                            |   0.003   0.000   0.000   0.127 |
## (7.8,11]                            |   0.046   0.013   0.001   1.837 |
## (11,16.6]                           |  -0.038   0.009   0.000  -1.508 |
## (16.6,46]                           |  -0.011   0.001   0.000  -0.446 |
## (-1,6]                              |   0.013   0.001   0.000   0.538 |
##                                      Dim.2     ctr    cos2  v.test
## Creative Mobile Technologies, LLC   0.023   0.003   0.000   0.826 |
## VeriFone Inc.                      -0.006   0.001   0.000  -0.826 |
## Standard rate                      -0.019   0.010   0.021 -10.099 |
## Special rate                        1.099   0.599   0.021  10.099 |
## Credit card                         0.476   3.232   0.218  32.563 |
## Cash                               -0.461   3.112   0.215 -32.374 |
## Other                              -0.188   0.007   0.000  -1.098 |
## Street-hail                        -0.018   0.010   0.020  -9.972 |
```

```
## Dispatch                            1.106   0.584   0.020   9.972 |
## night                               0.025   0.003   0.000   0.754 |
## morning                             0.000   0.000   0.000   0.004 |
## valley                             -0.043   0.015   0.001  -1.882 |
## afternoon                           0.023   0.005   0.000   1.183 |
## (0,1]                              -0.021   0.011   0.002  -3.412 |
## (1,6]                               0.115   0.059   0.002   3.412 |
## (0,1.01]                           -0.952   6.628   0.305 -38.491 |
## (1.01,1.8]                         -0.442   1.429   0.066 -17.869 |
## (1.8,3.31]                          0.340   0.834   0.038  13.620 |
## (3.31,11.5]                         1.065   8.242   0.378  42.883 |
## (-74.1,-73.96]                      0.750   4.105   0.189  30.287 |
## (-73.96,-73.95]                    -0.060   0.024   0.001  -2.317 |
## (-73.95,-73.92]                    -0.399   1.200   0.056 -16.466 |
## (-73.92,-73.79]                    -0.280   0.575   0.026 -11.349 |
## (40.5,40.7]                         0.705   3.627   0.167  28.462 |
## (40.7,40.74]                        0.295   0.591   0.027  11.365 |
## (40.74,40.8]                       -0.454   1.631   0.077 -19.362 |
## (40.8,40.92]                       -0.503   1.790   0.081 -19.897 |
## (-74.1,-73.97]                      0.925   5.705   0.255  35.201 |
## (-73.97,-73.94]                    -0.169   0.236   0.011  -7.420 |
## (-73.94,-73.91]                    -0.476   1.638   0.075 -19.096 |
## (-73.91,-73.75]                    -0.192   0.258   0.012  -7.529 |
## (40.57,40.7]                        0.576   2.535   0.118  23.988 |
## (40.7,40.75]                        0.223   0.398   0.019   9.582 |
## (40.75,40.79]                      -0.119   0.092   0.004  -4.440 |
## (40.79,40.92]                      -0.777   4.205   0.190 -30.421 |
## (-0.1,0.5]                         -0.002   0.000   0.000  -0.390 |
## (0.5,2]                             0.013   0.001   0.000   0.390 |
## (-0.1,0.4]                          1.098   0.598   0.021  10.090 |
## (0.4,0.5]                          -0.019   0.010   0.021 -10.090 |
## (-0.1,0.1]                          1.087   0.579   0.020   9.926 |
## (0.1,0.8]                          -0.019   0.010   0.020  -9.926 |
## (-0.1,1]                           -0.390   2.774   0.255 -35.231 |
## (1,22]                              0.653   4.643   0.255  35.231 |
## (-1,1]                             -0.018   0.010   0.024 -10.916 |
## (1,50]                              1.325   0.702   0.024  10.916 |
## (-1,7.8]                           -1.076   8.667   0.402 -44.200 |
## (7.8,11]                           -0.407   1.173   0.053 -16.117 |
## (11,16.6]                           0.349   0.875   0.040  13.949 |
## (16.6,46]                           1.155   9.737   0.447  46.647 |
## (-1,6]                             -0.966   6.819   0.313 -39.043 |
##                                     Dim.3     ctr    cos2  v.test
## Creative Mobile Technologies, LLC   0.174   0.265   0.008   6.346 |
## VeriFone Inc.                      -0.048   0.073   0.008  -6.346 |
## Standard rate                      -0.015   0.009   0.013  -7.980 |
## Special rate                        0.869   0.526   0.013   7.980 |
## Credit card                         0.101   0.206   0.010   6.935 |
## Cash                               -0.105   0.225   0.011  -7.344 |
## Other                               0.420   0.050   0.001   2.460 |
## Street-hail                        -0.015   0.009   0.014  -8.232 |
## Dispatch                            0.913   0.560   0.014   8.232 |
## night                              -0.340   0.765   0.022 -10.428 |
## morning                            -0.019   0.003   0.000  -0.674 |
```

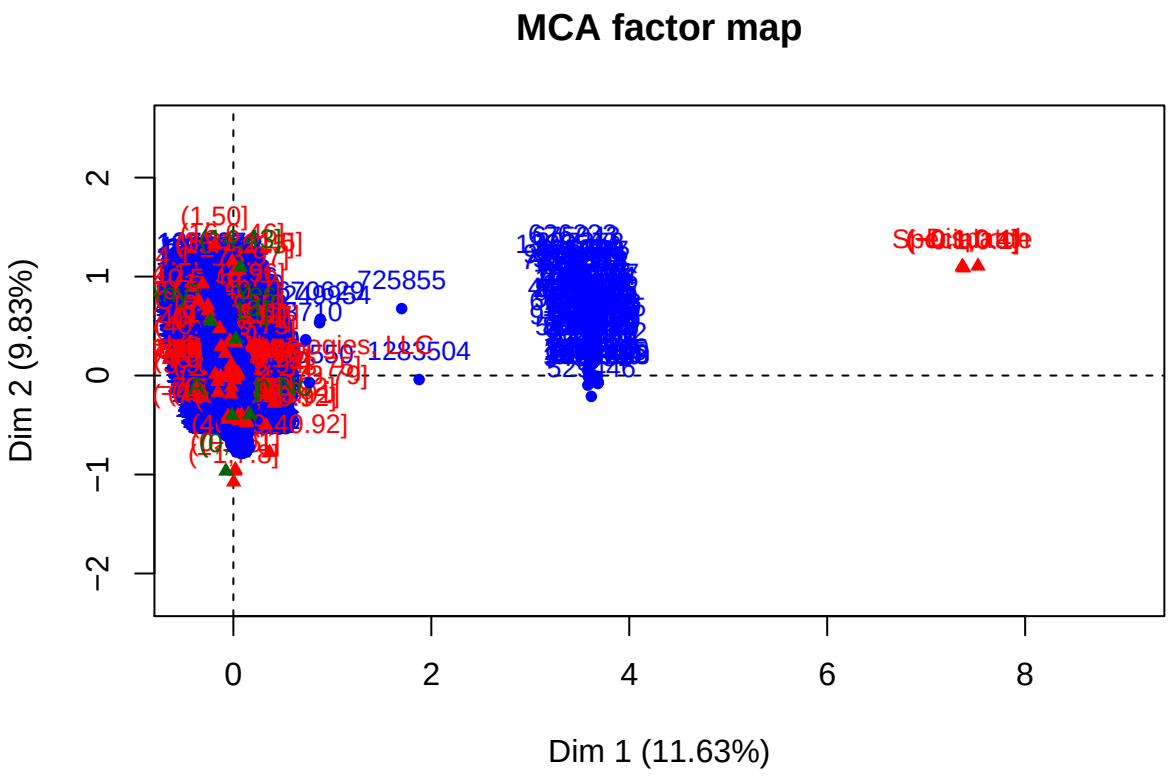
```
## valley                              0.071   0.058   0.002   3.091 |
## afternoon                           0.109   0.174   0.007   5.680 |
## (0,1]                               0.023   0.018   0.003   3.719 |
## (1,6]                              -0.125   0.098   0.003  -3.719 |
## (0,1.01]                            0.485   2.416   0.079  19.604 |
## (1.01,1.8]                          0.160   0.263   0.009   6.464 |
## (1.8,3.31]                         -0.069   0.048   0.002  -2.752 |
## (3.31,11.5]                        -0.581   3.441   0.112 -23.372 |
## (-74.1,-73.96]                      0.955   9.347   0.306  38.554 |
## (-73.96,-73.95]                     0.159   0.243   0.008   6.159 |
## (-73.95,-73.92]                    -0.037   0.014   0.000  -1.516 |
## (-73.92,-73.79]                    -1.060  11.592   0.380 -42.976 |
## (40.5,40.7]                         1.053  11.360   0.371  42.493 |
## (40.7,40.74]                       -0.381   1.386   0.044 -14.683 |
## (40.74,40.8]                       -0.846   7.953   0.267 -36.067 |
## (40.8,40.92]                        0.227   0.512   0.017   8.972 |
## (-74.1,-73.97]                      0.541   2.745   0.087  20.599 |
## (-73.97,-73.94]                     0.437   2.203   0.075  19.124 |
## (-73.94,-73.91]                     0.064   0.042   0.001   2.566 |
## (-73.91,-73.75]                    -1.101  11.853   0.381 -43.081 |
## (40.57,40.7]                        0.961   9.903   0.329  39.994 |
## (40.7,40.75]                       -0.530   3.151   0.106 -22.749 |
## (40.75,40.79]                      -0.793   5.727   0.180 -29.627 |
## (40.79,40.92]                       0.292   0.837   0.027  11.449 |
## (-0.1,0.5]                         -0.010   0.003   0.000  -1.546 |
## (0.5,2]                             0.051   0.017   0.000   1.546 |
## (-0.1,0.4]                          0.886   0.547   0.014   8.144 |
## (0.4,0.5]                          -0.015   0.010   0.014  -8.144 |
## (-0.1,0.1]                          0.920   0.582   0.014   8.397 |
## (0.1,0.8]                          -0.016   0.010   0.014  -8.397 |
## (-0.1,1]                           -0.051   0.066   0.004  -4.568 |
## (1,22]                              0.085   0.110   0.004   4.568 |
## (-1,1]                              0.020   0.016   0.029  11.784 |
## (1,50]                             -1.430   1.150   0.029 -11.784 |
## (-1,7.8]                            0.444   2.072   0.068  18.230 |
## (7.8,11]                            0.101   0.102   0.003   4.013 |
## (11,16.6]                          -0.036   0.013   0.000  -1.453 |
## (16.6,46]                          -0.518   2.749   0.090 -20.910 |
## (-1,6]                              0.479   2.359   0.077  19.370 |
##
## Categorical variables (eta2)
##                                    Dim.1 Dim.2 Dim.3
## VendorID                         | 0.001 0.000 0.008 |
## RateCodeID                       | 0.941 0.021 0.013 |
## Payment_type                     | 0.018 0.218 0.012 |
## Trip_type                        | 0.946 0.020 0.014 |
## pick_up_period                   | 0.002 0.001 0.024 |
## f.passenger                      | 0.000 0.002 0.003 |
## f.distance                       | 0.002 0.589 0.151 |
## f.pickup_longitude               | 0.081 0.203 0.519 |
## f.pickup_latitude                | 0.047 0.263 0.519 |
## f.dropoff_longitude              | 0.061 0.270 0.412 |
## f.dropoff_latitude               | 0.048 0.249 0.480 |
## f.extra                          | 0.006 0.000 0.000 |
```

```
## f.MTA_tax                          | 0.941 0.021 0.014 |
## f.Improvement_surcharge            | 0.933 0.020 0.014 |
## f.tip_amount                       | 0.040 0.255 0.004 |
## f.toll                             | 0.001 0.024 0.029 |
## f.total                            | 0.001 0.704 0.121 |
## f.ttime                            | 0.001 0.579 0.110 |
##
## Supplementary categories
##                                        Dim.1    cos2  v.test     Dim.2
## AnyTip No                          |   0.167   0.039  13.772 |  -0.393
## AnyTip Yes                         |  -0.234   0.039 -13.772 |   0.550
## (0,6]                              |  -0.075   0.002  -3.086 |  -0.966
## (6,9]                              |  -0.013   0.000  -0.542 |  -0.403
## (9,14]                             |   0.022   0.000   0.892 |   0.369
## (14,43]                            |   0.072   0.002   2.821 |   1.096
##                                       cos2  v.test     Dim.3    cos2
## AnyTip No                             0.216 -32.428 |  -0.074   0.008
## AnyTip Yes                            0.216  32.428 |   0.104   0.008
## (0,6]                                 0.323 -39.632 |   0.459   0.073
## (6,9]                                 0.056 -16.554 |   0.157   0.009
## (9,14]                                0.045  14.758 |  -0.107   0.004
## (14,43]                               0.375  42.725 |  -0.554   0.096
##                                    v.test
## AnyTip No                          -6.121 |
## AnyTip Yes                          6.121 |
## (0,6]                              18.825 |
## (6,9]                               6.433 |
## (9,14]                             -4.282 |
## (14,43]                           -21.589 |
##
## Supplementary categorical variables (eta2)
##                                     Dim.1 Dim.2 Dim.3
## AnyTip                             | 0.039 0.216 0.008 |
## f.fare_amount                      | 0.003 0.601 0.136 |
##
## Supplementary continuous variable
##                                       Dim.1    Dim.2    Dim.3
## Total_amount                       | -0.017 |  0.752 | -0.324 |
```

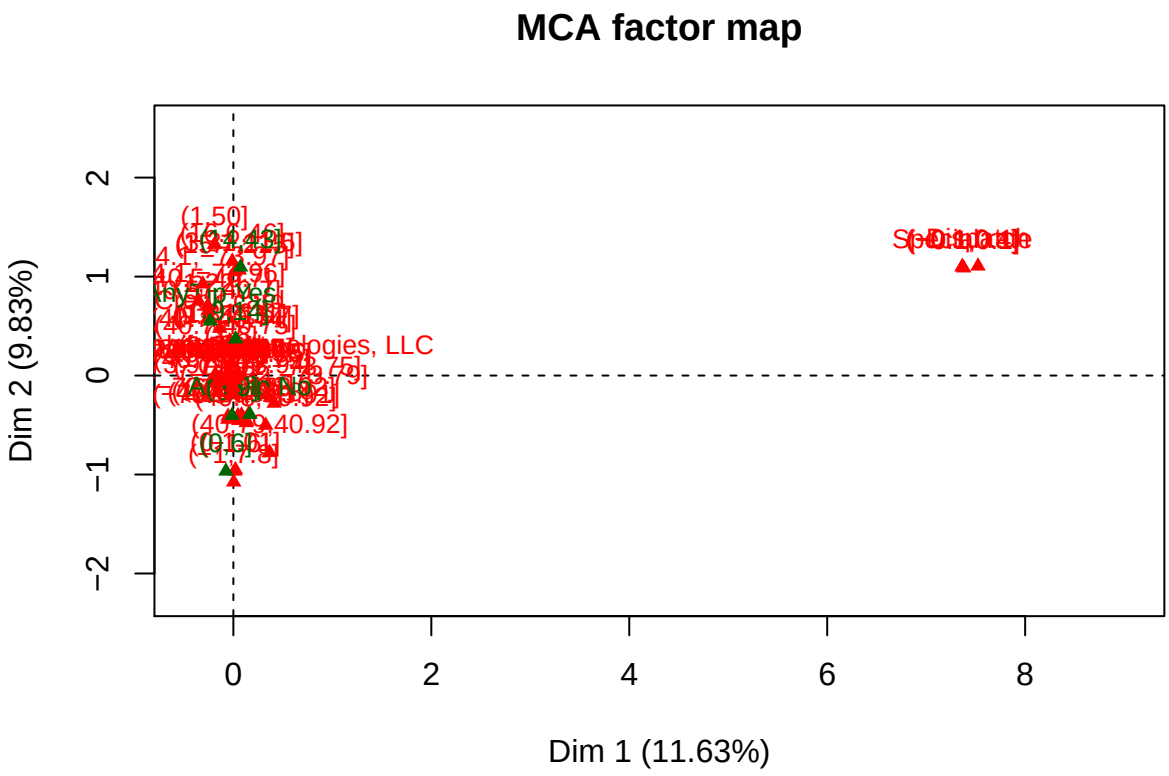
```
mean(res.mca$eig)
```

```
## [1] 26.04711
```

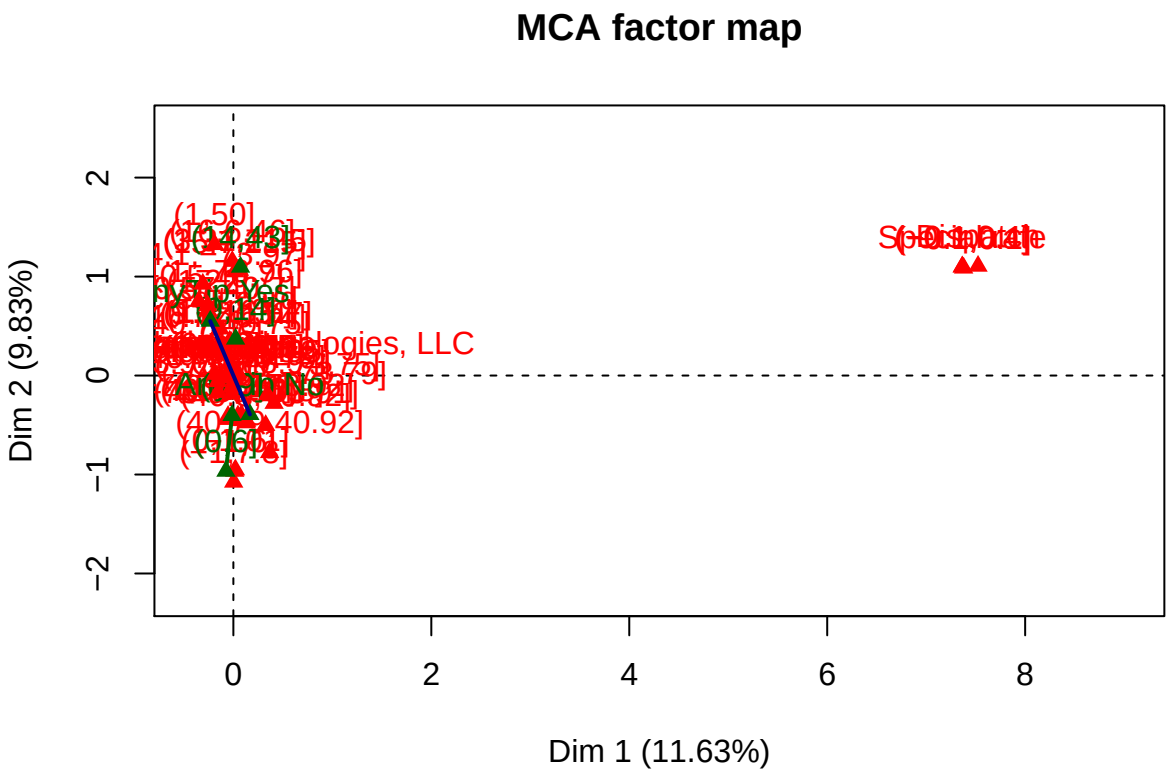
```
# Individual Representation
plot.MCA(res.mca,choix=c("ind"),cex=0.8)
```




```
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),cex=0.8)
```



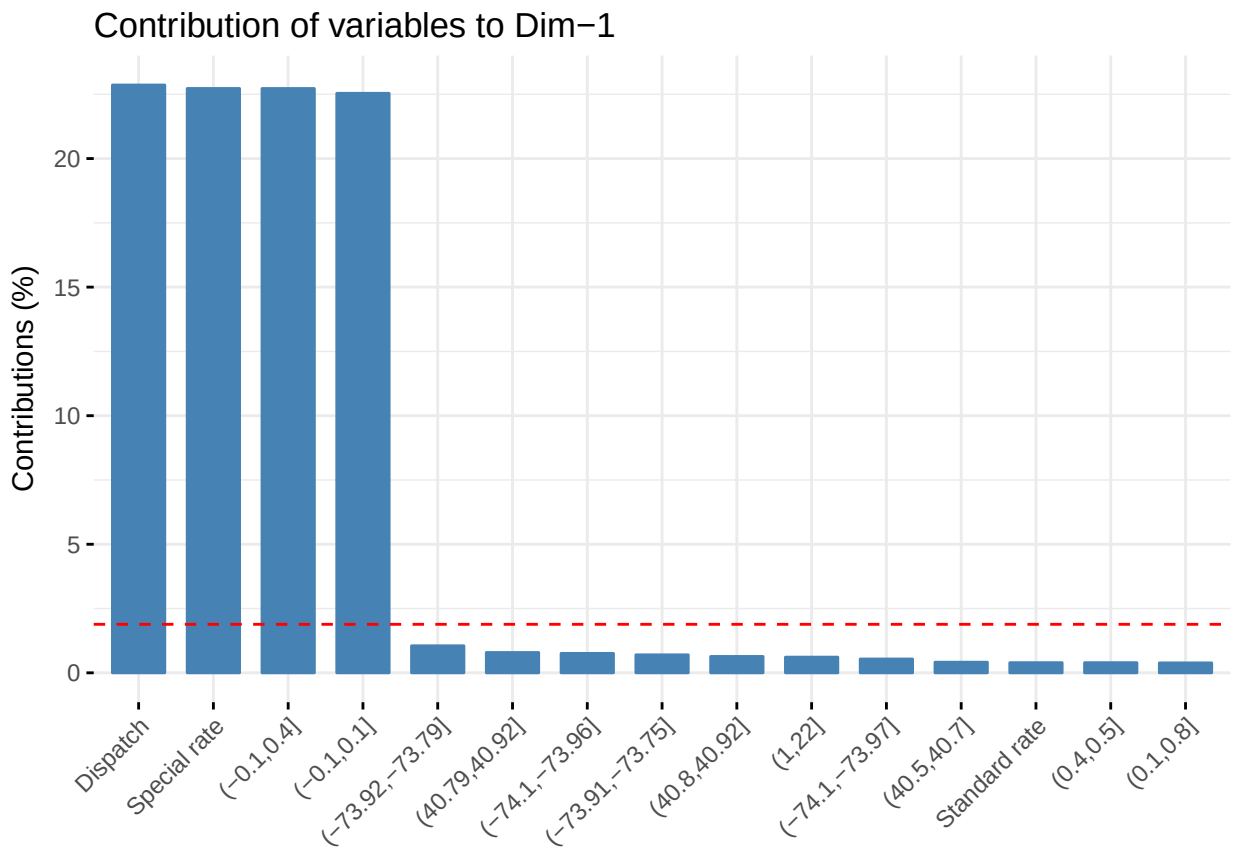

```
# Representation of categories
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),axes=c(1,2))
lines(res.mca$quali.sup$coord[3:4, 1],res.mca$quali.sup$coord[3:4,2],lwd=2,col="darkgreen") # Trencada
lines(res.mca$quali.sup$coord[1:2,1],res.mca$quali.sup$coord[1:2,2],lwd=2,col="darkblue") # Trencada An
```



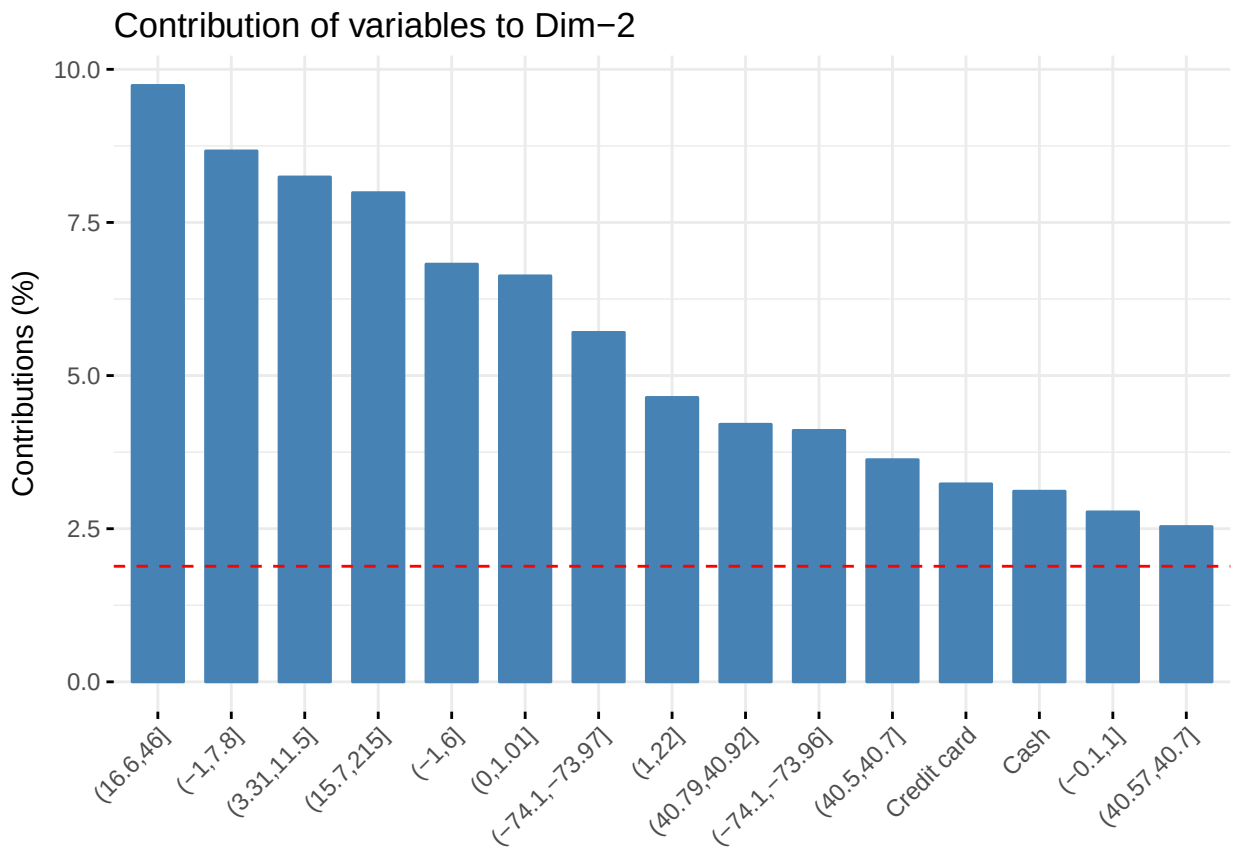

```
names(res.mca)
```

```
## [1] "eig"        "call"       "ind"        "var"        "svd"
## [6] "quali.sup"  "quanti.sup"
```

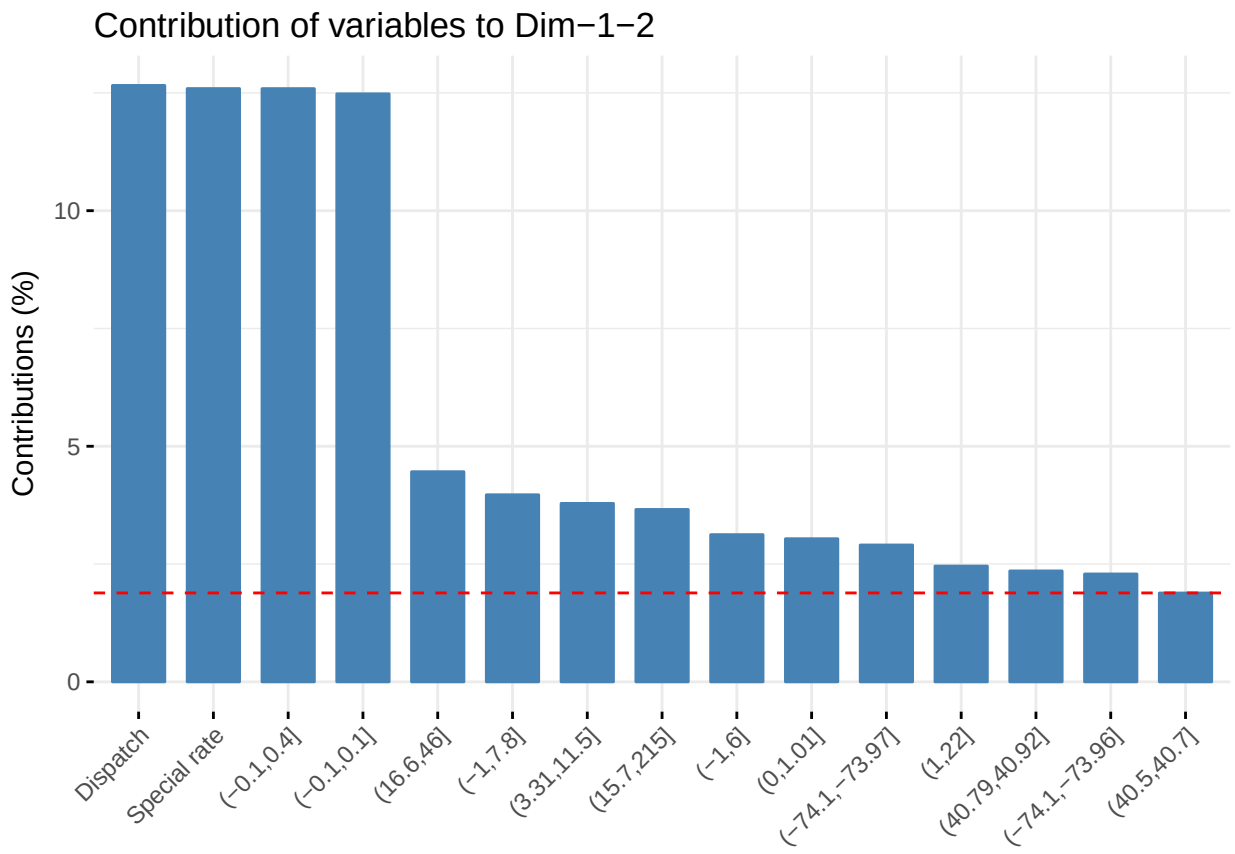
```
# Contributions of rows to dimension 1
fviz_contrib(res.mca, choice = "var", axes = 1, top = 15)
```




```
# Contributions of rows to dimension 2
fviz_contrib(res.mca, choice = "var", axes = 2, top = 15)
```



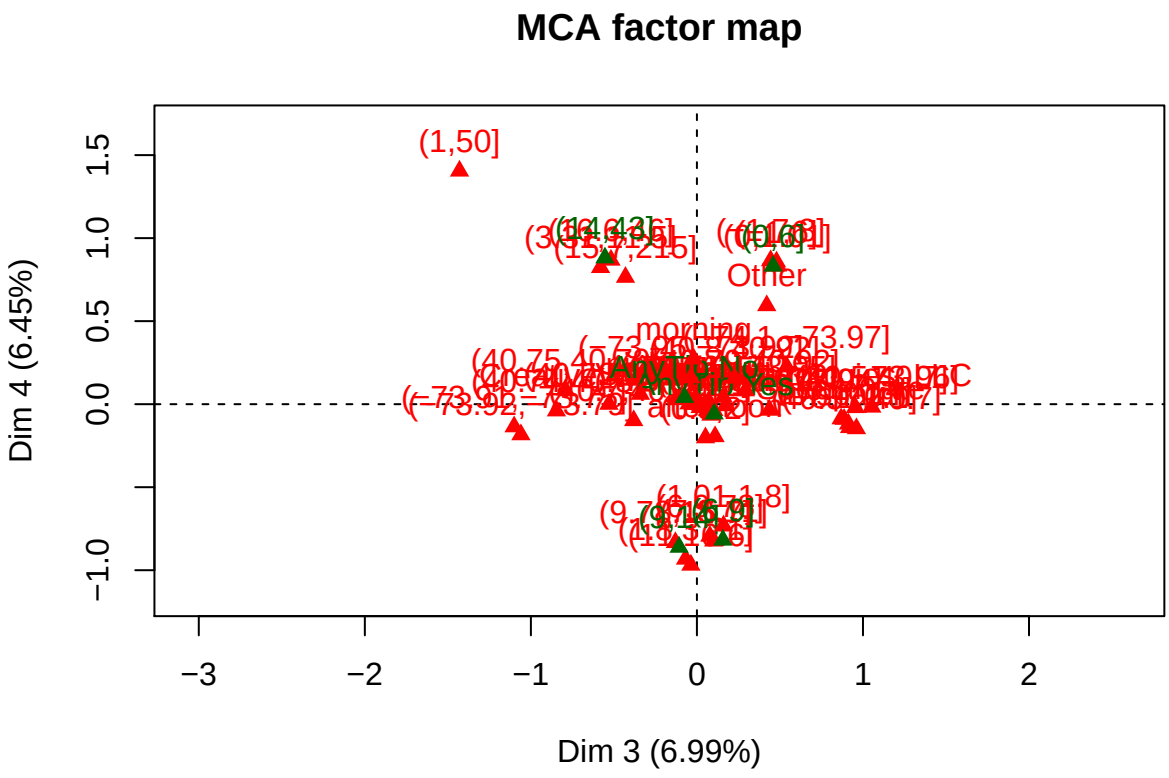

```
# Total contribution to dimension 1 and 2
fviz_contrib(res.mca, choice = "var", axes = 1:2, top = 15)
```



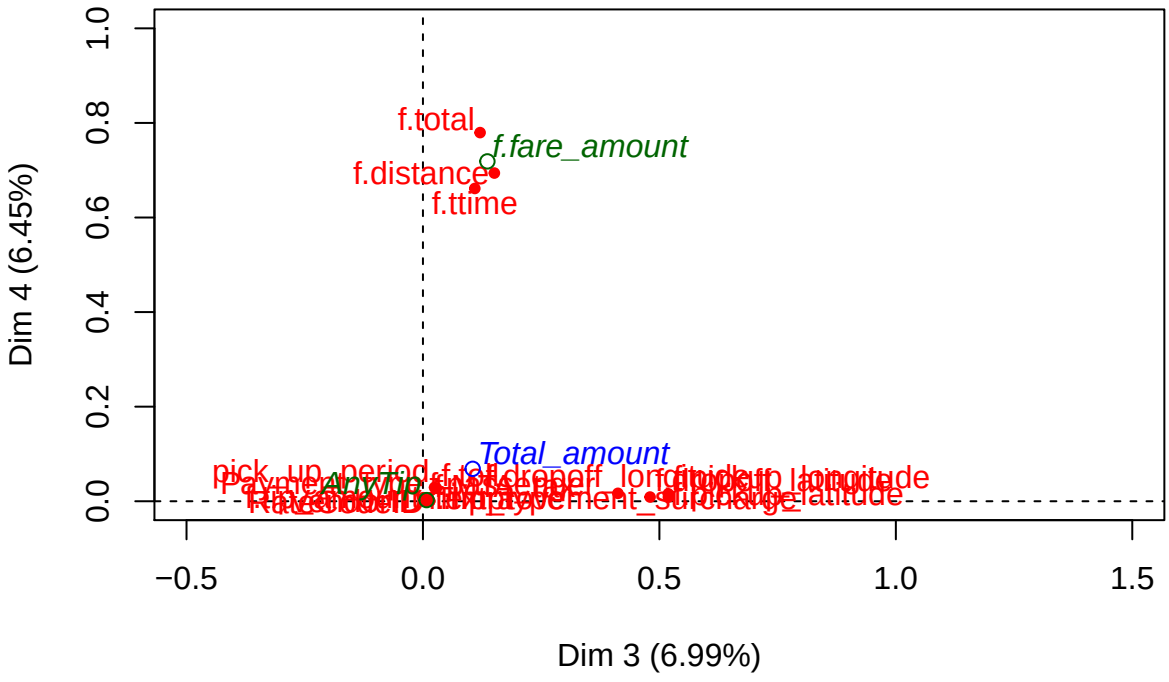

```
#we see that the the provider of the record,categories of the rate code are most contributive and the t
head(round(res.mca$var$contrib,2), 10)
```

```
##                                 Dim 1 Dim 2 Dim 3 Dim 4 Dim 5
## Creative Mobile Technologies, LLC  0.01  0.00  0.27  0.00  0.00
## VeriFone Inc.                      0.00  0.00  0.07  0.00  0.00
## Standard rate                      0.39  0.01  0.01  0.00  0.00
## Special rate                      22.72  0.60  0.53  0.01  0.07
## Credit card                        0.23  3.23  0.21  0.03  1.65
## Cash                               0.22  3.11  0.23  0.02  1.71
## Other                              0.00  0.01  0.05  0.11  0.13
## Street-hail                        0.38  0.01  0.01  0.00  0.00
## Dispatch                          22.86  0.58  0.56  0.01  0.07
## night                              0.01  0.00  0.77  0.03  0.64
```

```
plot.MCA(res.mca,choix=c("ind"),invisible=c("ind"),axes=c(3,4))
```



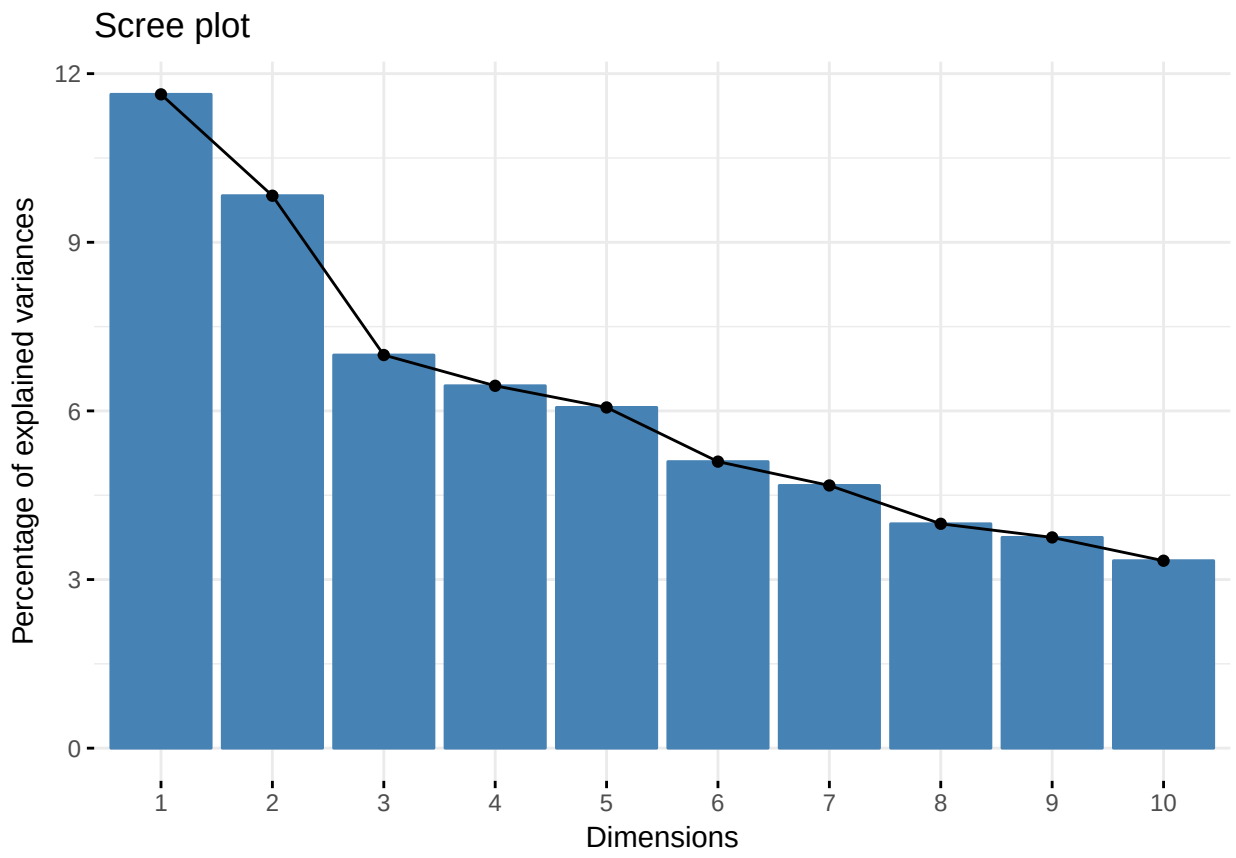

```
plot.MCA(res.mca,choix=c("var"),axes=c(3,4))
```



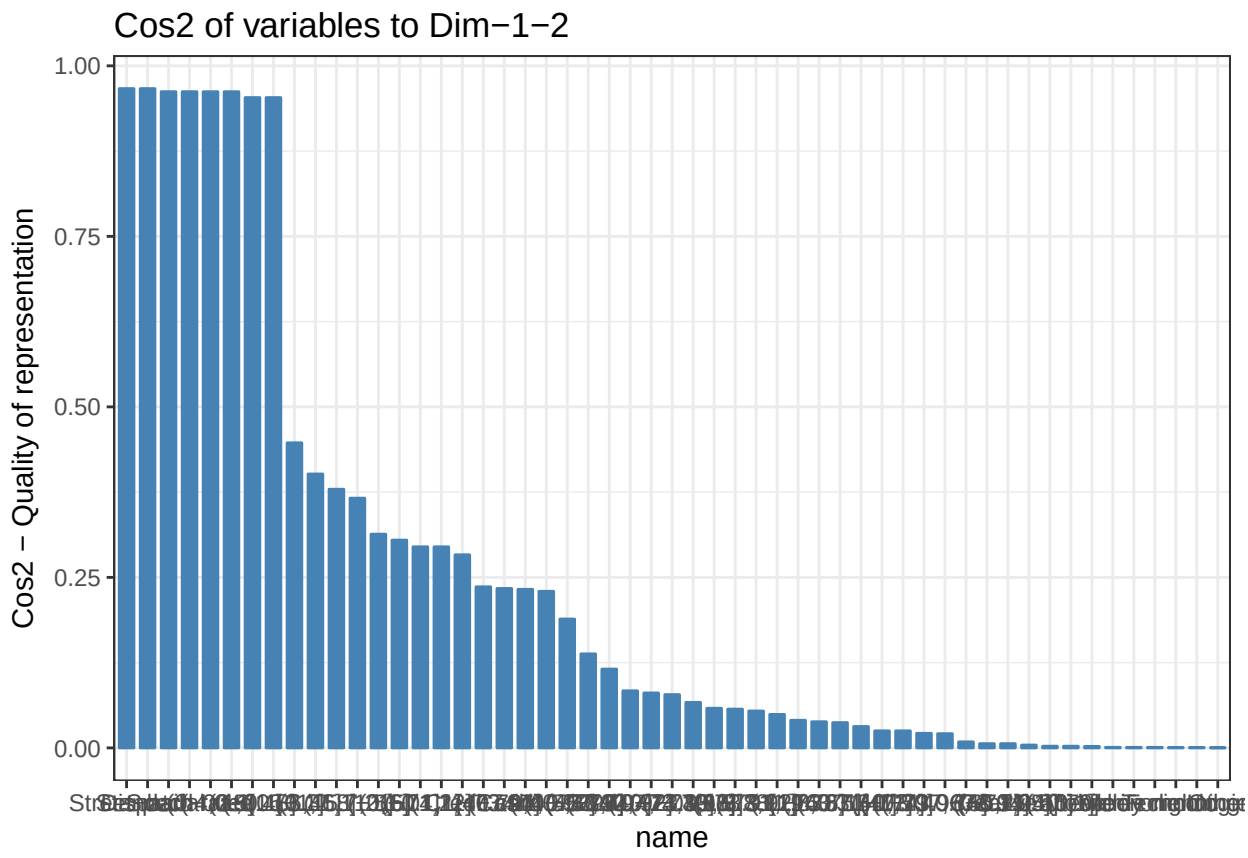

```
# Use modern ggplot facilities
names(res.mca)
```

```
## [1] "eig"        "call"       "ind"        "var"        "svd"
## [6] "quali.sup"  "quanti.sup"
```

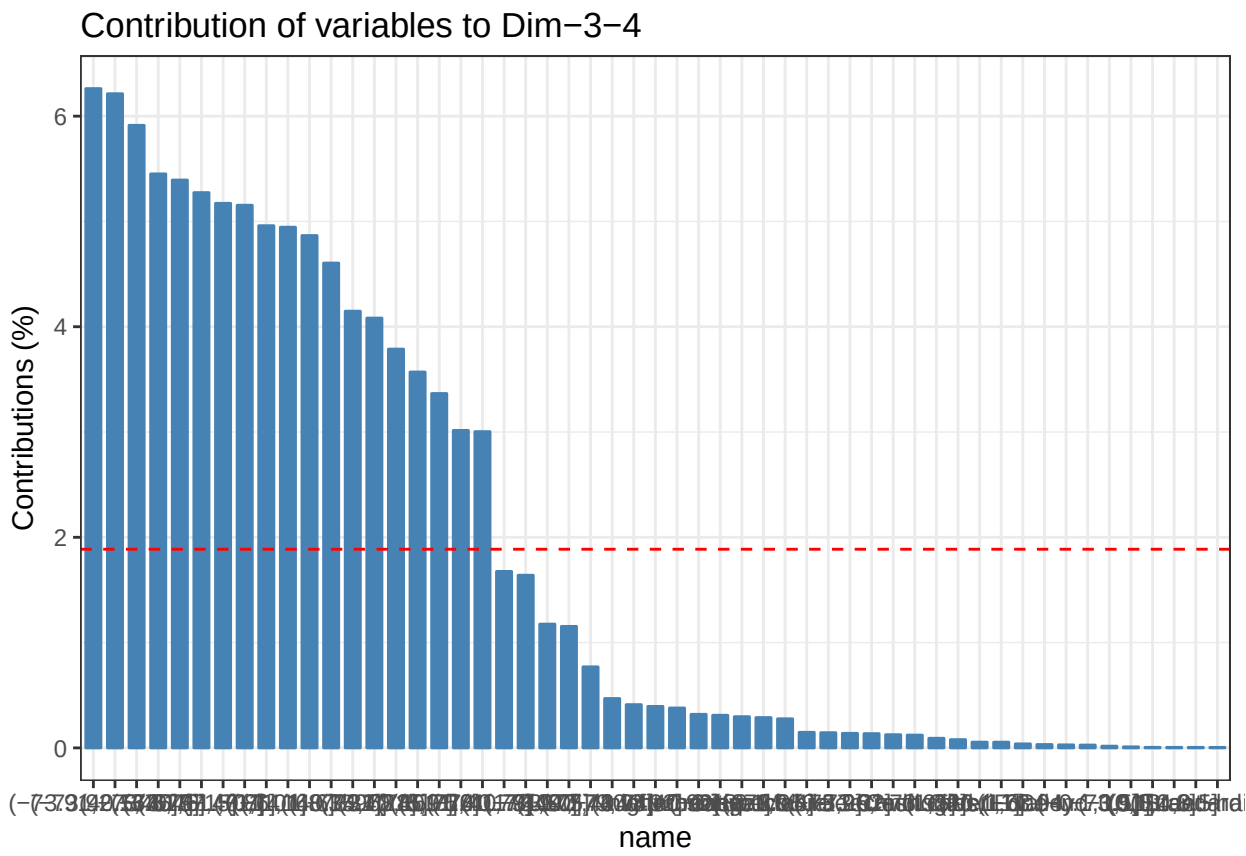
```
fviz_eig(res.mca)
```



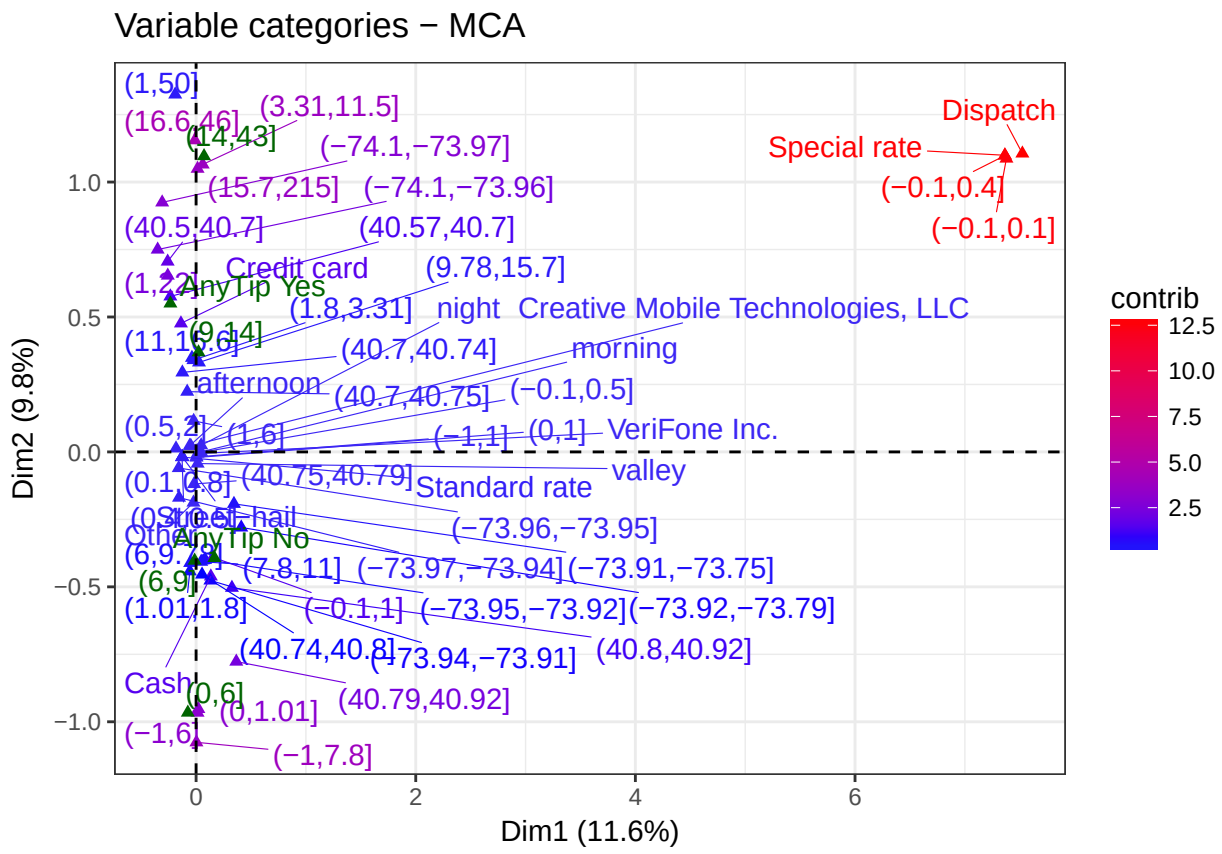

```
fviz_cos2(res.mca, choice = "var", axes = 1:2)+theme_bw()
```



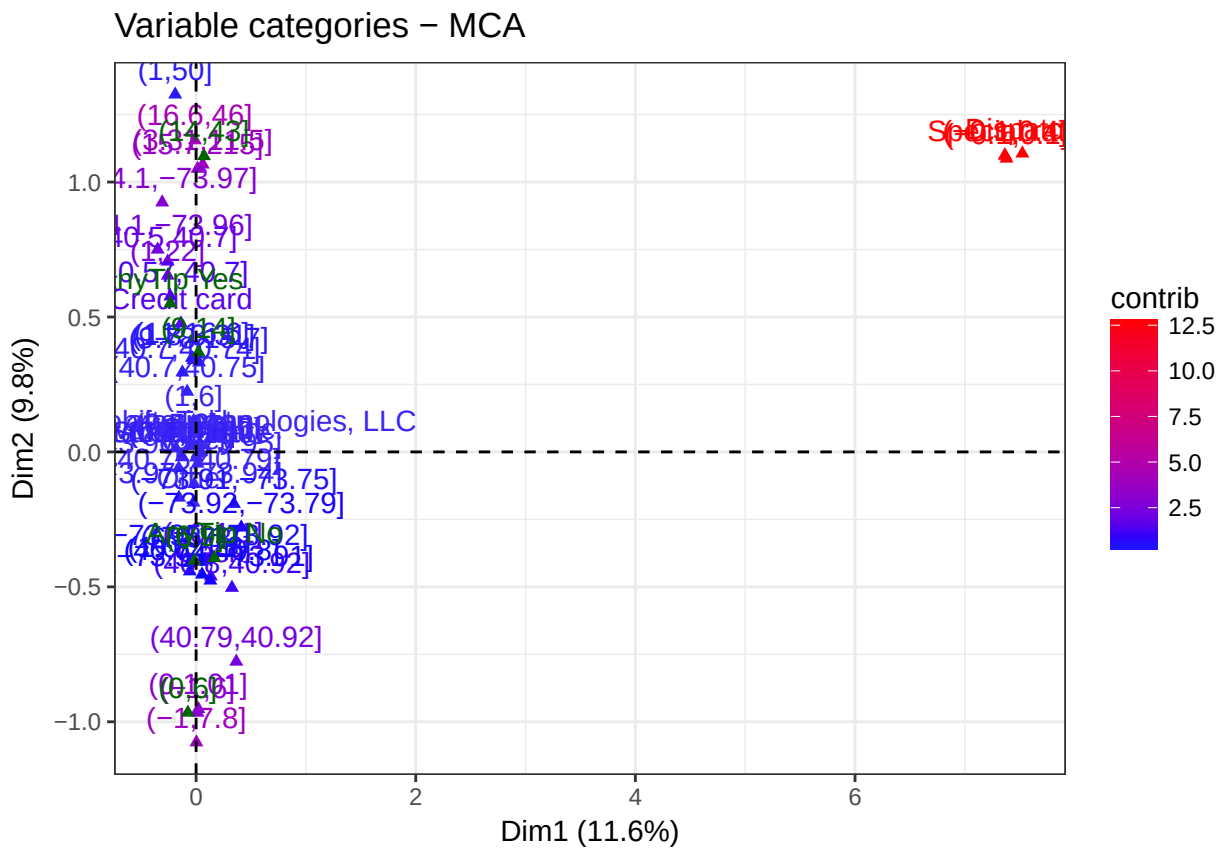

```
fviz_contrib(res.mca, choice = "var", axes = 3:4)+theme_bw()
```



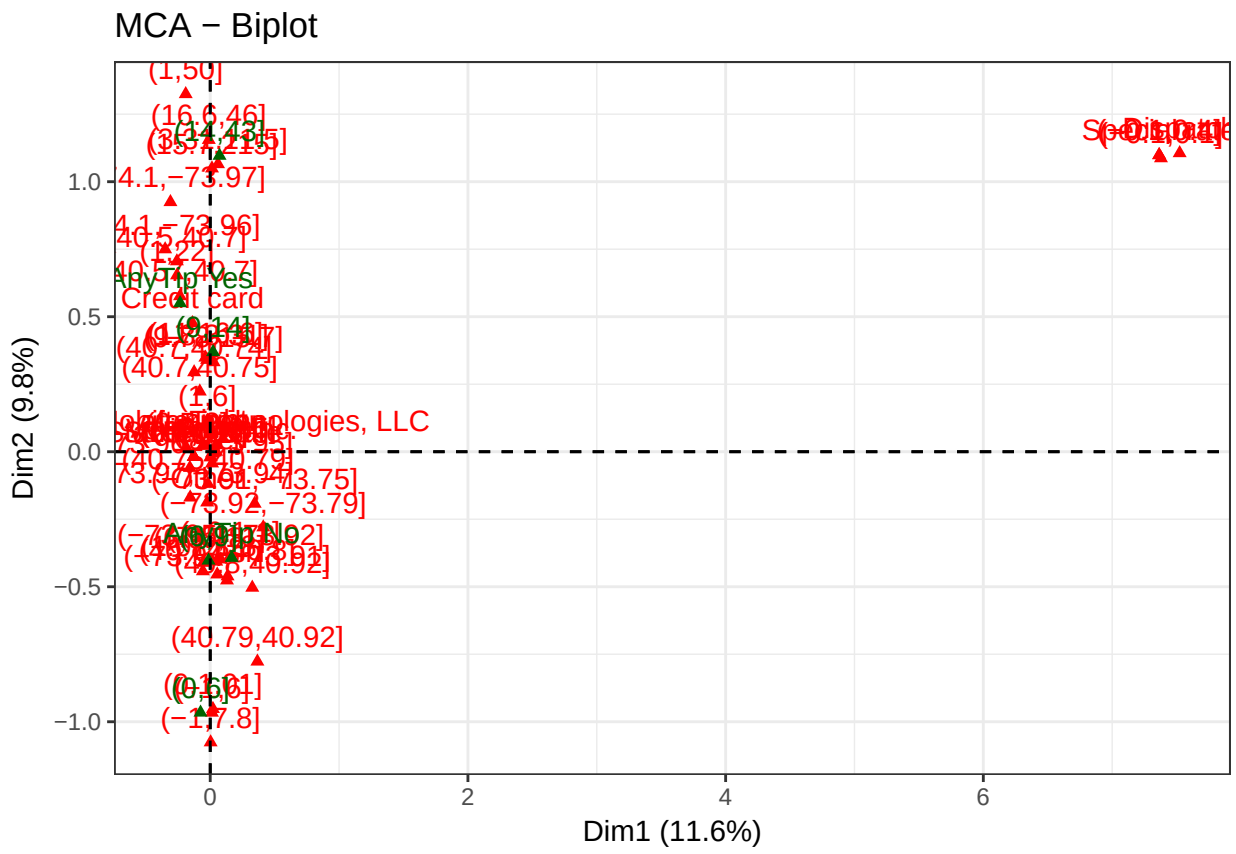

```
fviz_mca_var(res.mca, col.var="contrib",repel=TRUE)+
    scale_color_gradient2(low="green", mid="blue",
    high="red", midpoint=0.75)+theme_bw()
```




```
fviz_mca_var(res.mca, col.var="contrib",repel=F)+
    scale_color_gradient2(low="green", mid="blue",
    high="red", midpoint=0.75)+theme_bw()
```



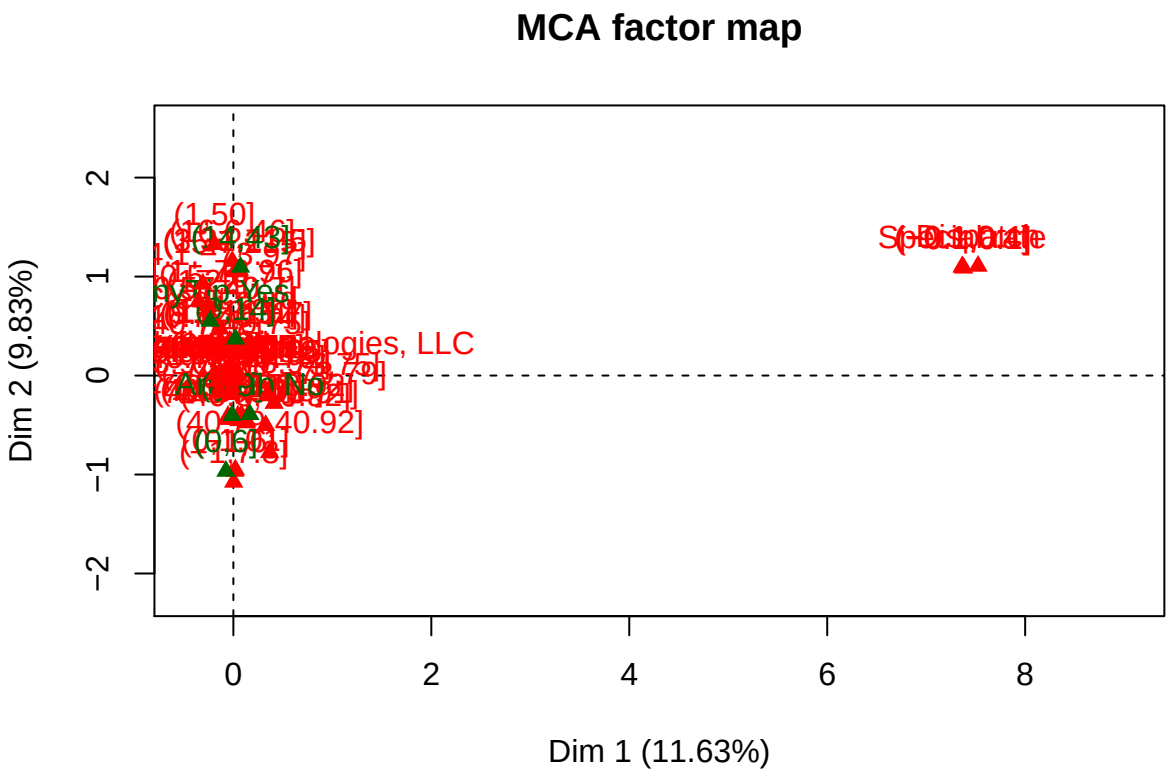

```
fviz_mca_biplot(res.mca, invisible="ind",axes=1:2,repel=FALSE)+theme_bw()
```

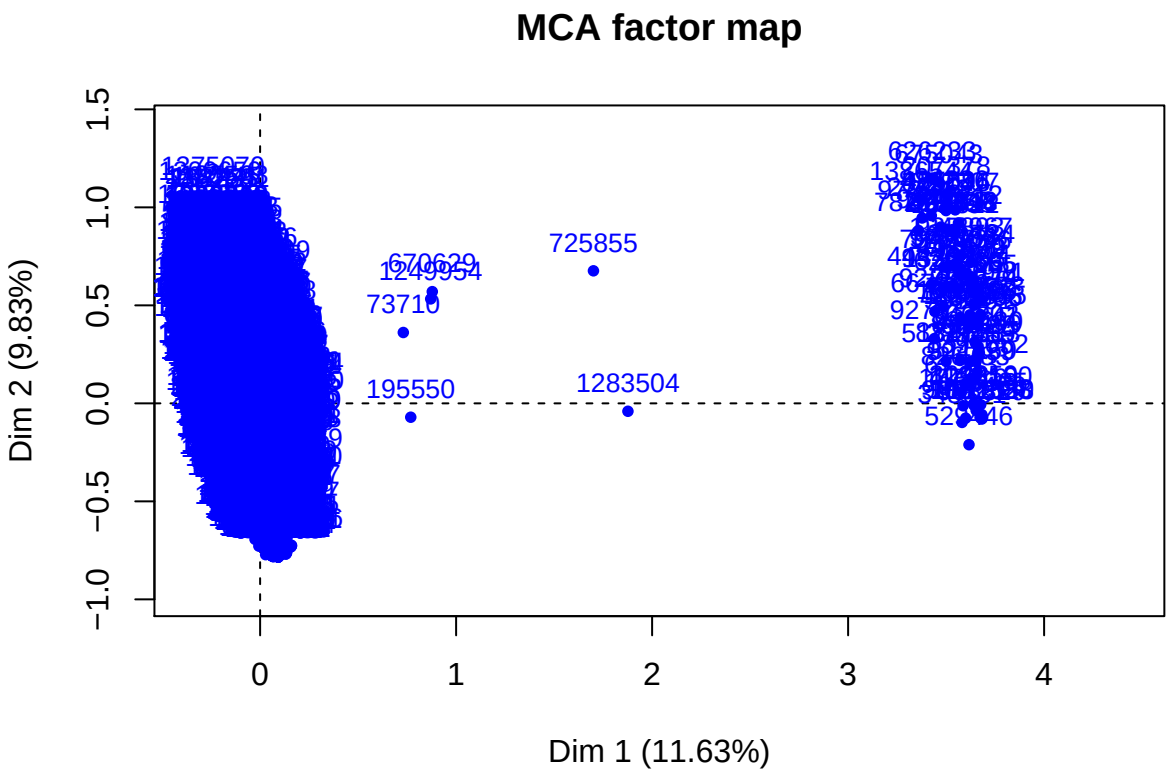


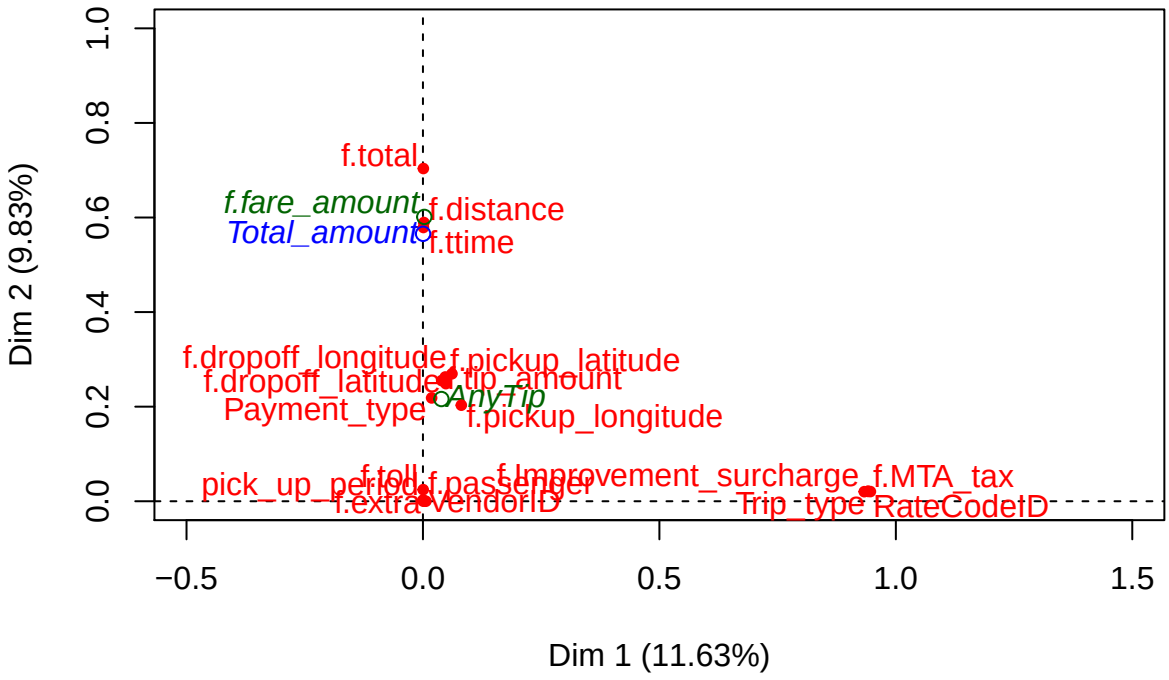

```
#fviz_mca_biplot(res.mca, axes=1:2,repel=TRUE,arrows=c(TRUE,FALSE))+theme_bw()
```

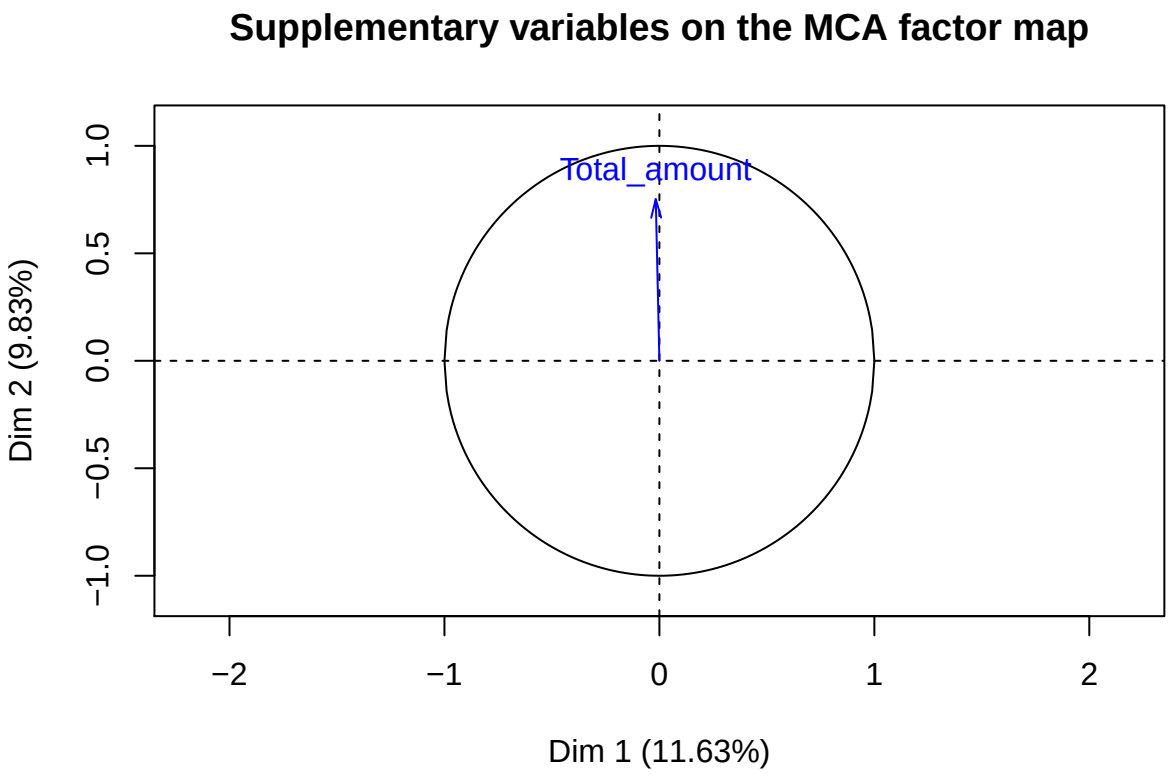
Synthesis through HCPC: Hierarchical Clustering

```
###
### Clustering the individuals
### Before, hou have to perform a MCA with the number of axes
### that you have decided to take into account (indicated through ncp=)
?HCPC
res.mca<-MCA(df[,c(vars_cat,"Total_amount")],quali.sup=c(5,13),quanti.sup=21,ncp=26)
```

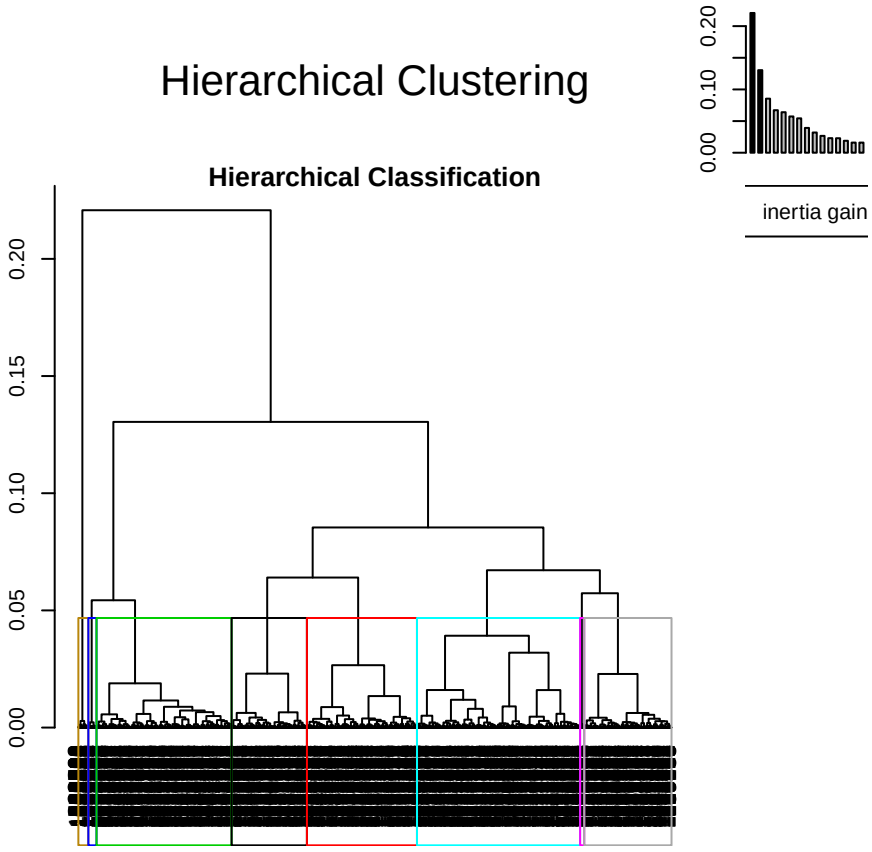

MCA factor map
Dim 2 (9.83%)
Dim 1 (11.63%)
2
1
0
−1
−2
0
2
4
6
8


MCA factor map
Dim 2 (9.83%)
Dim 1 (11.63%)
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
725855
670629
1249954
73710
195550
1283504

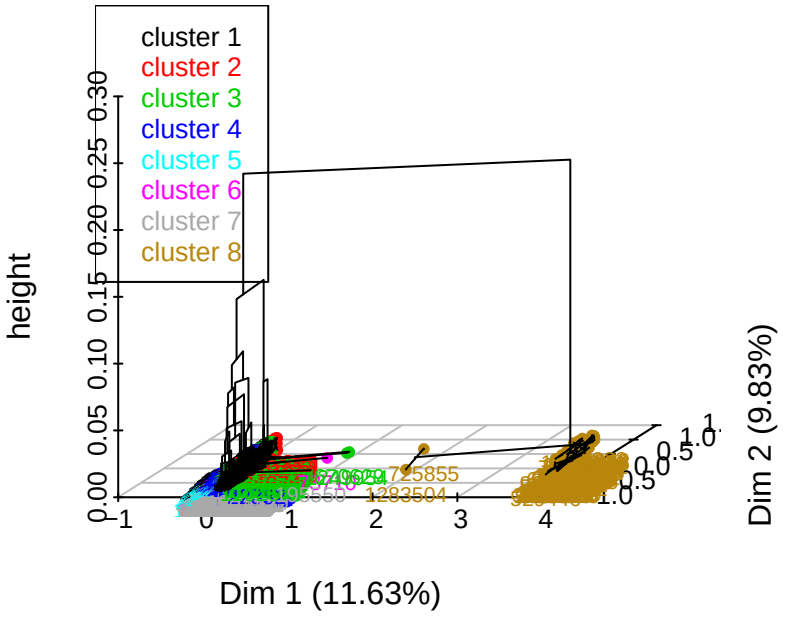

f.total
f.fare_amount
f.distance
Total_amount
f.ttime
f.dropoff_longitude
f.pickup_latitude
f.dropoff_latitude
tip_amount
cAnyTip
Payment_type
f.pickup_longitude
f.toll
f.Improvement_surcharge
f.MTA_tax
pick_up_period
f.passenger
f.extra
VendorID
Trip_type
RateCodeID
Dim 2 (9.83%)
Dim 1 (11.63%)
0.0
0.2
0.4
0.6
0.8
1.0
−0.5
0.0
0.5
1.0
1.5

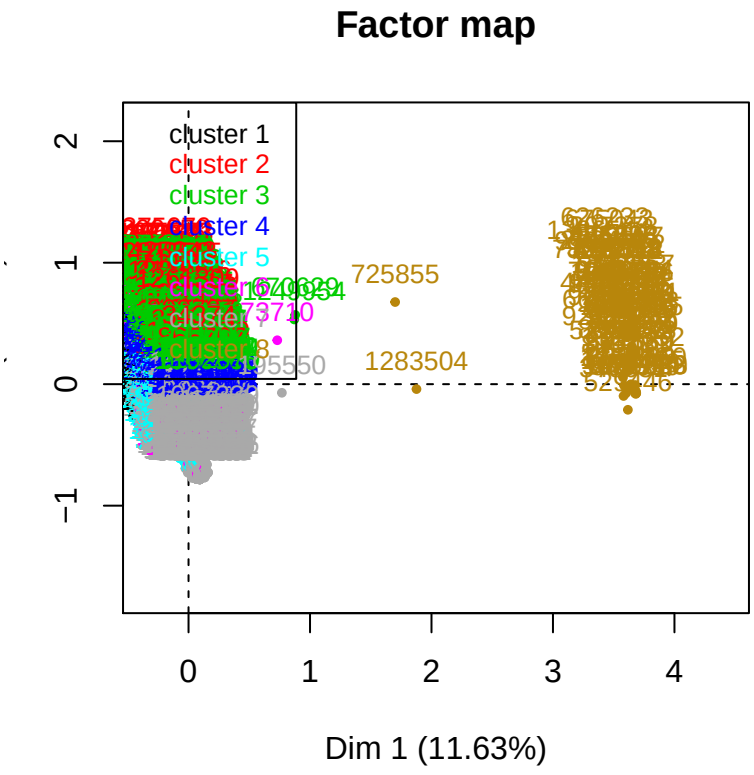



```
res.hcpc<-HCPC(res.mca,nb.clust=8,order=TRUE)
```


Hierarchical Clustering
Hierarchical Classification
0.00
0.05
0.10
0.15
0.20
0.00
0.10
0.20
inertia gain

Hierarchical clustering on the factor map

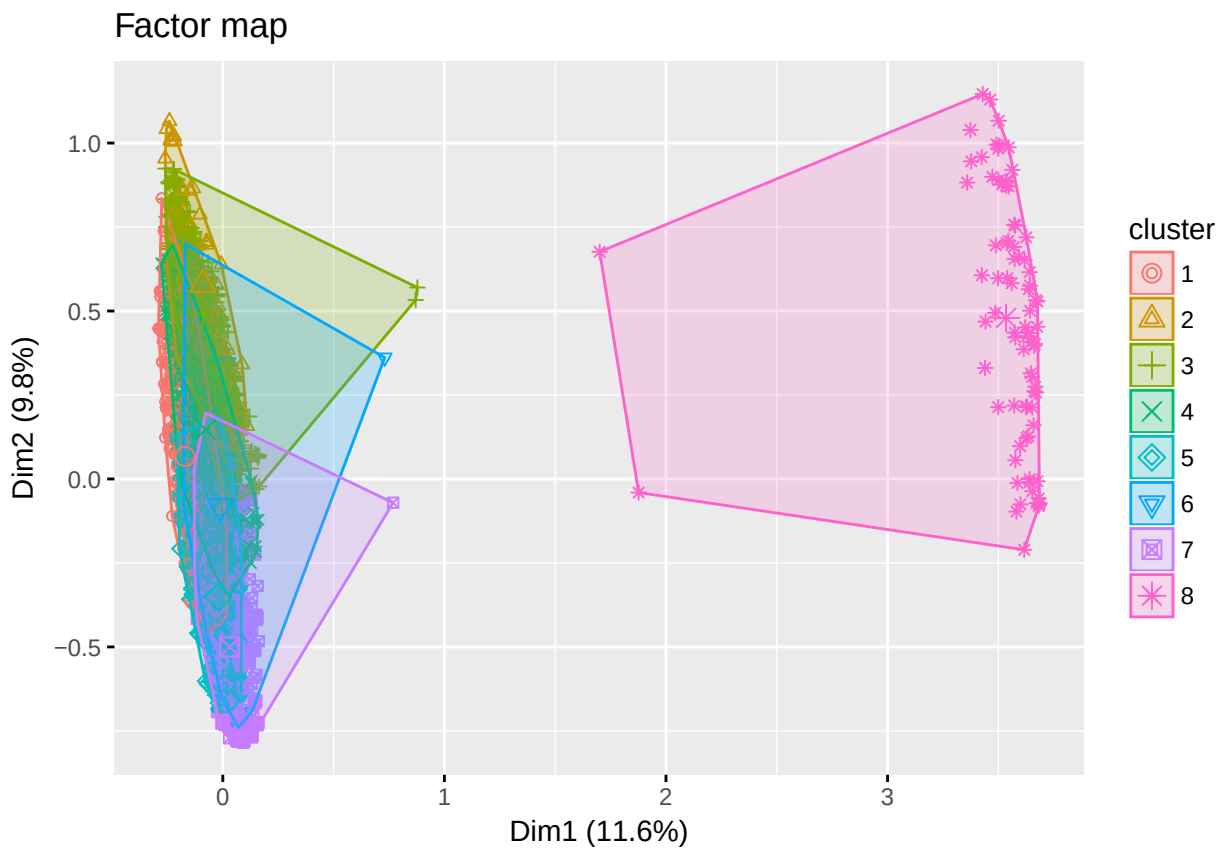




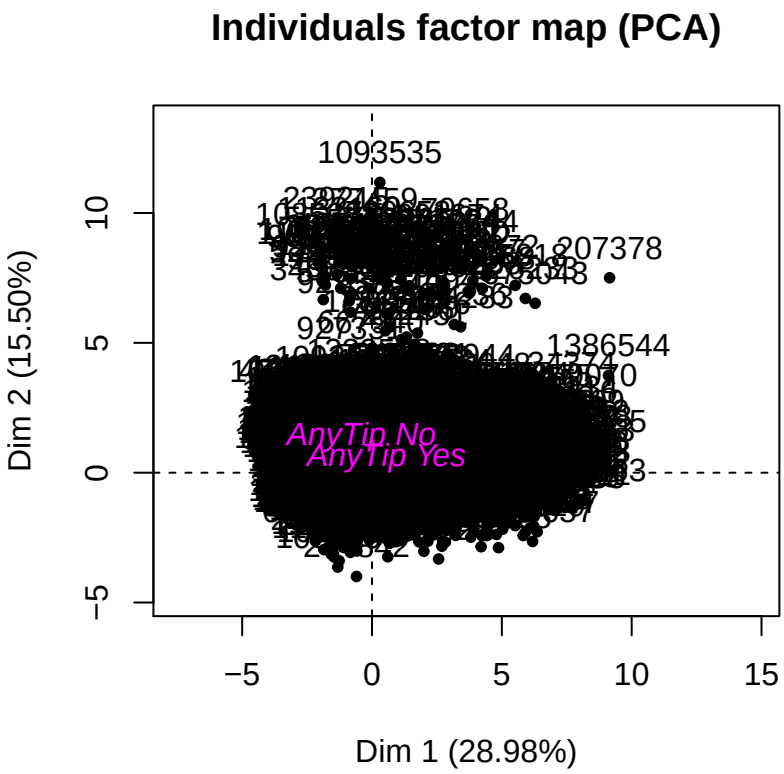

```
names(res.hcpc)
```

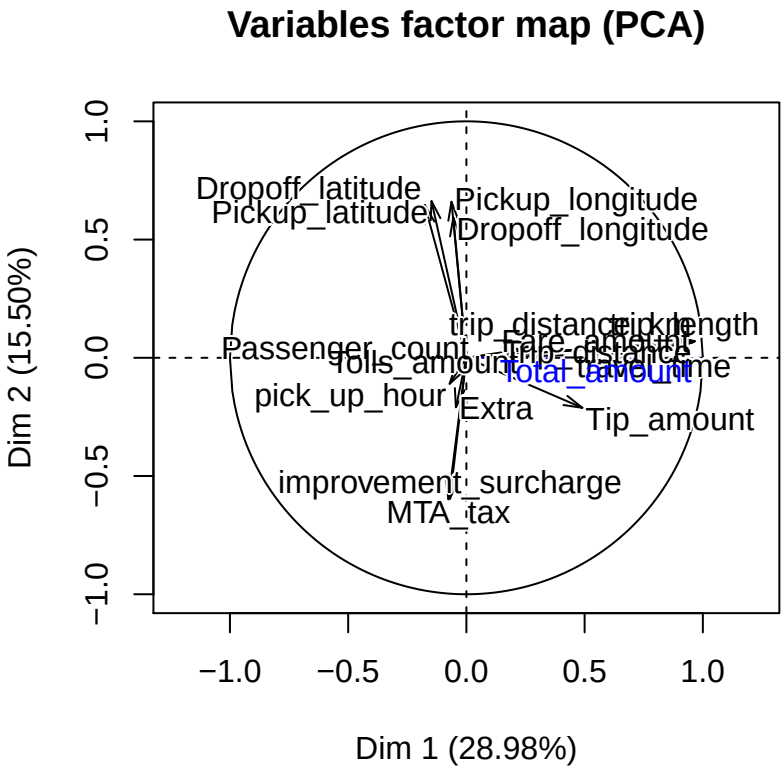
```
## [1] "data.clust" "desc.var"   "desc.axes"  "call"       "desc.ind"
```

```
fviz_cluster(res.hcpc, geom = "point", main = "Factor map")
```

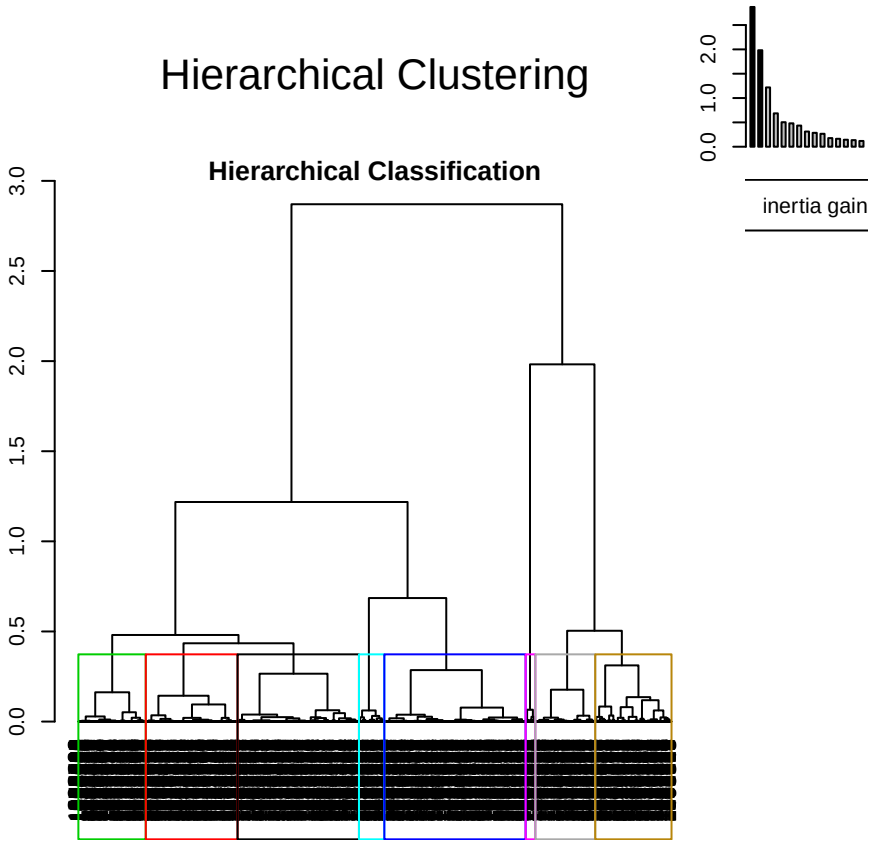



```
#Anytip
vars_con_pca<-c(6,7,8,9,10,11,12,13,14,15,16,17,18,22,23,24,25,26)
res.pca<-PCA(df[,vars_con_pca], quanti.sup = 13, quali.sup = 14, ncp = 6 ) # TotalAmount and AnyTip
```

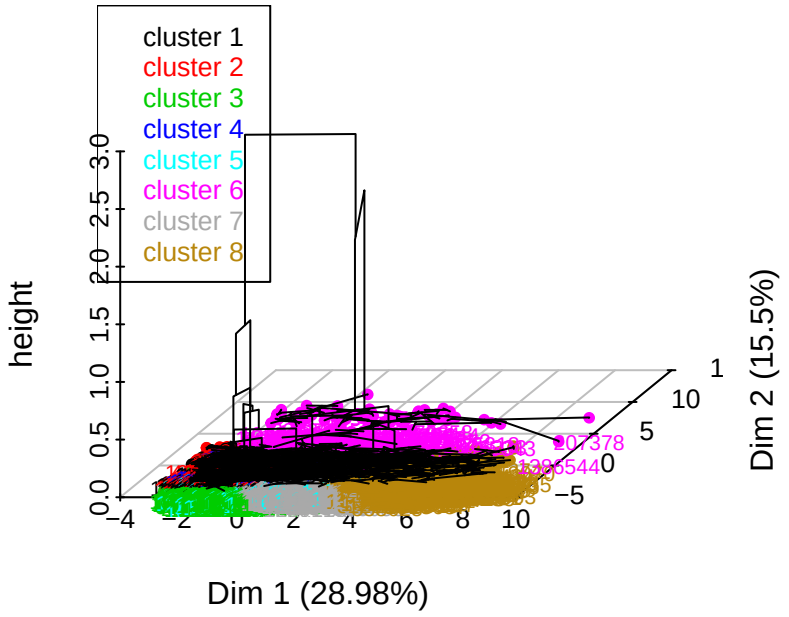

Individuals factor map (PCA)
1093535
207378
1386544
AnyTip No
AnyTip Yes
Dim 2 (15.50%)
Dim 1 (28.98%)
−5
0
5
10
15
−5
0
5
10

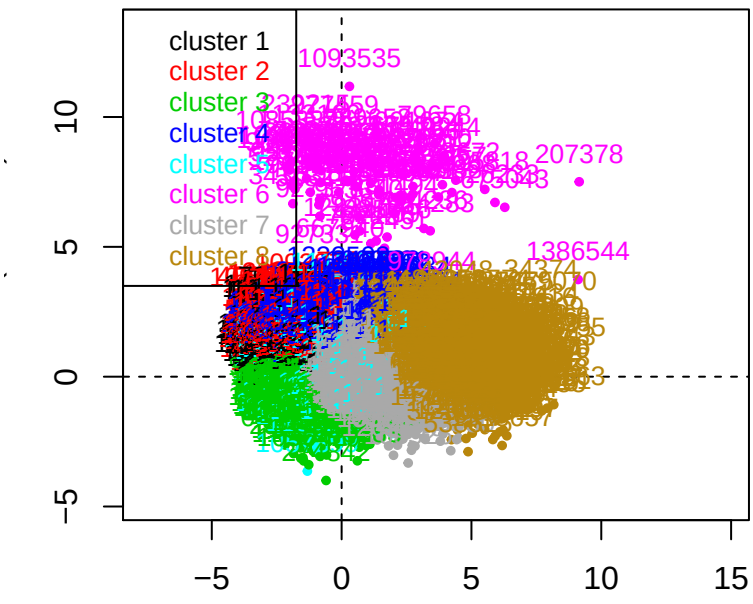



```
res.hcpcPCA <-HCPC(res.pca, nb.clust = 8, order=TRUE)
```


Hierarchical Clustering
Hierarchical Classification
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
1.0
2.0
inertia gain

Hierarchical clustering on the factor map






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

```
##
##       1   2   3   4   5   6   7   8
##   1   0   1 608   7  27   0  19   0
##   2   1   2   0   0   3   0  14  47
##   3  14  21   8  93  36   0 546 369
##   4 167 174 277 183  56   0 141   0
##   5 184 230 253 356  84   0   0   1
##   6   5  11   6   0   0   0   9   3
##   7 264 523   0   3  32   0   3   3
##   8   0   0   0   0   0  81   1   0
```

```
claHPPCA<-factor(res.hcpcPCA$data.clust$clust,labels=paste("kHP-",1:8))
claHP<-factor(res.hcpc$data.clust$clust,levels=c(5,7,1,4,2,8,3,6),labels=c("kKM-5","kKM-7","kKM-1","kKM
tt<-table(claHPPCA,claHP)
tt
```

```
##          claHP
## claHPPCA kKM-5 kKM-7 kKM-1 kKM-4 kKM-2 kKM-8 kKM-3 kKM-6
##   kHP- 1   184   264     0   167     1     0    14     5
##   kHP- 2   230   523     1   174     2     0    21    11
##   kHP- 3   253     0   608   277     0     0     8     6
##   kHP- 4   356     3     7   183     0     0    93     0
##   kHP- 5    84    32    27    56     3     0    36     0
##   kHP- 6     0     0     0     0     0    81     0     0
##   kHP- 7     0     3    19   141    14     1   546     9
```

```
##   kHP- 8     1     3     0     0    47     0   369     3
```

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

```
## [1] 0.4379367
```

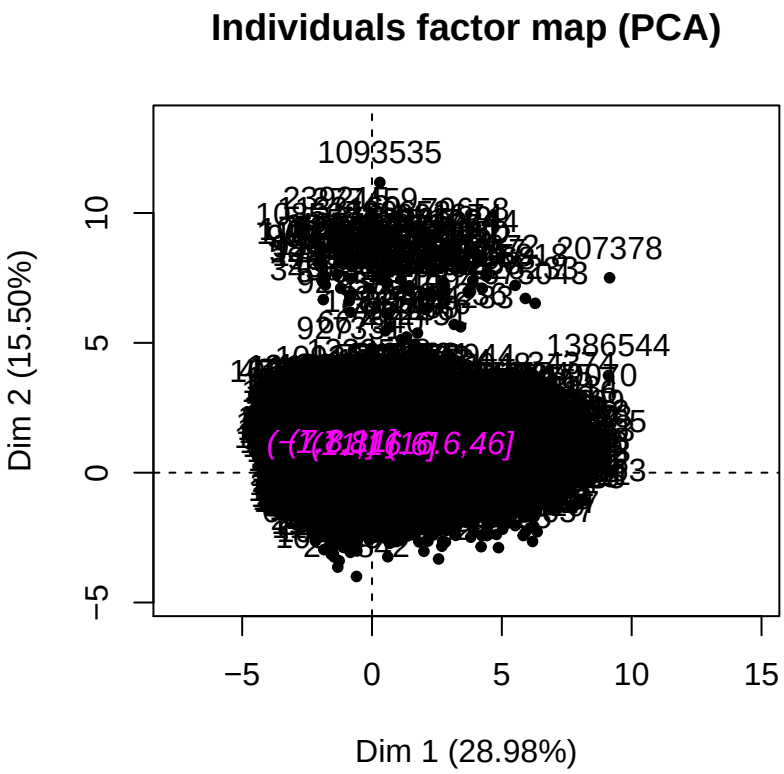
```
#F.total
vars_con_pca<-c(6,7,8,9,10,11,12,13,14,15,16,17,18,23,24,25,26,41)
names(df[,vars_con_pca])
```

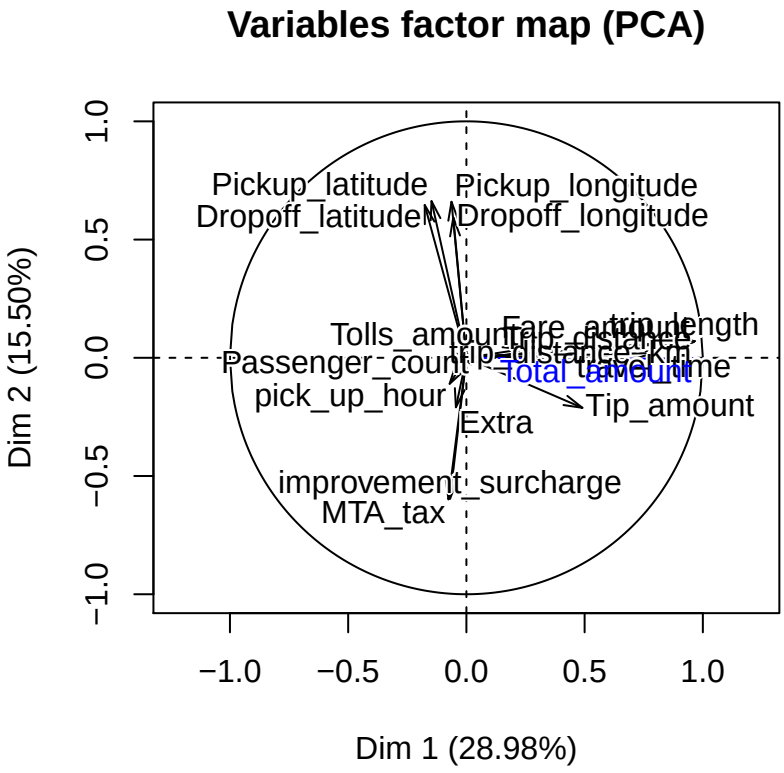
```
##  [1] "Pickup_longitude"      "Pickup_latitude"
##  [3] "Dropoff_longitude"     "Dropoff_latitude"
##  [5] "Passenger_count"       "Trip_distance"
##  [7] "Fare_amount"           "Extra"
##  [9] "MTA_tax"               "Tip_amount"
## [11] "Tolls_amount"          "improvement_surcharge"
## [13] "Total_amount"          "trip_length"
## [15] "trip_distance_km"      "travel_time"
## [17] "pick_up_hour"          "f.total"
```

```
names(df)
```

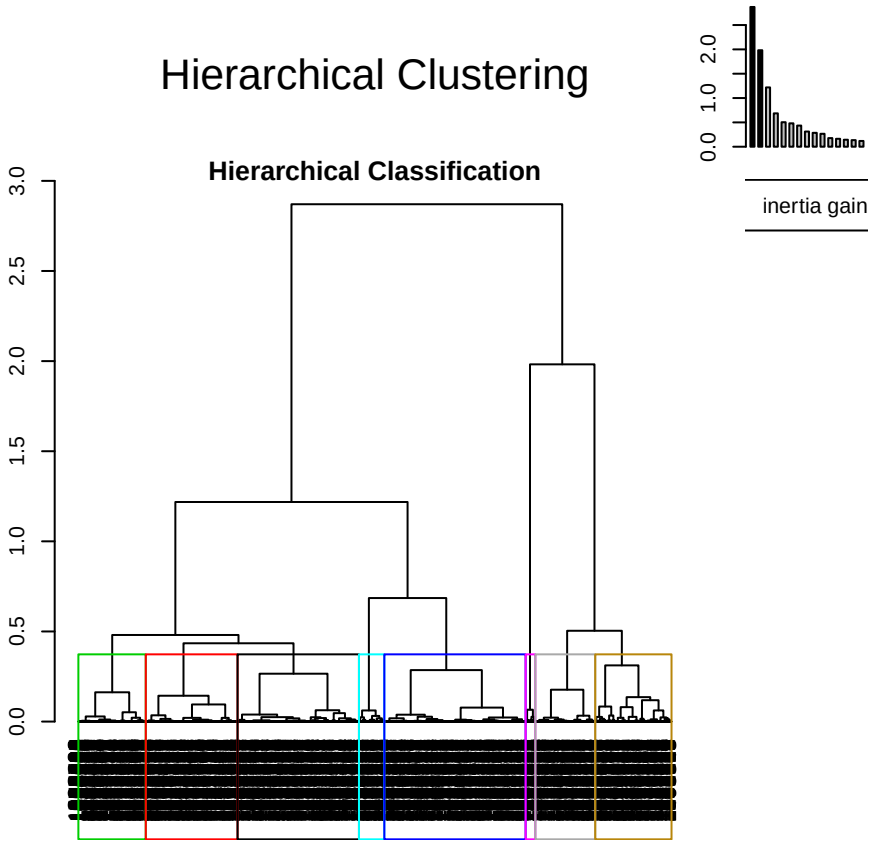
```
##  [1] "VendorID"                "lpep_pickup_datetime"
##  [3] "Lpep_dropoff_datetime"   "Store_and_fwd_flag"
##  [5] "RateCodeID"              "Pickup_longitude"
##  [7] "Pickup_latitude"         "Dropoff_longitude"
##  [9] "Dropoff_latitude"        "Passenger_count"
## [11] "Trip_distance"  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

```
res.pca<-PCA(df[,vars_con_pca], quanti.sup = 13, quali.sup = 18, ncp = 6 ) # TotalAmount and f.total
```

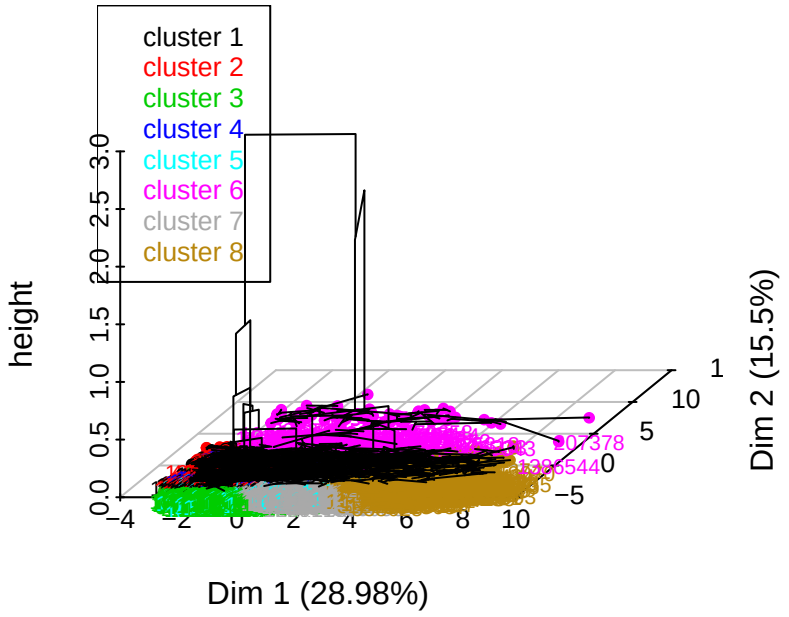

Individuals factor map (PCA)
1093535
207378
1386544
Dim 2 (15.50%)
Dim 1 (28.98%)
−5
0
5
10
15

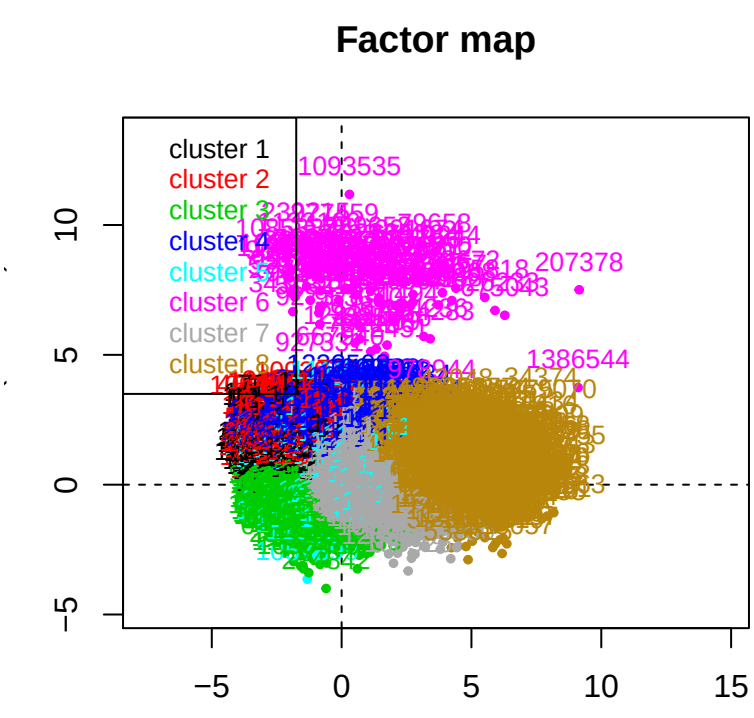



```
res.hcpcPCA <-HCPC(res.pca, nb.clust = 8, order=TRUE)
```


Hierarchical Clustering
Hierarchical Classification
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
1.0
2.0
inertia gain

Hierarchical clustering on the factor map






```
claHPPCA<-factor(res.hcpcPCA$data.clust$clust,labels=paste("kHP-",1:8))
claHP<-factor(res.hcpc$data.clust$clust,levels=c(4,7,1,5,2,8,3,6),labels=c("kKM-5","kKM-7","kKM-1","kKM
tt<-table(claHPPCA,claHP)
tt
```

```
##          claHP
## claHPPCA kKM-5 kKM-7 kKM-1 kKM-4 kKM-2 kKM-8 kKM-3 kKM-6
##   kHP- 1   167   264     0   184     1     0    14     5
##   kHP- 2   174   523     1   230     2     0    21    11
##   kHP- 3   277     0   608   253     0     0     8     6
##   kHP- 4   183     3     7   356     0     0    93     0
##   kHP- 5    56    32    27    84     3     0    36     0
##   kHP- 6     0     0     0     0     0    81     0     0
##   kHP- 7   141     3    19     0    14     1   546     9
##   kHP- 8     0     3     0     1    47     0   369     3
```

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

```
## [1] 0.4699959
```