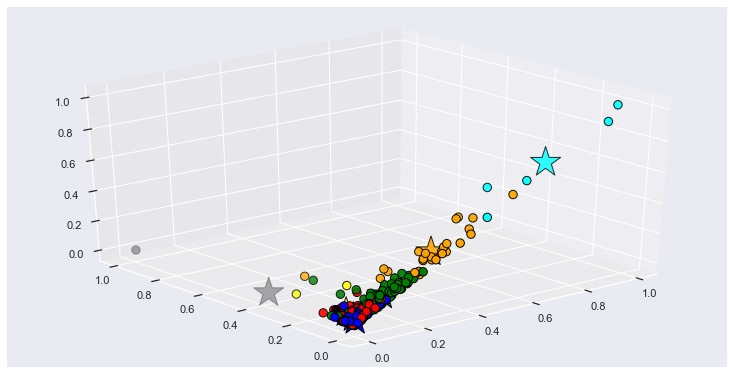
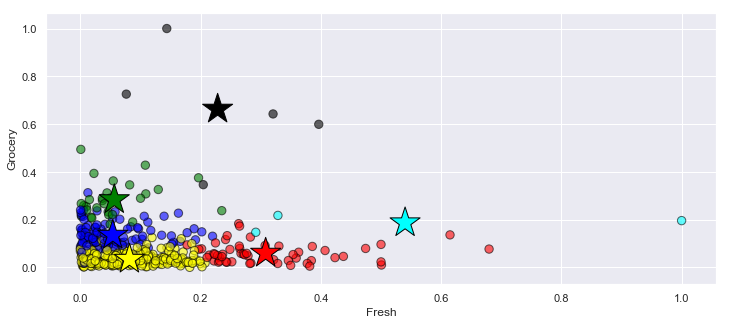
Memoria práctica individual: Clustering, técnicas de agrupamiento





Índice

[1. Análisis descriptivo de los datos 2](#_Toc11594674)

[2. Tratamiento de missing (si los hay) 2](#_Toc11594675)

[3. Tratamiento de variables categóricas 2](#_Toc11594676)

[4. Análisis gráfico de los datos 2](#_Toc11594677)

[5. Normalización y Separación de datos 2](#_Toc11594678)

[6. Agrupamiento mediante el algoritmo K-means 2](#_Toc11594679)

[7. Métricas de ajuste 2](#_Toc11594680)

[8. Análisis de resultados obtenidos 2](#_Toc11594681)

[9. Conclusiones 2](#_Toc11594682)

1. Análisis descriptivo de los datos

El conjunto de datos que vamos a analizar se refiere a los clientes de un distribuidor mayorista. Incluye el gasto anual en unidades monetarias (m.u.) en diversas categorías de productos.

Nuestro objetivo es identificar los grupos o clúster de clientes, para mejor entender que es lo que compran y lo que pagan por ello para poder hacer unas ventas más personalizadas maximizar beneficios. Si por ejemplo hay un grupo interesado en comprar exclusivamente leche, y compra en grandes cantidades, podríamos hacer una promoción dirigida a la compra de grandes cantidades.

Los atributos de los datos son los siguientes:

1. FRESH: gasto anual (m.u.) en productos frescos (Continuo);
2. MILK: gasto anual (m.u.) en productos lácteos (Continuo);
3. GROCERY: gasto anual (m.u.) en productos comestibles (Continuo);
4. FROZEN: gasto anual (m.u.) en productos congelados (Continuo)
5. DETERGENTS\_PAPER: gasto anual (m.u.) en detergentes y productos de papel (Continuo)
6. DELICATESSEN: gasto anual (m.u.) en productos delicatessen (Continuo);
7. CHANEL: canal de clientes - Horeca (hotel / restaurante / cafetería) o canal minorista (Nominal)
8. REGIÓN: Región de los clientes - Lisbon, Oporto u Otro (Nominal)

Si hacemos una descripción de los datos mediante el siguiente comando, podemos encontrar los siguientes resultados:



Vemos por ejemplo como las compras en productos comestibles son las que mas beneficios dan a la empresa, y además las que mas desviación tienen, es decir que hay clientes muy diversos.

1. Tratamiento de missing (si los hay)

Para el tratamiento de los missing llevamos a cabo el siguiente procedimiento:

1. Vemos si hay o no caracteres nulos o vacíos.
2. Vemos si hay o no Caracteres desconocidos.

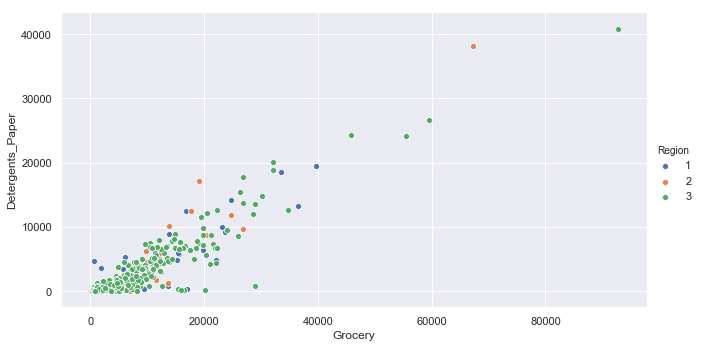
En nuestro caso **no aparecen caracteres desconocidos**, podemos ver el desarrollo en el código en el anexo.

1. Tratamiento de variables categóricas

En este caso todas las variables son enteras. Hay dos categorías discretas y el resto son continuas pero no nos importa de momento ya que estamos haciendo clustering.

1. Análisis gráfico de los datos

En esta sección hemos representado por parejas todos los datos gracias a la librería Seaborn. Una vez hemos representado todos hemos elegido los datos que mas nos han interesado. A continuación podemos ver una de estas gráficas, (podemos encontrar el resto en el código anexo).

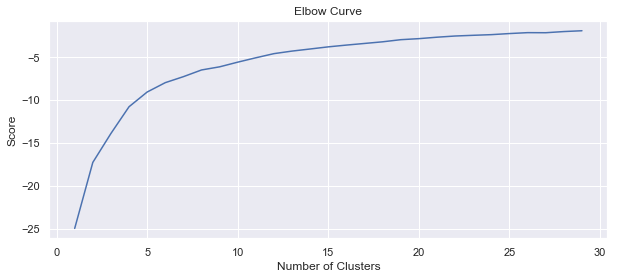


Pese a que no se encuentran muchas relaciones claras entre las categorías, si comparamos las compras alimentarias y los productos de limpieza se observa cierta relación directa entre las compras alimentarias y detergentes.

1. Normalización y Separación de datos
2. Normalización: nos hemos servido de la librería preprocessing.
3. Separación de datos: las dos primeras columnas ya son categóricas, por lo que no nos aportan información a la hora de realizar clustering. Las hemos quitado.
4. Agrupamiento mediante el algoritmo K-means

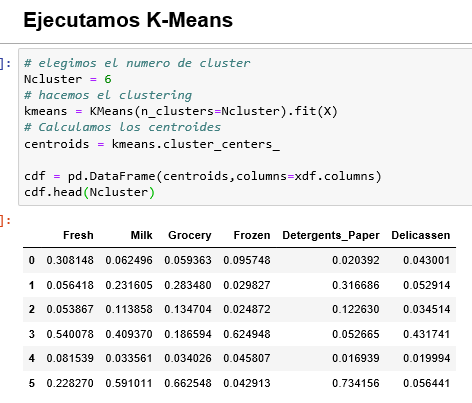
Hemos realizado los siguientes pasos:

1. Encontrar el valor de K: hemos visto el score según el numero de clusters y hemos elegido un número que se encuentre en el codo de la gráfica.

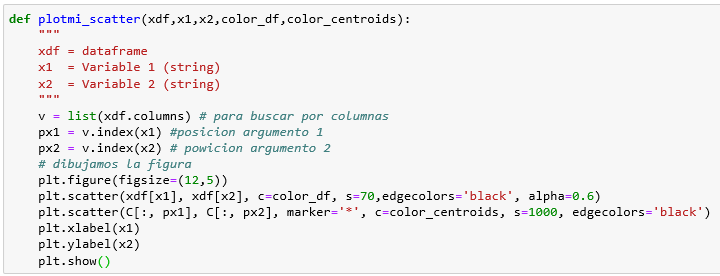


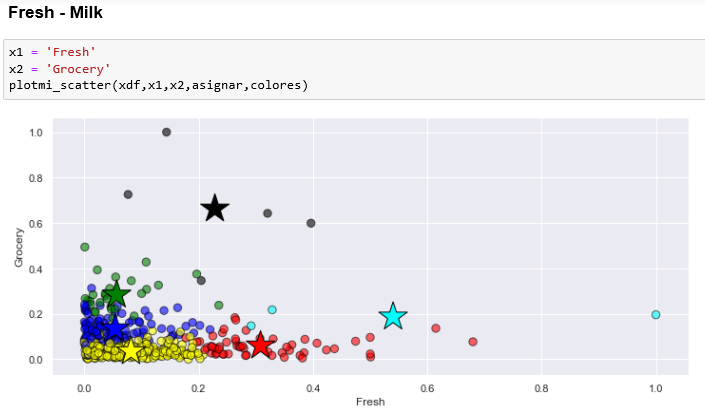
Se han elegido **6 clusters.**

1. Ejecución del algoritmo y calculo de los centroides: una vez elegido el valor de K ejecutamos el algoritmo. Hemos incluido en una tabla los centroides de las categorías según el grupo.



1. Representamos los resultados. Hemos representados de variables contra otras y una contra todas para entender mejor los grupos. Hemos creado una función (plotmi\_scatter) para ejecutarla con las variables que queramos. Aquí debajo la función y un ejemplo del clustering obtenido de los gastos productos frescos vs los lácteos.





1. Métricas de ajuste

Dado que estamos ante un aprendizaje puramente no supervisado, no tenemos referencias para hacer métricas. Las métricas de ajuste que hemos utilizado han sido métricas internas:

* Score de Kmeans: distancia media a los centroides de cada clase.
  + Score obtenido con 6 clusters: -7.96
* Silhouette: El valor de la silueta es una medida de qué tan similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación).
  + Resultado silhouette: 0.35

No son resultados increíbles, pero es un buen compromiso entre no utilizar muchos clusters para entender los sectores y su resultado. Utilizar un numero de agrupaciones mayor nos daría mejores resultados, pero sería más difícil interpretar el grupo y tomar decisiones acordes a ello.

1. Análisis de resultados obtenidos

Una vez hecho los 6 grupos podemos interpretar características comunes de dicho grupo. Los hemos clasificado por colores haciendo uso de pandas y una asignación de colores arbitraria. Tras analizar las distintas gráficas entre atributos se podemos describir la personalidad o tendencia de cada grupo:

|  |  |  |  |
| --- | --- | --- | --- |
| Nº | Nombre de Grupo o Color | Porcentaje del DataSet | Características del Grupo |
| 1 | Negro | 1.1% | El grupo negro se correspondería con un cliente centrado en la alimentación a gran escala, no se interesa mucho por otros productos. |
| 2 | Cyan | 0.7% | El grupo cian correspondería con un cliente mayorista, pero compra todo tipo de productos. Es el segundo con mayores beneficios a la empresa. |
| 3 | Verde | 6.6% | Este grupo está orientado en la alimentación, aunque con un volumen de compra menor, también interesado por productos de limpieza. No se interesa en delicatessen. |
| 4 | Rojo | 12.7% | El grupo rojo está bastante interesado en los productos frescos. Compra algún otro producto, pero a muy baja escala. |
| 5 | Azul | 23.6% | El grupo azul sigue bastante la tendencia del grupo verde, pero compra con un volumen mucho menor |
| 6 | Amarillo | 55.2% | El grupo amarillo corresponde con una amplia mayoría del dataset, compran de todo, pero en menor cantidad. Tienen mas afinidad por los productos congelados. |

1. Conclusiones