Centro de Investigación en Matemáticas

Unidad Monterrey

Proyecto No. 1.

Métodos multivariados de Análisis de Datos

Análisis de nutrientes en pizzas.

Ricardo Cruz Sánchez
Rolando Corona Jiménez

Índice

1. Introducción.

La bromatología es la ciencia encargada del análisis de los nutrientes contenidos en los alimentos. Los estudios relacionados con esta disciplina cobran importancia al considerar que existen cantidades recomendadas en la ingesta diaria de cualquier individuo y el incremento o decremento de las cantidades repercute directamente en la salud del la persona.

Las porciones de nutrientes que posee cierto alimento en particular, se determinan a través de diversas pruebas de laboratorio, las cuales buscan ser lo más precisas y por lo general reportan los nutrientes contenidos en 100 gramos del alimento en cuestión.

Los alimentos considerados como *comida rápida* suelen tener cantidades elevadas en los nutrientes, por lo que la ingesta de este tipo de alimentos suele sobrepasar o aportar considerablemente a los límites recomendados.

Particularmente, la pizza, tiende a ser uno de los alimentos con los cuales se sobrepasan los límites de nutrientes recomendados. Esto, principalmente, se debe a que es una mezcla de ingredientes cuya aportación nutrimental es elevada, a saber, harina, tomate, queso y carne.

En el presente trabajo, se considera una base de datos relativa a las pruebas nutrimentales realizadas en distintas pizzas y con base a estos datos se pretende realizar un análisis multivariado.

El documento se divide en 4 secciones, la primera de ellas realiza un análisis exploratorio de los datos, donde se presentan las variables y resumen de datos a través medidas en forma de gráficas. Esta exploración se hace con el fin de conocer mejor las variables y tener una idea a priori de los resultados esperados. Posteriormete, la segunda sección implementa los modelos de reducción de dimensiones, enfocados en trabajar en un espacio que permita una mejor visualización de los datos. La siguiente sección, presenta los modelos de clasificación, con los que se espera asignar una categoría a cada tupla de datos, considerados como variables predictivas. Por último, la última sección resume los resultados obtenidos y plantea las mejoras que se pueden realizar a este tipo de trabajos.

El documento, así como los códigos, datos, gráficas y resultados de este trabajo se encuentran en el siguiente repositorio: https://github.com/rolandocj/proyecto-pizzas/tree/develop

2. Análisis exploratorio.

La base de datos, *pizzas.xls*, proporcionada para el desarrollo de este proyecto consta de 9 variables y 379 registros. El significado de cada una de las variables se explica a continuación.

- **Ident:** Variable tipo numérica, la cual corresponde a un identificador para cada pizza.
- **HUMED:** Variable tipo numérica que indica el porcentaje de humedad contenido en la pizza.
- **PROTE:** Variable tipo numérica que indica la cantidad de gramos de proteina contenida en 100g de pizza.
- **GRASA:** Variable tipo numérica que indica la cantidad de gramos de grasa contenida en 100g de pizza.
- **CENIZA:** Variable tipo numérica que indica la cantidad de gramos de ceniza contenida en 100g de pizza.
- **SODIO:** Variable tipo numérica que indica la cantidad de gramos de sodio contenida en 100g de pizza.
- **CARBO:** Variable tipo numérica que indica la cantidad de gramos de carbohidratos contenida en 100g de pizza.
- **CALOR:** Variable tipo numérica que indica la cantidad de calorias contenida en 100g de pizza.
- **MARCA:** Varibale tipo string, la cual indica la marca que fabrico la pizza. Es una variable categórica.

Conociendo solo esto, una duda natural sería ¿Existe algún comportamiento significativo por marca? ya que MARCA es la única variable categórica. Para esto, primero se analizará dicha variable. En la figura 2, se muestra como la muestras, si no están balanceadas perfectamente, casi todas entran en el mismo rango de observaciones, variando desde 30 hasta 36 observaciones por marca.

Se continua con el análisis del resto de las variables y empiricamente se puede considerar que la variable Ident no aportará mucho considerando comportamientos de acuerdo a marca. La figura 2 muestra el scatterplot de esta variable considerando el etiquetado por marca. Lo rescatable de esta gráfica son dos saltos que se presentan, las cuales sugieren prestar atención en las marcas donde se presentan estos saltos.

A partir de este punto, las variables restantes son características del alimento en cuestión. Cabe resaltar que cada una de ellas se obtiene a través de diversos procesos químicos los cuales son importantes de especificar, ya que, para una sola característica pueden existir diversas metodologías para su medición y cada una de ellas posee su error de medición o simplemente

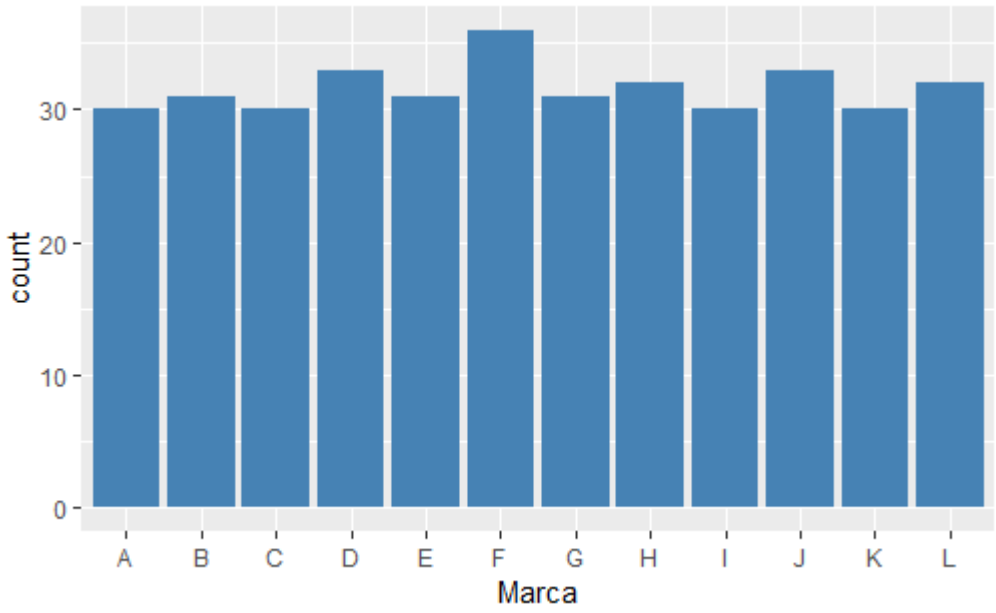

Figura 1: Conteo de registros por marca

su eficiencia se enfoca en cierto tipo de alimentos, prepaciones y/o presentaciones.

Sin embargo, para esta base de datos no fue posible recolectar este tipo de información. Por lo tanto se procede considerando que se implemento la medición adecuada con una formula estandar para todas las pizzas que permita la comparación y la ejecución de los procesos en ambientes similares.

Para cada una de las variables, se selecciono el gráfico *boxplot* como manera idonea de resumir el comportamiento. Esto debido a que el boxplot contiene varias medidas de interés y gráficamente al comparar dos o más se pueden distinguir comportamientos visuales importantes.

Comenzado con la variable HUMED, en la figura 2 se aprecia que existe un comportamiento homogeneo de humedad dentro de cada marca, no así entre marcas, pues se pueden distinguir 3 grupos de acuerdo al nivel de humedad presentado.

Para la variable PROTE, se aprecia un comportamiento similar, en la figura 2 se muestra un comportamiento elevado en las primeras 4 marcas en comparación al resto.

Para la variable GRASA, la figura 2 muestra como existen 2 grupos y particularmente la marca H presenta observaciones en los dos grupos marcados.

Para la variable CENIZ, la figura 2 muestra un efecto similar al de proteinas, pues las primeras 4 marcas presentan niveles elevados, lo cual es de interés, pues al menos en México, existen normas las cuales regulan que el nivel de sodio no exceda 3 g por cada 100 g del

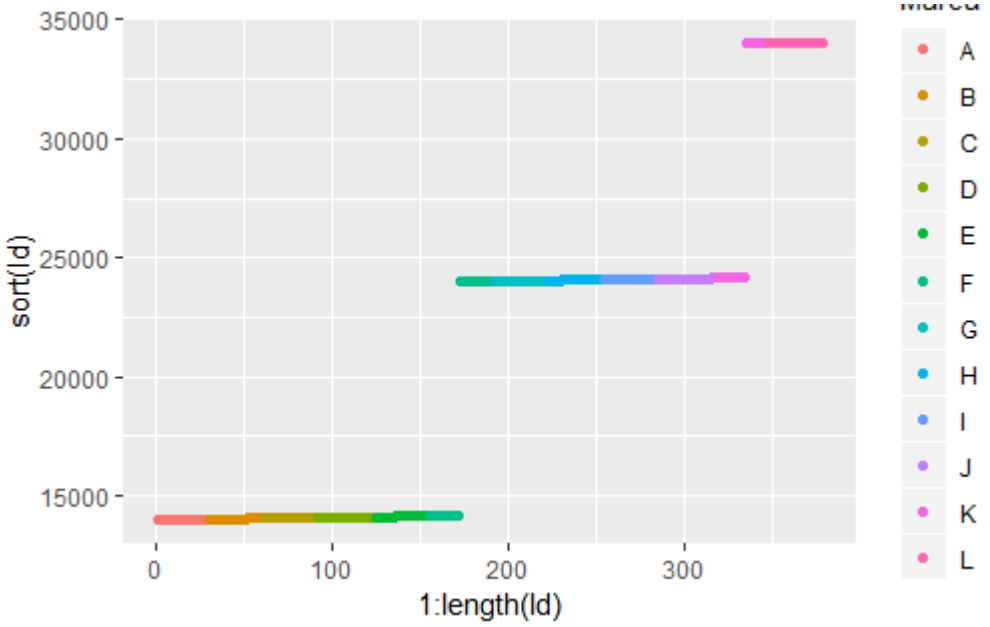

Figura 2: Comportamiento de la variable Ident de acuerdo a la marca

alimento. Sin embargo, no se posee información de la procedencia de la base de datos.

La variable SODIO, resumida en la figura 2, muestra niveles altos de sodio para la marca A y un comportamiento homogeneo en el resto de las marcas. En México, una rebanada de pizza mediana oscila entre los 75.7 y 100.7 g en promedio y el límite recomendado de sodio para una persona mayor de 19 años es de 2400mg, esto quiere decir que las pizzas de la marca A, contienen más de la mitad del sodio recomendado diario en una sola rebanada.

El comportamiento similar entre proteinas y sodio podría ser explicado por sus ingredientes si es que no todas las pizzas tienen una formula estandar, pues las pizzas que tienen más proteina podría ser explicado por la presencia de productos carnicos y tendría el efecto similar en el sodio.

Para la variable CARBO, la figura 2 presenta 3 grupos separados por el nivel de carbohidratos.

Para la variable CALOR, la figura 2 muestra 3 grupos de acuerdo al nivel de calorias contenidas.

Empiricamente, un nutriologo puede considerar que los principales nutrientes de un alimento son carbohidratos, grasas y proteinas. La suma del gramaje de cada una de estas componentes debe ser muy cercano al total de masa. En nuestro caso, deberían sumar alguna cantidad proxima a 100g. La figura 2 muestra que solo 2 marcas, G y H, tienen cantidades cercanas a 100g, por lo tanto, debe existir algo más en la composición cuyo gramaje sea significativo, ejemplos de esto, pueden ser fibras crudas contenidas en harinas y verduras.

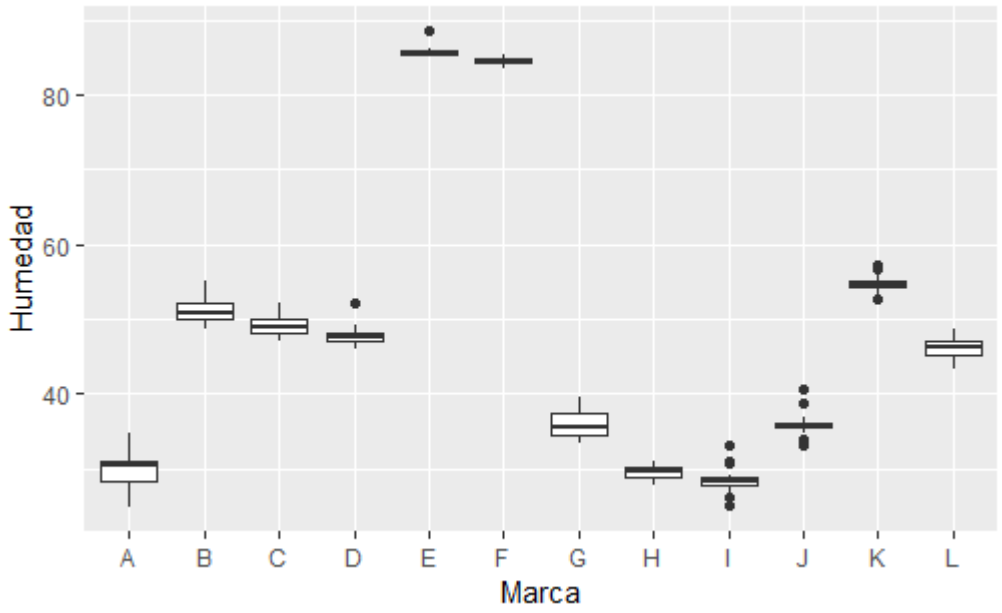

Figura 3: Comportamiento de la variable Humed de acuerdo a la marca

En la teoría de la bromatología, existen aproximaciones que se satisfacen con buena precisión entre las calorías y algunos nutrientes. Las aproximaciones se muestran a continuación:

- una caloría por cada 6 gramos carbohidratos
- una caloría por cada 9 gramos de grasas
- una caloría por cada 4 gramos de proteina

Si los datos se calcularon de manera correcta, la diferencia entre las calorías registradas debería ser próximo al nutriente dividido por su respectivo factor.

Los cálculos de las aproximaciones arrojaron los siguientes errores para cada uno de los nutrientes

Nutriente	Error promedio
Carbohidratos	1.37
Grasas	0.68
Proteina	0.04

Por último, se presenta un gráfico de correlación entre las características de los alimentos. La figura 2 muestra, al menos 4 patrones importantes, a saber, Humedad-calorias, Proteina-ceniza, Ceniza-sodio y Ceniza-carbohidratos

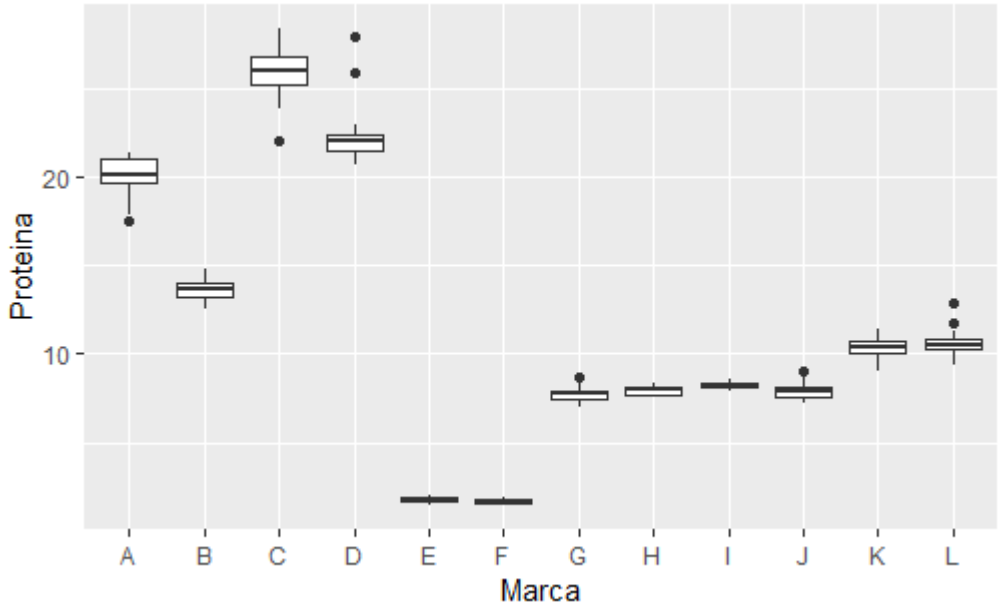

Figura 4: Comportamiento de la variable Prote de acuerdo a la marca

La relación entre humedad y calorías, tal vez corresponda a la presencia del fosfato, pues la presencia de fosfato modifica el valor de estas dos variables y se encuentra en alimentos como el queso y carnes.

La alta correlación entre Proteinas y cenizas puede ser explicado por la presencia de queso, pues es un alimento el cual es fuente de proteinas y minerales.

En el caso de la alta correlación entre sodio y cenizas, puede ser explicado por la presencia de tomate, ya que, este alimento eleva el valor de ambas variables.

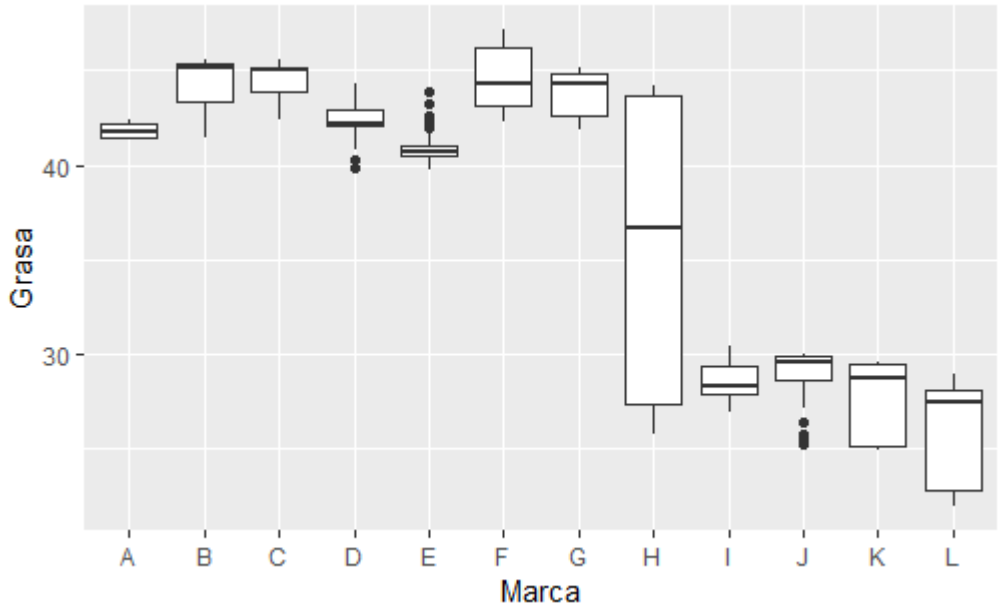

Figura 5: Comportamiento de la variable Grasa de acuerdo a la marca

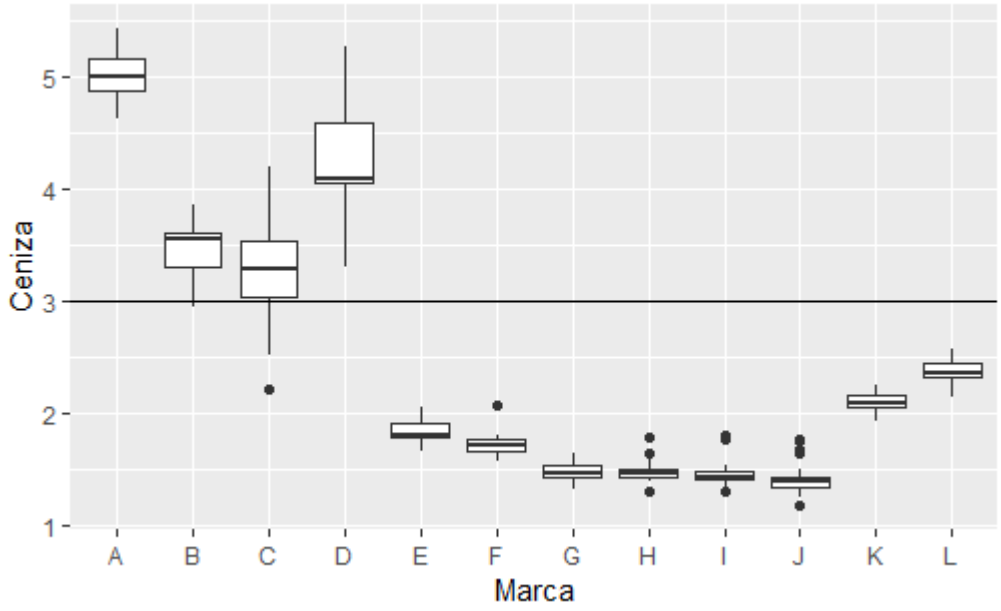

Figura 6: Comportamiento de la variable CENIZ de acuerdo a la marca

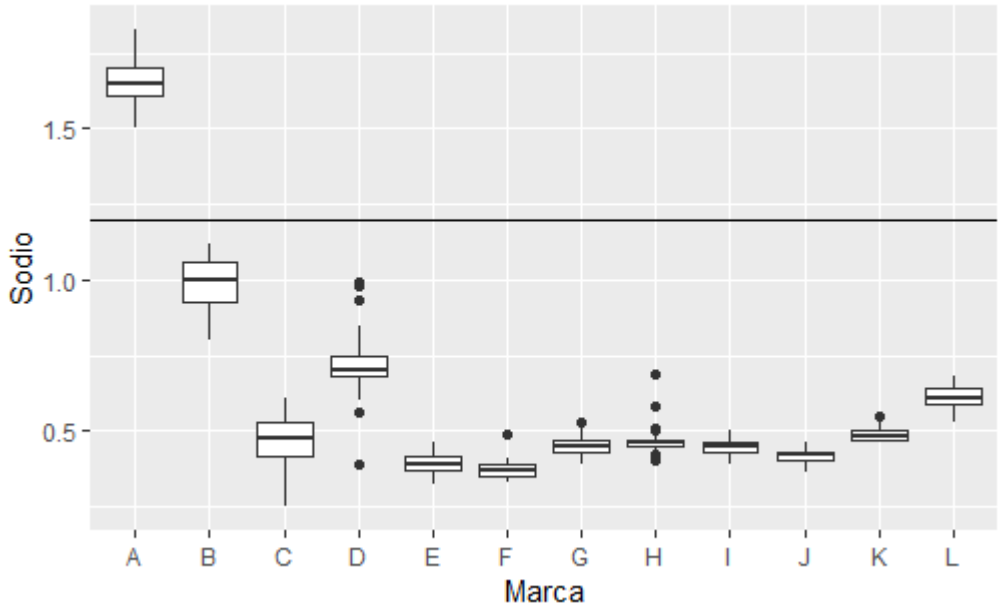

Figura 7: Comportamiento de la variable sodio de acuerdo a la marca

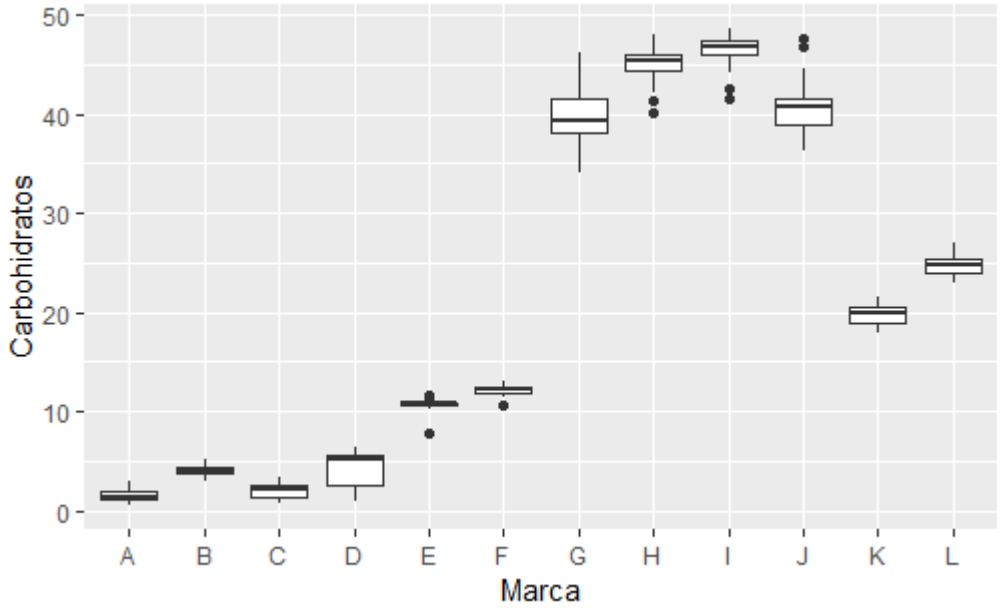

Figura 8: Comportamiento de la variable Carbo de acuerdo a la marca

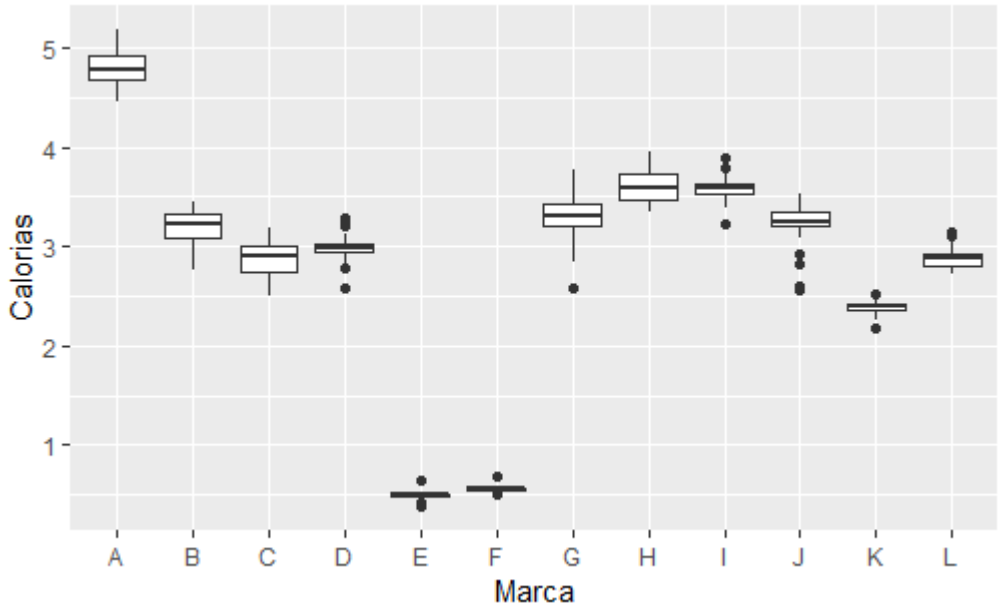

Figura 9: Comportamiento de la variable Calor de acuerdo a la marca

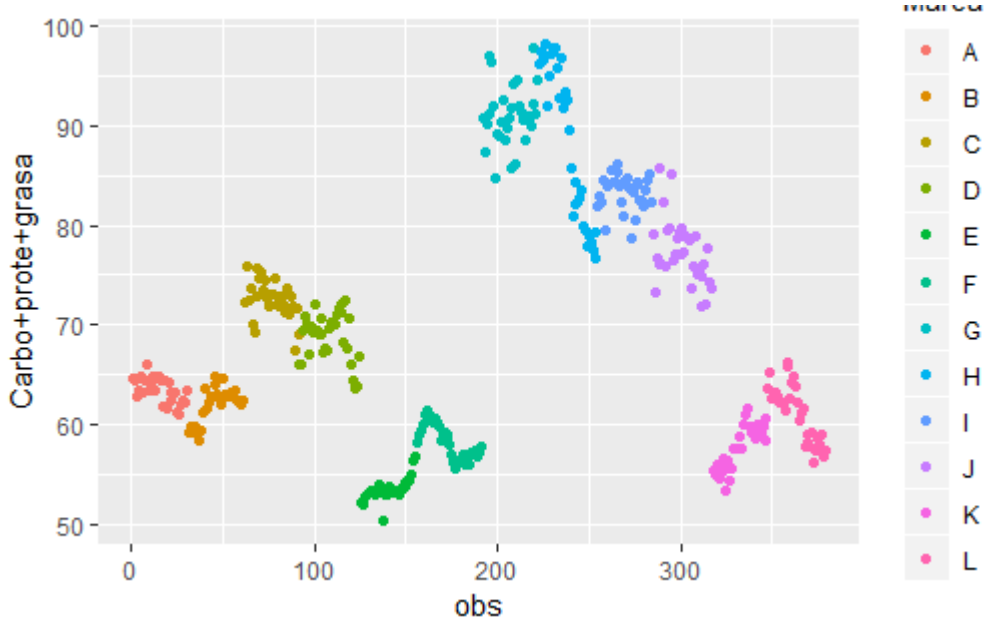

Figura 10: Grasa+Proteina+Carbohidratos para cada observación

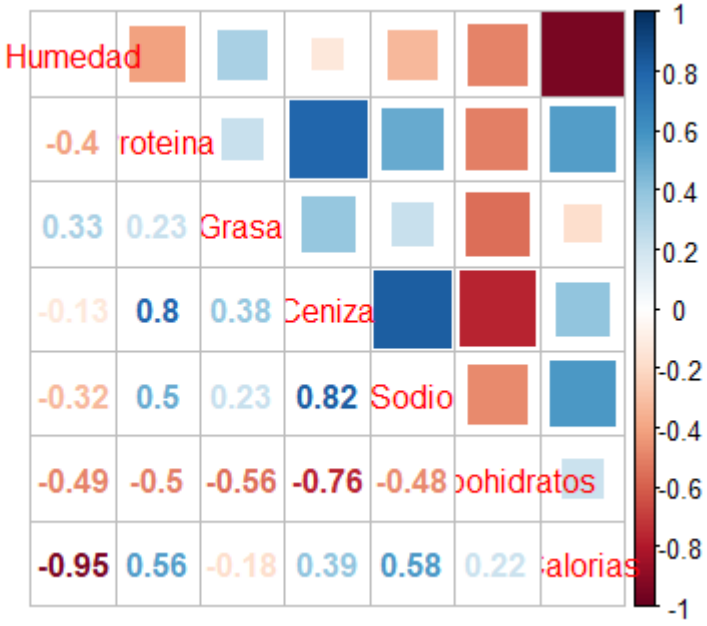

Figura 11: Correlación de las variables.

3. Modelos de reducción de dimensiones.

3.1. Análisis de componentes principales (PCA)

Al hacer el análisis de componentes principales a los datos se obtuvieron las siguientes cargas en las primeras dos componentes. La varianza acumulada en las primeras dos componentes es del 83.59 % y se considera suficiente para el análisis. Como se puede observar la primera componente está asociada con la ceniza, el sodio y la proteína, mientras que la segunda se asocia con la humedad, los carbohidratos y las calorías. Las componentes principales se obtuvieron usando la función prcomp de R con los datos normalizados, la normalización no afectó drásticamente la representación gráfica, sin embargo, permitió observar de forma más clara las relaciones entre las variables ya mencionadas.

	PC1	PC2
Humedad	0.21	0.58
Proteina	-0.47	-0.03
Grasa	-0.19	0.41
Ceniza	-0.51	0.15
Sodio	-0.47	-0.02
Carbohidratos	0.32	-0.49
Calorias	-0.34	-0.48
Varianza acumulada	48.64 %	83.59 %

Tabla 1: Pesos asociados a las primeras dos componentes principales.

3.1.1. Representación gráfica (biplot)

La figura 3.1.1 muestra la representación en las primeras dos componentes principales de los datos, en ella se aprecia una clara separación entre algunos grupos de pizzas, específicamente, la marca A se distingue por tener altos niveles de sodio y proteina, las marcas B, C y D se caracterizan por un alto nivel de ceniza, E y F por tener altos valores de humedad, finalmente el resto de marcas (G, H, I, J, K y L) se distinguen por su alto contenido en carbohidratos. También el biplot muestra la alta colinealidad existente entre las variables sodio y proteina, y sugiere un contraste entre la variables grasa y carbohidratos.

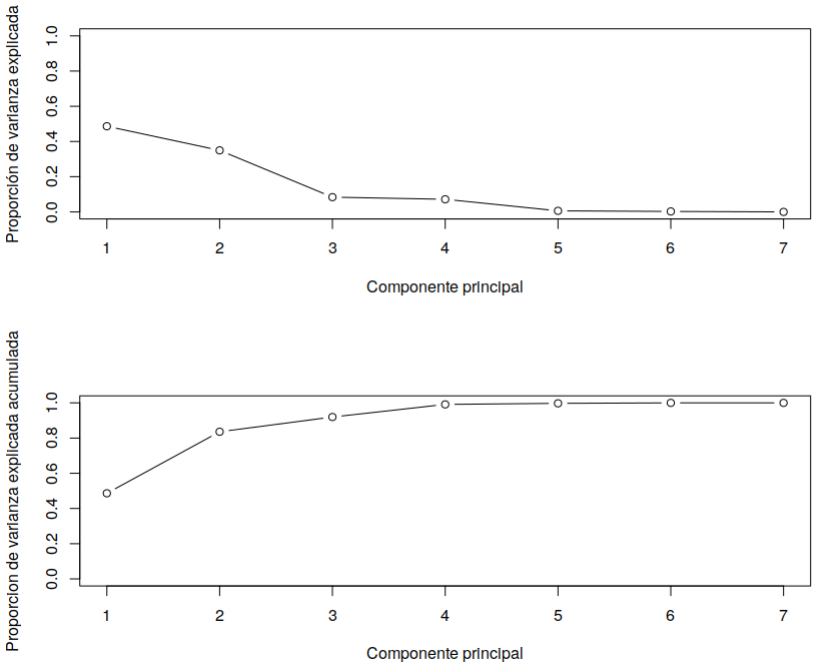

Figura 12: Varianza explicada por las componentes principales

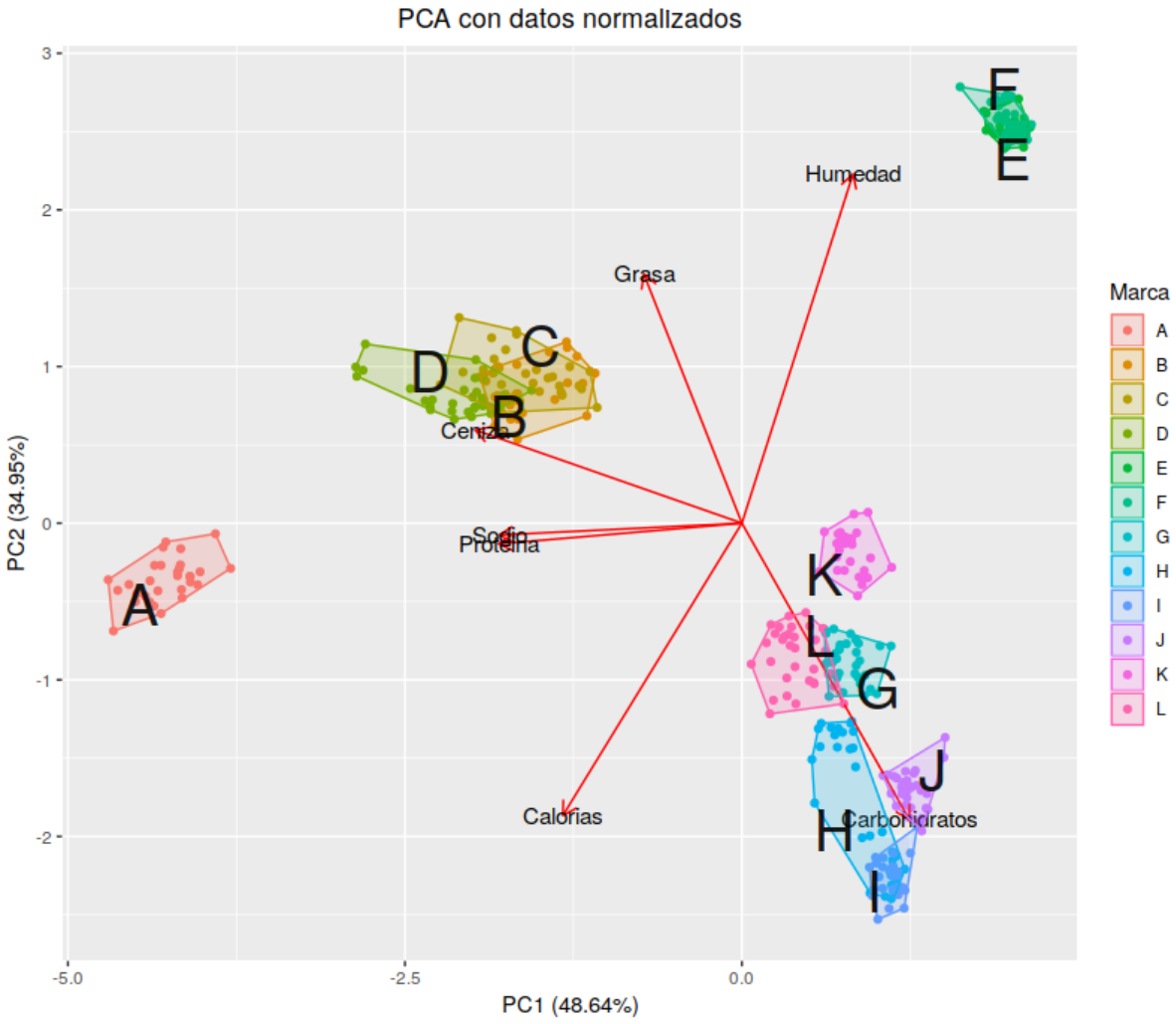

Figura 13: Biplot PCA

3.2. Análisis de factores

Se realizó el análisis de factores usando dos factores, esto debido a la baja dimensión del problema. Para ello se uso la función factanal de R, que estima los factores usando el método de máxima verosímilitud, además, se usó la rotación varimax.
El resumen de los resultados se muestra en la tabla 3.2. El valor del estadístico de prueba χ^2 (con 8 grados de libertad) para verificar la hipótesis nula que la matriz de correlación de datos puede ser representada usando dos factores es 1837.4, obteniendo además un $p-$valor de 0, por lo que la hipótesis nula se rechaza. Este resultado puede ser explicado por el hecho de que el determinante de la matriz de correlación es cercano a cero. A pesar de ello, se puede afirmar que la aproximación a la matriz de correlación a partir de la matriz $LL' + \Psi$ es adecuada, como se muestra en la tabla 3.2, la diferencia entre la matriz R y $LL' + \Psi$ es mínima, considerando un redondeo a tres dígitos. Esto en conjunto con el hecho que la proporción de varianza acumulada usando dos factores es del 80.9 %, permite afirmar que dos factores son suficientes para la representación de los datos.

	Factor1	Factor2	Varianza específica	Comunalidades
Humedad	0.06	-1.00	0.01	1.00
Proteina	0.76	0.44	0.23	0.77
Grasa	0.47	-0.30	0.69	0.31
Ceniza	0.94	0.19	0.08	0.92
Sodio	0.73	0.39	0.32	0.68
Carbohidratos	-0.90	0.44	0.01	1.00
Calorias	0.23	0.97	0.01	0.99
Proporción de Varianza	44 %	36.9 %		
Varianza acumulada	44 %	80.9 %		

Tabla 2: Resultados del análisis factorial.

	Humedad	Proteina	Grasa	Ceniza	Sodio	Carbohidratos	Calorias
Humedad	-0.00	-0.02	0.00	-0.00	0.01	-0.00	-0.00
Proteina	-0.02	0.00	-0.01	-0.01	-0.22	-0.01	-0.04
Grasa	0.00	-0.01	0.00	-0.01	-0.00	-0.00	0.00
Ceniza	-0.00	-0.01	-0.01	-0.00	0.06	0.00	-0.00
Sodio	0.01	-0.22	-0.00	0.06	-0.00	0.01	0.04
Carbohidratos	-0.00	-0.01	-0.00	0.00	0.01	-0.00	-0.00
Calorias	-0.00	-0.04	0.00	-0.00	0.04	-0.00	0.00

Tabla 3: Diferencia entre R y $LL' + \Psi$, con redondeo a tres dígitos.

3.2.1. Interpretación

El primer factor se relaciona principalmente con las variables ceniza, carbohidratos y proteina mientras que el segundo factor se asocia con la humedad y las calorías. Las varianzas específicas y las comunalidades indican que la varianza de la mayoría de las variables queda

bien explicada a partir de los factores, con excepción de la variable grasa, cuya varianza específica es de .69.

3.2.2. Representación gráfica (biplot)

Para obtener una representación gráfica en dos dimensiones, se realizó la estimación de los *scores* usando el método de Bartlett. El biplot resultante se muestra en la figura 3.2.2. El resultado es muy similar al obtenido usando componentes principales, pues se logran diferenciar los mismos grupos que las marcas forman de acuerdo a sus características; sin embargo, existe una pequeña diferencia y es que las marcas K y L se separan del grupo formado por las marcas G, H, I y J, para situarse más al centro de la gráfica, lo que indica que los niveles de nutrientes de las marcas K y L se sitúan cerca de la media.

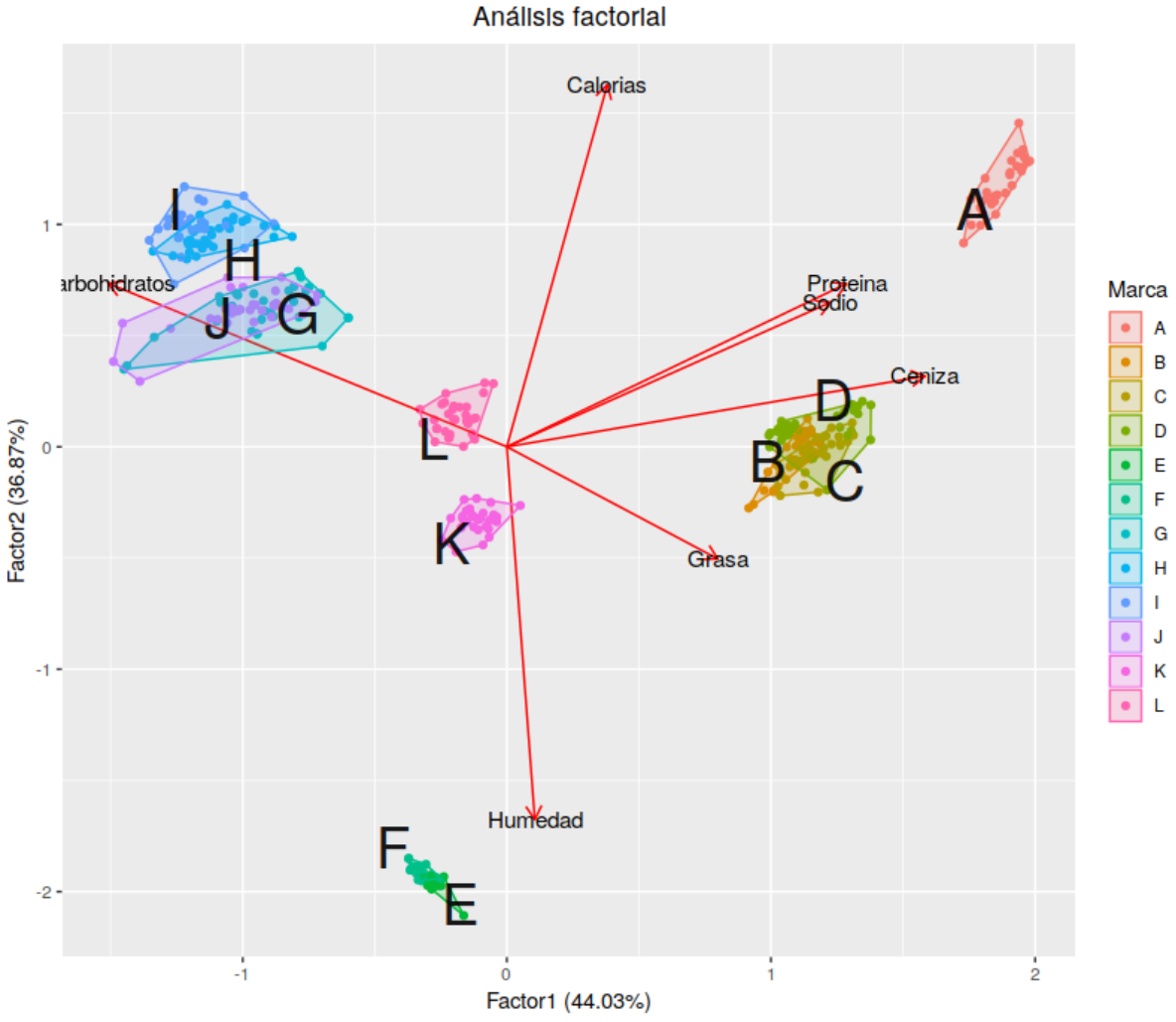

Figura 14: Biplot usando 2 factores

4. Análisis por agrupación.

Con el fin de identificar grupos con características similares de forma automática, se procedió a usar algoritmos de clustering. Específicamente, se hace clustering jerárquico completo usando la función agnes de la biblioteca cluster.

4.1. Elección del número de clusters

Existen criterios para la elección de número de clusters adecuado dado un conjunto de datos, los criterios utilizados en este problema son dos: suma de cuadrados entre clústers (wss) y el criterio de ancho de *silhouette*. En el primer caso, el criterio consiste en graficar la suma de cuadrados entre clústers para varios números de clústers y observar a partir de que número la mejora a la suma de cuadrados entre clústers es mínima. En una gráfica que compare a wss con el número de clusters, se espera que en el número óptimo se forme un *codo* que indique el cumplimiento del criterio. La figura 4.1 muestra el resultado de calcular wss para ciertos números de clústers. A partir de observar la gráfica, se determina que una cantidad adecuada para el número de clusters usando estre criterio es 5.

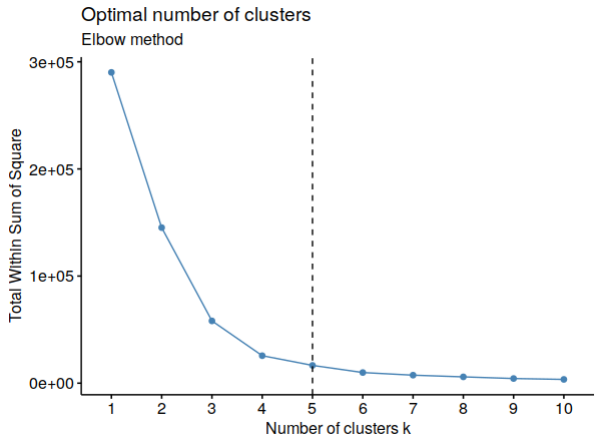

Figura 15: Suma total de cuadrados entre clústers vs número de clusters

El otro criterio muy popular en la literatura es el método silhouette, que es una medida de qué tan bien un individuo está asignado en un clúster a partir de calcular la distancia promedio a cada uno de los elementos del clúster al que fue asignado y la distancia promedio al clúster más cercano de entre aquellos a los que no pertenece. Mientras más alta sea la medida promedio silhouette, mejor asignados están los datos en sus respectivos clústers. La figura 4.1 muestra la comparativa de ancho de silhouette promedio para cierto número de clústers, y se observa que el número adecuado de clústers para los datos es 5. Para la creación de las gráficas de wss y silhouette se usó la función fviz_nbclust de la biblioteca factoextra.

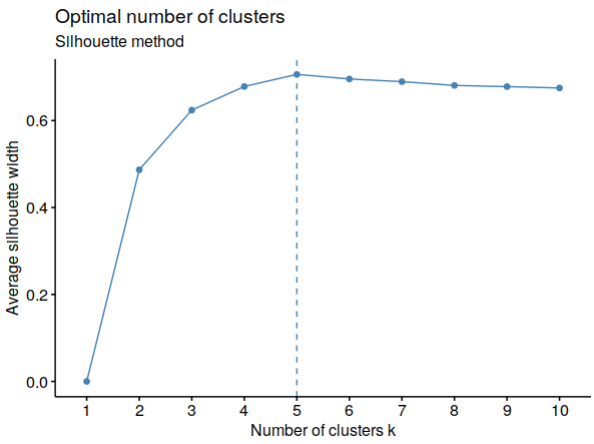

Figura 16: Silhouette promedio vs número de clusters

4.2. Asignación de clusters

Habiendo elegido el número de clústers por crear, se procedió a calcular la asignación de los datos a cada clúster, usando clustering jerárquico con la ayuda de la función agnes. Los clústers obtenidos se muestran en las representaciones en dos dimensiones de los datos (PCA y Factores) en las figuras 4.2 y 4.2. En ambos casos se observa que los grupos formados coinciden con los descritos en la sección de reducción de dimensión. Hay que notar que cada una de las marcas de pizza está contenida en un único clúster.

5. Modelos de clasificación.

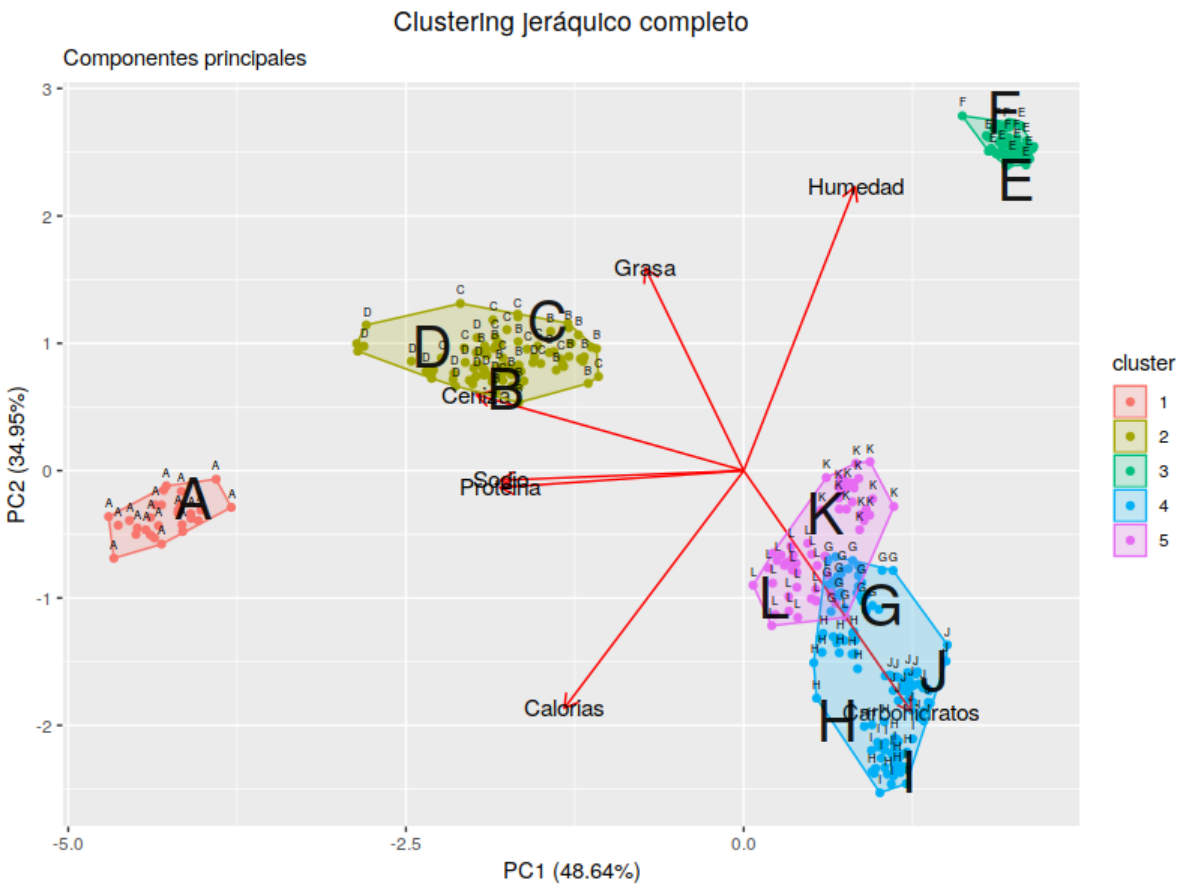

Figura 17: Clustering jerárquico completo (Representación: PCA)

6. Regresión Multivariada

6.1. MANOVA

El análisis de varianza múltiple tiene como objetivo comparar si los valores de un conjunto de datos numéricos son significativamente distintos al separarlos por clases.

En este caso, la variable que distingue que clases es Marca, la cual tiene 12 categorias y esperamos contrastar la hipotesis nula $H_0 : \mu_1 = \ldots = \mu_12$, es decir, que la media de todas las categorías es igual, contra la hipotesis alternativa que existe al menos un par de categorías en las cuales las medias no son iguales.

Antes de realizar esta prueba, se realiza el ANOVA para cada una de las 8 variables numéricas con una hipotesis similar pero en una sola variable. En las 8 pruebas ANOVA se rechazó la hipotesis nula.

De la misma manera se rechaza la hipotesis nula para MANOVA. Lo cual implica que existen al menos dos categorías con medias iguales considerando todas las variables.

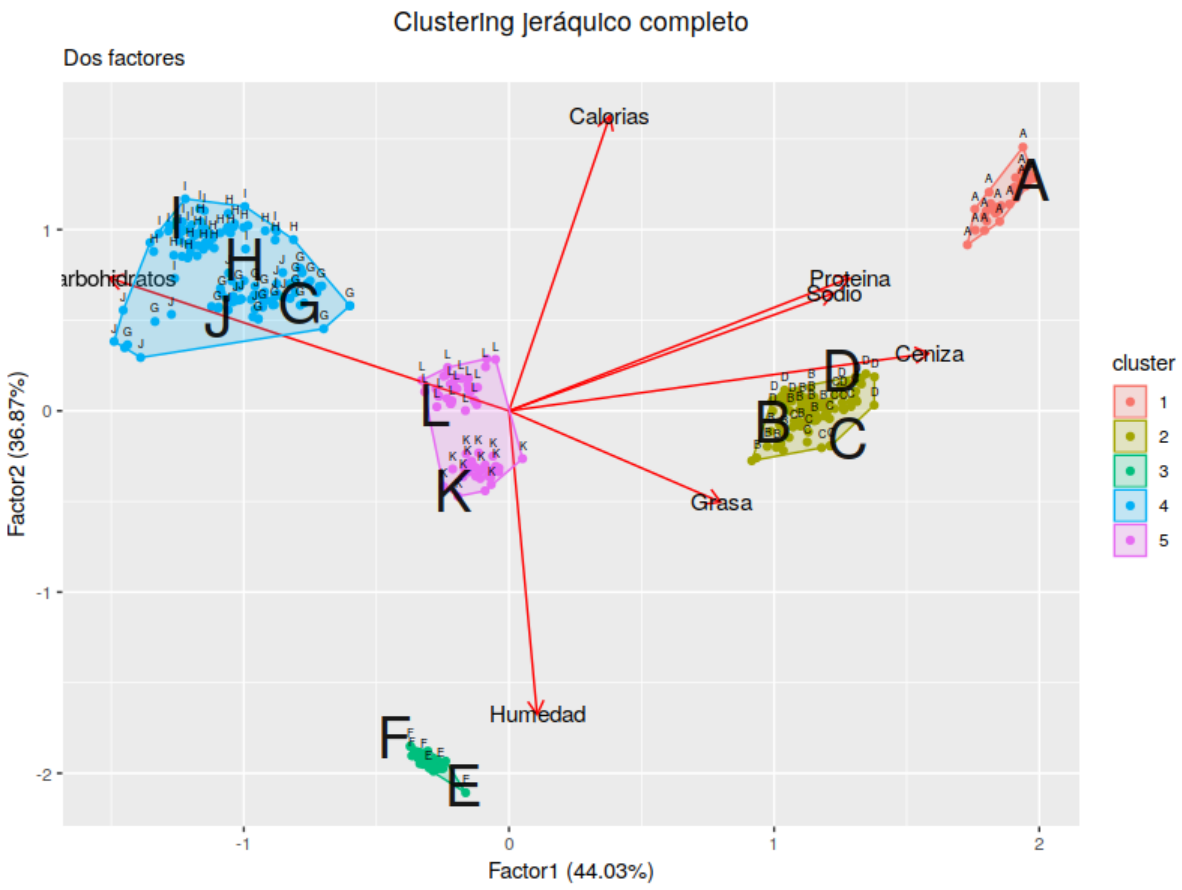

Figura 18: Clustering jerárquico completo (Representación: Factores)

El proceso para comprobar la hipotesis alternativa sería tomar los 2^{k-1} subconjuntos y realizar la prueba MANOVA.

Ejemplo de esto, podría ser el subconjunto que solo posee la categoría J y G. Realizando ANOVA por variables, se obtiene no existe evidencia suficiente para rechazar que las medias son iguales considerando las variables Humedad, Proteina, Calorias y Carbohidratos por separado y aun nivel de significancia del .05.

Sin embargo, al ejecutar el MANOVA para solo estas dos clases, se rechaza la hipotesis que sus medias sean iguales.

6.2. Regresión

Para ejecutar la Regresión Multivariada, se consideraron como variables regresoras a aquellas que son nutrientes, a saber, Proteinas, Grasas, Sodio y Carbohidratos. Mientras que las variables dependientes serán Humedad, Cenizas y Calorias.

Para las 3 variables dependientes, se encontró que la variables Grasa no es significativa

para el modelo. Por lo tanto, se elimina del modelo.

Con esto se consiguen estadisticos de R^2 de 0.96, 0.94 y 0.93 para las respectivas variables dependientes, lo cual era de esperarse, pues en la figura 11, se aprecia como las variables dependientes están muy correlacionadas con las variables regresoras.

7. Conclusiones.