Cajamar Tratamiento y Cluster

Contents

VARIABLES

```
## 'data.frame':    2241460 obs. of  15 variables:
##  $ X                : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ CP_CLIENTE       : int  30000 30000 30000 30000 30000 30000 30000 30000 30000 30000 ...
##  $ CP_COMERCIO      : int  30001 30001 30001 30001 30001 30001 30001 30001 30001 30001 ...
##  $ SECTOR           : Factor w/ 11 levels "ALIMENTACION",..: 1 9 9 6 5 6 6 7 9 1 ...
##  $ DIA              : Factor w/ 731 levels "2015-10-01","2015-10-02",..: 8 26 51 73 77 84 94 128 131 1
##  $ FRANJA_HORARIA   : Factor w/ 12 levels "00-02","02-04",..: 10 9 9 10 7 11 7 11 9 11 ...
##  $ IMPORTE          : num  16.8 373.7 134.5 25 131 ...
##  $ NUM_OP           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PROP_SEX         : num  0 0.99 0.99 0 0 0.99 0 0.99 0.99 0.99 ...
##  $ PROP_R_BAJA      : num  0.99 0.99 0.99 0 0.99 0.99 0 0 0.99 0.99 ...
##  $ PROP_R_MEDIA     : num  0 0 0 0.99 0 0 0.99 0 0 0 ...
##  $ PROP_R_ALTA      : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 1 0.01 0.01 ...
##  $ PROP_JOVEN       : num  0.99 0 0 0 0 0.99 0.99 0 0 0.99 ...
##  $ PROP_ADULTO      : num  0 0.99 0.99 0.99 0 0 0 0.99 0.99 0 ...
##  $ PROP_PENSIONISTA: num  0.01 0.01 0.01 0.01 1 0.01 0.01 0.01 0.01 0.01 ...
```

Descripción de las variables:

- CP_CLIENTE: variable categórica, código postal en el que reside el comprador (cliente), formato 30XXX. Hay un total de 200 códigos postales.
- CP_COMERCIO: variable categórica, código postal en el que está ubicado el comercio, formato 30XXX. Hay un total de 16 códigos postales.
- SECTOR: variable categórica, nombre del sector en el que está englobado el comercio. Hay un total de 10 sectores.
- DIA: variable tipo fecha, día del año en el que se ha realizado la compra, formato YYYY-MM-DD
- FRANJA_HORARIA: variable categórica, franja horaria en la que se realiza la compra, están definidas en periodo de 2 horas, formato hora inicio – hora fin (XX-XX). Definiendo así 12 franjas desde las 00 hasta las 24 horas.
- IMPORTE: variable real, importe total en euros de las compras realizadas por los clientes de un código postal en los comercios de un código postal y un sector durante una franja horaria.
- NUM_OP: variable entera, número de operaciones realizadas por los clientes de un código postal en los comercios de un código postal y un sector durante una franja horaria.
- PROP_SEX: variable real, proporción del número de operaciones realizadas por los clientes de sexo masculino.

- PROP_R_BAJA: variable real, proporción del número de operaciones realizadas por los clientes de renta baja.
- PROP_R_MEDIA: variable real, proporción del número de operaciones realizadas por los clientes de renta media.
- PROP_R_ALTA: variable real, proporción del número de operaciones realizadas por los clientes de renta alta.
- PROP_JOVEN: variable real, proporción del número de operaciones realizadas por los clientes de edad comprendida entre 18 y 35.
- PROP_ADULTO: variable real, proporción del número de operaciones realizadas por los clientes de edad comprendida entre 36 y 65.
- PROP_MAYOR: variable real, proporción del número de operaciones realizadas por los clientes de edad superior a 65.

DISTRIBUCION RENTAS

TIPO	PORCENTAJE_TOTAL
RENTA ALTA	0.13
RENTA MEDIA	0.35
RENTA BAJA	0.38
RENTA MEDIA-BAJA	0.07
TOTAL_CATEGORIAS	0.93

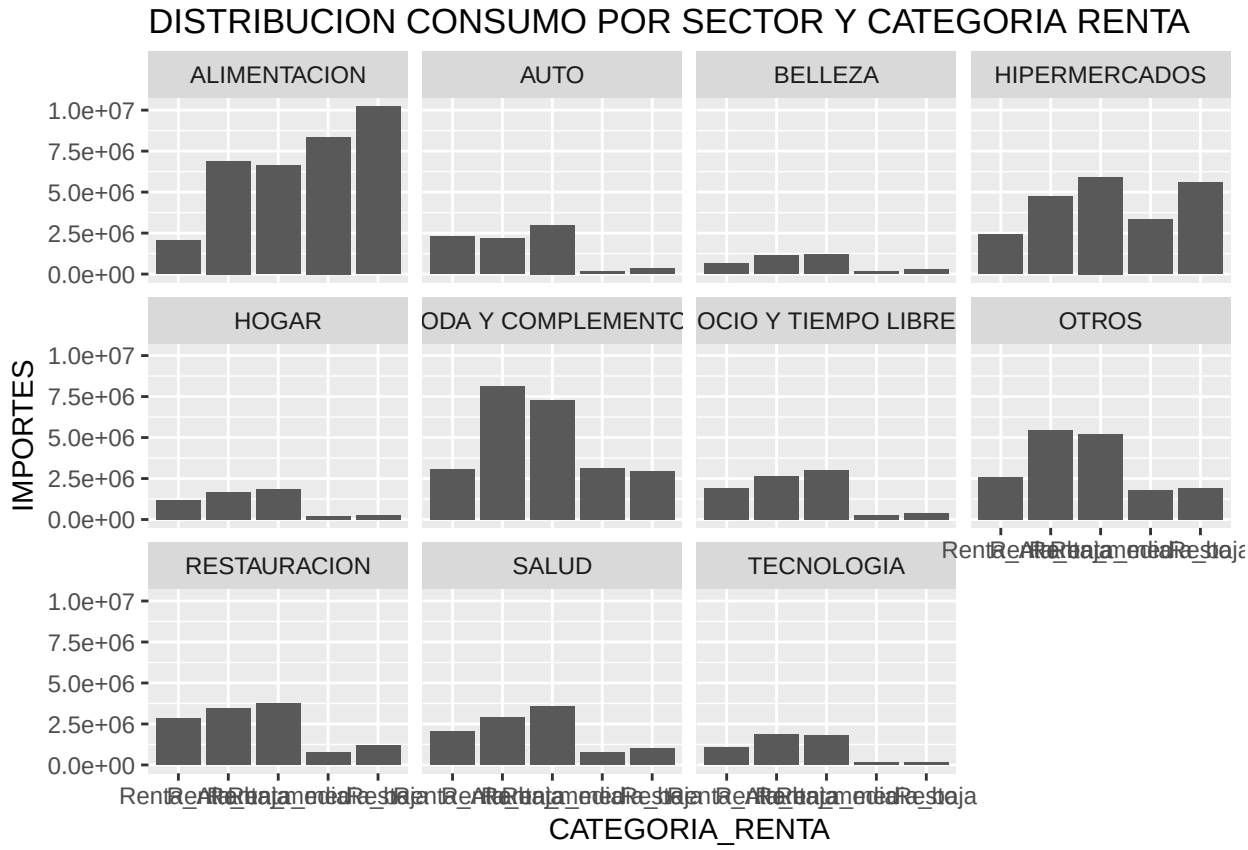
Renta Alta: - Proporción Renta Alta mayor igual a un 70%. - 13% TOTAL.

Renta Media: - Proporción Renta Media mayor igual a un 70%. - 35% TOTAL.

Renta Baja: - Proporción Renta Baja mayor igual a un 70%. - 38% TOTAL.

Renta Media-Baja: - Renta Proporción Media y Baja mayor o igual a un 40% - Proporción Renta Media y Baja superior a un 50% y 30% indistintamente. - 7% TOTAL

Resto: - Cualquiera no incluido dentro de las categorías anteriores. - 7% TOTAL.



DISTRIBUCION EDAD

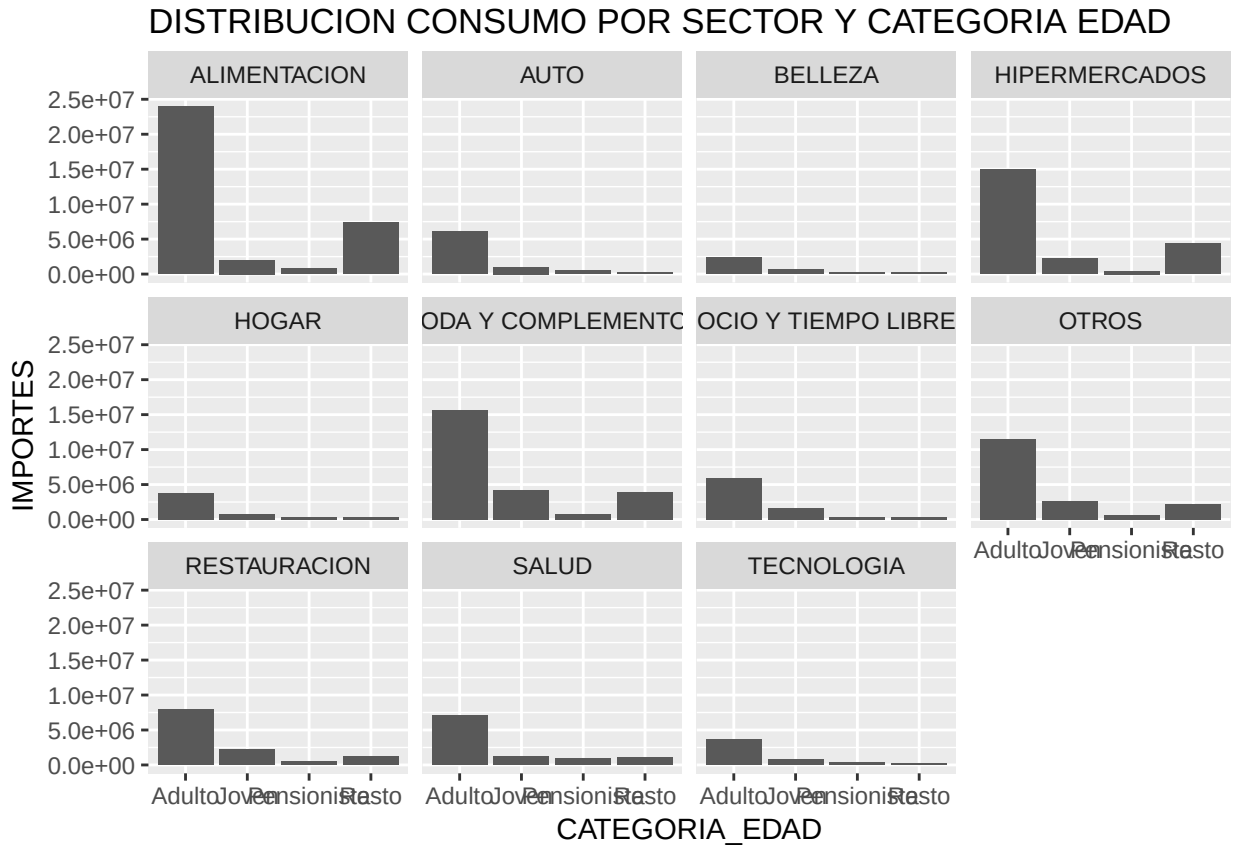
TIPO	PORCENTAJE_TOTAL
JOVEN	0.21
ADULTO	0.67
PENSIONISTA	0.04
TOTAL_CATEGORIAS	0.92

Renta joven: - Proporción Renta Alta mayor igual a un 70% - 21% TOTAL

Renta Adulto: - Proporción Renta Media mayor igual a un 70% - 67% TOTAL

Renta Pensionistas: - Proporción Renta Baja mayor igual a un 70%. - 4% TOTAL

Resto: - cualquiera no incluido dentro de las categorías anteriores. - 8% TOTAL.



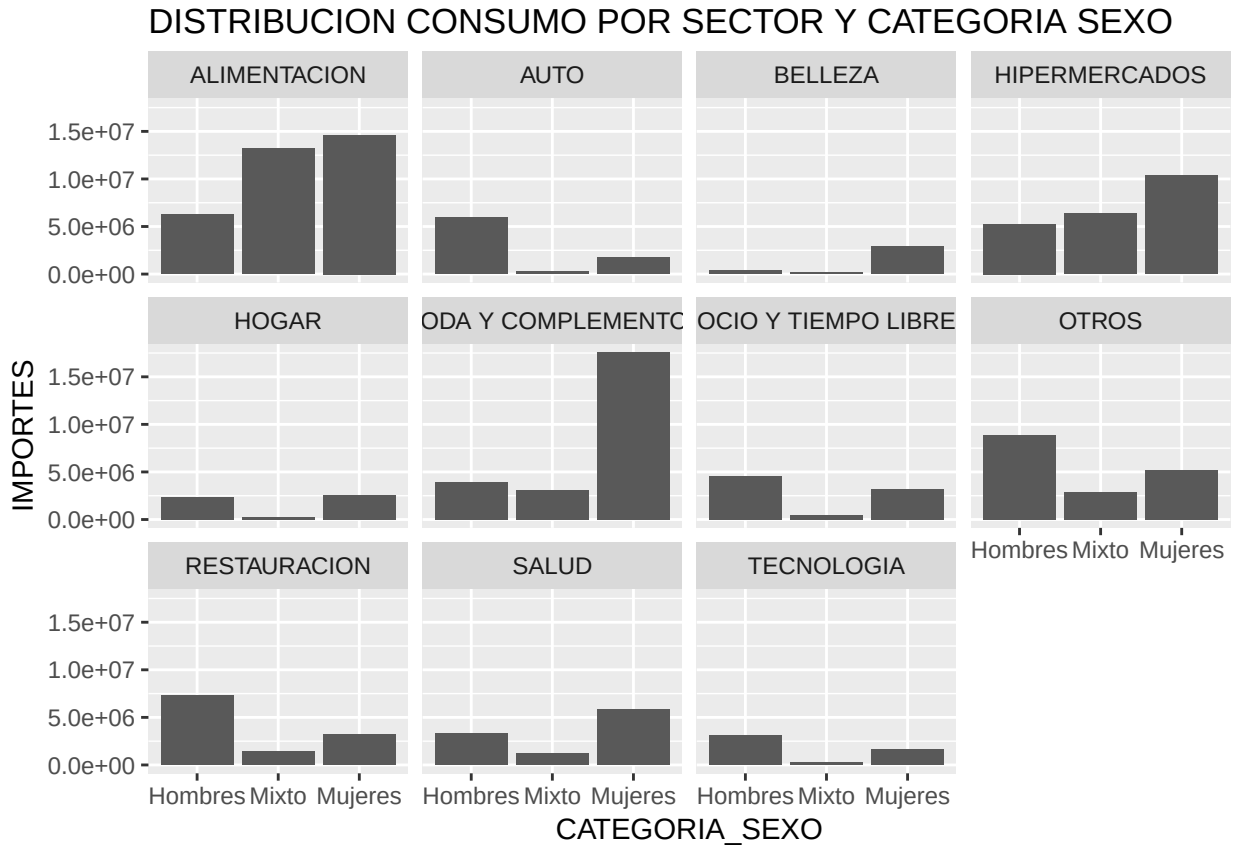
DISTRIBUCION SEXO

TIPO	PORCENTAJE_TOTAL
MUJER	0.54
MIXTO	0.09
HOMBRE	0.37
TOTAL_CATEGORIAS	1

Mujeres: - Probabilidad sexo entre 0 y 0.30. - 54% TOTAL.

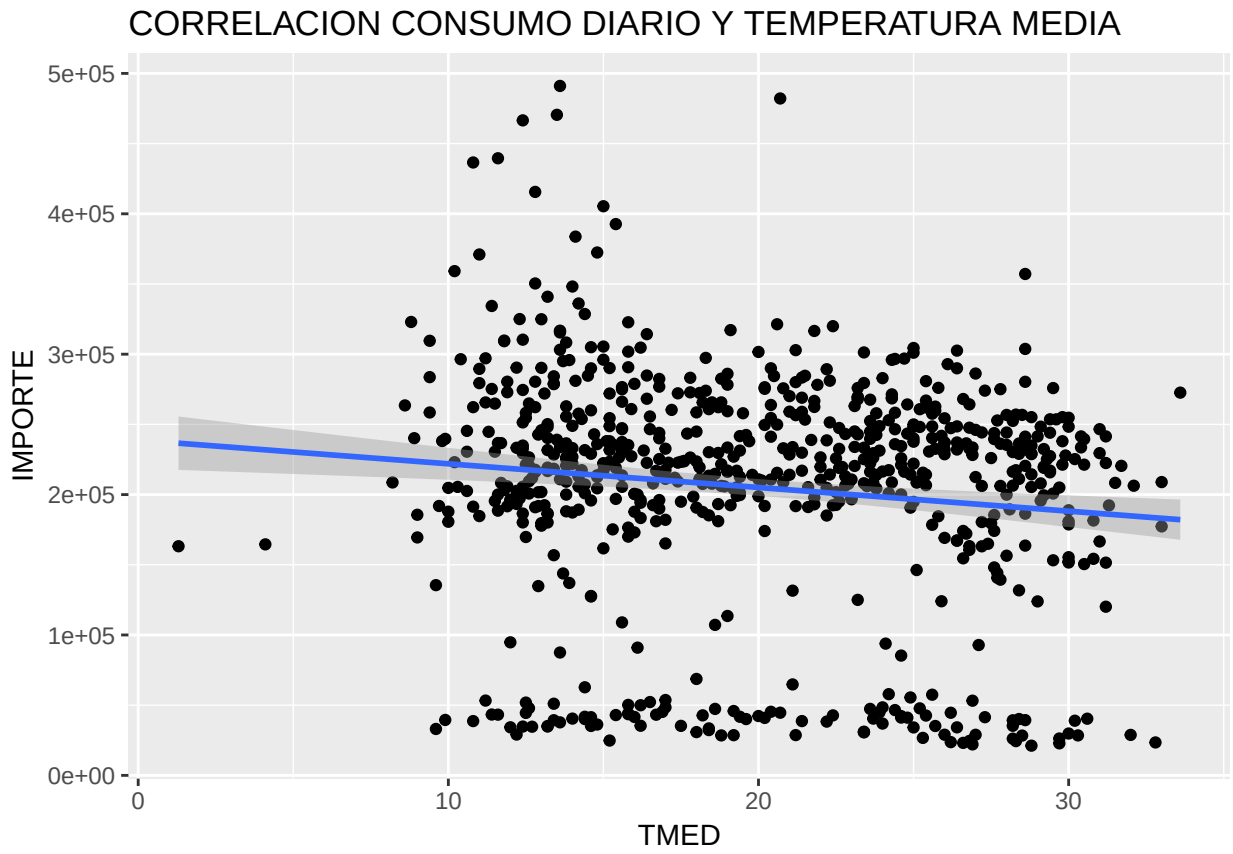
Mixto: - Probabilidad sexo entre 0.30 y 0.70 - 9% TOTAL

Hombres: - Probabilidad superior a 0.70 - 37% TOTAL

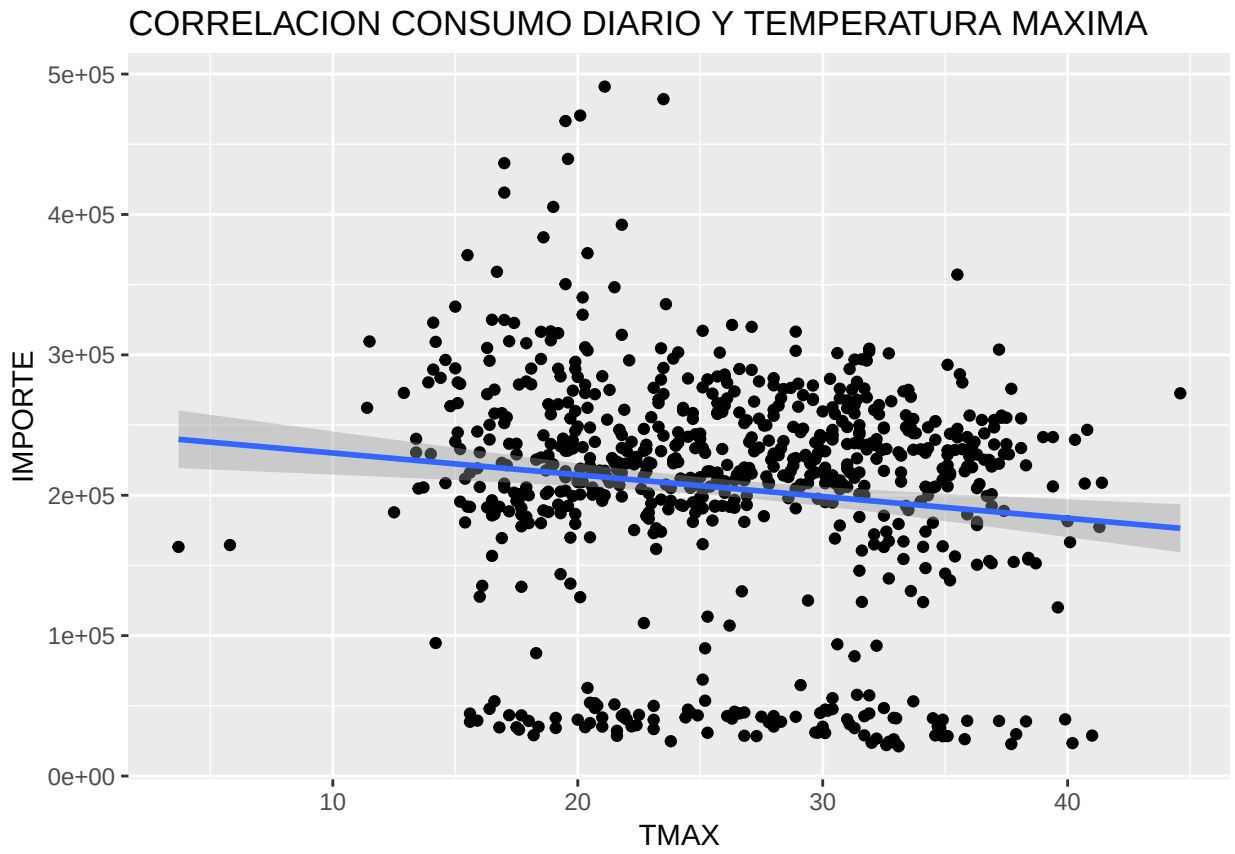


Como se puede apreciar las mujeres gastan mucho mas que los hombres en Moda y Complementos, Belleza e Hipermercados. Sin embargo los hombres gastan mas en tecnología, Otros, Automoción y Restauración.

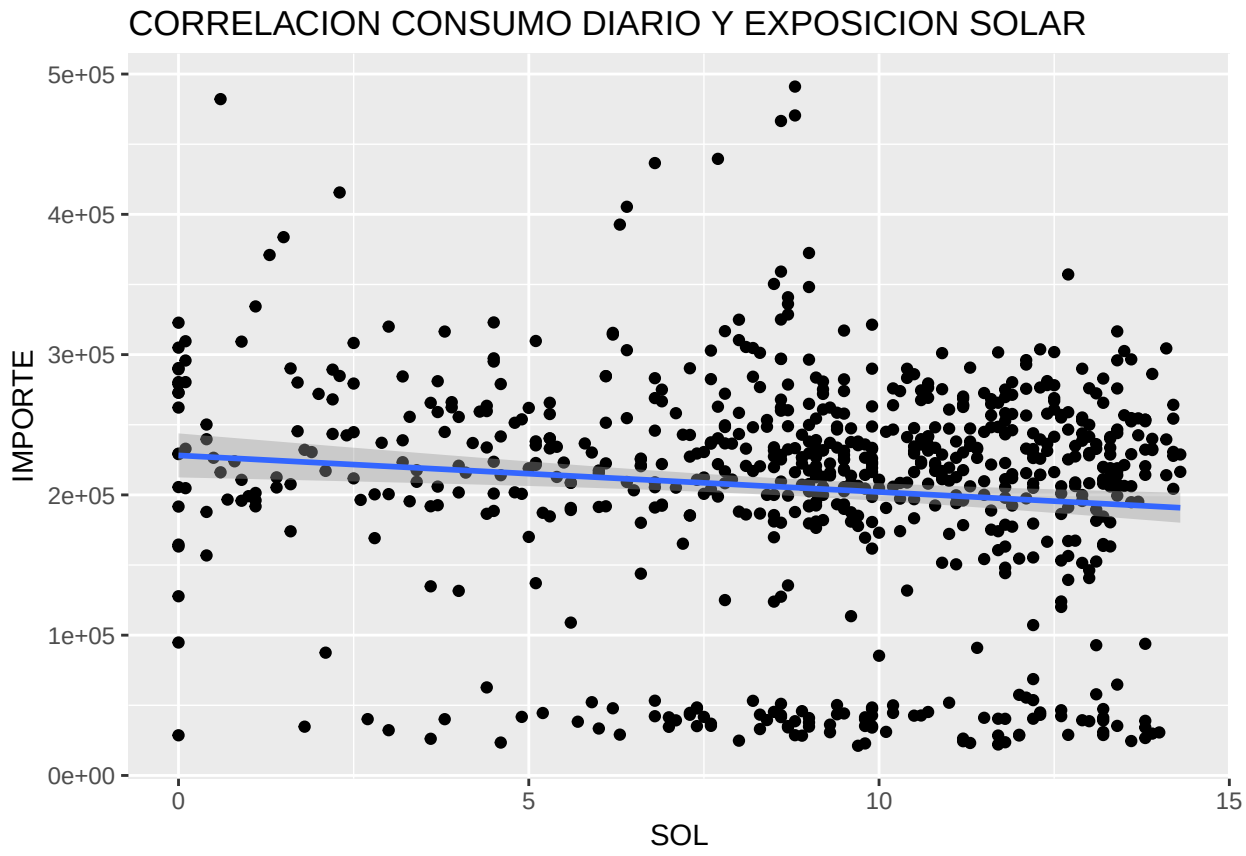
ANALISIS CLIMA



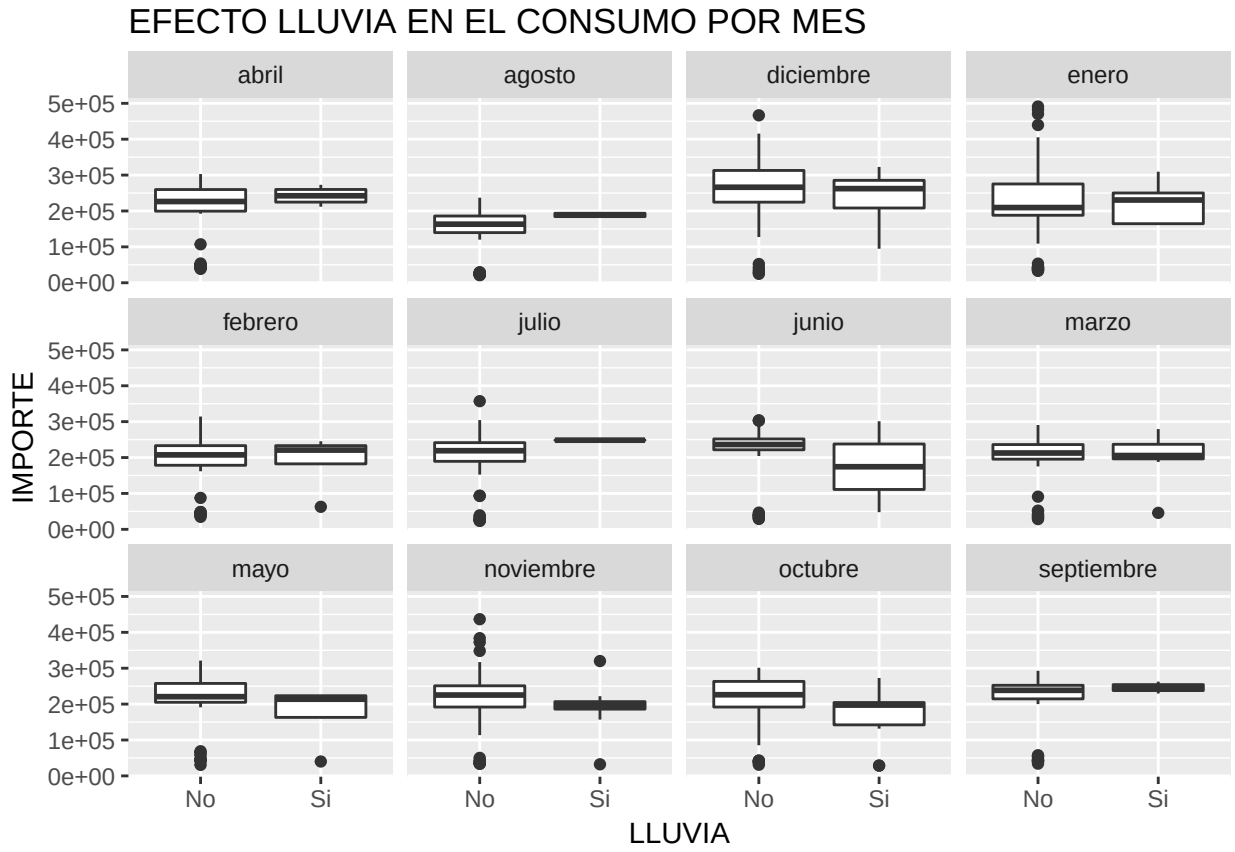

```
## [1] "La correlación de Temperatura media y el consumo diario es de un  -0.13"
```



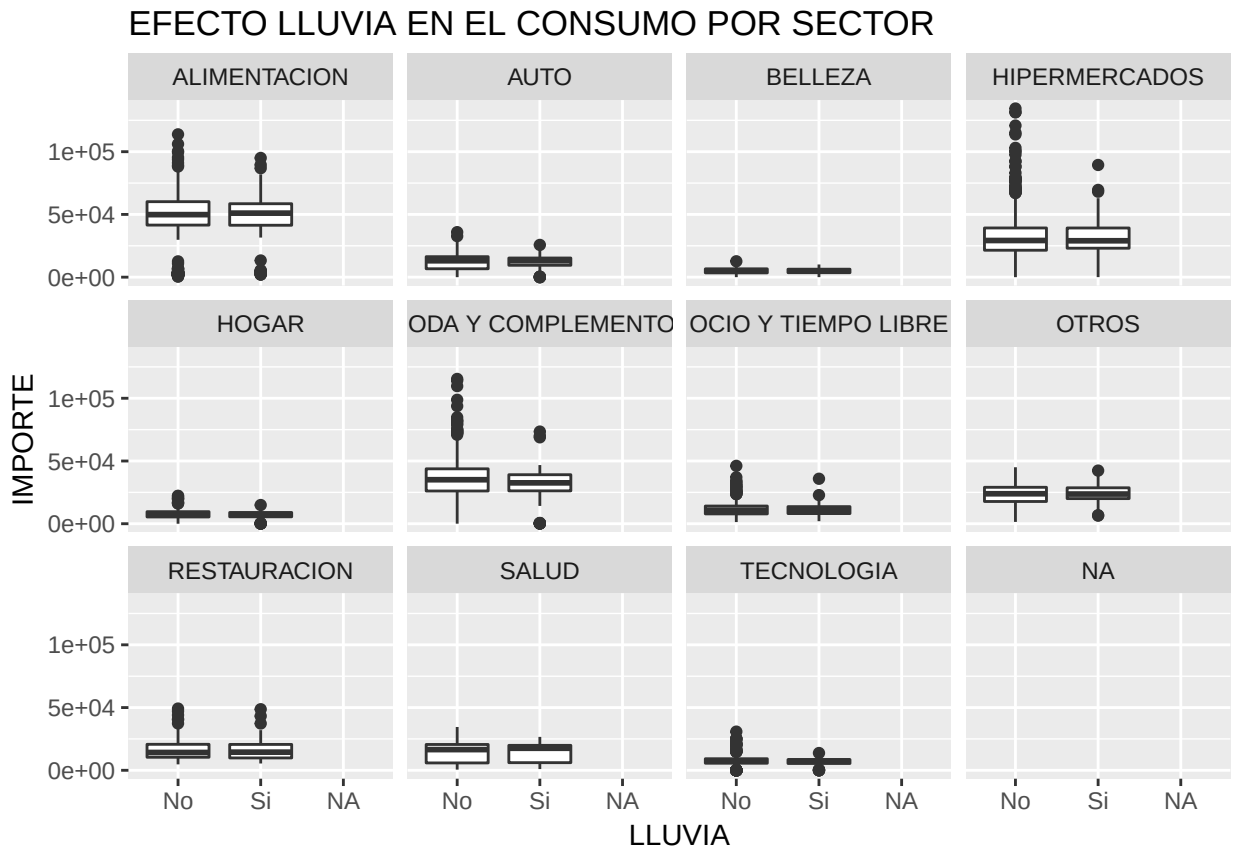

```
## [1] "La correlación de Temperatura máxima y el consumo diario es de un  -0.13"
```



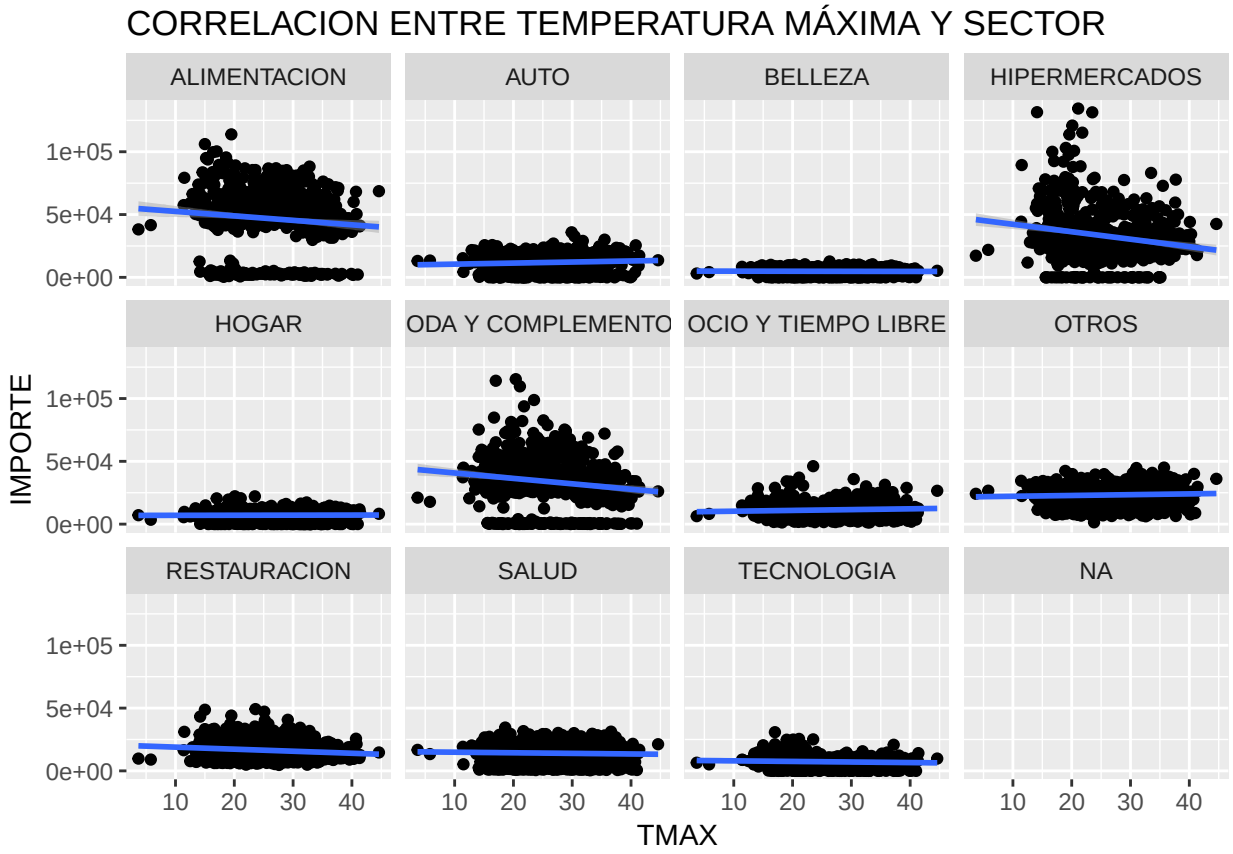

```
## [1] "La correlación de la exposición solar y el consumo diario es de un  -0.12"
```



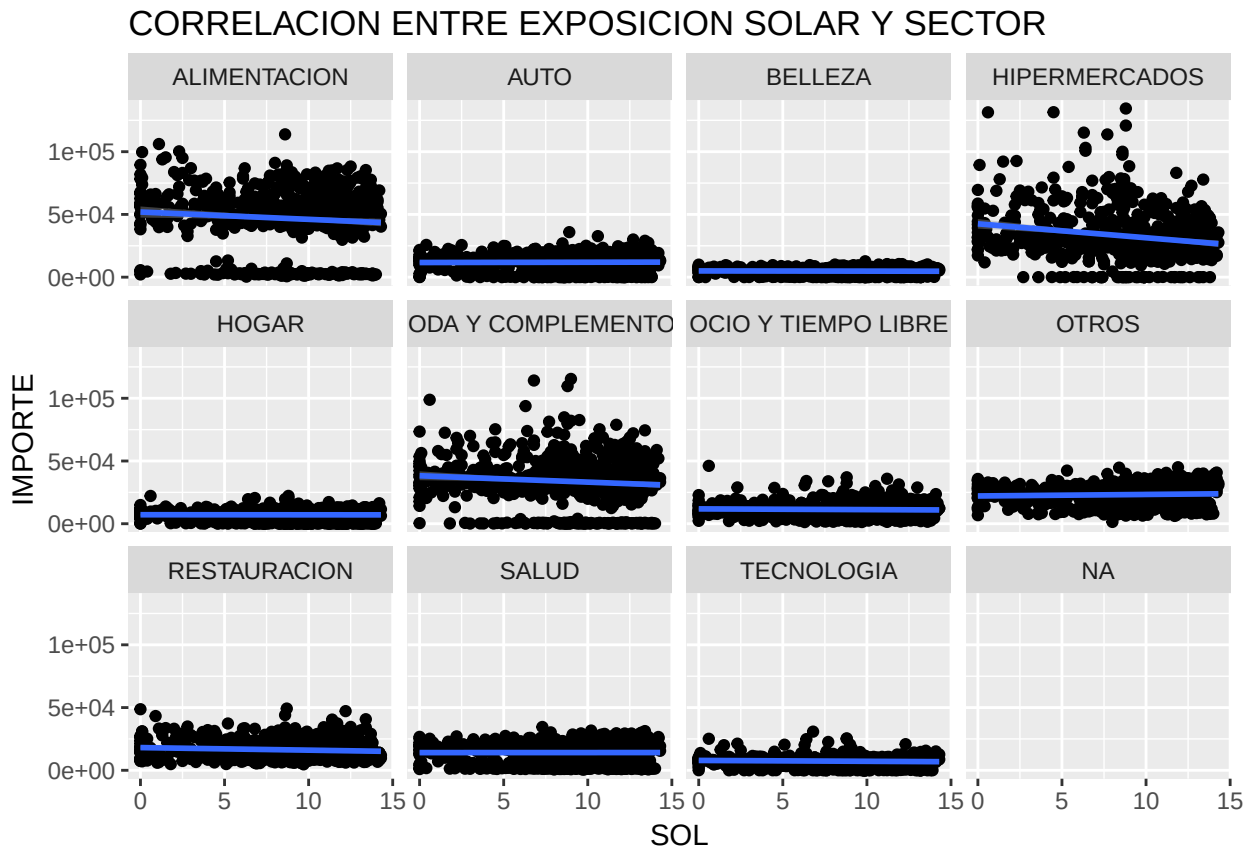
No apreciamos ninguna relación significativa.



No apreciamos ninguna relación significativa



Como se puede apreciar los sectores que se ven ligeramente afectados por la exposición la temperatura máxima son la alimentación, Hipermercados y moda y complementos. Con una relación inversa, a mayor temperatura máxima menos consumo en dichos sectores.

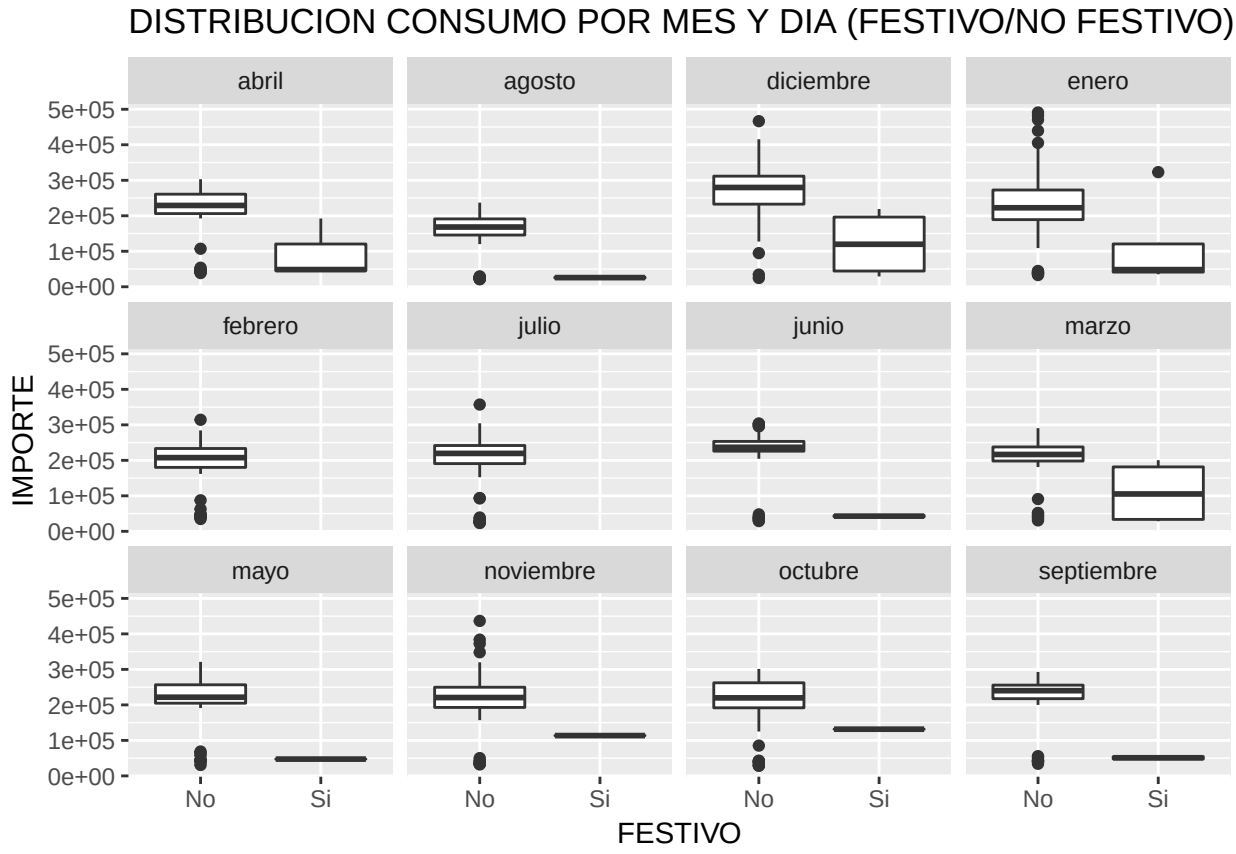


Como se puede apreciar los sectores que se ven ligeramente afectados por la exposición solar son la alimentación, Hipermercados y moda y complementos. Con una relación inversa, a mayor temperatura máxima menos consumo en dichos sectores.

ANALISIS CLUSTER

TRATAMIENTO PREVIO

Representación grafica del consumo sectorial, distinguiendo entre festivos y no festivos. COmo se puede apreciar en los meses de enero, marzo y abril hay valores extremos o outliers en los días festivos que coinciden con la navidad y semana santa.



Hacemos la agrupación por los siguientes campos:

- CP_CLIENTE
- CATEGORIA DE EDAD
- CATEGORIA DE RENTA
- CATEGORIA DE SEXO
- SECTOR

De esta manera de cada codigo postal de clientes, tendremos el consumo agrupado por todas las posibles combinaciones de categorias de edad, Renta, sexo y sector. En total son unas 178 categorías.

```
## Warning: Setting row names on a tibble is deprecated.
```

Exploramos los valores de los estadísticos de las variables

```
##                                          Min    Med   Mean      SD
## Adulto_Renta_Alta_Hombres_ALIMENTACION     0 1223.2 5480.4 12924.9
## Adulto_Renta_Alta_Hombres_AUTO             0 2016.6 9221.1 17636.8
## Adulto_Renta_Alta_Hombres_BELLEZA          0   14.0  509.6  1255.4
## Adulto_Renta_Alta_Hombres_HIPERMERCADOS    0 2290.1 5469.8  7126.7
## Adulto_Renta_Alta_Hombres_HOGAR            0  504.4 3331.4  7492.7
## Adulto_Renta_Alta_Hombres_MODA Y COMPLEMENTOS   0 1356.9 4176.2  6329.2
##                                             Max
## Adulto_Renta_Alta_Hombres_ALIMENTACION    141435.5
## Adulto_Renta_Alta_Hombres_AUTO            128395.9
## Adulto_Renta_Alta_Hombres_BELLEZA          10153.1
## Adulto_Renta_Alta_Hombres_HIPERMERCADOS    36624.0
## Adulto_Renta_Alta_Hombres_HOGAR            54013.3
## Adulto_Renta_Alta_Hombres_MODA Y COMPLEMENTOS  36091.7
```

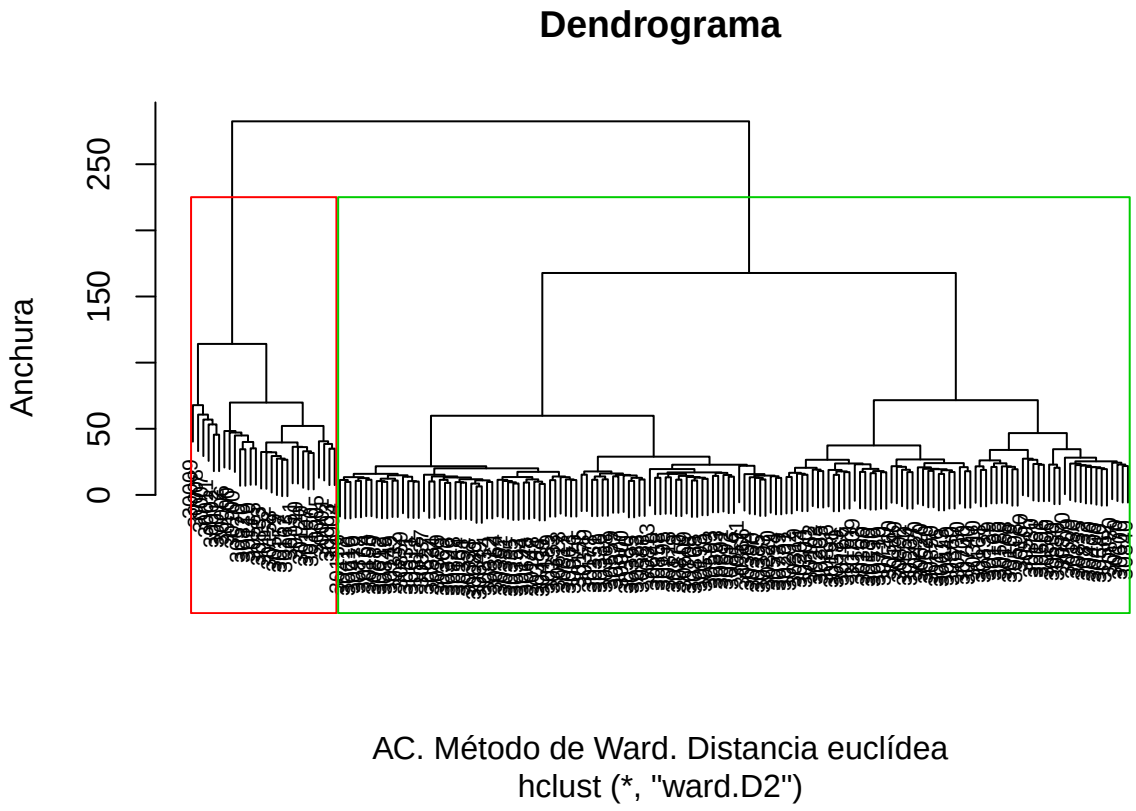
Como se puede apreciar hay bastante dispersión en los datos, una alternativa para reducir la dispersión es realizar una tranformación logarítmica de los datos.

```
##                                              Min Med Mean  SD  Max
## Adulto_Renta_Alta_Hombres_ALIMENTACION         0 7.1  6.2 3.3 11.9
## Adulto_Renta_Alta_Hombres_AUTO                 0 7.6  6.4 3.8 11.8
## Adulto_Renta_Alta_Hombres_BELLEZA              0 2.7  3.0 3.2  9.2
## Adulto_Renta_Alta_Hombres_HIPERMERCADOS        0 7.7  6.6 3.2 10.5
## Adulto_Renta_Alta_Hombres_HOGAR                0 6.2  5.1 3.5 10.9
## Adulto_Renta_Alta_Hombres_MODA Y COMPLEMENTOS  0 7.2  6.4 3.1 10.5
```

Como se puede apreciar, con la tranformación logarítmica se ha reducido considerablemente la dispersión en los datos.

CLUSTER

CLUSTER JERÁRQUICO: DENDOGRAMA

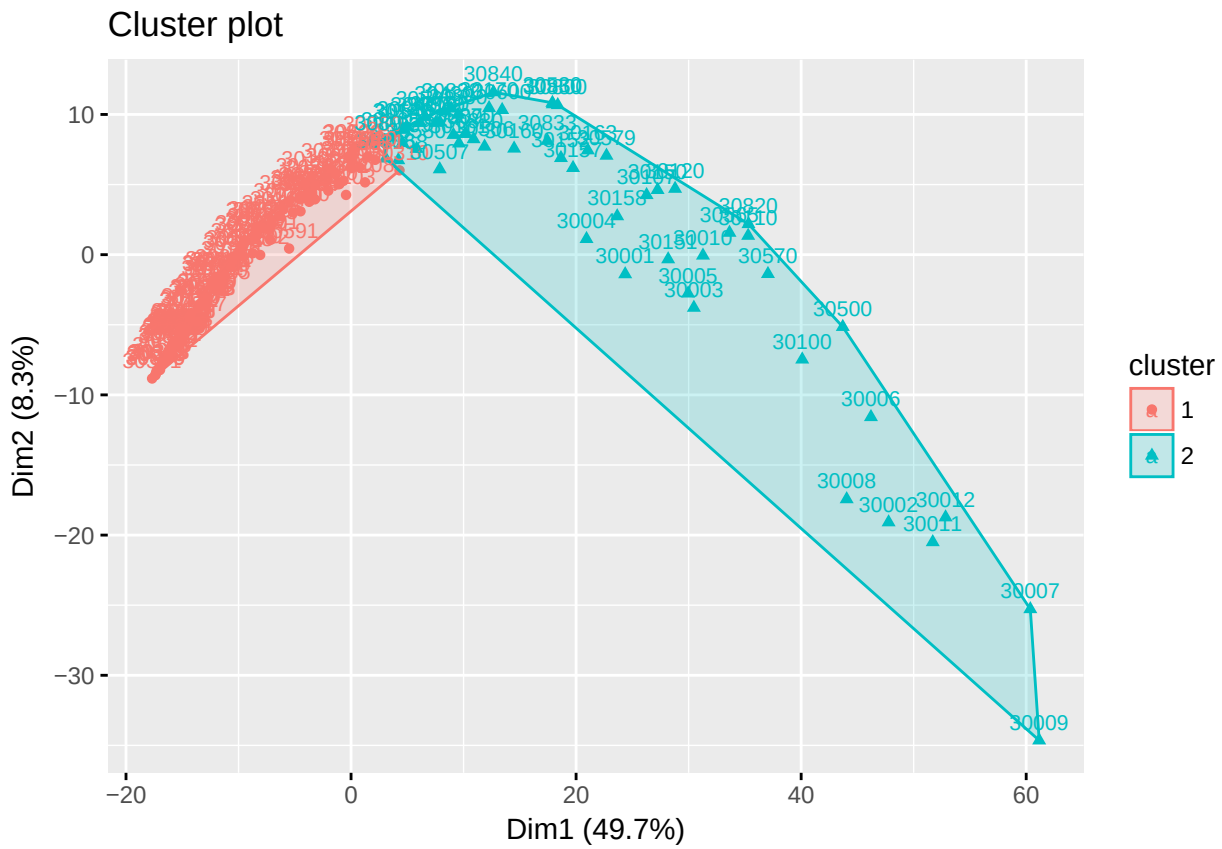


De acuerdo con lo oservado en el dendograma dividiremos las observaciones en 2 grupos.

CLUSTER NO JERÁRQUICO: PAM

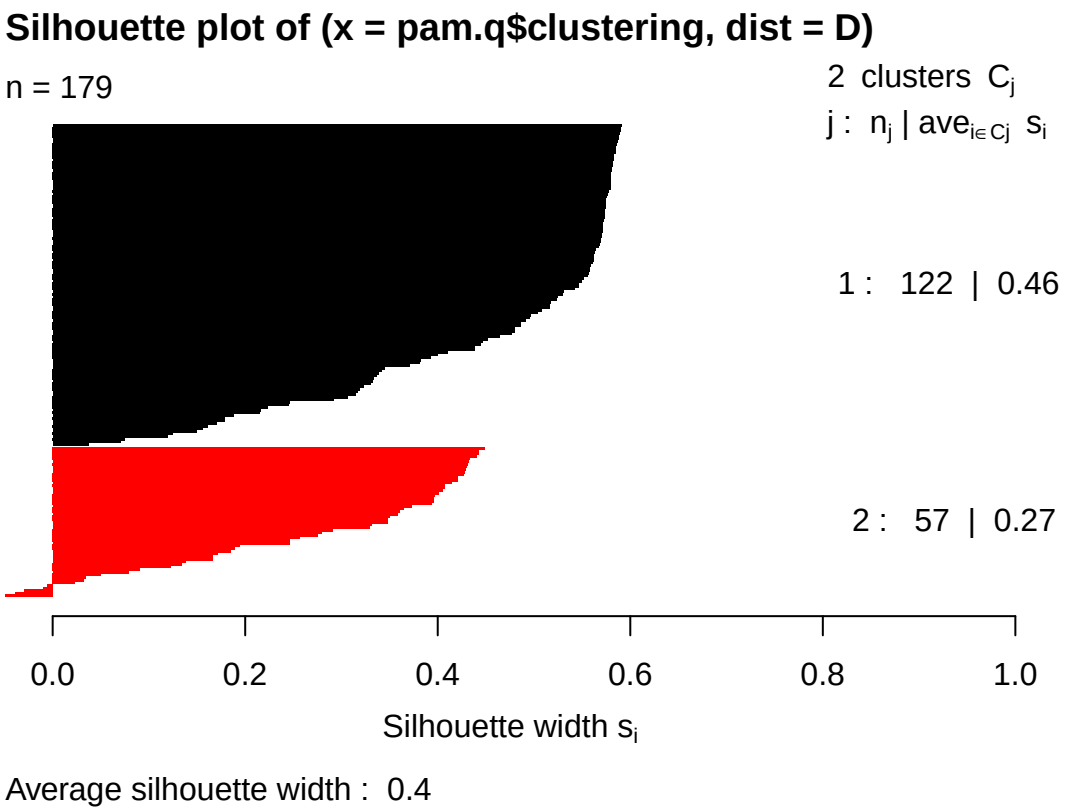
Una vez aceptada la conveniencia de llevar a cabo el análisis cluster, vamos a realizar clustering no jerárquico. Los métodos de clustering no jerárquico a diferencia de los métodos de clustering jerárquico, hay que establecer el número de clusters a calcular.

Tras el estudio de la muestra, viendo que las variables explicativas son de caracter cuantitativo y habiendo hecho las pruebas que permiten establacer el numero adecuado de grupos antes de proceder a la segmetancion, procederemos a aplicar un método no jerárquico. Concretamente, aplicaremos el algoritmo *PAM*.



Vemos com los dos grupos que están claramente diferenciados. Por debajo del valor cero en la primera dimensión formarán parte del primer cluster y por encima al segundo cluster.

```
## Warning in daisy(df_cluster_cp_cliente_t): binary variable(s) 229, 293,
## 314, 323, 336, 366, 428, 460, 474, 497, 506, 596 treated as interval scaled
```



El promedio de la silueta presenta un valor razonable. Con muy pocas observaciones clasificadas incorrectamente en el cluster 2, esto se debe a que son ligeramente diferentes que las observaciones de su propio grupo.

```
## Warning in write.csv2(df_cluster, "df_cluster.csv", sep = "|", row.names =
## FALSE): attempt to set 'sep' ignored
```