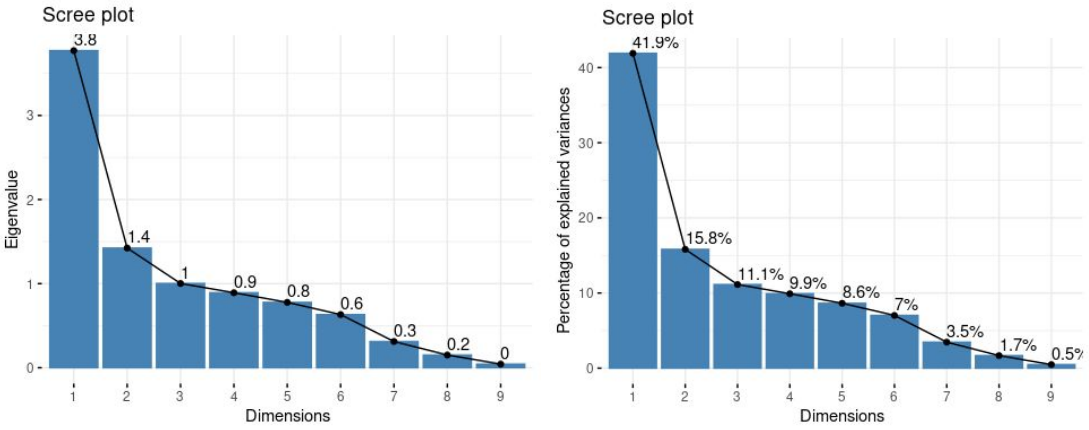
Deliverable II

Carles Capilla Cànovas
Jesús Molina Roldán

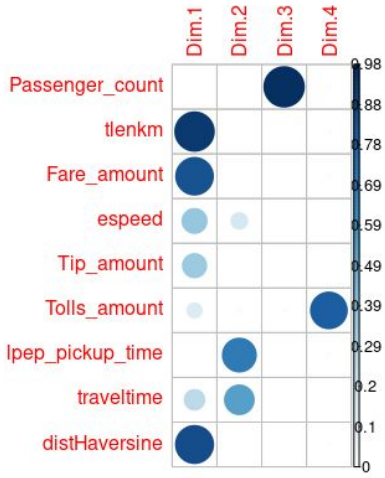
Principal Component Analysis (PCA)

Eigenvalues and dominant axes analysis.



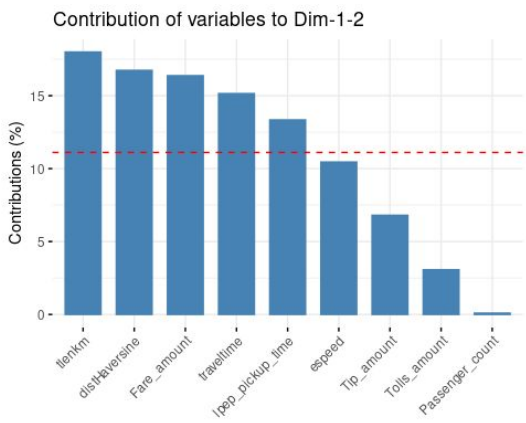
En esta imagen podemos ver los valores propios. Como podemos ver hasta la tercera dimensión tenemos un valor propio igual a 1. Según el criterio de Kaiser deberíamos eliminar todas las componentes con valor propio por debajo de 1, lo que significa que deberíamos coger hasta la tercera dimensión. Según la regla de Elbow, debemos coger hasta que no haya un descenso significativo, lo que significa que también se debería coger hasta la tercera dimensión. A pesar de todo, hemos decidido incluir hasta la cuarta dimensión, ya que nos facilita el estudio. Como podemos ver hasta la cuarta dimensión encontramos una varianza acumulada del 78.75%. También podemos admirar como la primera dimensión contribuye mucho en el PCA, explicando un 41.9% de la varianza.

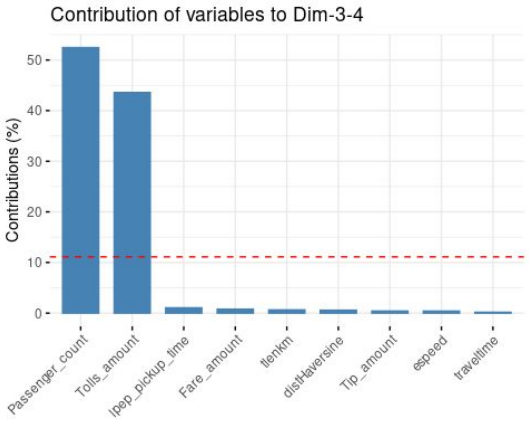
Quality of representation

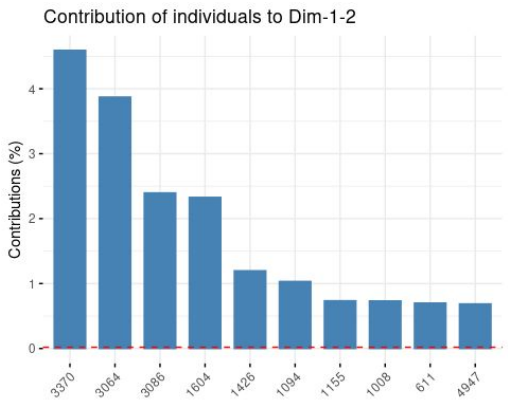


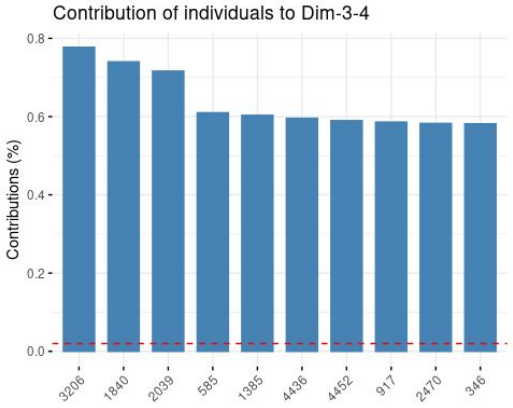
A partir de la suma del coseno al cuadrado de la primera dimensión más el coseno al cuadrado de la segunda dimensión podemos obtener cualidad de las variables en el primer plano factorial. Como podemos ver tlenkm y distHaversine son las dos variables que mejor se representan en la primera dimensión.

Contribution





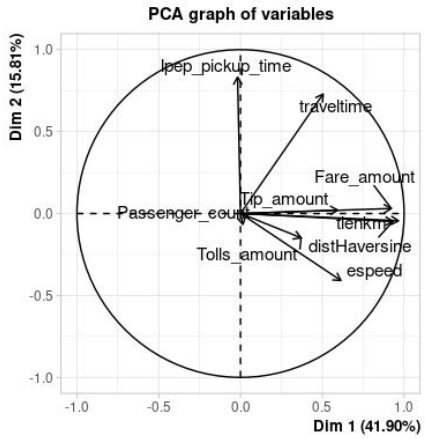




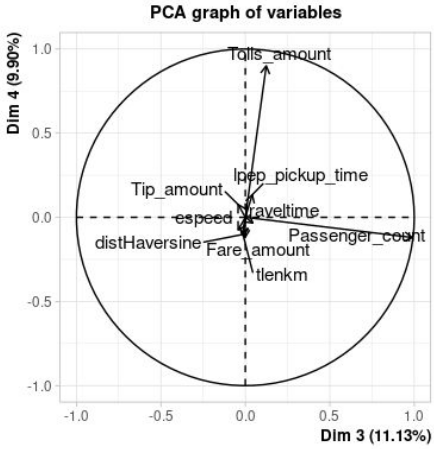
En este plot encontramos las 10 variables que contribuyen más en el primer plano factorial. Como podemos ver la variable tlenkm es la variable que contribuye más junto con la variable distHaversine. Si vemos el segundo plano factorial, vemos que el número total de pasajeros y el número total de pagos por peajes influyen bastante en el segundo plano factorial. En este plot encontramos los 10 individuos que contribuyen más en el primer plano factorial. Como podemos ver el individuo 3370 es el individuo que contribuye más, junto con el individuo 3064 al primer plano factorial. Si vemos el segundo plano factorial, vemos que el individuo 3206 y el individuo 1840, son los individuos que contribuyen más el segundo plano factorial.

Interpreting the axes

Active numerical variables

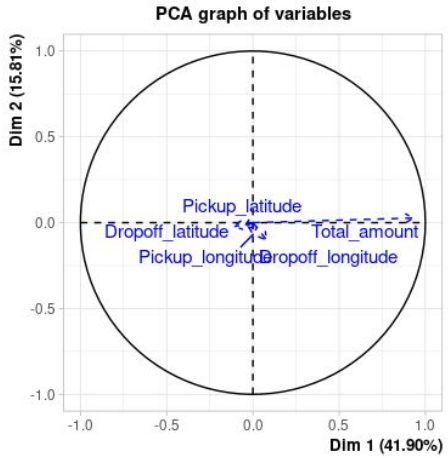


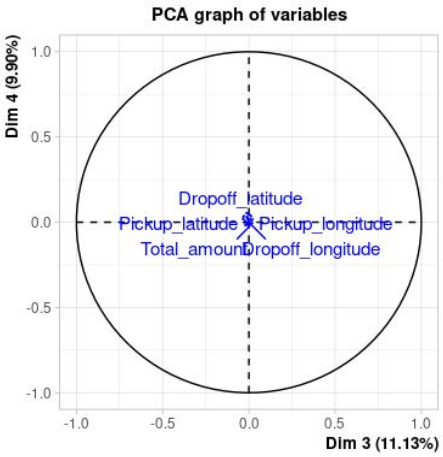
Si vemos el primer plano factorial podemos ver que la variable mejor representadas en la primera dimensión es la cantidad de km recorridos en el taxi y la tarifa pagada. Como podemos ver las variables tlenkm, Tip_amount y Fare_amount están agrupadas, lo que significa que están positivamente correlacionadas.



Si observamos el segundo plano factorial, podemos de nuevo que la variable Passenger contribuye mucho en la dimensión 3 y que la variable Tolls_amount contribuye mucho en la dimensión 4.

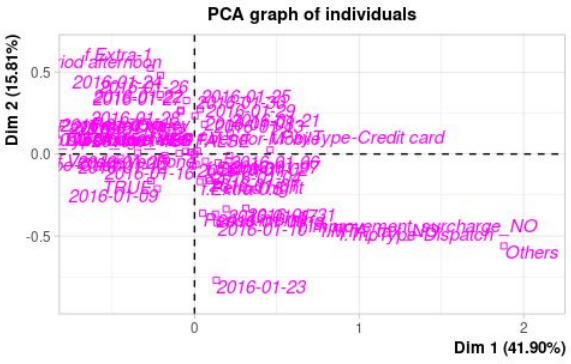
Supplementary numerical variables

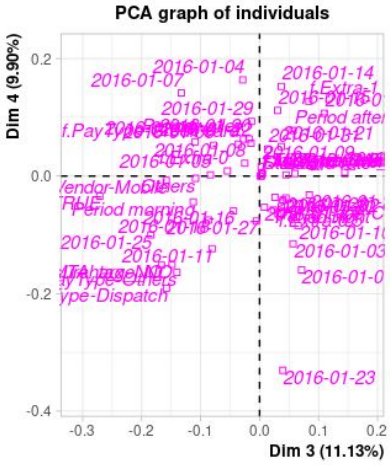




Como podemos ver, la única variable numérica complementaria que contribuye en la construcción del primer plano factorial es Total_amount.

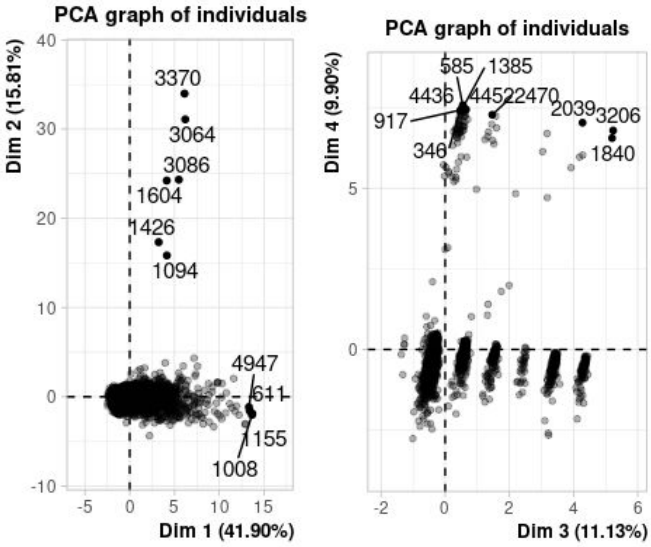
Supplementary categorical variables





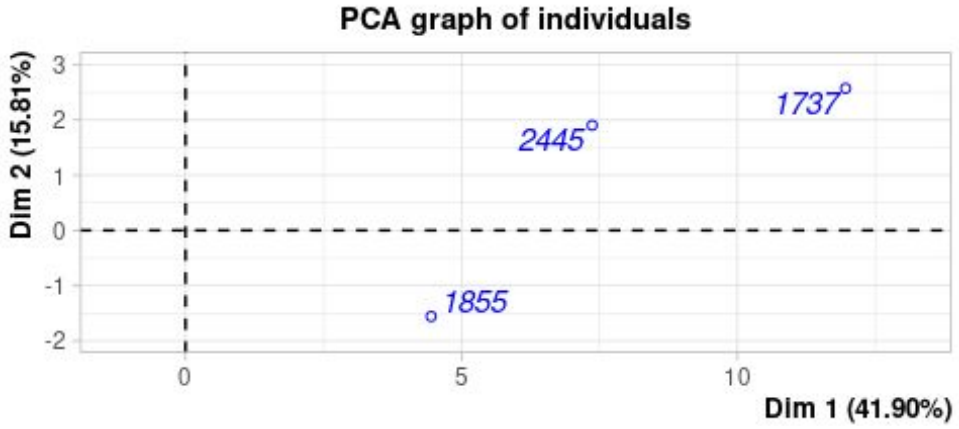
En estas dos imágenes nos encontramos las variables categóricas suplementarias situadas en las dos dimensiones más importantes. Si nos fijamos, los viajes realizados el día 23 de enero del 2016 se encuentran siempre alejados del conjunto. Las categorías son difíciles de distinguir pero si nos fijamos, en ambos planos factoriales los individuos con f.Extra-1 están cerca de los individuos con Period afternoon.

Individuals



Si analizamos el primer plano factorial vemos que los individuos que contribuyen más en el plano forma dos grupos. El primer grupo está formado por los individuos 4947, 611, 1155 y 1008. El segundo grupo está formado por los individuos 3064, 3086, 1604, 1426 y 1094. Si vemos en el dataset las características de dichos individuos encontramos que son viajes de taxis con un largo recorrido en km, más de 30 km, tal como se puede analizar con el plano factorial de las variables activas numéricas. Si vemos los individuos que han contribuido al segundo plano factorial, vemos que son individuos con que han pagado 5.54 dólares en peajes.

Supplementary individuals

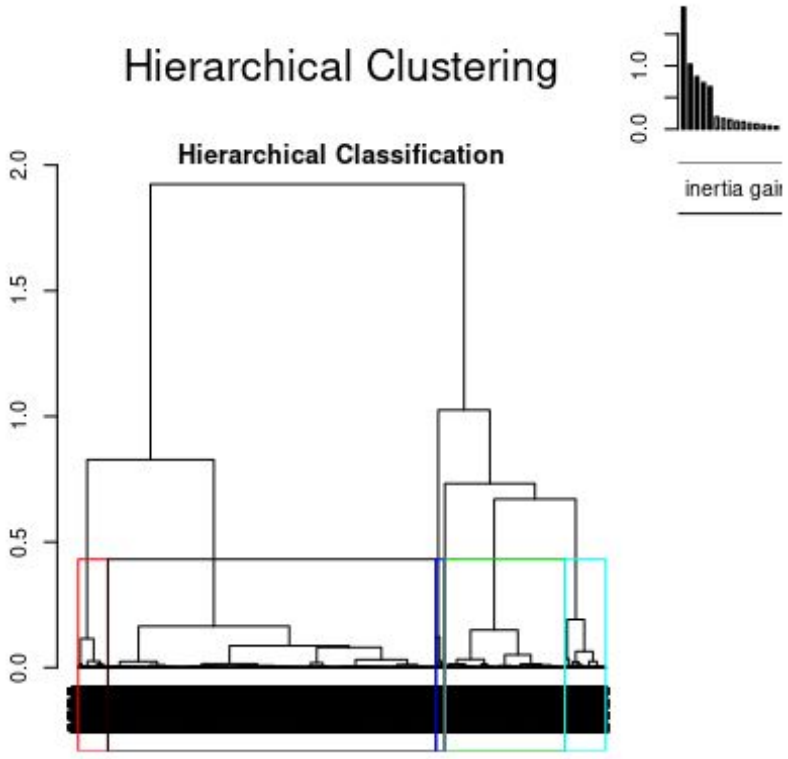



```
> df[c("1855","2445", "1737"), ]
             VendorID          Payment_type Store_and_fwd_flag    RateCodeID   f.Extra     f.MTA_tax     f.Improvement_surcharge
1855 f.Vendor-VeriFone f.PayType-Credit card              FALSE        Others f.Extra-0  f.MTA_tax_NO  f.Improvement_surcharge_NO
2445 f.Vendor-VeriFone        f.PayType-Cash              FALSE        Others f.Extra-0  f.MTA_tax_NO  f.Improvement_surcharge_NO
1737   f.Vendor-Mobile      f.PayType-Others              FALSE Standard rate f.Extra-0 f.MTA_tax_YES f.Improvement_surcharge_YES
     lpep_pickup_period             Trip_type lpep_pickup_date multiouts          f.espeed          f.tlenkm          f.traveltime
1855     Period morning    f.TripType-Dispatch       2016-01-12      TRUE f.espeed-(25,130]    f.tlenkm-[0,1]   f.traveltime-[0,10]
2445       Period night    f.TripType-Dispatch       2016-01-15      TRUE f.espeed-(25,130] f.tlenkm-(5,67.9] f.traveltime-(40,548]
1737      Period valley f.TripType-Street-Hail       2016-01-11      TRUE f.espeed-(25,130] f.tlenkm-(5,67.9] f.traveltime-(40,548]
          f.distHaversine     AnyToll             f.Fare_amount        f.Passenger_count          f.Total_amount Passenger_count     tlenkm
1855  f.distHaversine-[0,5] AnyToll Yes f.Fare_amount-(14.5,71.5]      f.Passenger_count-2 f.Total_amount-(40,95.5]               2  0.0804672
2445  f.distHaversine-[0,5] AnyToll Yes    f.Fare_amount-(9,14.5] f.Passenger_count-Others    f.Total_amount-(8,11]               3 56.0856400
1737 f.distHaversine-(5,10]  AnyToll No f.Fare_amount-(14.5,71.5]      f.Passenger_count-1 f.Total_amount-(40,95.5]               1 67.9143168
     Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude Fare_amount   espeed Tip_amount Tolls_amount lpep_pickup_time
1855        -74.08338        40.64220         -74.08366         40.64284    65.00000 48.28032         13      0.00000                9
2445        -74.00069        40.59997         -73.96132         40.62992    10.00000 40.04528          0      0.00000               20
1737        -73.86628        40.87276         -73.86414         40.82361    61.50166 37.35547          0      0.22224               15
     traveltime distHaversine     AnyTip Total_amount  hcpck claKM
1855    0.10000    0.07562334 AnyTip Yes     78.00000 kHP- 4 kKM-1
2445   84.03333    4.70946300  AnyTip No     10.00000 kHP- 4 kKM-3
1737  109.08333    5.47499787  AnyTip No     70.61709 kHP- 4 kKM-3
```

```
> res.pca$ind$contrib[c("1855","2445", "1737"), ]
         Dim.1      Dim.2        Dim.3       Dim.4
1855 0.1058658 0.03712658 0.0003443667 0.006059698
2445 0.2744526 0.05496966 0.0474553876 0.083996708
1737 0.7316331 0.09785061 0.0016757330 0.152349937
```

Si vemos los individuos suplementarios observamos que el individuo 1737 es un individuo que contribuye más en el primer plano factorial. En contraposición el individuo que contribuye más al segundo plano factorial es el 2445. Si observamos las características de dichos individuos vemos que la distancia recorrida de ambos viajes es superior a 50 km.

HCPC



Viendo la inertia gain (pérdida importante de ir entre n clusters a n+1 clusters) y aplicando Kaiser Rule podemos ver que el número de clusters óptimo es 6.

Description of the clusters by the variables

Categorical variables which characterizes the clusters

```
##                              p.value df
## Payment_type            8.703621e-41 10
## RateCodeID              7.823482e-15  5
## VendorID                9.081168e-11  5
## f.Extra                 7.914188e-08 10
## lpep_pickup_period      5.249054e-06 15
## Trip_type               9.975894e-06  5
## f.MTA_tax               2.966099e-05  5
## f.Improvement_surcharge 2.904864e-04  5
```

En esta imagen podemos encontrar las variables categóricas que contribuyen más en los clusters ordenadas por importancia. Podemos ver que las variables categóricas Payment_type, RateCodeID y VendorID son muy importantes en la construcción de los clusters.

Description of each cluster by the categories

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Payment_type=f.PayType-Cash	70.64220	56.817453	50.14	3.893540e-34	12.181644
RateCodeID=Standard rate	62.94009	98.075072	97.14	6.026448e-07	4.990368
f.Extra=f.Extra-1	68.17043	17.452679	15.96	1.855149e-04	3.737966
lpep_pickup_period=Period afternoon	67.31813	19.890921	18.42	5.129457e-04	3.473903
Trip_type=f.TripType-Street-Hail	62.64097	98.010908	97.54	6.556979e-03	2.718545
lpep_pickup_date=2016-01-05	71.53285	3.144049	2.74	2.274727e-02	2.277653
f.MTA_tax=f.MTA_tax_YES	62.57707	97.690087	97.32	3.959358e-02	2.057964
lpep_pickup_date=2016-01-19	70.13889	3.240295	2.88	4.822322e-02	1.975396
lpep_pickup_date=2016-01-31	55.31915	3.336542	3.76	4.502308e-02	-2.004439
f.MTA_tax=f.MTA_tax_NO	53.73134	2.309913	2.68	3.959358e-02	-2.057964
lpep_pickup_date=2016-01-17	54.16667	2.919474	3.36	2.799497e-02	-2.197357
Trip_type=f.TripType-Dispatch	50.40650	1.989092	2.46	6.556979e-03	-2.718545
f.Extra=f.Extra-0.5	59.43945	36.060314	37.82	9.855191e-04	-3.294628
lpep_pickup_period=Period night	58.53211	40.936798	43.60	1.065019e-06	-4.879228
RateCodeID=Others	41.95804	1.924928	2.86	6.026448e-07	-4.990368
Payment_type=f.PayType-Credit card	53.89610	42.605069	49.28	4.274103e-34	-12.174037

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
VendorID=f.Vendor-VeriFone	6.954067	94.7552448	77.94	8.701382e-16	8.043914
f.Extra=f.Extra-0.5	6.980434	46.1538462	37.82	3.089559e-03	2.958684
lpep_pickup_period=Period night	6.697248	51.0489510	43.60	9.268990e-03	2.601970
f.Improvement_surcharge=f.Improvement_surcharge_YES	5.841218	99.3006993	97.24	1.470478e-02	2.439569
Payment_type=f.PayType-Cash	6.501795	56.9930070	50.14	1.703725e-02	2.385903
f.MTA_tax=f.MTA_tax_YES	5.836416	99.3006993	97.32	1.784021e-02	2.368918
Trip_type=f.TripType-Street-Hail	5.823252	99.3006993	97.54	3.013982e-02	2.168248
f.Extra=f.Extra-1	7.393484	20.6293706	15.96	3.102468e-02	2.156756
RateCodeID=Standard rate	5.826642	98.9510490	97.14	4.158928e-02	2.037607
lpep_pickup_date=2016-01-07	2.142857	1.0489510	2.80	4.729447e-02	-1.983654
RateCodeID=Others	2.097902	1.0489510	2.86	4.158928e-02	-2.037607
lpep_pickup_period=Period morning	3.953871	8.3916084	12.14	3.913031e-02	-2.062814
Trip_type=f.TripType-Dispatch	1.626016	0.6993007	2.46	3.013982e-02	-2.168248
lpep_pickup_date=2016-01-25	1.986755	1.0489510	3.02	2.938094e-02	-2.178337
Payment_type=f.PayType-Credit card	4.951299	42.6573427	49.28	2.107704e-02	-2.306602
f.MTA_tax=f.MTA_tax_NO	1.492537	0.6993007	2.68	1.784021e-02	-2.368918
lpep_pickup_period=Period valley	4.411765	19.9300699	25.84	1.654204e-02	-2.396732
f.Improvement_surcharge=f.Improvement_surcharge_NO	1.449275	0.6993007	2.76	1.470478e-02	-2.439569
f.Extra=f.Extra-0	4.110775	33.2167832	46.22	4.527082e-06	-4.585589
VendorID=f.Vendor-Mobile	1.359927	5.2447552	22.06	8.701382e-16	-8.043914

$`3`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Payment_type=f.PayType-Credit card	29.82955	60.693642	49.28	5.929786e-20	9.145603
lpep_pickup_period=Period night	26.60550	47.894302	43.60	5.533695e-04	3.453497
f.Extra=f.Extra-0.5	26.22951	40.957886	37.82	9.942040e-03	2.577839
lpep_pickup_date=2016-01-01	29.95595	5.615194	4.54	4.304996e-02	2.023225
RateCodeID=Others	31.46853	3.715937	2.86	4.549479e-02	2.000051
lpep_pickup_date=2016-01-19	17.36111	2.064410	2.88	4.637866e-02	-1.991931
RateCodeID=Standard rate	24.00659	96.284063	97.14	4.549479e-02	-2.000051
lpep_pickup_date=2016-01-12	16.99346	2.146986	3.06	2.980772e-02	-2.172636
lpep_pickup_date=2016-01-14	16.86747	2.312139	3.32	2.089799e-02	-2.309822
lpep_pickup_period=Period afternoon	20.52117	15.606936	18.42	3.327487e-03	-2.935744
f.Extra=f.Extra-1	18.67168	12.303881	15.96	4.570701e-05	-4.076559
Payment_type=f.PayType-Cash	18.74751	38.810900	50.14	1.065964e-19	-9.082001

```
$`4`
                                                 Cla/Mod   Mod/Cla Global      p.value    v.test
Payment_type=f.PayType-Credit card                8.847403 72.909699  49.28 1.216693e-17  8.551338
RateCodeID=Others                                17.482517  8.361204   2.86 1.043849e-06  4.883187
Trip_type=f.TripType-Dispatch                    16.260163  6.688963   2.46 4.211775e-05  4.095544
f.MTA_tax=f.MTA_tax_NO                           15.671642  7.023411   2.68 4.779382e-05  4.066161
f.Improvement_surcharge=f.Improvement_surcharge_NO 14.492754  6.688963   2.76 2.297538e-04  3.683832
f.Extra=f.Extra-1                                 4.260652 11.371237  15.96 2.134395e-02 -2.301844
f.Improvement_surcharge=f.Improvement_surcharge_YES  5.738379 93.311037  97.24 2.297538e-04 -3.683832
f.MTA_tax=f.MTA_tax_YES                           5.713111 92.976589  97.32 4.779382e-05 -4.066161
Trip_type=f.TripType-Street-Hail                  5.720730 93.311037  97.54 4.211775e-05 -4.095544
RateCodeID=Standard rate                          5.641342 91.638796  97.14 1.043849e-06 -4.883187
Payment_type=f.PayType-Cash                       3.071400 25.752508  50.14 8.539854e-19 -8.852737

$`5`
                                 Cla/Mod   Mod/Cla Global     p.value   v.test
lpep_pickup_date=2016-01-01     1.3215859  50.00000   4.54 0.001725940 3.133739
lpep_pickup_period=Period night 0.2752294 100.00000  43.60 0.006842763 2.704401
lpep_pickup_date=2016-01-30     0.8403361  33.33333   4.76 0.031721963 2.147897

$`6`
                                   Cla/Mod   Mod/Cla Global      p.value    v.test
Payment_type=f.PayType-Credit card 2.3944805 72.839506  49.28 1.580726e-05  4.317128
RateCodeID=Others                  6.9930070 12.345679   2.86 1.039172e-04  3.881259
f.Extra=f.Extra-0                  2.1202942 60.493827  46.22 9.859179e-03  2.580730
lpep_pickup_date=2016-01-04        4.0816327  7.407407   2.94 4.036183e-02  2.050027
lpep_pickup_date=2016-01-31        3.7234043  8.641975   3.76 4.227456e-02  2.030807
lpep_pickup_period=Period valley   2.2445820 35.802469  25.84 4.573573e-02  1.997824
f.Extra=f.Extra-0.5                1.1105235 25.925926  37.82 2.397750e-02 -2.257489
RateCodeID=Standard rate           1.4618077 87.654321  97.14 1.039172e-04 -3.881259
Payment_type=f.PayType-Cash        0.8775429 27.160494  50.14 2.509186e-05 -4.213972
```

Cluster 1

Por lo que hace a las categorías pertenecientes a las variables de tipo factor pertenecientes al cluster 1, y dado el tratamiento que hemos aplicado a nuestros datos, podemos observar como las dos categorías más características de este serían, por un lado el pago mediante efectivo. Este pago en efectivo tiene una representación global de un 50.15% cuando en el cluster 1 esta es de 56.78% por lo que podemos decir que esta categoría aparece sobrerrepresentada en este cluster en concreto. Además, vemos como un 70.63% de todas las observaciones donde el pago es en efectivo, aparecen en el cluster 1, esto también podría justificarse por el gran número de observaciones pertenecientes a este cluster en comparación a los demás.

Debido a esta sobrerrepresentación, podemos ver que, por el contrario, el pago con tarjeta está infrarrepresentado pasando de una representación global del 49.29% al 42.64% dentro de este cluster. Y además, tenemos que del total de observaciones donde el pago se realiza mediante tarjeta de crédito, cerca de un 54% están en este cluster.

No solo son estas las categorías caracterizadas por este cluster ya que todas las que aparecen en el output de res.hcpc *desc.var* category, aun así, son las más destacables.

Cluster 2

Para el cluster 2, la categoría que podemos destacar debido a su sobrerrepresentación és la del VendorID = Verizone. Esta categoría representa un 77.95% de las observaciones globales cuando en este cluster número 2 estas observaciones ascienden a un 94.76%. Por el contrario, seguramente debido a las pocas observaciones pertenecientes a este cluster, las

pertenecientes al cluster 2 con VendorID = Verizone solo representan cerca del 7% de las observaciones totales con este VendorID.

Por otro lado, como categoría infrarepresentada, como es obvio ya que se trata de un factor binario, tendríamos la perteneciente a las observaciones donde VendorID = Mobile que representan el complementario en cuanto a observaciones globales, un 22.05% así como a las pertenecientes al propio cluster, un 5.24%. Además podemos ver como las observaciones de esta categoría pertenecientes a este cluster, solo suponen un 1.36% de las observaciones totales.

Cluster 3

A diferencia del cluster 1, esta vez tenemos sobrerrepresentada la categoría de pago mediante tarjeta de crédito. Esta presenta como ya hemos dicho un 49.29% de las observaciones globales y, en cambio, en este cluster ascienden hasta un 60.66% suponiendo un 29.8% las observaciones pertenecientes a este cluster del total de observaciones donde la tarjeta de crédito aparece como método de pago.

Por otro lado y como podemos suponer, tendremos como categoría infrarrepresentada la que denomina los pagos en efectivo de un 38.84% de observaciones dentro del cluster frente a un 50.15% de las observaciones globales. Finalmente mencionar que estas observaciones suponen un 18.75% de las observaciones totales de nuestra muestra donde los pagos son en efectivo.

Cluster 4

Debido al pequeño número de observaciones que tiene el cluster 4, cualquier categoría podría llegar a considerarse caracterizada. En este caso tendríamos la fecha de recogida 1/1/2016 que sufre una sobrerrepresentación del 50% en este cluster frente al 4.54% de representación global que tiene. Las observaciones de esta fecha en el cluster 4 suponen un 1,32% de sus observaciones totales.

Después tendríamos el periodo de recogida de noche que también aparece sobrerrepresentado en este cluster de un 43.61% de observaciones en las que aparece a un 100% dentro de este cluster número 4, por lo que todas las observaciones dentro de este cluster tendrán como período de recogida, que este ha sido por la noche. Debido a que este periodo supone unas 2180 observaciones de nuestra muestra, por mucho porcentaje de estas que aparezca en el cluster, el número de observaciones que lo componen es mínimo, por lo que solo acaba representando un 0.84% de las observaciones donde este período es la noche.

Y para terminar la fecha 30/1/2016 aparece también sobrerrepresentada aumentando de un 4.76% de representación global hasta un 33.3% dentro de este cluster. Vemos también como la participación de esta fecha en el cluster número 4 sólo supone un 0.84% del total de observaciones donde la fecha es la indicada.

Cluster 5

Para el cluster 5, las categorías más caracterizadas serían la de pago mediante tarjeta y mediante efectivo de nuevo. El pago con tarjeta está sobrerrepresentado en este cluster(73.06%) respecto al porcentaje de observaciones totales de la misma categoría(49.29%). Además, las observaciones con la categoría Tarjeta de crédito dentro del cluster 5 suponen un 8.81% de las observaciones totales del tipo de pago Credit card. Y, por otro lado, el pago en efectivo se ve infrarrepresentado pasando de un 50.15% de

observaciones globales en toda nuestra muestra a un 25.93% dentro del cluster, lo que supone un 3.07% de las observaciones globales donde Tipo de pago es efectivo.

Cluster 6

Y para finalizar con el cluster 6 no tenemos dos/tres categorías que destaquen sobre el resto de las que nos presenta el desc.var$category. Vemos como el pago por tarjeta, el RateCodeID = Otros, el Extra 0, las fechas 4 y 31 de enero de 2016 así como el periodo de recogida del mediodía se ven sobrerrepresentadas dentro de este último cluster. Y, por otro lado, el Extra = 0.5, la tarifa estándar y el tipo de pago en efectivo están infrarrepresentadas dentro de este cluster número 6.

Quantitative variables which characterizes the clusters

```
##                      Eta2       P-value
## Passenger_count   0.798826768  0.000000e+00
## tlenkm            0.701244237  0.000000e+00
## Fare_amount       0.669365956  0.000000e+00
## espeed            0.316696185  0.000000e+00
## Tip_amount        0.272609209  0.000000e+00
## Tolls_amount      0.990410306  0.000000e+00
## traveltime        0.773024273  0.000000e+00
## distHaversine     0.677446046  0.000000e+00
## Total_amount      0.677248346  0.000000e+00
## lpep_pickup_time  0.241495539 2.011289e-296
## Dropoff_longitude 0.015093868  5.818893e-15
## Dropoff_latitude  0.012869428  1.289390e-12
## Pickup_latitude   0.010550394  3.388122e-10
## Pickup_longitude  0.003714435  2.291007e-03
```

En el output de quanti.var podemos observar que dentro de las variables numéricas más asociadas a la muestra globalmente, las que predominan serían:

Passenger_count

Tip_amount

tlenkm Tolls_amount

Fare_amount

Traveltime

espeed

distHaversine

Total_amount

lpep_pickup_time

Description of each cluster by the quantitative variables

$`1`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dropoff_latitude	6.594838	40.7486743	40.74449280	0.0566303	0.05767784	4.257204e-11
Pickup_latitude	5.278030	40.7497472	40.74645114	0.0559643	0.05680712	1.305799e-07
Tolls_amount	-11.891577	0.0000000	0.09183507	0.0000000	0.70251226	1.309107e-32
Passenger_count	-22.260872	1.1154957	1.37460000	0.3826614	1.05880822	8.850835e-110
traveltime	-26.744977	8.0004720	13.00095517	4.4246451	17.00805091	1.412584e-157
Tip_amount	-27.556518	0.6337164	1.23015723	0.9610917	1.96891907	3.697028e-167
espeed	-30.846707	17.8086743	20.97100764	5.4707493	9.32574766	6.201323e-209
Total_amount	-45.656047	9.3705000	14.54115814	3.4276610	10.30225573	0.000000e+00
distHaversine	-45.769105	1.6903393	3.21086535	0.8923275	3.02208211	0.000000e+00
tlenkm	-45.784707	2.2503168	4.58146198	1.1219635	4.63162829	0.000000e+00
Fare_amount	-46.386473	7.5429276	11.99993894	2.8486735	8.74051843	0.000000e+00

$`2`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Passenger_count	63.023610	5.2062937	1.37460000	0.59961722	1.05880822	0.000000000
Pickup_latitude	-2.269941	40.7390468	40.74645114	0.04718364	0.05680712	0.023211182
Tolls_amount	-2.276588	0.0000000	0.09183507	0.00000000	0.70251226	0.022810838
Total_amount	-2.311406	13.1738112	14.54115814	7.61900952	10.30225573	0.020810452
tlenkm	-2.387720	3.9464421	4.58146198	3.11639710	4.63162829	0.016953252
distHaversine	-2.470950	2.7820795	3.21086535	2.00794747	3.02208211	0.013475468
Tip_amount	-2.839549	0.9091259	1.23015723	1.42164079	1.96891907	0.004517734

$`3`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Fare_amount	24.305876	17.314863749	11.99993894	5.64395224	8.74051843	1.699059e-130
distHaversine	23.075943	4.955539729	3.21086535	1.75914169	3.02208211	8.076955e-118
Total_amount	22.699730	20.391775392	14.54115814	6.23967100	10.30225573	4.507511e-114
tlenkm	22.069923	7.138770810	4.58146198	2.28777504	4.63162829	6.149616e-108
espeed	15.722019	24.639103159	20.97100764	9.41146415	9.32574766	1.068635e-55
traveltime	14.118847	19.008574895	13.00095517	8.36123658	17.00805091	2.907098e-45
Tip_amount	13.046692	1.872810834	1.23015723	2.01981565	1.96891907	6.636683e-39
Pickup_longitude	-3.756357	-73.939864649	-73.93586028	0.04267852	0.04261074	1.724048e-04
Pickup_latitude	-3.938254	40.740854145	40.74645114	0.05809040	0.05680712	8.207682e-05
Dropoff_latitude	-4.694268	40.737719103	40.74449280	0.05783748	0.05767784	2.675637e-06
Tolls_amount	-4.958860	0.004681657	0.09183507	0.10449893	0.70251226	7.090812e-07
Dropoff_longitude	-6.669336	-73.943347103	-73.93489420	0.05234703	0.05066108	2.569625e-11
Passenger_count	-8.592768	1.146985962	1.37460000	0.43564971	1.05880822	8.489897e-18

$`4`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
tlenkm	46.112918	16.559185	4.581462	6.79891777	4.63162829	0.000000e+00
distHaversine	44.253180	10.710994	3.210865	3.78962368	3.02208211	0.000000e+00
Fare_amount	43.840623	33.489714	11.999939	10.72037383	8.74051843	0.000000e+00
Total_amount	43.009015	39.390208	14.541158	13.15926915	10.30225573	0.000000e+00
espeed	29.466785	36.382134	20.971008	14.56778400	9.32574766	7.673989e-191
Tip_amount	28.593030	4.387386	1.230157	3.88197499	1.96891907	8.202595e-180
traveltime	19.402917	31.508105	13.000955	17.57953510	17.00805091	7.291369e-84
Dropoff_longitude	5.704871	-73.918686	-73.934894	0.09120726	0.05066108	1.164315e-08
lpep_pickup_time	-3.121624	12.265912	13.633786	6.97177044	7.81353684	1.798568e-03
Pickup_latitude	-3.472807	40.735387	40.746451	0.06566803	0.05680712	5.150458e-04
Dropoff_latitude	-4.479670	40.730003	40.744493	0.06888997	0.05767784	7.475845e-06

$`5`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
traveltime	55.286972	396.69444	13.00096	104.81932037	17.00805091	0.000000e+00
lpep_pickup_time	34.472241	123.54051	13.63379	31.04230913	7.81353684	2.091014e-260
Dropoff_latitude	2.302676	40.79869	40.74449	0.04491073	0.05767784	2.129710e-02
espeed	-5.510714	0.00100	20.97101	0.00000000	9.32574766	3.573815e-08

$`6`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Tolls_amount	70.362620	5.540000	0.09183507	0.00000000	0.70251226	0.000000e+00
Total_amount	21.758712	39.248148	14.54115814	15.31579586	10.30225573	5.713080e-105
distHaversine	18.784138	9.467659	3.21086535	6.11971400	3.02208211	1.018264e-78
tlenkm	18.653711	14.104012	4.58146198	7.96807766	4.63162829	1.178083e-77
Fare_amount	16.796420	28.181047	11.99993894	12.83914942	8.74051843	2.592256e-63
espeed	13.854973	35.212112	20.97100764	14.02499384	9.32574766	1.187017e-43
Tip_amount	13.187988	4.092099	1.23015723	3.83249289	1.96891907	1.028943e-39
traveltime	5.944239	24.144033	13.00095517	9.11695699	17.00805091	2.777437e-09
Pickup_latitude	3.273619	40.766948	40.74645114	0.04679207	0.05680712	1.061797e-03
Dropoff_longitude	2.600295	-73.920375	-73.93489420	0.06862972	0.05066108	9.314357e-03
lpep_pickup_time	-2.212494	11.728395	13.63378632	5.85422153	7.81353684	2.693253e-02

Cluster 1 Las variables más asociadas significativamente a este cluster son tlenkm, Fare_amount, distHaversine y Total_amount con medias inferiores en el propio cluster que las suyas en el global de las observaciones además de unas desviaciones inferiores también dentro del cluster que en general. Como ejemplo podemos considerar la variable de tlenkm cuya media global es de 4.56 vs la media dentro del cluster 1 que sería de 2.25, no solo significativa porque aparece en la lista sino que, observando la desviación estándar global que es 4.49 y la del provio cluster que sería de 1.12, está diferencia entre las medias de 4.55-2.25 = 2.3 representa más del 50% de la desviación total que sufre la variable en todas sus observaciones.

Cluster 2 La variable más asociada a este cluster es el número de pasajeros, también podríamos considerar el importe de propinas o incluso el distHaversine pero solo comentaremos la más significativa. Su media global se situa en 1.37 por lo que vemos que predominan los viajes con un solo pasajero, en cambio, en este cluster vemos que la media asciende hasta los 5.21 por lo que podemos considerar que si no son todas, la mayor parte de observaciones donde los pasajeros sean 6 y 5 estarán en este cluster.

Cluster 3 Las variables más asociadas al cluster 3 serían el Fare_amount seguida del distHaversine, el importe total y la distancia en km del trayecto. Observando la que nos indica el p valor que estaría más asociada(Fare_amount) podemos ver como existe una diferencia considerable entre la media global de sus observaciones(11.98) y la media dentro del cluster(17.32), de esto podemos deducir que seguramente el total_amount también tendrá una media mayor en este cluster y, por lo que podemos observar en el output del quanti para este cluster, esto es así y por tanto, podemos afirmar que los trayectos pertenecientes a este cluster son más caros que la media.

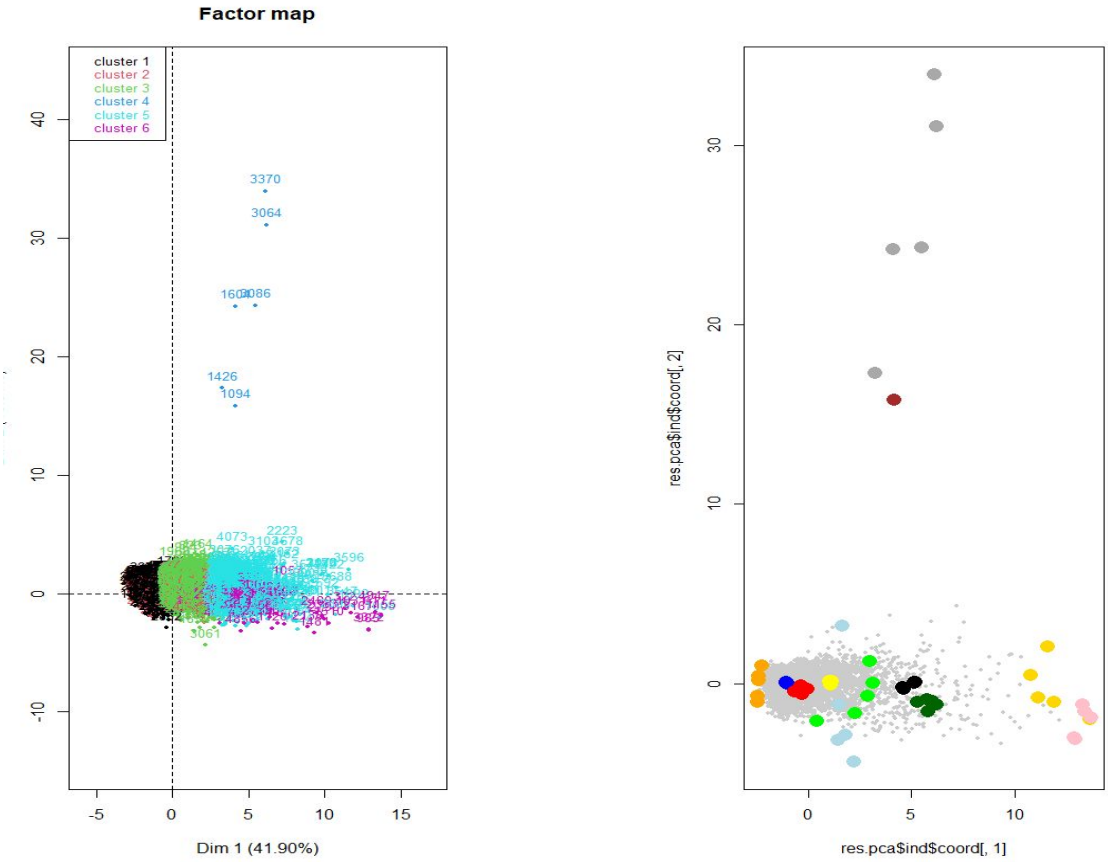
Cluster 4 Para este cluster, las variables más asociadas son la del tiempo de viaje y la de la hora de recogida. Destacar sobretodo la diferencia entre la media global de 12.97 y la media dentro del cluster de 396.69 del tiempo de trayecto, lo que nos muestra que los trayectos dentro de este cluster son los que a primera vista llevarán más tiempo de todos los demás.

Cluster 5 Las variables numéricas más asociadas a este cluster serían la distancia del trayecto en km, el distHaversine, el Fare_amount y el total_amount, seguidas por la velocidad efectiva y el importe de las propinas. Todas ellas con una media superior dentro del cluster respecto a su media global.

Cluster 6 La variable numérica más asociada dentro de este cluster sería la de Tolls_amount seguida del importe total. La media de la primera es de 0.092 mientras que dentro del cluster es de 5.54 siendo este el valor máximo de la misma, lo que da a entender que todas sus observaciones tendrán ese valor para dicha variable lo que da a entender que sean las mismas rutas, o almenos pasen por los mismos peajes durante todo el trayecto.

Description of the clusters by the individuals

Characteristic individuals



En el gráfico anterior podemos observar los individuos más representativos de cada uno de los clusters en color azul a la vez que sus más distantes en color naranja.

Para el cluster número 1 vemos como sus individuos más representativos o parangones son los situados a la izquierda del 0 en el eje de las x que corresponde con donde están la mayoría de los individuos. Por otro lado vemos como los elementos más distintivos son los situados a la izquierda del todo del gráfico, lo más separado al centro de gravedad del propio cluster. Podemos ver que sus parangones se identifican por pertenecer al mismo VendorID(Verifone), pagar todos en efectivo, no tener como cierto el Store_and_fwd,por tanto tener conexión para guardar el trayecto, pagar la tarifa estandard con un extra de 0 y pagar la tasa MTA a la vez que el sobrecargo de mejora. Por tanto podremos afirmar que la mayoría de los individuos seguirán estas características. Referente a los individuos distintivos vemos como cada uno de ellos muestra alguna o bién varias diferencias con las características mencionadas por lo que resultan distintos a esta mayoría representativa.

Los individuos más representativos del cluster 2 están en la misma nube de puntos mencionada anteriormente para los pertenecientes al cluster 1 pero pintados de rojo, hecho que tiene sentido por la proximidad de este cluster al anterior y su centro de gravedad debido a su dispersión entre los clusters 1 y 3. En cambio, sus individuos más distintivos se sitúan sobre x= 2.5 formando un cuarto de círculo rodeando el extremo derecho de este cluster que parece coincidir con el 3 coloreados de color verde. Para este cluster vemos que las variables con observaciones semejantes varían sutilmente respecto al cluster 1.Estas serían el tipo de VendorID, que es el mismo que para el cluster 1, al Store_and_fwd, también falso, el pago de la tarifa estandard así como el pago de la tasa MTA y el improvement surcharge. Además tenemos como el Tolls_amount en todos es 0 y todos pertenecen al tipo de trayecto Street-Hail. Para los individuos distintivos vemos como predominan los 6 pasajeros en lugar de 5 de los parangones así como que se efectúan más pagos en efectivo que en tarjeta.

Los parangones del cluster 3 aparecen como puntos amarillos a la derecha de la nube de puntos azules y rojos que mencionamos coincidiendo como es lógico con el centro de gravedad de dicho cluster. Sus individuos más distintivos son situados por encima y debajo junto a la nube gris que contiene todos los individuos a la altura de las x indicada por sus individuos más representativos en este caso pintados de un azul muy claro. Hecho que caracterizaría este cluster respecto a los anteriores según sus parangones sería que estos muestran observaciones de 1 pasajero por trayecto así como la no predominancia de un tipo concreto de VendorID ni de tipo de pago en específico.

Para el cluster 4 observamos solo un individuo como parangón coloreado de marrón siendo este el punto más cercano al eje de las y del gráfico. Se consideran todos los demás individuos pertenecientes al cluster como individuos distintivos del mismo y podemos verlos de color gris. En este cluster podemos ver trayectos de un solo viajero y nocturnos, de muy larga duración y de importe elevado como más característicos además observamos velocidades efectivas suficientemente altas como para considerar que son trayectos fuera de la ciudad

El cluster 5 tiene representados sus individuos más representativos de color negro, situados donde podemos observar una mayor cantidad de ellos y, de color dorado, al otro extremo(derecho) sus puntos más distintivos, los más lejanos a su centro de gravedad y por tanto, serán menos representativos del cluster. En este caso observamos trayectos más cortos y con velocidades efectivas más controladas suponiendo que se trata de trayectos dentro de la propia ciudad donde también predominan los trayectos nocturnos e individuales.

Para el cluster 6 vemos como los puntos coloreados de rosa, los situados más a la derecha del gráfico, aparecen como sus individuos más distintivos al situarse estos muy alejados de la cantidad principal de estos que forman el cluster. Sus parangones, por otro lado, podemos verlos en verde oscuro cerca del centro de gravedad del cluster 5. En este caso no vemos que predominen absolutamente los trayectos individuales, aún así podemos asumir que son trayectos más situados por el centro de la ciudad debido a las distancias recorridas, su velocidad efectiva así como el tiempo empleado donde estos se distribuyen entre el periodo del medio día y por la noche.

Hierarchical tree result

Ratio between within inertias

```
res.hcpc$call$t$quot[1:res.hcpc$call$t$nb.clust]
```

```
## [1] 0.7999712 0.7981521 0.7760833 0.7358976 0.8974407 0.9018483
```

Si vemos la relación entre inercias dentro de un mismo cluster vemos que es bastante grande, es decir la inercia intra cluster es elevada.

Inertia gain

```
res.hcpc$call$t$inert.gain[1:res.hcpc$call$t$nb.clust]
```

```
## [1] 1.9240191 1.0252503 0.8276295 0.7327972 0.6707766 0.1916894
```

Si vemos la inertia gain vemos que es bastante alta des del cluster 1 al 5. Si vemos la pérdida entre pasar de 5 cluster a 6 cluster vemos que es bastante bajo con un valor de 0.1916, lo que significa que seguramente la elección de un hierarchical tree con más de 7 cluster no tendría sentido.

Partition quality

```
(res.hcpc$call$t$within[1]-res.hcpc$call$t$within[res.hcpc$call$t$nb.clust])/res.hcpc$call$t$within[1]
```

```
## [1] 0.7348677
```

Podemos ver como confirmando que el número de clusters óptimo es 6 obtenido con la llamada a “res.hcpc$call$t$nb.clust” tenemos una calidad de esta división en 6 clusters del 73.49%.

K-Means: Partitioning in k=6

Profiling KM

```
K-means clustering with 6 clusters of sizes 2374, 81, 991, 286, 250, 1018

##Within cluster sum of squares by cluster:

##[1] 1406.1036 1013.3541  811.5994  788.6252 1590.3512 5559.9415

 ##(between_SS / total_SS =  68.3 %)
```

Para explicar la clasificación obtenida con el Kmeans hemos decidido coger 6 clusters. Este número de clusters nos deja una explicación del 68.3% con las medidas que podemos ver en la primera línea del output.

Global association variables numeric

```
##                         Eta2       P-value
## Passenger_count   0.444694791  0.000000e+00
## tlenkm            0.603831600  0.000000e+00
## Fare_amount       0.576696172  0.000000e+00
## espeed            0.263490766  0.000000e+00
## Tolls_amount      0.619167093  0.000000e+00
## distHaversine     0.564957345  0.000000e+00
## Total_amount      0.595040329  0.000000e+00
## traveltime        0.233684977 2.464593e-285
## lpep_pickup_time  0.230551028 6.439587e-281
## Tip_amount        0.222218401 2.920709e-269
## Dropoff_longitude 0.021909039  2.966999e-22
## Dropoff_latitude  0.008728726  2.548210e-08
## Pickup_latitude   0.006045956  1.290650e-05
## Pickup_longitude  0.004354320  5.715964e-04
```

Como podemos ver según el p valor de cada una de las variables numéricas, las que están más asociadas globalmente a la muestra son las que hemos marcado en azul, las cuales coinciden con las que habíamos encontrado mediante el hierarchical clustering aunque no con el mismo p valor lo que significa que estas asociaciones han variado.

```
## $`1`
##                      v.test Mean in category Overall mean sd in category
## tlenkm           -14.494179        2.1488108   4.58146198     1.14987360
## distHaversine    -14.623896        1.6093841   3.21086535     0.88623093
## Total_amount     -16.052113        8.5485336  14.54115814     3.66900369
## Fare_amount      -16.658218        6.7237753  11.99993894     2.49596451
## lpep_pickup_time -31.568368        4.6955400  13.63378632     5.06942330
##                  Overall sd       p.value
## tlenkm            4.63162829  1.318696e-47
## distHaversine     3.02208211  1.977536e-48
## Total_amount     10.30225573  5.525040e-58
## Fare_amount       8.74051843  2.637671e-62
## lpep_pickup_time  7.81353684 1.003687e-218
##
## $`2`
```

```
##                    v.test Mean in category Overall mean sd in category
## espeed          -28.488256       16.6990429  20.97100764     4.78955497
## Total_amount    -33.761969        8.9482429  14.54115814     2.82948643
## Fare_amount     -34.220223        7.1904585  11.99993894     2.37583417
## tlenkm          -34.234381        2.0318487   4.58146198     0.90988893
## distHaversine   -34.790326        1.5202574   3.21086535     0.72944165
##                  Overall sd       p.value
## espeed           9.32574766 1.637649e-178
## Total_amount    10.30225573 7.134769e-250
## Fare_amount      8.74051843 1.209845e-256
## tlenkm           4.63162829 7.448958e-257
## distHaversine    3.02208211 3.406549e-265
##
## $`3`
##                    v.test Mean in category Overall mean sd in category
## Passenger_count  46.567793        3.5220126   1.37460000     2.10570696
## Fare_amount      22.254258       20.4714885  11.99993894    13.07651024
## distHaversine    21.227733        6.0048397   3.21086535     4.36243875
## Total_amount     21.207526       24.0567296  14.54115814    14.81389736
## tlenkm           21.203364        8.8585778   4.58146198     6.44443468
## espeed           15.801924       27.3891029  20.97100764    14.77674556
##                  Overall sd       p.value
## Passenger_count    1.05880822  0.000000e+00
## Fare_amount        8.74051843 1.025760e-109
## distHaversine      3.02208211 5.295615e-100
## Total_amount      10.30225573 8.138191e-100
## tlenkm             4.63162829 8.890774e-100
## espeed             9.32574766  3.017646e-56
##
## $`4`
##                    v.test Mean in category Overall mean sd in category
## distHaversine     8.368848     3.913570e+00   3.21086535     1.11636717
## Fare_amount       8.027703     1.394947e+01  11.99993894     3.33137049
## Total_amount      7.468417     1.667893e+01  14.54115814     4.16121161
## traveltime        6.496199     1.607079e+01  13.00095517     5.77794647
## Passenger_count  -8.967635     1.110787e+00   1.37460000     0.35730644
##                  Overall sd      p.value
## distHaversine      3.02208211 5.818427e-17
## Fare_amount        8.74051843 9.931486e-16
## Total_amount      10.30225573 8.116542e-14
## traveltime        17.00805091 8.237461e-11
## Passenger_count    1.05880822 3.029493e-19
##
## $`5`
##                    v.test Mean in category Overall mean sd in category
## Tolls_amount     55.633797        3.461877   0.09183507     2.65949782
## tlenkm           34.575806       18.390031   4.58146198    10.96442081
## Total_amount     33.806675       44.572688  14.54115814    18.77537258
## distHaversine    30.937505       11.272706   3.21086535     6.54220495
## Fare_amount      29.587600       34.299155  11.99993894    16.05801588

##                  Overall sd       p.value
## Tolls_amount       0.70251226  0.000000e+00
## tlenkm             4.63162829 5.837763e-262
## Total_amount      10.30225573 1.573501e-250
## distHaversine      3.02208211 3.741411e-210
```

```
## Fare_amount        8.74051843 2.157664e-192

##
## $`6`
##                    v.test Mean in category Overall mean sd in category
## Fare_amount       26.523537        21.613820  11.99993894     5.95924104
## tlenkm            25.381329         9.456502   4.58146198     2.16340894
## distHaversine     25.023467         6.346922   3.21086535     2.06401744
## Total_amount      24.648529        25.071766  14.54115814     6.17563347

##                   Overall sd       p.value
## Fare_amount        8.74051843 5.188538e-155
## tlenkm             4.63162829 4.054420e-142
## distHaversine      3.02208211 3.395953e-138
## Total_amount      10.30225573 3.816230e-134

res.cat$test.chi2

##                            p.value  df
## f.tlenkm                0.000000e+00  10
## f.traveltime            0.000000e+00  20
## f.distHaversine         0.000000e+00  10
## AnyToll                 0.000000e+00   5
## f.Fare_amount           0.000000e+00  15
## f.Passenger_count       0.000000e+00  10
## f.Total_amount          0.000000e+00  35
## hcpck                   0.000000e+00  25
## f.espeed                1.998801e-282  10
## lpep_pickup_period      6.478927e-115  15
## f.Extra                 1.929912e-86  10
## AnyTip                  2.328424e-42   5
## Payment_type            1.311420e-35  10
## lpep_pickup_date        8.734495e-21 150
## RateCodeID              9.513616e-19   5
## multiouts               7.109065e-10   5
## VendorID                2.397221e-05   5
## Trip_type               1.137969e-04   5
## f.MTA_tax               3.822004e-04   5
## f.Improvement_surcharge 2.412695e-03   5
```

Para el cluster 1 vemos que la variable más caracterizada dentro del propio cluster es la hora de recogida con un p valor muy pequeño en comparación a las demás, la cual observamos que tiene una media inferior a la global por lo que asumimos que contendrá trayectos con recogidas temprano.

Referente al cluster 2 tenemos las variables distHaversine, distancia en kilómetros, importe de la tarifa, importe total y velocidad efectiva como más caracterizadas dentro del cluster. Estas tienen una media inferior dentro del cluster que la que les corresponde al total de observaciones por lo que podremos asumir que se trata de un cluster compuesto por trayectos cortos, lentos y baratos, es decir, trayectos urbanos.

Por lo que hace al cluster 3 vemos caracterizadas, sobretodo, las variables del número de pasajeros, el importe de la tarifa y total, el distHaversine así como la distancia en kilómetros y la velocidad efectiva. En este caso estas tienen una media superior dentro

del cluster que en el total de observaciones por lo que podemos inducir que se trata de trayectos interurbanos que en consecuencia de salir del centro pueden circular a mayor velocidad al evitar atascos.

Para el cluster 4 tendríamos caracterizadas las variables distHaversine así como el importe total del trayecto y su tarifa, el número de pasajeros, estos con una media menor dentro del cluster, y finalmente el tiempo de viaje. En este caso no están tan caracterizadas como en los clusters anteriores por lo que suposiciones que hagamos pueden no ser tan acertadas al ser un cluster más equilibrado. En este caso podríamos llegar a asumir que se trata de trayectos de larga distancia que, en consecuencia, derivan en caros.

Referente al cluster 5 tenemos caracterizados el importe por peajes, el de la tarifa y el total, así como las distancias en km, la distHaversine así como el tiempo de trayecto. Todas estas con una media mayor dentro del cluster muy caracterizada por lo que podemos asumir que este cluster también lo formarán principalmente trayectos interurbanos que pasen por carreteras de pago y cuya tarifa tenga un coste mayor.

Y finalmente, para el cluster 6 vemos caracterizadas con una media superior dentro del cluster las variables referentes a la tarifa del trayecto y la total, y la distHaversine y la distancia en kilómetros. Lo que nos llevaría a asumir que se trata de un cluster formado por trayectos largos.

Si analizamos la variables categóricas que contribuye más en la construcción de Hierarchical clustering, vemos que los factores tlenkm, traveltime y distHaversine tienen una chi2 igual a 0 y por lo tanto son las variables que caracterizan los clustering. En la figura marcada en azul podemos ver las variables categóricas que caracterizan los clustering ordenadas según la probabilidad value (p-value).

Global association category categoricas

	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.distHaversine=f.distHaversine-[0,5]	52.9786712	100.00000000	81.58	4.703457e-261	34.515457
f.tlenkm=f.tlenkm-(1,5]	61.1935484	87.78343360	62.00	1.757096e-255	34.142008
f.traveltime=f.traveltime-[0,10]	64.2004773	74.68764461	50.28	3.383813e-206	30.641942
f.espeed=f.espeed-(1,25]	53.1096774	95.23368811	77.50	2.629290e-174	28.146709
f.espeed=f.espeed-(25,130]	9.2046470	4.76631189	22.38	1.382287e-172	-28.005765
f.distHaversine=f.distHaversine-(5,10]	0.0000000	0.00000000	14.54	1.215477e-199	-30.145857
f.Fare_amount=f.Fare_amount-(14.5,71.5]	0.7311129	0.41647385	24.62	0.000000e+00	-Inf
f.tlenkm=f.tlenkm-(5,67.9]	0.3369272	0.23137436	29.68	0.000000e+00	-Inf

Para el cluster 1 vemos que tenemos sobrerrepresentadas las categorías de distancia y tiempo cercanas a 0 e infrarrepresentadas, por otro lado las de tiempo de trayecto prolongado así como su distancia. También vemos sobrerrepresentadas las

velocidades efectivas inferiores a 25 km/h que junto al análisis de las variables numéricas de cada cluster nos podría indicar que se trata de trayectos urbanos que suceden por la mañana.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.Passenger_count=f.Passenger_count-Others	66.1971831	59.1194969	8.52	3.320893e-217	31.457441

Las categorías del cluster 2 solo nos ofrecen información destacable sobre que el número de pasajeros suele ser mayor a 2 debido a la sobrerrepresentación de la misma.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
AnyToll=AnyToll No	96.55172414	64.1221374	1.74	1.316343e-143	25.515807

El cluster 3 nos indica que existe una gran cantidad de las observaciones que no han pasado por ningún peaje. Así que podemos definir que la mayoría de trayectos interurbanos que se producen dentro de este cluster siguen rutas por las que no se suele pasar por carreteras de pago.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.Fare_amount=f.Fare_amount-(14.5,71.5]	39.31762794	93.4362934	24.62	3.235575e-282	35.898399
f.tlenkm=f.tlenkm-(5,67.9]	34.23180593	98.0694981	29.68	1.469359e-280	35.792030
f.distHaversine=f.distHaversine-(5,10]	54.74552957	76.8339768	14.54	7.087502e-270	35.098572
f.Total_amount=f.Total_amount-(20,30]	60.00000000	68.9189189	11.90	5.842961e-251	33.835940

Para el cluster 4 tenemos sobrerrepresentadas las categorías con una tarifa elevada a la vez que trayectos con distancias largas que concuerda con las observaciones anteriores definiendo un cluster de trayectos de largo recorrido.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.Total_amount=f.Total_amount-(11,18]	50.4261364	69.40371457	28.16	4.969131e-217	31.444640
f.Fare_amount=f.Fare_amount-(9,14.5]	50.7923930	62.65884653	25.24	4.679811e-187	29.169849
f.traveltime=f.traveltime-(10,20]	41.3004214	67.05767351	33.22	5.378902e-139	25.096881
f.tlenkm=f.tlenkm-(5,67.9]	38.1401617	55.32746823	29.68	5.573471e-84	19.416723

El cluster 5 por otro lado tiene unos importes totales y tiempos de trayecto más contenidos aún siendo estas distancias en kilómetros considerables por lo que suponemos que se trata de trayectos interurbanos o más bien por las afueras de la ciudad.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.traveltime=f.traveltime-[0,10]	24.1447892	87.9710145	50.28	6.699286e-112	22.478754
lpep_pickup_period=Period night	24.0366972	75.9420290	43.60	3.892522e-77	18.589713

En el cluster 6 vemos sobrerrepresentados los trayectos con un tiempo muy bajo y nocturnos, que junto al análisis anterior nos mostraría como incierta esta suposición y no nos permitiría llegar a una suposición del tipo de trayectos que caracterizan este cluster.

Confusion Table

	kKM-3	kKM-6	kKM-2	kKM-1	kKM-5	kKM-4
kHP- 1	0	661	0	2161	295	0
kHP- 2	0	17	263	0	0	6
kHP- 3	0	12	9	0	728	462
kHP- 4	44	0	205	0	0	50
kHP- 5	6	0	0	0	0	0
kHP- 6	81	0	0	0	0	0

```
> (2161+263+0+50)/(661+2161+295+17+263+6+12+9+728+462+44+205+50+6+81)
[1] 0.4948
```

A partir de la tabla de confusión, podemos ver que muchos individuos del cluster 1 del kKM y del kHP coinciden. Si observamos las características que coinciden entre el primer cluster de ambos métodos, vemos que los dos clusters tienen 3km de media aproximadamente y 7 dólares de Fare Amount. Si vemos el resultado del kHP la variable duration time es la sexta variable más importante y en kKM la variable duration time es la octava variable más importante.

Si nos fijamos, el cluster 5 tiene sólo 6 individuos en contraposición el cluster 5 tiene aproximadamente 1000, lo que significa que no hay ningún tipo de relación.

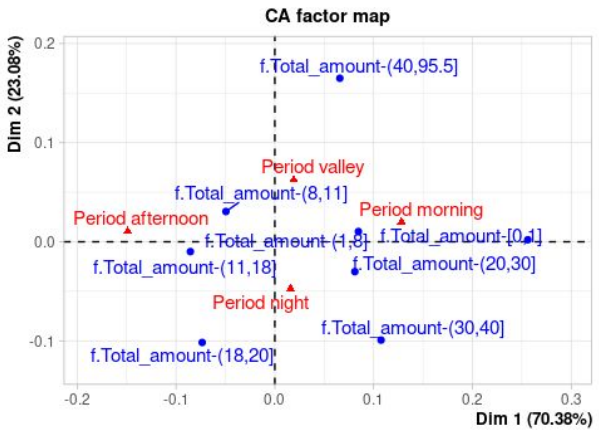
Si comparamos la cantidad de miembros que coincide entre clusters con el mismo índice vemos que solo el 50% de miembros coinciden.

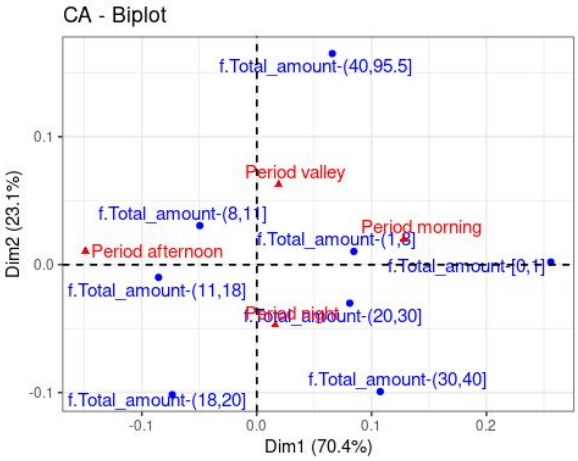
Correspondence Analysis (CA)

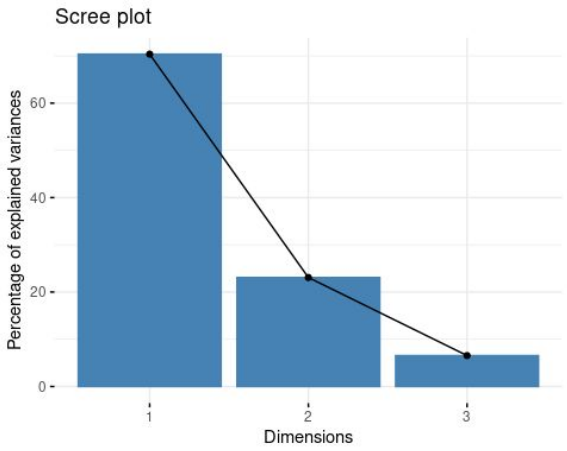
CA in Total amount and Pick up period

```
chisq.test(tt)
## data:  tt
## X-squared = 44.632, df = 21, p-value = 0.001935
```

Como podemos ver el p-value es 0.001935, menor que 0.05, por lo tanto es estadísticamente significativa y podríamos rechazar la hipótesis nula: filas y columnas son independientes. Como podemos ver en la tabla de contingencia hay valores inferiores a 5, por lo que la aproximación de Pearson's Chi-squares test puede ser incorrecta.







Viendo la tabla de contingencias, podemos observar como las categorías de importe total de 40-95.5$, de 30-40$, DE 18-20$ así como los trayectos nocturnos, serían los menos presentes en las observaciones al estar más alejados del centro de gravedad(de los ejes).

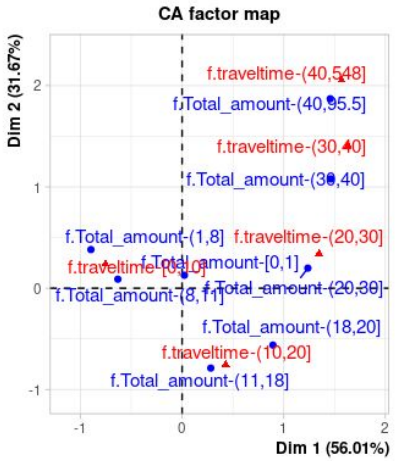
Podemos observar como los trayectos nocturnos están muy relacionados con costar entre 20$ y 30$ y, en cambio, los trayectos que son por la mañana, entre 1$ y 8$.

Como conclusión podemos asumir como falsa la suposición de independencia entre categorías de importe total y periodo de recogida.

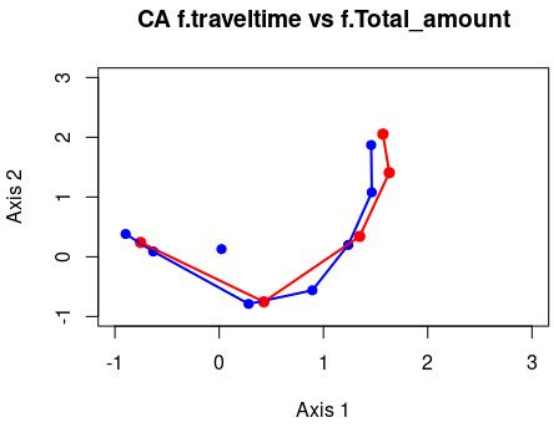
CA in Total amount and Travel time

```
chisq.test(tt)
## X-squared = 6148.5, df = 28, p-value < 2.2e-16
```

Como podemos ver en la tabla de contingencia hay valores inferiores a 5, por lo que la aproximación de Pearson's Chi-squares test puede ser incorrecta. Si observamos el p-value 2.2e-16, es inferior a 0.05 por lo tanto, podríamos rechazar la hipótesis nula.



Si vemos el primer plano factorial vemos como los factores están muy bien representados. Vemos que aquellos viajes que duran entre 40 y 548 minutos tienen un coste entre 40 y 95.5. Los viajes entre 11 y 40 minutos tienen un coste mediano, entre 15 y 50. Y los viajes con coste bajo tienen una duración entre 0 y 15 minutos.



En esta imagen podemos ver perfectamente el efecto Guttman que se forma a partir de aplicar CA.El cual también nos indica la clara relación entre filas y columnas.

```
summary(res.ca)

##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 6148.456 (p-value =  0 ).
##
```

```
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4
## Variance               0.689   0.389   0.144   0.007
## % of var.             56.013  31.666  11.720   0.601
## Cumulative % of var.  56.013  87.679  99.399 100.000
```

Como podemos ver el estadístico de la chi square es igual a 6148.456 lo que significa que tienen una gran relación las dos variables.

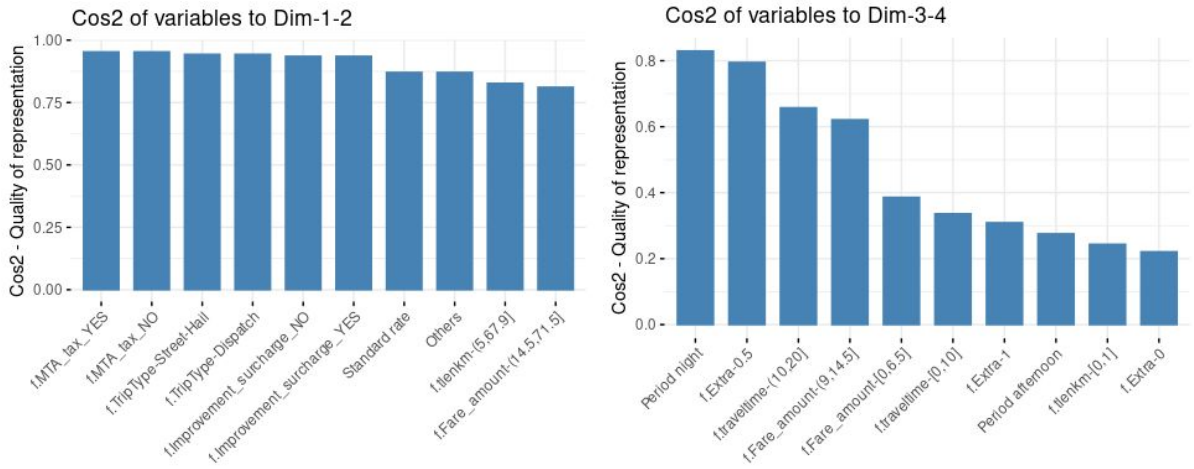
```
mean(res.ca$eig[,1])

## [1] 0.3074228
```

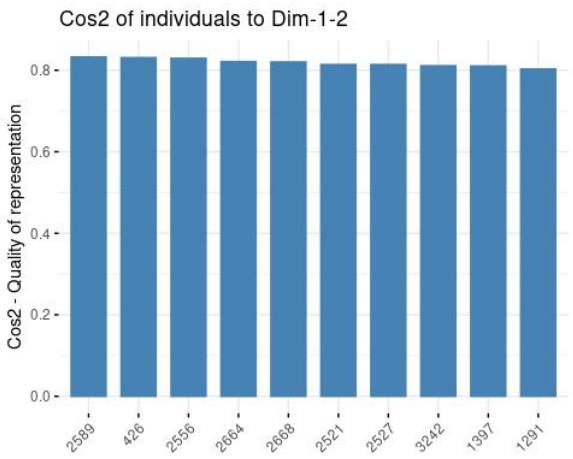
Según el criterio de Kaiser deberíamos coger aquellas dimensiones que tienen un valor propio superior a 0.3074228. Por lo tanto, deberíamos coger hasta la segunda dimensión. Con estas dos dimensiones tendríamos explicado un 87.68% de la muestra.

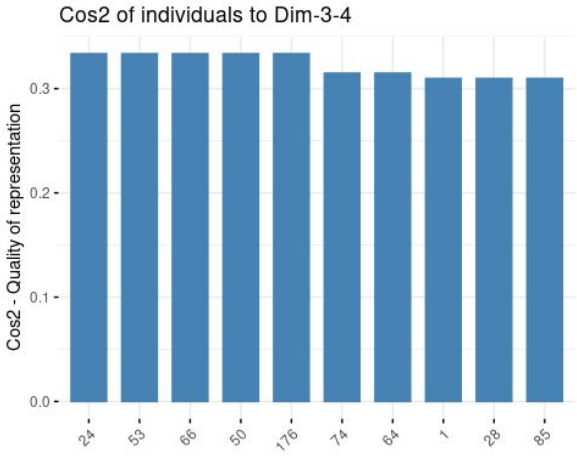
MCA analysis

Quality of representation



Tras realizar el análisis de correspondencias múltiples, podemos ver que la variables MTA_tax, TripType y Improvement_surcharge contribuye en la construcción del primer plano factorial. Podemos ver que los viajes de taxi realizados por la noche son los mejor representados en el segundo plano factorial. También podemos ver que los individuos que han hecho un viaje entre 10 y 20 minutos están bien representados en el segundo plano factorial.



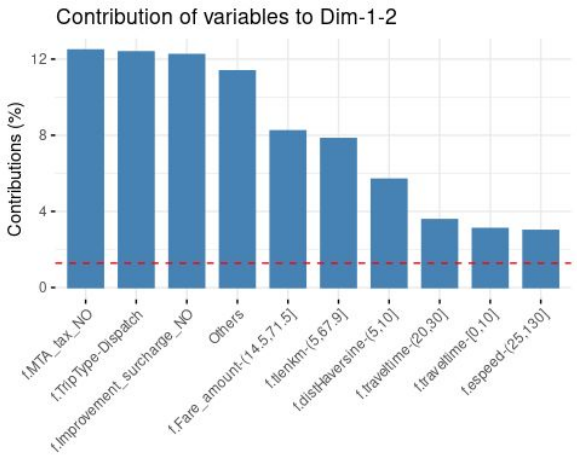


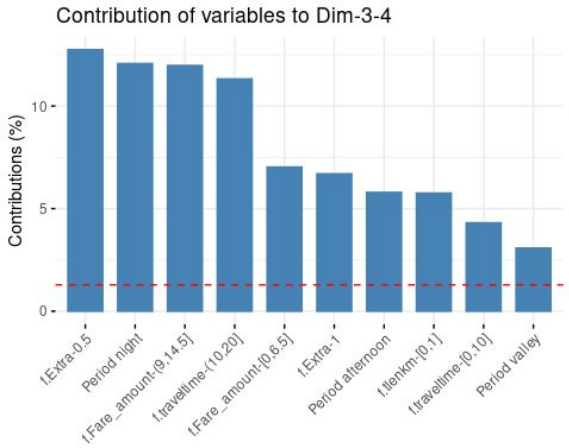
	VendorID	Payment_type	Store_and_fwd_flag	RateCodeID	f.Extra	f.MTA_tax	f.Improvement_surcharge
24	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
53	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
66	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
50	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
176	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
74	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
64	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
1	f.Vendor-VeriFone	f.PayType-Cash	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
28	f.Vendor-VeriFone	f.PayType-Cash	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES
85	f.Vendor-VeriFone	f.PayType-Cash	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES

	lpep_pickup_period	Trip_type	lpep_pickup_date	multiouts	f.espeed	f.tlenkm	f.traveltime
24	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
53	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
66	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
50	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
176	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
74	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(25,130]	f.tlenkm-(5,67.9]	f.traveltime-(10,20]
64	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(25,130]	f.tlenkm-(5,67.9]	f.traveltime-(10,20]
1	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
28	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]
85	Period night	f.TripType-Street-Hail	2016-01-01	FALSE	f.espeed-(1,25]	f.tlenkm-(1,5]	f.traveltime-(10,20]

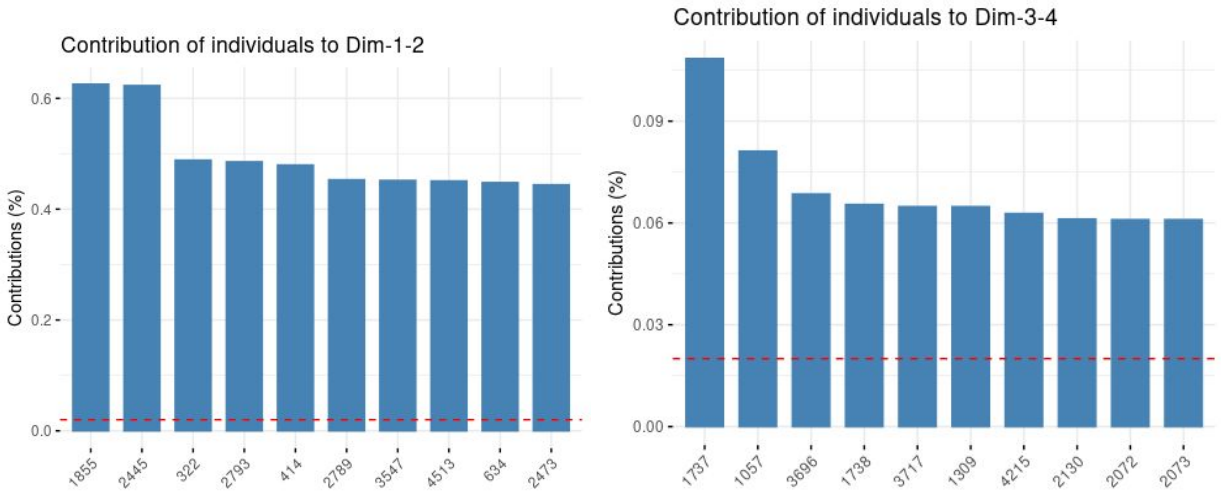
Como podemos ver, los individuos mejor representados en el segundo plano factorial son viajes de taxi realizados por la noche, pagados con extra de 0.5 y con una duración de viaje entre 10 y 20 minutos. Si nos fijamos todas estas características son las mejor representadas en el segundo plano factorial.

Contribution





Como vemos, los individuos que contribuyen más en el primer plano factorial no han pagado la tasa MTA y no han pagado la tarifa improvement surcharge. En contraposición los individuos que contribuyen más han pagado 0.5 de extra y han hecho el viaje por la noche.



	VendorID	Payment_type	Store_and_fwd_flag	RateCodeID	f.Extra	f.MTA_tax	f.Improvement_surcharge
1855	f.Vendor-VeriFone	f.PayType-Credit card	FALSE	Others	f.Extra-0	f.MTA_tax_NO	f.Improvement_surcharge_NO
2445	f.Vendor-VeriFone	f.PayType-Cash	FALSE	Others	f.Extra-0	f.MTA_tax_NO	f.Improvement_surcharge_NO
322	f.Vendor-VeriFone	f.PayType-Cash	FALSE	Others	f.Extra-0	f.MTA_tax_NO	f.Improvement_surcharge_NO
1737	f.Vendor-Mobile	f.PayType-Others	FALSE	Standard rate	f.Extra-0	f.MTA_tax_YES	f.Improvement_surcharge_YES
1057	f.Vendor-VeriFone	f.PayType-Cash	FALSE	Standard rate	f.Extra-1	f.MTA_tax_YES	f.Improvement_surcharge_YES
3696	f.Vendor-Mobile	f.PayType-Credit card	FALSE	Standard rate	f.Extra-0.5	f.MTA_tax_YES	f.Improvement_surcharge_YES

	lpep_pickup_period	Trip_type	lpep_pickup_date	multiouts	f.espeed	f.tlenkm	f.traveltime
1855	Period morning	f.TripType-Dispatch	2016-01-12	TRUE	f.espeed-(25,130]	f.tlenkm-[0,1]	f.traveltime-[0,10]
2445	Period night	f.TripType-Dispatch	2016-01-15	TRUE	f.espeed-(25,130]	f.tlenkm-(5,67.9]	f.traveltime-(40,548]
322	Period afternoon	f.TripType-Dispatch	2016-01-02	FALSE	f.espeed-(25,130]	f.tlenkm-(5,67.9]	f.traveltime-(30,40]
1737	Period valley	f.TripType-Street-Hail	2016-01-11	TRUE	f.espeed-(25,130]	f.tlenkm-(5,67.9]	f.traveltime-(40,548]
1057	Period valley	f.TripType-Street-Hail	2016-01-07	FALSE	f.espeed-(1,25]	f.tlenkm-(5,67.9]	f.traveltime-(40,548]
3696	Period night	f.TripType-Street-Hail	2016-01-23	FALSE	f.espeed-(1,25]	f.tlenkm-[0,1]	f.traveltime-[0,10]

	f.distHaversine	AnyToll	f.Fare_amount	f.Passenger_count	f.Total_amount	Passenger_count
1855	f.distHaversine-[0,5]	AnyToll Yes	f.Fare_amount-(14.5,71.5]	f.Passenger_count-2	f.Total_amount-(40,95.5]	2
2445	f.distHaversine-[0,5]	AnyToll Yes	f.Fare_amount-(9,14.5]	f.Passenger_count-Others	f.Total_amount-(8,11]	3
322	f.distHaversine-(10,26.1]	AnyToll No	f.Fare_amount-(14.5,71.5]	f.Passenger_count-Others	f.Total_amount-(40,95.5]	4
1737	f.distHaversine-(5,10]	AnyToll No	f.Fare_amount-(14.5,71.5]	f.Passenger_count-1	f.Total_amount-(40,95.5]	1
1057	f.distHaversine-(10,26.1]	AnyToll No	f.Fare_amount-(14.5,71.5]	f.Passenger_count-1	f.Total_amount-(40,95.5]	1
3696	f.distHaversine-[0,5]	AnyToll Yes	f.Fare_amount-[0,6.5]	f.Passenger_count-Others	f.Total_amount-(1,8]	3

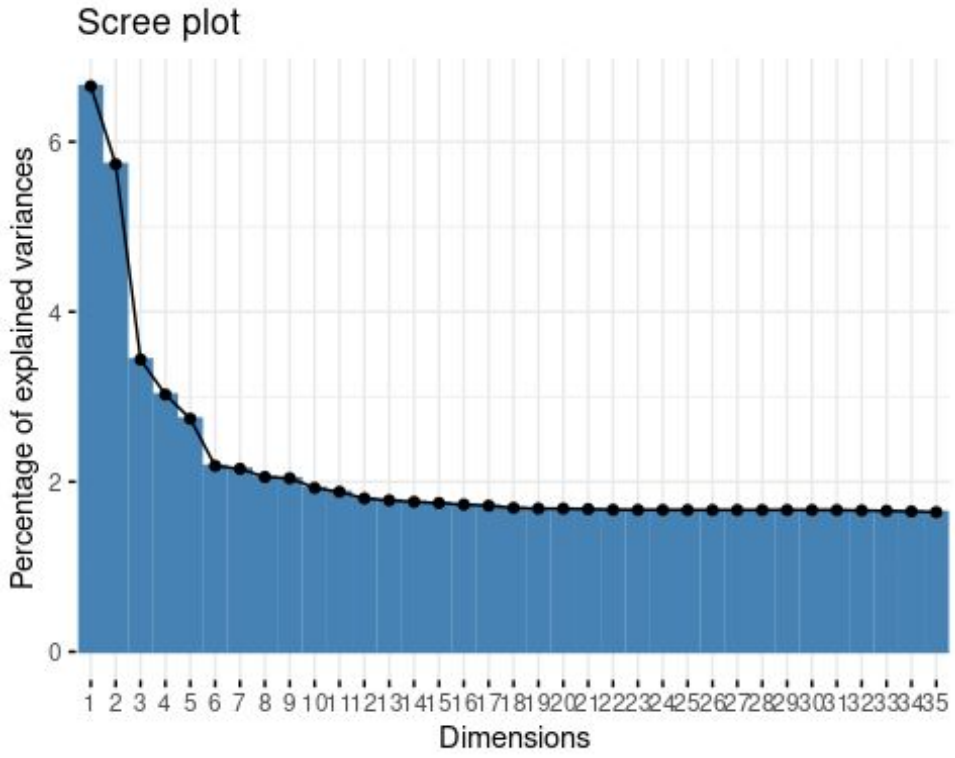
	tlenkm	Pickup_longitude	Pickup_latitude	Dropoff_longitude	Dropoff_latitude	Fare_amount	espeed	Tip_amount	Tolls_amount
1855	0.0804672	-74.08338	40.64220	-74.08366	40.64284	65.00000	48.28032	13.00	0.00000
2445	56.0856400	-74.00069	40.59997	-73.96132	40.62992	10.00000	40.04528	0.00	0.00000
322	32.7018701	-73.91185	40.82726	-73.78302	40.64885	57.66478	50.33208	0.00	5.54000
1737	67.9143168	-73.86628	40.87276	-73.86414	40.82361	61.50166	37.35547	0.00	0.22224
1057	22.0641062	-73.86156	40.73050	-73.95473	40.81818	51.00000	20.73369	0.00	5.54000
3696	0.9656064	-73.96154	40.71416	-73.95172	40.71512	5.00000	10.89713	1.25	0.00000

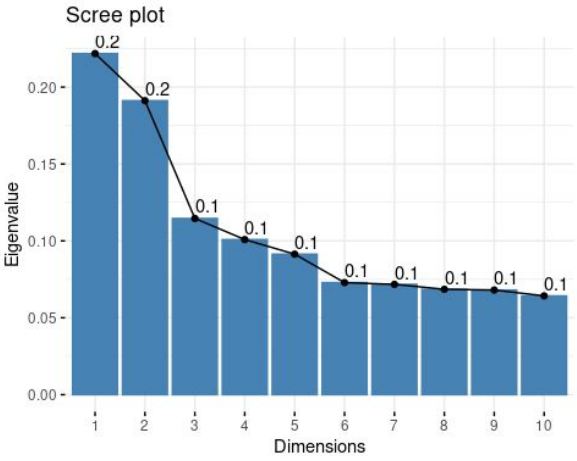
	lpep_pickup_time	traveltime	distHaversine	AnyTip	Total_amount	hcpck	claKM
1855	9	0.100000	0.07562334	AnyTip Yes	78.00000	kHP- 4	kKM-1
2445	20	84.033330	4.70946300	AnyTip No	10.00000	kHP- 4	kKM-3
322	18	38.983333	22.63957168	AnyTip No	95.54000	kHP- 6	kKM-3
1737	15	109.083333	5.47499787	AnyTip No	70.61709	kHP- 4	kKM-3
1057	16	63.850000	12.52834389	AnyTip No	58.34000	kHP- 6	kKM-3
3696	0	5.316667	0.83537860	AnyTip Yes	7.55000	kHP- 1	kKM-4

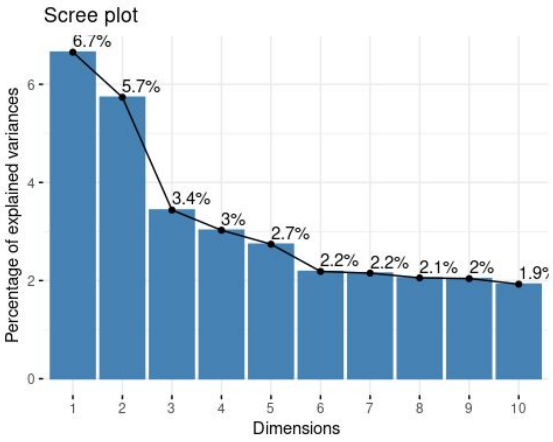
Como podemos ver los 3 primeros individuos mejor representados en el primer plano factorial no han pagado la tasa MTA y no han pagado la tarifa improvement surcharge. Estas características no se encuentran en los individuos del segundo plano factorial. Los individuos del segundo plano factorial son viajes realizados en Period valley y Period night. También podemos ver que en el segundo plano factorial encontramos que los tres individuos que contribuyen más en el segundo plano factorial han pagado

diferentes extra, por lo que podríamos decir que en el segundo plano factorial se divide los individuos según el tipo de extra pagado y el periodo en el que se realizó el viaje.

Eigenvalues and dominant axes analysis.







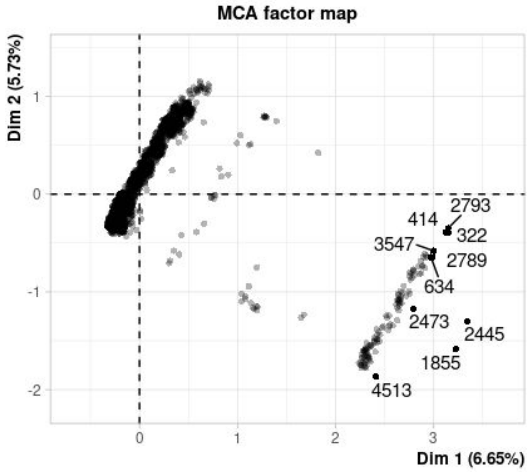

```
mean(res.mca$eig[,1])
```

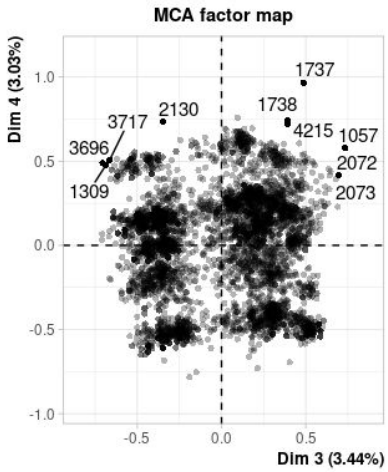
```
## [1] 0.05555556
```

Según el criterio de Kaiser generalizado deberíamos coger aquellas dimensiones cuyo valor propio sea superior a 0.055. Lo que significa que tendríamos que coger hasta la dimensión 35, esta dimensión proporciona un 74.4% de varianza acumulada. En

contraposición, si aplicamos la regla de Elbow deberíamos coger hasta la dimensión 6 pero solo nos proporciona un 23.78% de varianza acumulada.

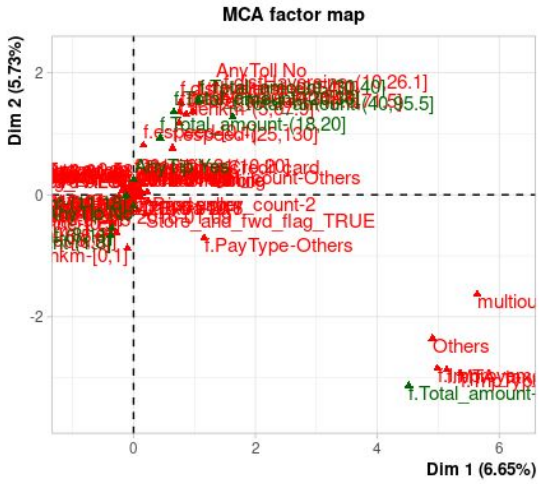
Individuals

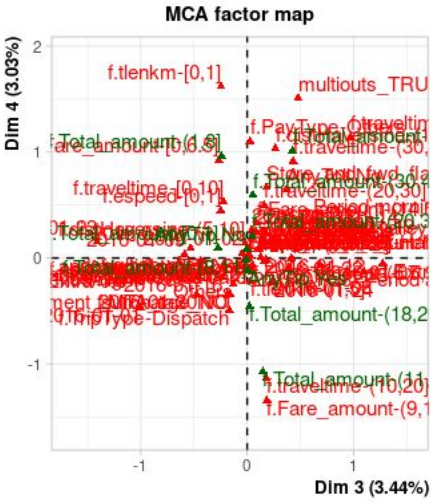


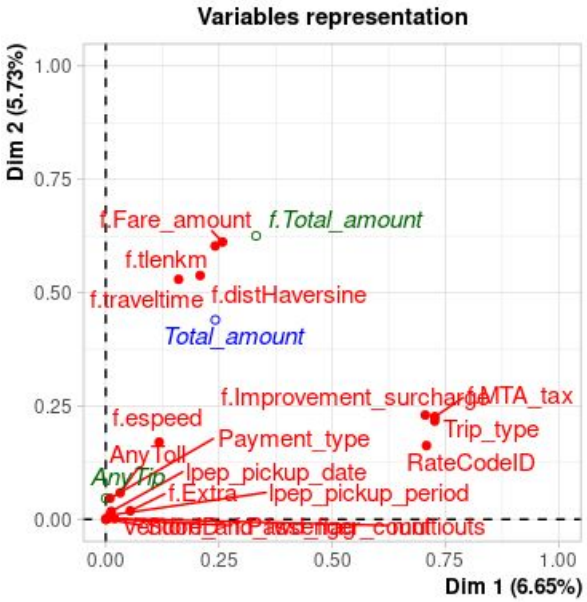


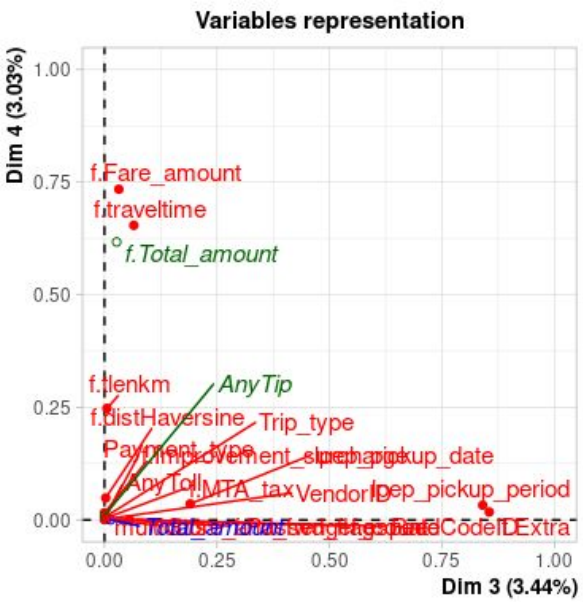
En las imágenes podemos observar los individuos reflejados en los planos factoriales. Si observamos el primer plano nos encontramos que los individuos están separados en 2 conjuntos, pero a pesar de eso los individuos que contribuyen se encuentran en un solo conjunto. En el segundo plano factorial podemos admirar cómo los individuos forman una especie efecto Guttman. También se puede observar que en el segundo plano factorial se forma dos conjuntos pero está vez, estos dos conjuntos se encuentran más juntos.

Categorical variables, supplementary numerical variables and supplementary categorical variables





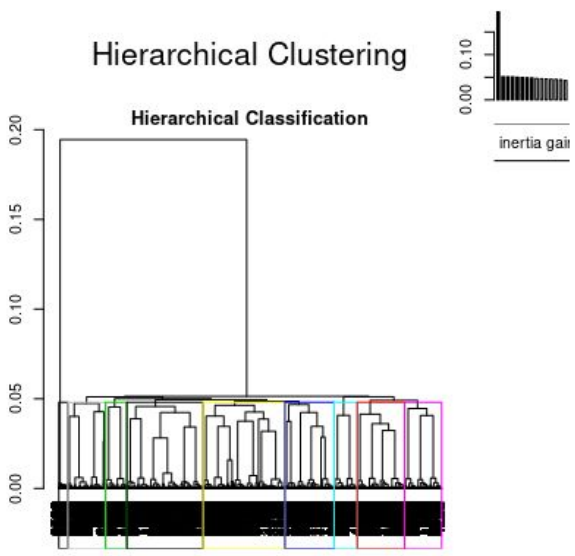


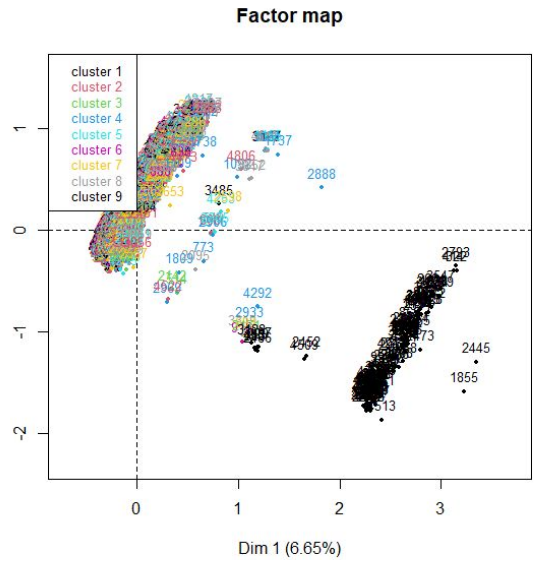


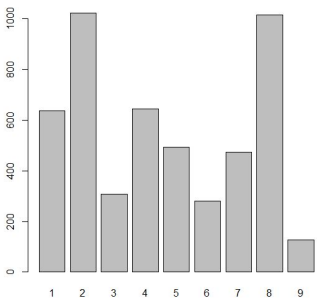
Como podemos ver nos vuelve a aparecer que el total amount está relacionado con el travel time tal como nos apareció haciendo CA. Además podemos observar cómo estas variables están muy relacionadas con las dimensiones 2 y 4 debido a su posición en los ejes.

A parte también podemos observar como la distancia del trayecto está relacionada con estas dos variables, hecho que tiene sentido ya que cuanto más largo sea un trayecto más tiempo hará falta para llevarlo a cabo y, por tanto, más caro será.

Hierarchical Clustering (from MCA)







En este gráfico podemos observar la distribución de individuos entre cada uno de los clusters donde los clusters 2 y 8 son los que contienen un mayor número respecto a los demás. Y en el Factor map anterior podemos ver como la separación entre clusters no queda nada bien definida a excepción del cluster 9, el cual aparece muy distante del resto.

Categorical variables which characterizes the clusters

```
res.hcpc$desc.var$test.chi2

##                          p.value  df
## RateCodeID             0.000000e+00   8
## f.MTA_tax              0.000000e+00   8
## f.Improvement_surcharge 0.000000e+00   8
## Trip_type              0.000000e+00   8
## lpep_pickup_date       0.000000e+00 240
## f.Total_amount         2.175017e-88  56
## f.Extra                1.081069e-56  16
## Payment_type           2.772605e-12  16
## multiouts              1.917731e-08   8
## f.tlenkm               3.940819e-07  16
## AnyTip                 5.725674e-06   8
## f.espeed               1.363908e-05  16
## f.Fare_amount          1.610297e-04  24
## f.distHaversine        3.387161e-03  16
## lpep_pickup_period     1.879570e-02  24
## f.traveltime           2.752741e-02  32
```

Como vemos, las variables categóricas que caracterizan los clusters, según el test de chi cuadrado, serían las anteriores, de las cuales podríamos destacar debido a su p.valor el RateCodeID, las tasas MTA, el tipo de trayecto y la fecha de recogida principalmente.

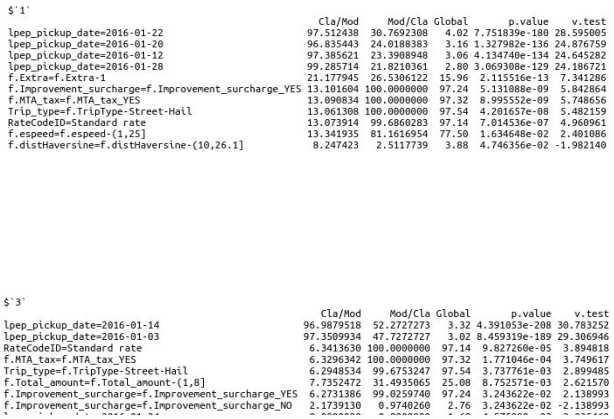
Numerical variables which characterizes the clusters

```
$`8`
              v.test Mean in category Overall mean sd in category Overall sd    p.value
Total_amount 2.358116       15.22199     14.54116      10.31011   10.30226 0.01836796

$`9`
              v.test Mean in category Overall mean sd in category Overall sd      p.value
Total_amount 5.861209       19.81014     14.54116      19.65805   10.30226 4.595097e-09
```

La variable de importe total podemos ver como aparece con una media superior respecto a la global en el cluster 8 así como en el 9, siendo en este más significativa la diferencia.

Description of each cluster by the categories



$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-22	97.512438	30.7692308	4.02	7.751839e-180	28.595005
lpep_pickup_date=2016-01-20	96.835443	24.0188383	3.16	1.327982e-136	24.876759
lpep_pickup_date=2016-01-12	97.385621	23.3908948	3.06	4.134740e-134	24.645282
lpep_pickup_date=2016-01-28	99.285714	21.8210361	2.80	3.069308e-129	24.186721
f.Extra=f.Extra-1	21.177945	26.5306122	15.96	2.115516e-13	7.341286
f.Improvement_surcharge=f.Improvement_surcharge_YES	13.101604	100.0000000	97.24	5.131088e-09	5.842864
f.MTA_tax=f.MTA_tax_YES	13.090834	100.0000000	97.32	8.995552e-09	5.748656
Trip_type=f.TripType-Street-Hail	13.061308	100.0000000	97.54	4.201657e-08	5.482159
RateCodeID=Standard rate	13.073914	99.6860283	97.14	7.014536e-07	4.960961
f.espeed=f.espeed-(1,25]	13.341935	81.1616954	77.50	1.634648e-02	2.401086
f.distHaversine=f.distHaversine-(10,26.1]	8.247423	2.5117739	3.88	4.746356e-02	-1.982140

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-30	98.7394958	22.99412916	4.76	6.393184e-167	27.536661
lpep_pickup_date=2016-01-15	96.1956522	17.31898239	3.68	6.455276e-117	22.985865
lpep_pickup_date=2016-01-10	98.8165680	16.34050881	3.38	6.518664e-117	22.985441
lpep_pickup_date=2016-01-29	96.5317919	16.34050881	3.46	7.258449e-111	22.372714
lpep_pickup_date=2016-01-19	97.9166667	13.79647750	2.88	4.716405e-96	20.795878
lpep_pickup_date=2016-01-05	98.5401460	13.20939335	2.74	3.024228e-93	20.483456
f.MTA_tax=f.MTA_tax_YES	21.0028771	100.00000000	97.32	3.091449e-14	7.594432
Trip_type=f.TripType-Street-Hail	20.9555054	100.00000000	97.54	4.119787e-13	7.251566
f.Improvement_surcharge=f.Improvement_surcharge_YES	20.9995886	99.90215264	97.24	4.659979e-13	7.234863
RateCodeID=Standard rate	21.0006177	99.80430528	97.14	2.949846e-12	6.980093
f.Fare_amount=f.Fare_amount-(6.5,9]	23.3365477	23.67906067	20.74	1.019080e-02	2.569287
f.distHaversine=f.distHaversine-[0,5]	21.0835989	84.14872798	81.58	1.636199e-02	2.400739
f.Total_amount=f.Total_amount-(8,11]	22.5303293	25.44031311	23.08	4.620768e-02	1.993491
lpep_pickup_period=Period afternoon	18.0238871	16.24266145	18.42	4.241879e-02	-2.029388
f.tlenkm=f.tlenkm-(5,67.9]	18.5309973	26.90802348	29.68	2.886766e-02	-2.185289
f.distHaversine=f.distHaversine-(5,10]	17.3314993	12.32876712	14.54	2.284340e-02	-2.276044
f.Fare_amount=f.Fare_amount-(14.5,71.5]	18.1153534	21.81996086	24.62	1.893066e-02	-2.346894

$`3`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-14	96.9879518	52.2727273	3.32	4.391053e-208	30.783252
lpep_pickup_date=2016-01-03	97.3509934	47.7272727	3.02	8.459319e-189	29.306946
RateCodeID=Standard rate	6.3413630	100.0000000	97.14	9.827260e-05	3.894818
f.MTA_tax=f.MTA_tax_YES	6.3296342	100.0000000	97.32	1.771046e-04	3.749617
Trip_type=f.TripType-Street-Hail	6.2948534	99.6753247	97.54	3.737761e-03	2.899485
f.Total_amount=f.Total_amount-(1,8]	7.7352472	31.4935065	25.08	8.752571e-03	2.621570
f.Improvement_surcharge=f.Improvement_surcharge_YES	6.2731386	99.0259740	97.24	3.243622e-02	2.138993
f.Improvement_surcharge=f.Improvement_surcharge_NO	2.1739130	0.9740260	2.76	3.243622e-02	-2.138993

$`4`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-09	95.698925	27.682737	3.72	1.527403e-156	26.655937
lpep_pickup_date=2016-01-11	98.816568	25.972006	3.38	1.696790e-154	26.478890
lpep_pickup_date=2016-01-07	98.571429	21.461897	2.80	7.988389e-126	23.859924
lpep_pickup_date=2016-01-18	98.360656	18.662519	2.44	2.253776e-108	22.115266
lpep_pickup_date=2016-01-23	97.560976	6.220840	0.82	2.877537e-35	12.392277
Trip_type=f.TripType-Street-Hail	13.184335	100.000000	97.54	3.538842e-08	5.512444
f.Improvement_surcharge=f.Improvement_surcharge_YES	13.163307	99.533437	97.24	8.419061e-06	4.454241
f.MTA_tax=f.MTA_tax_YES	13.152487	99.533437	97.32	1.365545e-05	4.349329
RateCodeID=Standard rate	13.094503	98.911353	97.14	1.552561e-03	3.164675
VendorID=f.Vendor-VeriFone	13.471901	81.648523	77.94	1.373415e-02	2.464142
f.traveltime=f.traveltime-[0,10]	13.922037	54.432348	50.28	2.412855e-02	2.255076
f.Total_amount=f.Total_amount-(30,40]	7.906977	2.643857	4.30	2.050218e-02	-2.317029
VendorID=f.Vendor-Mobile	10.698096	18.351477	22.06	1.373415e-02	-2.464142

$`5`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-08	97.714286	34.6153846	3.50	1.853040e-178	28.483924
lpep_pickup_date=2016-01-17	97.023810	32.9959514	3.36	1.075258e-167	27.601239
lpep_pickup_date=2016-01-27	97.560976	32.3886640	3.28	9.766050e-166	27.437609
f.Improvement_surcharge=f.Improvement_surcharge_YES	10.160428	100.0000000	97.24	4.715499e-07	5.037541
f.MTA_tax=f.MTA_tax_YES	10.152076	100.0000000	97.32	7.236958e-07	4.954894
Trip_type=f.TripType-Street-Hail	10.129178	100.0000000	97.54	2.345865e-06	4.721088
RateCodeID=Standard rate	10.088532	99.1902834	97.14	1.206959e-03	3.237230
f.Total_amount=f.Total_amount-(18,20]	14.427861	5.8704453	4.02	3.579829e-02	2.099211

$`6`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-26	99.2957746	50.3571429	2.84	4.922040e-193	29.637450
lpep_pickup_date=2016-01-06	96.5277778	49.6428571	2.88	1.775953e-183	28.886273
f.Extra=f.Extra-1	9.1478697	26.0714286	15.96	8.253107e-06	4.458511
RateCodeID=Standard rate	5.7648754	100.0000000	97.14	2.331491e-04	3.680093
Trip_type=f.TripType-Street-Hail	5.7412344	100.0000000	97.54	7.624867e-04	3.366065
f.Improvement_surcharge=f.Improvement_surcharge_YES	5.7383793	99.6428571	97.24	3.272053e-03	2.940952
f.MTA_tax=f.MTA_tax_YES	5.7336621	99.6428571	97.32	4.047670e-03	2.874423

$`7`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-16	94.5945946	44.3974630	4.44	7.741813e-219	31.576583
lpep_pickup_date=2016-01-02	96.7567568	37.8435518	3.70	5.158844e-189	29.323796
lpep_pickup_date=2016-01-24	100.0000000	17.7589852	1.68	7.441096e-90	20.099587
f.Extra=f.Extra-0	12.1592384	59.4080338	46.22	1.590821e-09	6.034879
Trip_type=f.TripType-Street-Hail	9.6985852	100.0000000	97.54	4.188845e-06	4.601785
RateCodeID=Standard rate	9.7179329	99.7885835	97.14	9.447959e-06	4.429437
f.Improvement_surcharge=f.Improvement_surcharge_YES	9.7079391	99.7885835	97.24	1.526126e-05	4.324883

$`8`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
lpep_pickup_date=2016-01-01	97.7973568	21.87192118	4.54	2.303963e-154	26.467352
lpep_pickup_date=2016-01-31	97.8723404	18.12807882	3.76	5.690690e-127	23.970196
lpep_pickup_date=2016-01-21	95.4285714	16.45320197	3.50	7.780280e-109	22.163210
lpep_pickup_date=2016-01-25	98.0132450	14.58128079	3.02	1.156642e-101	21.406724
lpep_pickup_date=2016-01-13	96.7948718	14.87684729	3.12	5.742471e-101	21.331902
lpep_pickup_date=2016-01-04	97.2789116	14.08866995	2.94	1.705184e-96	20.844631
Trip_type=f.TripType-Street-Hail	20.8119746	100.00000000	97.54	5.131686e-13	7.221766
f.Improvement_surcharge=f.Improvement_surcharge_YES	20.8556150	99.90147783	97.24	5.916016e-13	7.202404
f.MTA_tax=f.MTA_tax_YES	20.8384710	99.90147783	97.32	1.465677e-12	7.077699
RateCodeID=Standard rate	20.7535516	99.31034483	97.14	1.276323e-07	5.282213
f.tlenkm=f.tlenkm-(5,67.9]	22.5741240	33.00492611	29.68	9.911001e-03	2.578919

$`9`

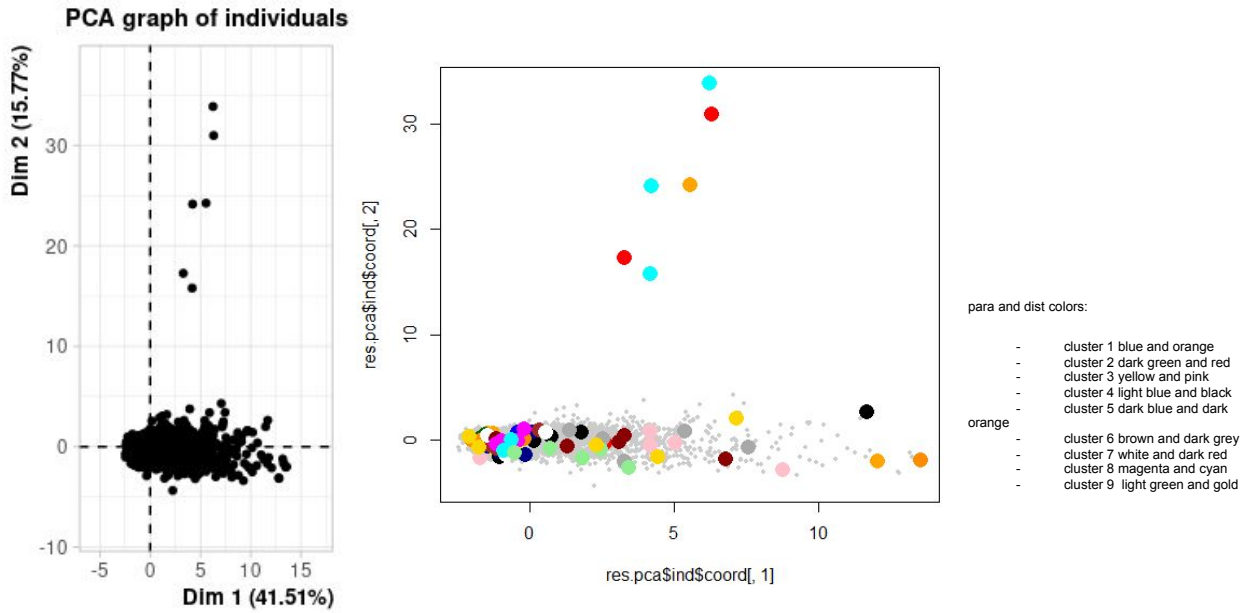
	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.MTA_tax=f.MTA_tax_NO	95.52238806	100.00000	2.68	4.853662e-248	33.636857
f.Improvement_surcharge=f.Improvement_surcharge_NO	92.75362319	100.00000	2.76	3.342807e-243	33.304324
Trip_type=f.TripType-Dispatch	99.18699187	95.31250	2.46	1.549633e-236	32.840636
RateCodeID=Others	83.91608392	93.75000	2.86	1.151660e-206	30.677058
f.Extra=f.Extra-0	5.45218520	98.43750	46.22	2.266220e-40	13.301582
f.Total_amount=f.Total_amount-[0,1]	83.33333333	11.71875	0.36	4.444113e-22	9.660334

Los clusters del 1 al 8 presentan una sobrerrepresentación muy clara de ciertas fechas diferentes según el cluster, constituyendo casi la totalidad de observaciones globales de dichas fechas pertenecientes a cada uno de los clusters.

Por otro lado, en el cluster 9 vemos una sobrerrepresentación de trayectos donde no aparece la tasa MTA ni un sobrecargo de mejora, además también lo están el tipo de trayecto Dispatch así como el RateCodeID de otros. A excepción del RatecodeID que supone un 83% del total de observaciones, el resto de categorías superan una

representación del 90% de la totalidad de las observaciones. Es decir, más del 90% de los trayectos que tienen estas categorías, están dentro del cluster 9.

Description of the clusters by the individuals



Los elementos más distintivos del clustering podemos ver que serían los del cluster 1,2,4,5 y 8 al ser los más alejados de la formación principal de clusters dentro del gráfico.

El del cluster 5 por ejemplo(naranja oscuro) se caracteriza por ser un trayecto con una velocidad efectiva muy alta y una distancia a la par con un tiempo de trayecto muy superior a la media. Por los mismos valores se rige el individuo distintivo del cluster 4(negro) situado más a la derecha.

Los individuos distintivos del cluster 8, por otro lado, vienen caracterizados por tener una velocidad efectiva muy alta, sin ser la distancia de estos muy elevada. Además tenemos como el importe de estos es razonablemente elevado y su RateCodeID pertenece al de otros.

Por lo que hace a los del cluster 1, uno de ellos tiene un tiempo de trayecto excesivamente largo aún teniendo el mismo rango de distancia de trayecto que el otro. Este segundo pese a ser de una duración más reducida parece tener un coste muy superior al primero teniendo los dos el mismo rango de tarifa del trayecto y no haber pasado por ningún peaje. Esto puede deberse a la presencia de una propina, aunque esta debería ser muy alta para marcar una diferencia tan significativa del individuo.

Referente a los individuos a destacar del cluster 2, tenemos dos trayectos cuya duración es muy larga siendo ínfimas sus velocidades efectivas y sus distancias. Tienen un coste final muy similar y los diferencian que uno es efectuado por la noche y otro por el medio día.

Hierarchical tree result

Ratio between within inertias

```
res.hcpc$call$t$quot[1:res.hcpc$call$t$nb.clust]

## [1] 0.9775104 0.9771113 0.9766060 0.9765191 0.9762269 0.9758285 0.9754768
## [8] 0.9754845        NA
```

Si vemos la relación entre inercias dentro de un mismo cluster vemos que es bastante grande, es decir la inercia intra cluster es elevada.

Inertia gain

```
res.hcpc$call$t$inert.gain[1:res.hcpc$call$t$nb.clust]

## [1] 0.19449854 0.05140255 0.05113830 0.05107095 0.05006133 0.04949422 0.04912733
## [8] 0.04863734 0.04742985
```

Si vemos la inertia gain vemos que es bastante baja des del cluster 1 al 9. La pérdida de pasar de n clusters a n+1 clusters es bastante baja.

Partition quality

```
(res.hcpc$call$t$within[1]-res.hcpc$call$t$within[res.hcpc$call$t$nb.clust])/res.hcpc$call$t$within[1]

## [1] 0.2199214
```

Podemos ver que cogiendo 9 cluster obtenemos una calidad del 21.99%, lo que significa que el número de cluster óptimos debería de ser superior a 9. Según hemos observado, para obtener una calidad del 80% deberíamos coger 200 clusters aproximadamente. Estos resultados demuestran que cada cluster divide significa una pequeña población de la muestra y por lo tanto, el clustering no es del todo correcto.