Data Visualization - Trabajo III

Karen Romero, Ari Romero, Nicole Lastra

26-01-2022

Ejercicio 1

Tenemos un dataset que contiene una muestra de 300 pizzas de 10 marcas diferentes y se desea caracterizarlas en cuanto a sus atributos nutricionales, además de conocer el grado de similaridad entre las marcas.

El dataset se encuentra en https://query.data.world/s/adyg4hhiz4zleyq3uufqld6t7fku2j y puede cargarse directamente desde esa URL.

Los atributos descritos en la metadata del dataset son los siguientes:

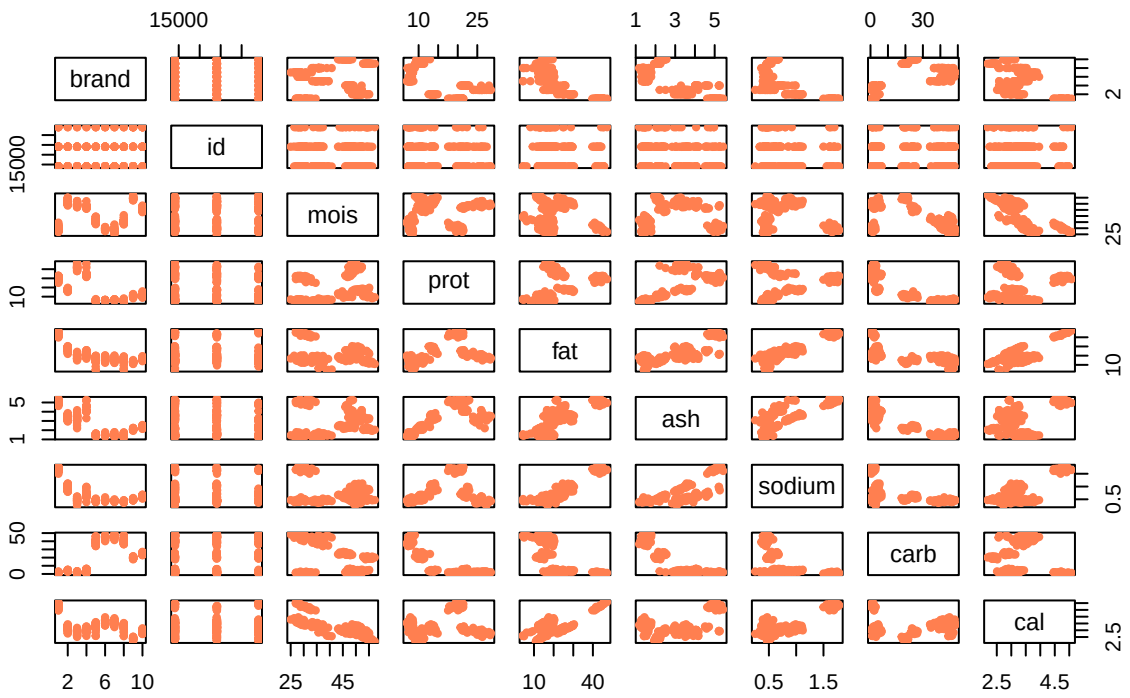
- brand: Pizza brand (class label)
- id: Muestra analizada
- mois: Cantidad de agua por 100 gramos de la muestra
- prot: Cantidad de proteina por 100 gramos de la muestra
- fat: Cantidad de grasa por 100 gramos de la muestra
- ash: Cantidad de cenizas por 100 gramos de la muestra
- sodium: Cantidad de sodio por 100 gramos de la muestra
- carb: Cantidad de carbohidratos por 100 gramos de la muestra
- cal: Cntidad de calorías por 100 gramos de la muestra

Por favor responda las siguientes preguntas realizando un análisis de correlación entre variables y una reducción dimensional mediante el método PCA (Principal Component Analysis).

a) ¿Qué variables considera usted que podría eliminarse a priori del análisis, ya que no aportan nueva información para el análisis solicitado? Justifique su respuesta. Se comienza cargando los datos y revisando que no existan incongruencias que puedan significar un trabajo extra, previo al analisis a realizar.

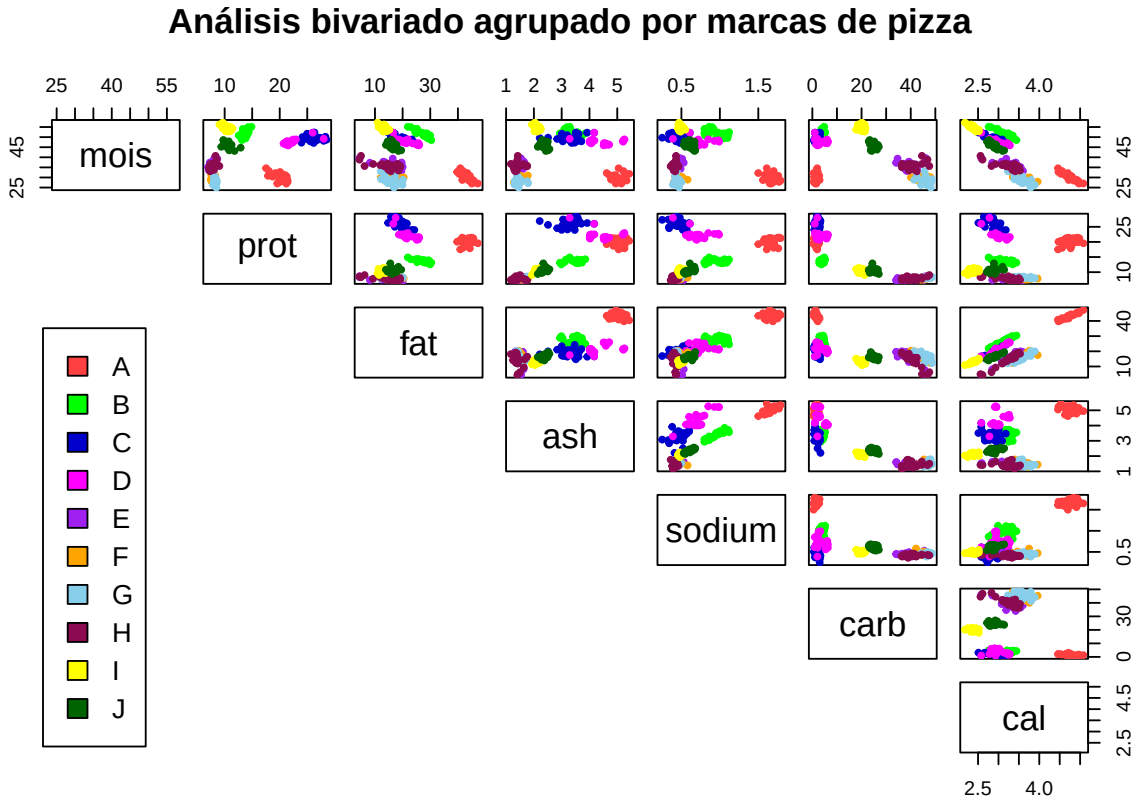
```
##   brand    id  mois  prot   fat  ash sodium carb  cal
## 1     A 14069 27.82 21.43 44.87 5.11   1.77 0.77 4.93
## 2     A 14053 28.49 21.26 43.89 5.34   1.79 1.02 4.84
## 3     A 14025 28.35 19.99 45.78 5.08   1.63 0.80 4.95
## 4     A 14016 30.55 20.15 43.13 4.79   1.61 1.38 4.74
## 5     A 14005 30.49 21.28 41.65 4.82   1.64 1.76 4.67
## 6     A 14075 31.14 20.23 42.31 4.92   1.65 1.40 4.67
```

Con los datos revisados, se procede a graficar la "Matriz de correlación" para ver si existe evidencias suficientes, desde lo visual, para descartar alguna de las variables.



Al observar el gráfico a priori, se pueden descartar rápidamente la variable "id", ya que no entrega infomación relevante para efectos de este estudio. Además, se observa una posible correlación entre las variables "fat" y "sodium", pero al no contar con más informción que la visual, se considera pertinente no descartarlas hasta una etapa futura, en donde se cuente con pruebas concretas. Lo mismo se aplica para el par de variables "fat" y "cal".

b) Genere un análisis visual bivariado (pares de variables) donde cada marca de pizza tenga un color de identificación diferente. Se utiliza la función "pairs" para generar la visualización, teniendo cuidado de asignar a cada marca un color específico. Se seleccionan las variables mois, prot, ash, sodium, carb y cal, ya que en el apartado anterior, pese a identificar algunas posibles candidatas a correlación, no se descartó ninguna (no se considera la variable brand, ya que es una variable no numérica, sin embargo la agrupación se realiza dado esta).



c) Ejecute una reducción de dimensionalidad mediante PCA e indique qué porcentaje de la varianza pueden explicar las 2 primeras dimensiones del resultado obtenido. Como observación, se debe recordar que en procedimientos generales se debería primero estandarizar las variables, pero en este caso no será necesario realizar este proceso, ya que todas las variables están medidas por cada "100 gramos de muestra de las pizzas en estudio" (mismo estandar).

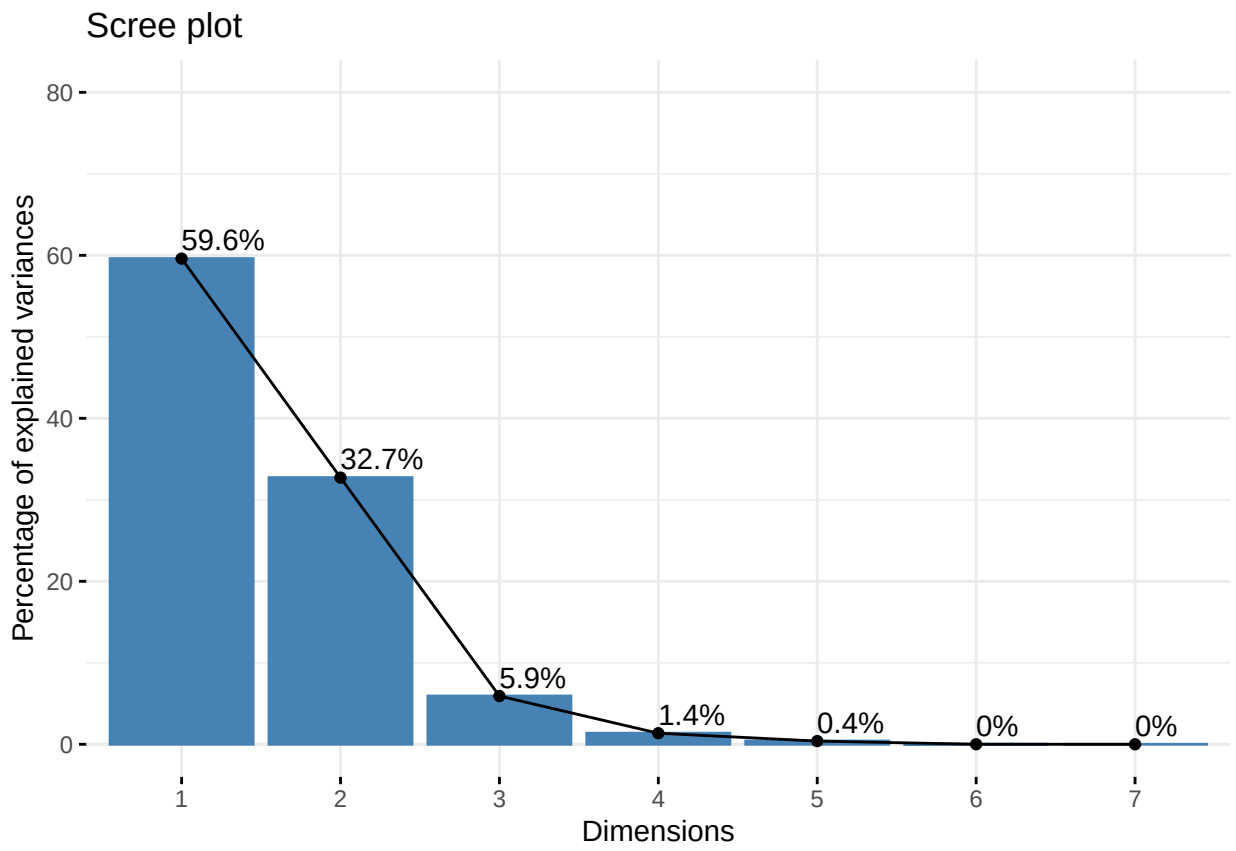
```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 300 individuals, described by 7 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
```

```
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"


##
## Call:
## PCA(X = pizza[2:8], scale.unit = TRUE, ncp = 7, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               4.172   2.290   0.415   0.095   0.028   0.000   0.000
## % of var.             59.597  32.721   5.922   1.360   0.395   0.005   0.000
## Cumulative % of var.  59.597  92.318  98.240  99.600  99.995 100.000 100.000
##
## Individuals (the 10 first)
##            Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 1      |  5.691 |  5.010  2.006  0.775 |  2.679  1.045  0.222 |  0.039  0.001
## 2      |  5.641 |  5.024  2.017  0.793 |  2.529  0.931  0.201 |  0.097  0.008
## 3      |  5.501 |  4.805  1.845  0.763 |  2.674  1.040  0.236 |  0.075  0.005
## 4      |  5.025 |  4.470  1.596  0.791 |  2.285  0.760  0.207 |  0.120  0.012
## 5      |  4.977 |  4.472  1.598  0.807 |  2.159  0.678  0.188 |  0.001  0.000
## 6      |  5.008 |  4.505  1.621  0.809 |  2.168  0.684  0.187 |  0.175  0.025
## 7      |  4.788 |  4.315  1.488  0.812 |  2.057  0.616  0.185 |  0.000  0.000
## 8      |  5.337 |  4.758  1.809  0.795 |  2.353  0.806  0.194 |  0.010  0.000
## 9      |  5.549 |  4.855  1.883  0.765 |  2.681  1.046  0.233 |  0.101  0.008
## 10     |  5.590 |  4.916  1.931  0.774 |  2.659  1.029  0.226 | -0.069  0.004
##          cos2
## 1       0.000 |
## 2       0.000 |
## 3       0.000 |
## 4       0.001 |
## 5       0.000 |
## 6       0.001 |
## 7       0.000 |
## 8       0.000 |
## 9       0.000 |
## 10      0.000 |
##
## Variables
##           Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
## mois   |  0.132  0.419  0.017 | -0.951 39.473  0.904 |  0.271 17.780  0.074 |
## prot   |  0.774 14.346  0.598 | -0.408  7.274  0.167 | -0.480 55.656  0.231 |
## fat    |  0.912 19.951  0.832 |  0.355  5.493  0.126 |  0.128  3.972  0.016 |
## ash    |  0.964 22.268  0.929 | -0.168  1.232  0.028 | -0.036  0.317  0.001 |
## sodium |  0.890 18.984  0.792 |  0.305  4.067  0.093 |  0.293 20.718  0.086 |
## carb   | -0.868 18.055  0.753 |  0.485 10.260  0.235 | -0.034  0.273  0.001 |
## cal    |  0.499  5.977  0.249 |  0.859 32.201  0.738 | -0.073  1.284  0.005 |
```

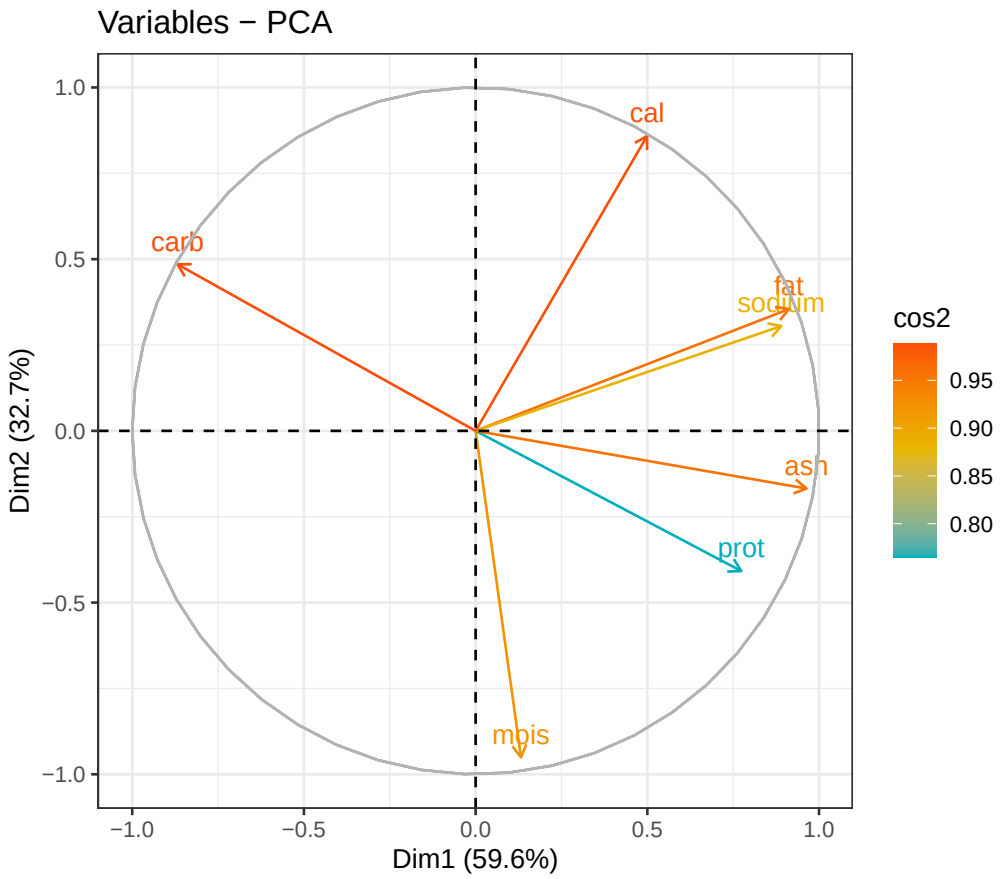
Como se puede observar en los resultados acumulados de la dim2, el porcentaje de la varianza explicado por las 2 primeras dimensiones es de un $92,318\%$.

Un método alternativo para evidenciar el porcentaje explicado de la varianza por cada dimensión, es observar un "Scree Plot" usando fviz_eig(), el cual gráfica los valores propios ordenados de mayor a menor, tal como se aprecia a continuación.



Efectivamente se corroboran los resultados, evidenciando desde el "Scree Plot" que las primermas dos dimensiones explican un $92,318\%$ de la variación.

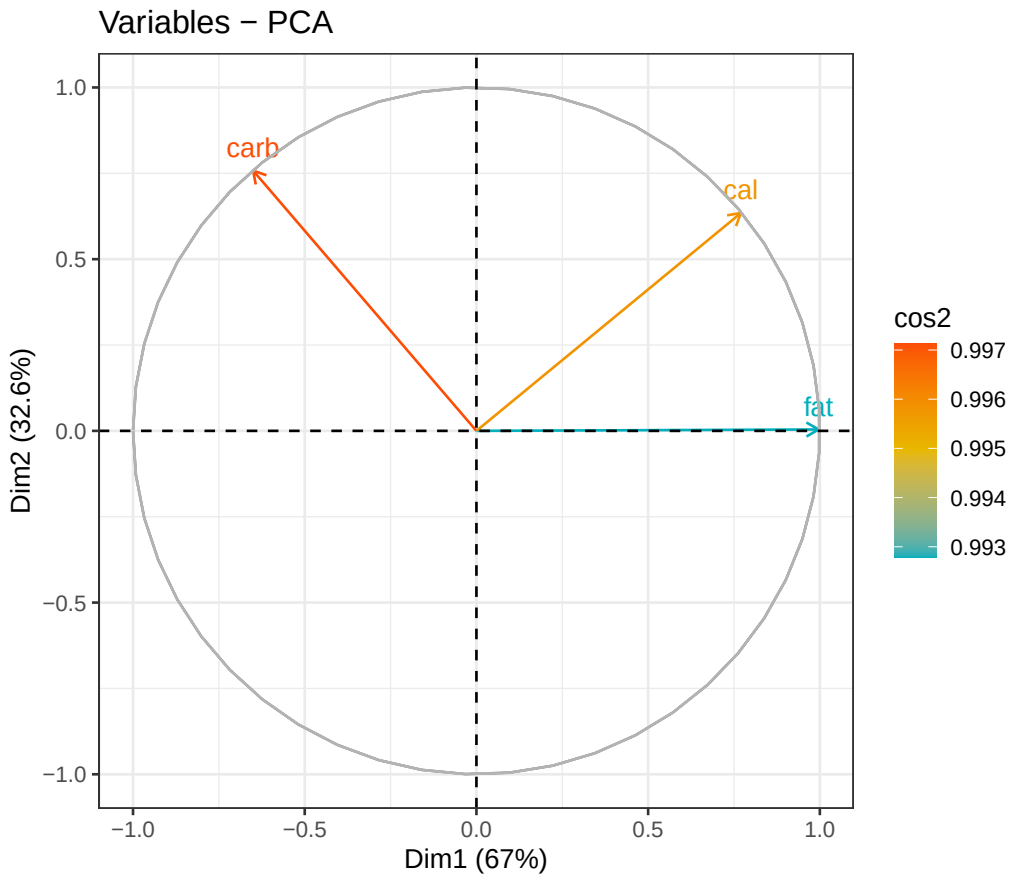
d) Genere el gráfico de circunferencia1 con las variables originales del dataset. ¿Es consistente esta visualización con la(s) variable(s) que propuso eliminar en la pregunta a)? Justifique su respuesta.



Para evidenciar mejor la información, se utilizó una gradiente entre tonos anaranjados y turqueza: - Mientras más cerca del azul esté la flecha, menos aporta a explicar la varianza, como es el caso de la variable “prot” la cual está relacionada con las proteínas en las muestras por cada 100 gramos de estas. - Mientras más cerca del anaranjado esté la flecha, más aporta a explicar la varianza, como es el caso de las variables “cal”, relacionada a las calorías y “carb”, relacionada a los carbohidratos.

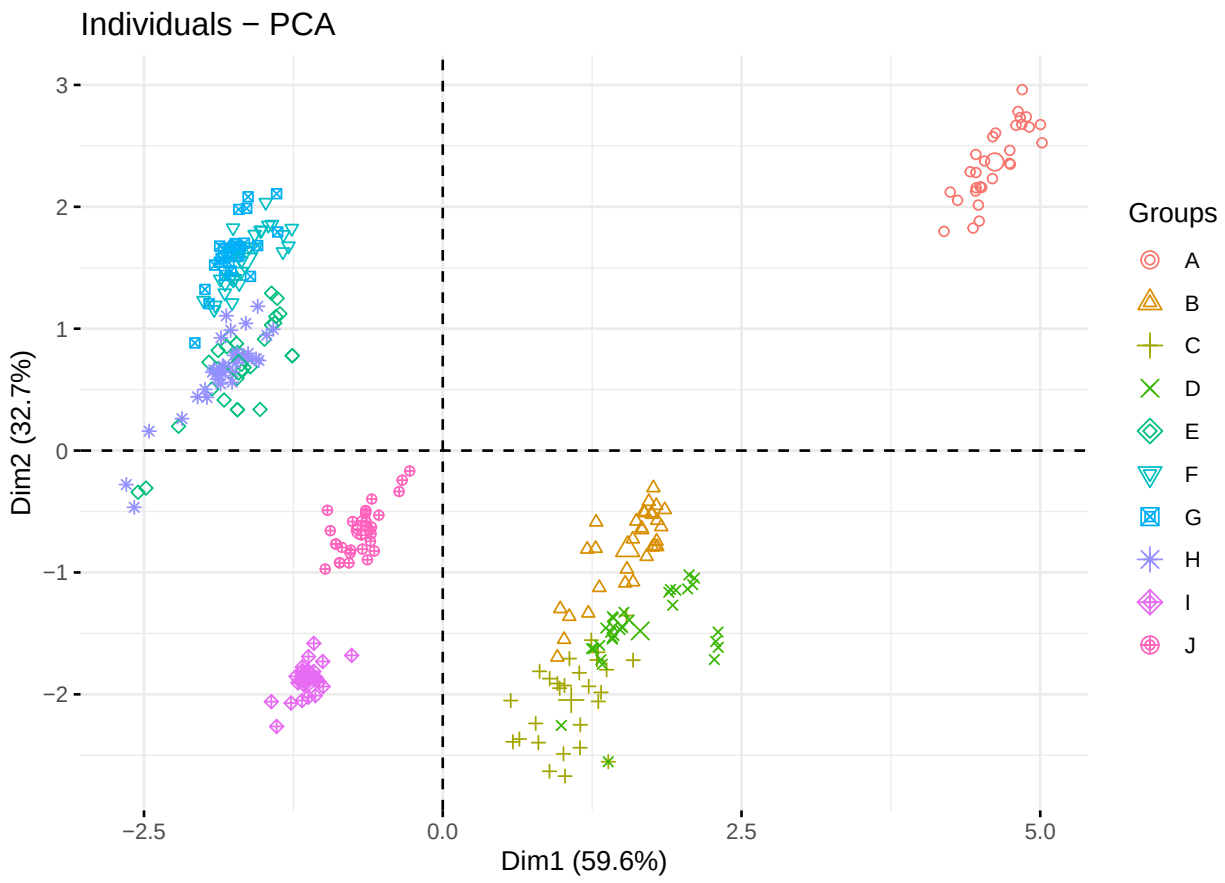
Respecto a las correlaciones, se observa que las variables “fat” y “sodium” estan directamente correlacionadas, mientras que en el caso de “carb” y “prot” esta correlación es carácter negativo. Al comparar estas concusines, respecto a lo observad a priori en el primer apartado de este trabajo, sólo las conjeturas respecto a las variables “fat” y “sodium” etaban correctas.

Finalmente, basado en todas las apreciaciaciones anteriores, sería oportuno descartar las variables “prot”, “mois”, sodium“, y también”ash" ya que se consideraán las res variables que más explican el modelo, es decir, se deberían considerar “carb”, “cal” y “fat” para efectos de estudio.

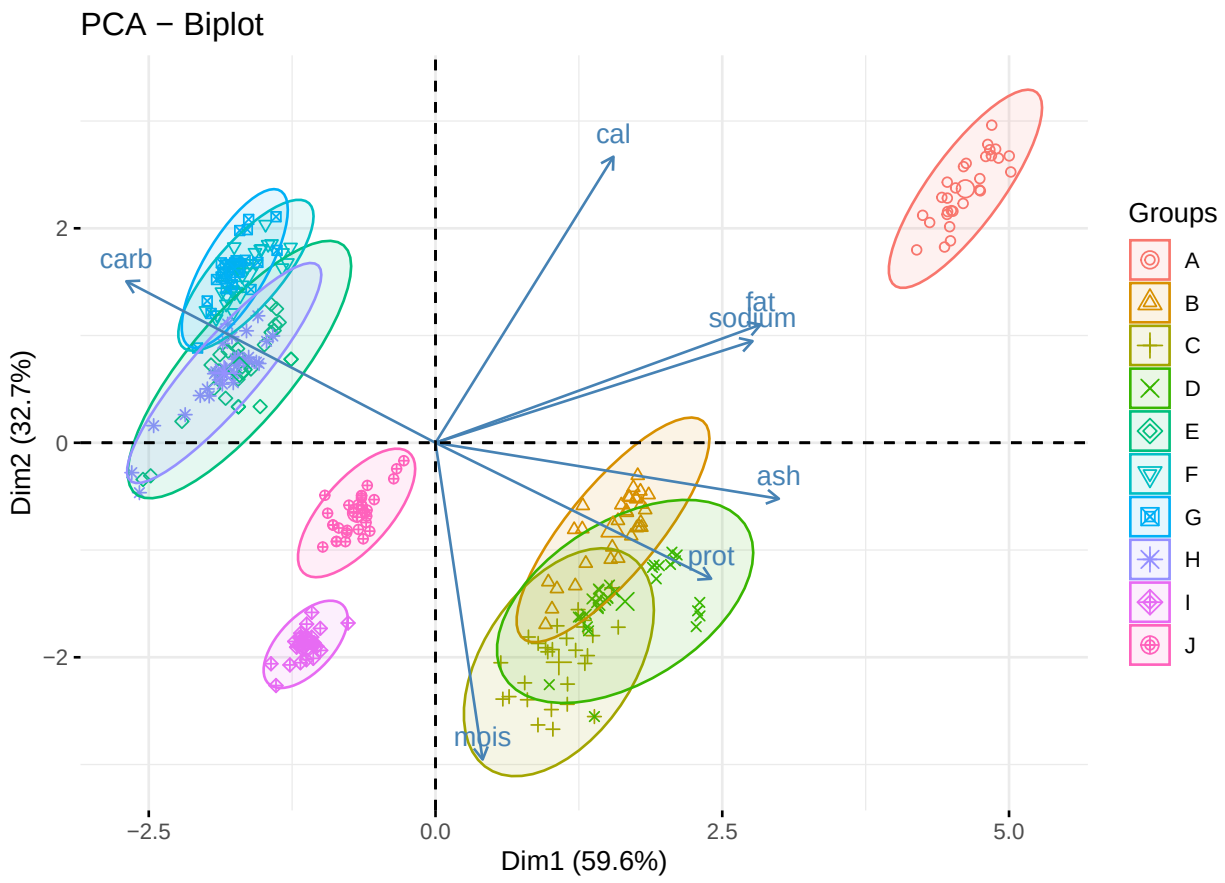


Se puede observar que al hacer la reducción de dimensión, estas variables explican un $99,6\%$ de la varianza.

e) Genere el gráfico de circunferencia con los 300 puntos de la muestra, diferenciando por color las 10 marcas presentes.+



f) Genere el biplot con las variables y muestras de pizzas.



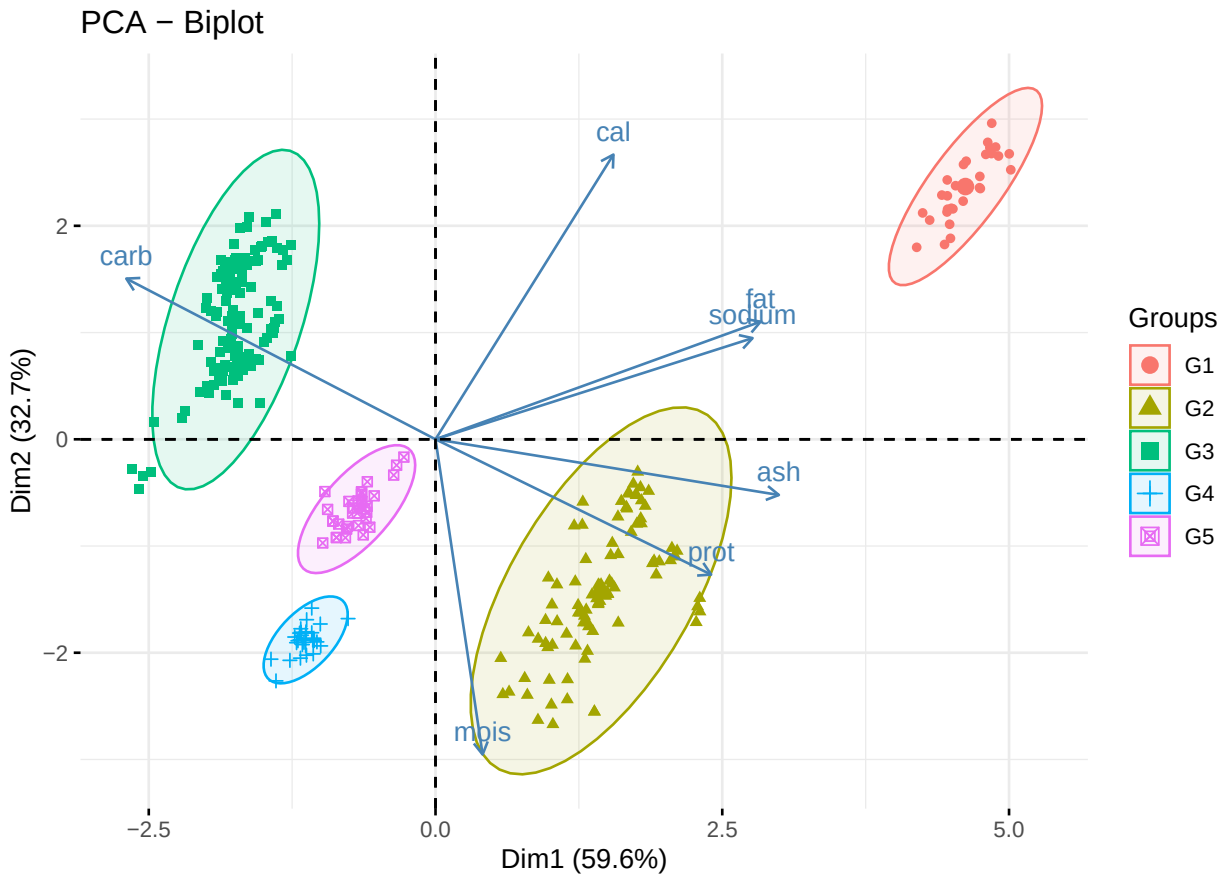
g) A partir de la visualización, es decir, sin ejecutar métodos de clustering, agrupe las marcas en 5 grupos (G1 a G5), de acuerdo con su similaridad nutricional. ¿Qué marcas componen cada uno de los 5 grupos y cuál es el número de muestras en cada uno? Sin realizar ningún método de clustering, se puede observar áreas específicas delimitadas en el gráfico anterior, las cuales están relacionadas de forma específica con ciertas marcas, lo que lo hace muy conveniente para efectos de separación.

Los grupos bajo este concepto quedarían como: - G1: Marca A - G2: Marcas B, C y D - G3: Marcas E, F, G y H - G4: Marca I - G5: Marca J

Con la ayuda de las funciones `revalue` y `mutate`, se define el respectivo data frame con la nueva distribución de grupos, tal como se muestra a continuación:

```
##   brand  mois  prot   fat  ash sodium carb  cal
## 1    G1 27.82 21.43 44.87 5.11   1.77 0.77 4.93
## 2    G1 28.49 21.26 43.89 5.34   1.79 1.02 4.84
## 3    G1 28.35 19.99 45.78 5.08   1.63 0.80 4.95
## 4    G1 30.55 20.15 43.13 4.79   1.61 1.38 4.74
## 5    G1 30.49 21.28 41.65 4.82   1.64 1.76 4.67
## 6    G1 31.14 20.23 42.31 4.92   1.65 1.40 4.67
```

h) Marque los centroides (puntos promedio) de cada grupo en el biplot. Dado a que no se realizó trabajo de clustering por medio de `kmeans`, simplemente se debe realizar un `biplot` de los grupos ya formados en el apartado anterior, ya que esta visualización incluye lo solicitado (cabe señala que como son muchos datos, el centro puede tender a perderse entre ellos, aún teniendo un tamaño más grande que el resto de las figuras).



i) ¿Qué grupo debe evitar una persona qué está con una dieta baja en carbohidratos? ¿baja en sodio? De acuerdo a lo observado en el gráfico del apartado anterior, una persona que debe seguir una dieta baja en carbohidratos debe evitar el consumo de pizzas del grupo G3, esto quiere decir, pizzas de las marcas `E`, `F`, `G` y `H`. Para el caso de la persona con dieta baja en sodio, el grupo a evitar sería G1, es decir las pizzas de la marca `A`.