

Análisis de relaciones de interdependencia

Paula Corbatón Álvarez

2023-01-26

Ejercicio 1: Método de agrupamiento no jerárquico K-means

a. Primero comprobamos que variables de las que nos interesan son de cadena y las convertimos.

```
# utilizo la función: glimpse(Ej1) para ver las variables del dataset
```

Educación: donde PG = 0 y UG = 1

```
Ej1$education <- factor(Ej1$education, labels = c(0, 1))  
Ej1$education <- as.numeric(Ej1$education)  
Ej1$education[Ej1$education == 1] <- 0  
Ej1$education[Ej1$education == 2] <- 1
```

Localización: donde City = 0 y Suburb = 1

```
Ej1$location <- factor(Ej1$location, labels = c(0, 1))  
Ej1$location <- as.numeric(Ej1$location)  
Ej1$location[Ej1$location == 1] <- 0  
Ej1$location[Ej1$location == 2] <- 1
```

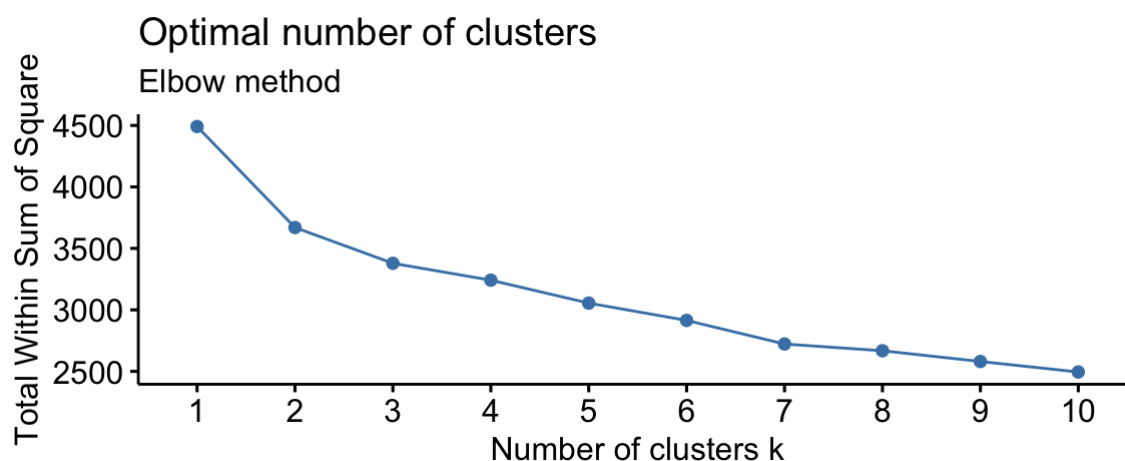
b. Ahora vamos a calcular el número de clústers que necesitamos. Para ello primero escalaremos nuestras variables numéricas:

b.1 Escalamos las variables:

```
Ej1.data <- Ej1[-7][,5:13]  
Ej1.scaled.data <- as.data.frame(scale(Ej1.data), center = TRUE, scale=TRUE)
```

b.2 Una vez escaladas las variables podemos determinar el numero de clusters óptimo: Para ello voy a utilizar “Elbow method” donde calculamos la suma de la distancia al cuadrado entre cada punto y el centroide en un grupo y lo comparamos con el número de clusters

```
fviz_nbclust(Ej1.scaled.data, kmeans, method = "wss") +  
labs(subtitle = "Elbow method")
```



Considero que el número óptimo de cústers es 3 puesto que me parece que a partir de este punto el gráfico comienza a moverse casi paralelo al eje X.

Ahora vamos a crear los clusters por el método K-means:

```
Ej1.km <- kmeans(Ej1.scaled.data, centers = 3) # Realizamos clustering
print(paste("between_SS / total_SS = ", (Ej1.km$betweenss/Ej1.km$totss)*100, "%"))
```

```
## [1] "between_SS / total_SS = 23.4708046474785 %"
```

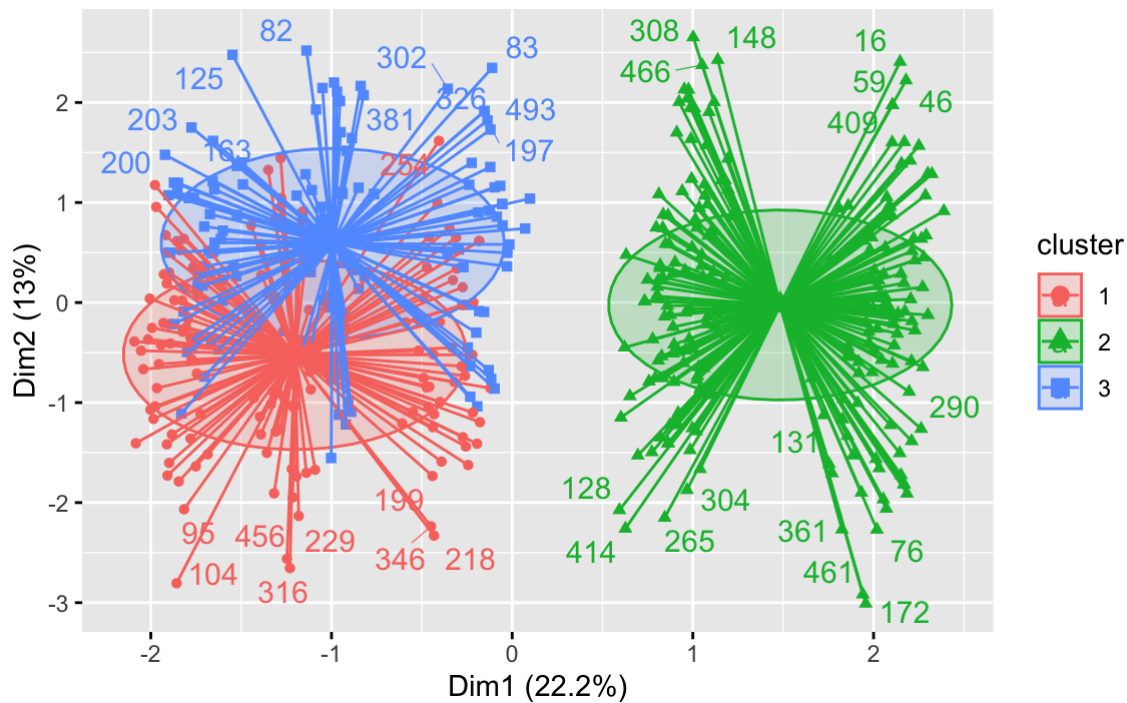
Aquí podemos ver a que grupo pertenece cada observación. Como podemos observar $\text{between_SS} / \text{total_SS} = 25\%$ teniendo en cuenta que hemos reducido 9 variables a 3, la explicación del 25% de las mismas es un porcentaje bastante aceptable.

Podemos visualizar los resultados:

```
Ej1.clusters <- Ej1.km$cluster
fviz_cluster(list(data = Ej1.scaled.data, cluster = Ej1.clusters), show.clust.cent
= T,
              ellipse.type = "euclid", star.plot= T, repel = T)
```

```
## Warning: ggrepel: 463 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

Cluster plot



Como podemos observar, la primera dimension explica el 22.2% de los casos y la segunda dimension explica el 13% de los casos.

c. Ahora vamos a interpretar cada uno de los clústeres en función de los valores de las variables de sus miembros. Para ello primero vamos a dividir los datos de las variables en función del cluster al que pertenecen

```
data.cluster1 <- Ej1.data %>%filter (Ej1.clusters==1)
data.cluster2 <- Ej1.data %>%filter (Ej1.clusters==2)
data.cluster3 <- Ej1.data %>%filter (Ej1.clusters==3)
#summary(Ej1.data)
```

```
summary(data.cluster1)
```

```
##      location      education      job_level      rating
## Min.      :0.0000   Min.      :0.000   Min.      :1.000   Min.      :1.000
## 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:1.000
## Median :0.0000   Median :0.000   Median :2.000   Median :2.000
## Mean      :0.4865   Mean      :0.473   Mean      :1.959   Mean      :1.784
## 3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:2.000
## Max.      :1.0000   Max.      :1.000   Max.      :3.000   Max.      :3.000
##      onsite      awards      certifications      salary
## Min.      :0.0000   Min.      :0.000   Min.      :0.0000   Min.      :24076
## 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:24076
## Median :0.0000   Median :5.000   Median :1.0000   Median :29805
## Mean      :0.4932   Mean      :4.601   Mean      :0.5473   Mean      :31620
## 3rd Qu.:1.0000   3rd Qu.:7.000   3rd Qu.:1.0000   3rd Qu.:42419
## Max.      :1.0000   Max.      :9.000   Max.      :1.0000   Max.      :42419
##      satisfied
## Min.      :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean      :0.3784
## 3rd Qu.:1.0000
## Max.      :1.0000
```

```
summary(data.cluster2)
```

```
##      location      education      job_level      rating
## Min.      :0.0000   Min.      :0.0000   Min.      :4.000   Min.      :1.00
## 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:2.00
## Median :0.0000   Median :0.0000   Median :4.000   Median :3.00
## Mean      :0.4836   Mean      :0.4977   Mean      :4.484   Mean      :3.23
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:5.000   3rd Qu.:4.00
## Max.      :1.0000   Max.      :1.0000   Max.      :5.000   Max.      :5.00
##      onsite      awards      certifications      salary
## Min.      :0.0000   Min.      :0.000   Min.      :0.000   Min.      :65715
## 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:65715
## Median :0.0000   Median :4.000   Median :0.000   Median :65715
## Mean      :0.4413   Mean      :4.479   Mean      :0.493   Mean      :75887
## 3rd Qu.:1.0000   3rd Qu.:7.000   3rd Qu.:1.000   3rd Qu.:86750
## Max.      :1.0000   Max.      :9.000   Max.      :1.000   Max.      :86750
##      satisfied
## Min.      :0.0000
## 1st Qu.:0.0000
## Median :1.0000
## Mean      :0.5399
## 3rd Qu.:1.0000
## Max.      :1.0000
```

```
summary(data.cluster3)
```

```
##      location      education      job_level      rating
##  Min.      :0.0000   Min.      :0.0000   Min.      :1.00   Min.      :3.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:4.000
##  Median :0.0000   Median :1.0000   Median :2.00   Median :4.000
##  Mean    :0.4748   Mean    :0.5036   Mean     :1.95   Mean     :4.201
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.00   3rd Qu.:5.000
##  Max.    :1.0000   Max.    :1.0000   Max.     :3.00   Max.     :5.000
##      onsite      awards      certifications      salary
##  Min.      :0.0000   Min.      :0.000   Min.      :0.000   Min.      :24076
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:24076
##  Median :0.0000   Median :4.000   Median :0.000   Median :29805
##  Mean    :0.4748   Mean    :4.561   Mean     :0.446   Mean     :31399
##  3rd Qu.:1.0000   3rd Qu.:7.000   3rd Qu.:1.000   3rd Qu.:42419
##  Max.    :1.0000   Max.     :9.000   Max.     :1.000   Max.     :42419
##      satisfied
##  Min.      :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean    :0.6619
##  3rd Qu.:1.0000
##  Max.    :1.0000
```

Para más información sobre los estadísticos de cada clúster ver Anexo (1c)

Cluster 1: contiene valores que se sitúan en la **media** del dataset de las variables: location, salary; valores que se se sitúan por **debajo de la media** del dataset en las variables: education, job_level, rating, satisfied; valores que se sitúan por **encima de la media** del dataset en la variable: certifications, onsite, Awards

Cluster 2: contiene valores que se sitúan en la **media** del dataset de las variables: location, education, rating, certifications, satisfied ; valores que se se sitúan por **debajo de la media** del dataset en las variables: onsite, awards; valores que se sitúan por **encima de la media** del dataset en la variable: job_level, salary

Cluster 3: contiene valores que se sitúan en la **media** del dataset de las variables: location, onsite, awards, salary; valores que se se sitúan por **debajo de la media** del dataset en las variables: job_level, certifications; valores que se sitúan por **encima de la media** del dataset en la variable: education, rating, satisfied

(si entendemos, por la descripción del data set que: - onsite: 0 significa que trabajan desde casa y 1 trabajan presencial - awards: hace referencia al nº de premios del empleado - certifications: 0 significa que no están certificados y 1 significa que si lo están)

Si quisiésemos etiquetar cada cluster, diríamos que el clúster 1 contiene mayoritariamente empleados certificados, que trabajan por lo general de forma presencial, y con un número de premios alto. Con un nivel de trabajo, calificación y satisfacción bajos, cuya educación es principalmente PG. El cluster 2 contiene empleados que por lo general trabajan desde casa y con pocos premios. Su nivel de trabajo y su salario son altos. El cluster 3 contiene empleados por lo general sin certificar y con un nivel de trabajo bajo. Tienen una calificación alta y están satisfechos en su mayoría, su educación es principalmente UG

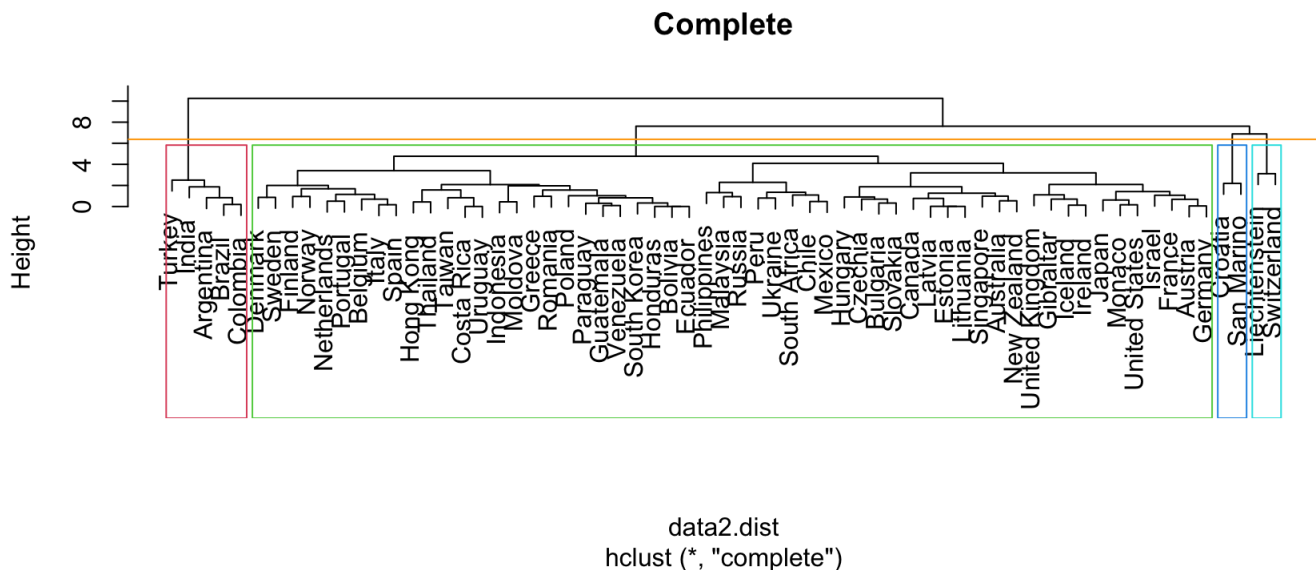
Ejercicio 2: Método de agrupamiento jerárquico

a. estandarizar los valores de las variables

```
Ej2.data <- Ej2[4:8]
data2 <- scale(Ej2.data, center = TRUE, scale=TRUE)
data2 <- as.data.frame(data2)
rownames(data2) <- Ej2$Country # asignamos a cada fila el nombre del país al que pertenece
```

b. Escogemos el número óptimo de clústers, para ello: Primero creamos la matriz de distancias y, a partir de ella los dendogramas:

```
data2.dist <- dist(data2, method = 'euclidean')
# plot(cs <- hclust(data2.dist, method = "single"), main = "Single") código para crear el dendograma usando el método simple
# plot(cs <- hclust(data2.dist, method = "average"), main = "Average") código para crear el dendograma usando el método average
hclust.ej2 <- hclust(data2.dist, method = "complete")
plot(hclust.ej2, main = "Complete")
rect.hclust(hclust.ej2, h = 6.375, border = 2:6)
abline(h=6.375, col="orange")
```



Considero que de todos los dendrogramas el mas equilibrado es el obtenido usando el método “Complete” por lo que utilizaré este para desarrollar mis conclusiones. **Para escoger el número optimo de clusters, tengo en cuenta que, cuanto mayor sea la distancia entre dos links mayores serán las diferencias entre sus características**

considero que el punto en el que se produce un salto más brusco es entre 2.25 y 10.25. Por lo tanto, cortaremos el dendograma en el punto 6.375. Resultando 4 como número óptimo de clusters

c. interpretar cada uno de los clústeres en función de los valores de las variables de sus miembros.

```
Ej2.clusters <- cutree(hclust.ej2, k = 4) #Dividimos el dataframe según los clústers
Ej2.grupo1 <- Ej2.data %>%filter (Ej2.clusters==1)
Ej2.grupo2 <- Ej2.data %>%filter (Ej2.clusters==2)
Ej2.grupo3 <- Ej2.data %>%filter (Ej2.clusters==3)
Ej2.grupo4 <- Ej2.data %>%filter (Ej2.clusters==4)
```

```
summary(Ej2.grupo1)
```

##	No. of TV Shows	No. of Movies	Cost Per Month Basic	Cost Per Month Standard
##	Min. :2930	Min. :1606	Min. :1.970	Min. :3.000
##	1st Qu.:3154	1st Qu.:1709	1st Qu.:2.640	1st Qu.:6.300
##	Median :3156	Median :1810	Median :3.740	Median :6.610
##	Mean :3224	Mean :1817	Mean :3.454	Mean :5.976
##	3rd Qu.:3162	3rd Qu.:1835	3rd Qu.:4.310	3rd Qu.:6.860
##	Max. :3718	Max. :2125	Max. :4.610	Max. :7.110
##	Cost Per Month Premium			
##	Min. :4.020			
##	1st Qu.:8.600			
##	Median :9.260			
##	Mean :8.354			
##	3rd Qu.:9.930			
##	Max. :9.960			

summary(Ej2.grupo2)

##	No. of TV Shows	No. of Movies	Cost Per Month Basic	Cost Per Month Standard
##	Min. :2449	Min. :1388	Min. : 5.640	Min. : 8.46
##	1st Qu.:3155	1st Qu.:1641	1st Qu.: 7.990	1st Qu.:10.99
##	Median :3555	Median :1858	Median : 8.990	Median :11.68
##	Mean :3636	Mean :1848	Mean : 8.623	Mean :12.19
##	3rd Qu.:4057	3rd Qu.:1994	3rd Qu.: 9.030	3rd Qu.:13.54
##	Max. :5234	Max. :2387	Max. :12.000	Max. :15.24
##	Cost Per Month Premium			
##	Min. :10.93			
##	1st Qu.:13.54			
##	Median :14.72			
##	Mean :15.81			
##	3rd Qu.:18.06			
##	Max. :20.32			

summary(Ej2.grupo3)

##	No. of TV Shows	No. of Movies	Cost Per Month Basic	Cost Per Month Standard
##	Min. :1675	Min. :373.0	Min. :9.03	Min. :11.29
##	1st Qu.:1740	1st Qu.:429.5	1st Qu.:9.03	1st Qu.:12.13
##	Median :1806	Median :486.0	Median :9.03	Median :12.98
##	Mean :1806	Mean :486.0	Mean :9.03	Mean :12.98
##	3rd Qu.:1872	3rd Qu.:542.5	3rd Qu.:9.03	3rd Qu.:13.82
##	Max. :1937	Max. :599.0	Max. :9.03	Max. :14.67
##	Cost Per Month Premium			
##	Min. :13.54			
##	1st Qu.:15.23			
##	Median :16.93			
##	Mean :16.93			
##	3rd Qu.:18.62			
##	Max. :20.32			

summary(Ej2.grupo4)

##	No. of TV Shows	No. of Movies	Cost Per Month Basic	Cost Per Month Standard
##	Min. :1712	Min. :1336	Min. :12.88	Min. :20.46
##	1st Qu.:2198	1st Qu.:1465	1st Qu.:12.88	1st Qu.:20.46
##	Median :2683	Median :1594	Median :12.88	Median :20.46
##	Mean :2683	Mean :1594	Mean :12.88	Mean :20.46
##	3rd Qu.:3168	3rd Qu.:1723	3rd Qu.:12.88	3rd Qu.:20.46
##	Max. :3654	Max. :1852	Max. :12.88	Max. :20.46
##	Cost Per Month Premium			
##	Min. :26.96			
##	1st Qu.:26.96			
##	Median :26.96			
##	Mean :26.96			
##	3rd Qu.:26.96			
##	Max. :26.96			

Grupo 1: junto con el grupo 2 posee el mayor número de **TV shows** (entre 2930 y 3718 con una media de 3224) y de **Movies** (entre 1606 y 2125 con una media de 1817), el precio de suscripción **Basic** oscila entre 1.970 y 4.610 con una media de 3.454, el **Standard** varía entre 3.000 y 7.110 con una media de 5.976 y el **Premium** oscila entre 4.020 y 9.960 con una media de 8.354.

Grupo 2: posee el menor número de **TV shows** (entre 2449 y 5234 con una media de 3636) y de **Movies** (entre 1388 y 2387 con una media de 1848), el precio de suscripción **Basic** oscila entre 5.640 y 12.000 con una media de 8.623, el **Standard** varía entre 8.46 y 15.24 con una media de 12.19 y el **Premium** oscila entre 10.93 y 20.32 con una media de 15.81

Grupo 3: junto con el grupo 1 posee el mayor número de **TV shows** (entre 1675 y 1937 con una media de 1806) y de **Movies** (entre 373.0 y 599.0 con una media de 486.0), el precio de suscripción **Basic** es de 9.03, el **Standard** varía entre 11.29 y 14.67 con una media de 12.98 y el **Premium** oscila entre 13.54 y 20.32 con una media de 16.93

Grupo 4: con relación a los demás grupos, posee un número medio de **TV shows** (entre 1712 y 3654 con una media de 2683) y de **Movies** (entre 1336 y 1852 con una media de 1594), el precio de suscripción **Basic** es de 12.88, el **Standard** es de 20.46 y el **Premium** es de 26.96

Destacar que en el grupo 1 y 2, a pesar de tener más o menos el mismo número de Tv Shows y Movies, la suscripciones de los países del grupo 2 son mucho más caras que las de los del grupo 1. El grupo 3 a pesar de ser el grupo con menos películas y series posee un rango de precio similar al grupo 2. El grupo 3 tiene un valor intermedio (entre los grupos (1,2) y (3)) de películas y series pero es el más caro.

Grupo 1 tiene mucha variedad de películas y series a un precio bajo

Grupo 2 tiene mucha variedad de películas y series a un precio alto

Grupo 3 tiene poca variedad de películas y series a un precio alto

Grupo 4 tiene un número medio de películas y series a un precio muy alto

Ejercicio 3: Análisis de componentes principales

- a. Vamos a escoger el número óptimo de componentes (1 o 2). Para ello primero escalaremos los datos que vayamos a utilizar, y aplicamos el análisis de componentes principales:

```
data.Ej3 <- scale(Ej3[3:8], center = TRUE, scale=TRUE)# selecciono las variables que
me pide el ejercicio y las escalo (center = TRUE (cada variable será restada a la med
ia de cada una de las mismas) y scale = TRUE (divide por la desviación típica))
# scree(data.Ej3)
acp <- prcomp(data.Ej3) # hacemos el análisis de componentes principales
summary(acp)
```



```
## Importance of components:
##               PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation    1.6263  1.3048  0.8603  0.75082  0.53449  0.2509
## Proportion of Variance 0.4408  0.2838  0.1233  0.09396  0.04761  0.0105
## Cumulative Proportion 0.4408  0.7246  0.8479  0.94189  0.98950  1.0000
```

Como podemos observar en la tabla Importance of components Si escogemos PC1 podremos explicar el 44.08% de la variabilidad de los datos mientras que si escogemos PC1 y PC2 podemos explicar el 72.46%. Por este motivo considero que es mejor escoger dos componentes.

```
#print(acp)
```

b. valoración de las condiciones de aplicación: Vamos a realizar el test de KMO a nuestros datos:

```
KMO(data.Ej3)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data.Ej3)
## Overall MSA = 0.66
## MSA for each item =
##           Fresh           Milk           Grocery           Frozen
##           0.66           0.85           0.62           0.66
## Detergents_Paper    Delicassen
##           0.61           0.59
```

Como podemos observar, el valor de Overall MSA = 0.66, por lo tanto, la realización del ACP es justificada (ya que es mayor de 0.5).

c. valorar si es necesario rotar los componentes

Imprimimos la matriz de rotación:

```
acp$rotation
```

```
##               PC1      PC2      PC3      PC4      PC5
## Fresh      -0.04288396 -0.52793212 -0.81225657 -0.23668559  0.04868278
## Milk       -0.54511832 -0.08316765  0.06038798 -0.08718991 -0.82657929
## Grocery    -0.57925635  0.14608818 -0.10838401  0.10598745  0.31499943
## Frozen     -0.05118859 -0.61127764  0.17838615  0.76868266  0.02793224
## Detergents_Paper -0.54864020  0.25523316 -0.13619225  0.17174406  0.33964012
## Delicassen -0.24868198 -0.50420705  0.52390412 -0.55206472  0.31470051
##           PC6
## Fresh      0.03602539
## Milk       0.03804019
## Grocery    -0.72174458
## Frozen     0.01563715
## Detergents_Paper 0.68589373
## Delicassen 0.07513412
```

Esta tabla nos muestra el grado de relación de cada variable con cada PC. Por ejemplo Milk toma un valor absoluto de 0.54511832 para PC1 mientras que para PC2 tiene un valor absoluto de 0.08316765. Lo contrario pasa con Fresh, que para PC1 toma un valor absoluto de 0.04288396 y para PC2 tiene un valor absoluto de 0.52793212

En la siguiente tabla podemos observar la correlación entre las variables y los componentes principales:

```
cor(Ej3[3:8], acp$x[,1:2])
```

##	PC1	PC2
## Fresh	-0.06974378	-0.6888622
## Milk	-0.88654635	-0.1085197
## Grocery	-0.94206631	0.1906204
## Frozen	-0.08324992	-0.7976140
## Detergents_Paper	-0.89227411	0.3330361
## Delicassen	-0.40444082	-0.6579050

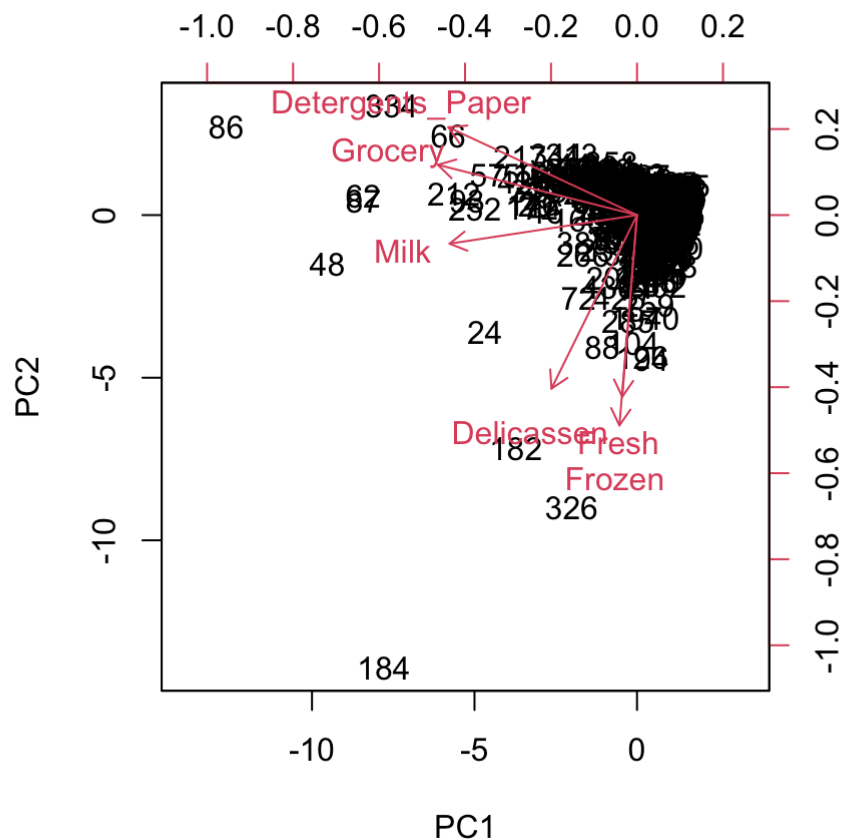
Considero que:

- Cada factor tiene unos pocos pesos altos y los demás próximos a cero.
- Cada variable no esta saturada más que en un factor.
- No deben existen factores con la misma distribución, es decir, dos factores distintos presentan distribuciones diferentes de cargas altas y bajas.

Por lo tanto no es necesario rotar los componentes principales.

d. etiquetar los componentes:

```
biplot(acp, scale = 0)
```



En este gráfico podemos observar como Detergents_paper, Grocery y Milk se mueven más en el eje de PC1 mientras que Delicatessen, Fresh y Frozen se mueven más en el eje de PC2. Esto nos indica que PC1 explica mejor las variables Detergents_paper, Grocery y Milk, y PC2 Delicatessen, Fresh y Frozen. Esto también lo podemos observar con la matriz de rotación (ver apartado c)

Podemos crear un nuevo dataset con estos dos componentes que representaría en un 72.46% el dataset "Ej3"

```
acp.Ej3 <- as.data.frame(acp$x[,1:2])
head(acp.Ej3)
```

```
##           PC1           PC2
## 1 -0.1930708  0.3047531
## 2 -0.4339260  0.3280392
## 3 -0.8102210 -0.8141689
## 4  0.7777625 -0.6520115
## 5 -0.1660982 -1.2699881
## 6  0.1559924  0.2948054
```

Anexo

1. Ejercicio 1c

- Cluster 1: contiene una mezcla más o menos uniforme de **location** con unas pocas más observaciones de Ciudad que de Suburbio, **education** también se sitúa entorno a la media con un poco más de PG que de UG. Las puntuaciones de **job_level** se sitúan entre 1 y 3 con una media aproximada de 2 (esto está por debajo de la media de las observaciones del dataset, que se sitúa entre 1 y 5 y tiene una media de 3). En cuanto a **rating**, al igual que el nivel de trabajo se sitúa por debajo de la media de las observaciones, con una puntuación aproximada de 2 y unos valores que varían entre 1 y 3. La variable **onsite** nos muestra que en este cluster hay una mezcla más o menos uniforme de personas que trabajan presencial y personas que trabajan online. **Awards** tiene una media de 4.6, con valores que oscilan entre 0 y 9 por lo que también parece que hay una mezcla uniforme de estos datos. **Certifications** tiene una media de 0.55 por lo que los empleados de este cluster están mayormente certificados. **Salary** varía entre 24076 y 42419 con una media 31620 y una mediana 29805 ambas bastante por debajo del conjunto de observaciones. **Satisfied** tiene una media de 0.3784 lo que indica que los individuos de este cluster están mayormente insatisfechos.
- Cluster 2: la variable **location** contiene valores que se sitúan en la media con unas pocas más observaciones de Ciudad que de Suburbio. **Education** también se sitúa entorno a la media con un poco más de PG que de UG. Los valores de **job_level** varían entre 4 y 5 con una media de 4.48 por lo tanto se sitúan por encima de la media del dataset. **Rating** se sitúa entre 1 y 5 con una media aproximada de 3, situándose en la media de las observaciones. **onsite** tiene una media de 0.44 (ligeramente inferior a la media del dataset). **Awards** toma valores entre 0 y 9 con una media de 4.479, lo que nos indica que en este cluster predominan valores medios de esta variable (un poco más bajos que en los clusters 1 y 3). **Certifications** tiene una media de 0.49 por lo que en este cluster hay más o menos el mismo número de empleados con y sin certificado. **salary** oscila entre 65715 y 86750 con una media de 75887 y una mediana de 65715, ambas superiores a las del dataset. **satisfied** toma una media de 0.54, lo que nos indica que los empleados están mayormente satisfechos.
- Cluster 3: Contiene una mezcla más o menos uniforme de las variables **location** (con una media aproximada de 0.47, que nos indica unas pocas más observaciones pertenecientes a Ciudad en comparación con Suburbio) y **education** (con una media de 0.53 lo que nos indica unas pocas más observaciones de UG que de PG). **job_level** se sitúa entre 1 y 3 con una media de aproximada de 2 (por debajo de la media de las observaciones del dataset). **rating** se sitúa entre 3 y 5, con una media aproximada de 4, (por encima de la media de las observaciones). **onsite** tiene una media de 0.47 (ligeramente superior a la del dataset). **awards** toma valores entre 0 y 9 con una media de 4.561, igual a la del dataset. **certifications** tiene una media de 0.45, por lo que este cluster es el que tiene más

individuos sin certificado. **salary** oscila entre 24076 y 42419 con una media de 31399 y una mediana de 29805, ambas en la media del dataset. **satisfied** tiene una media de 0.6619, lo que nos indica que los empleados de este cluster son los más satisfechos.