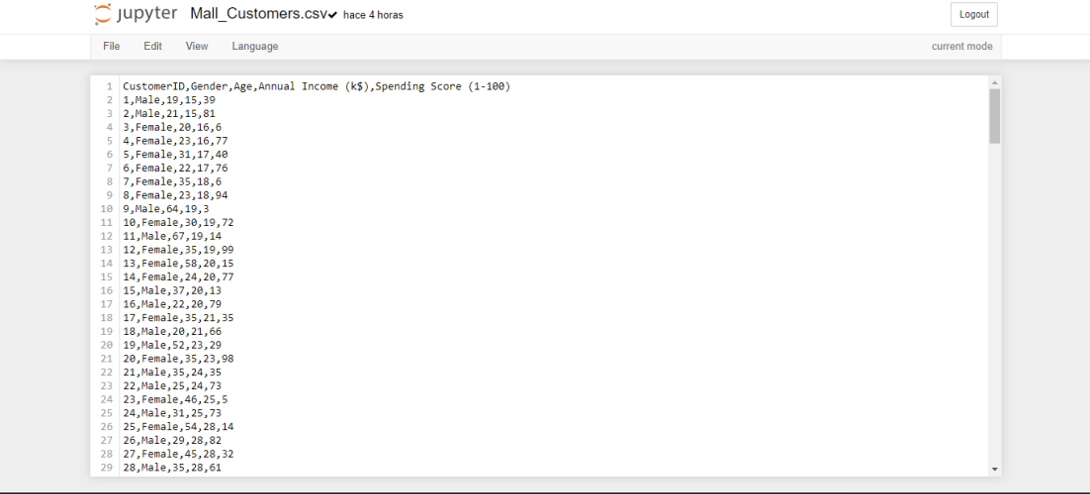
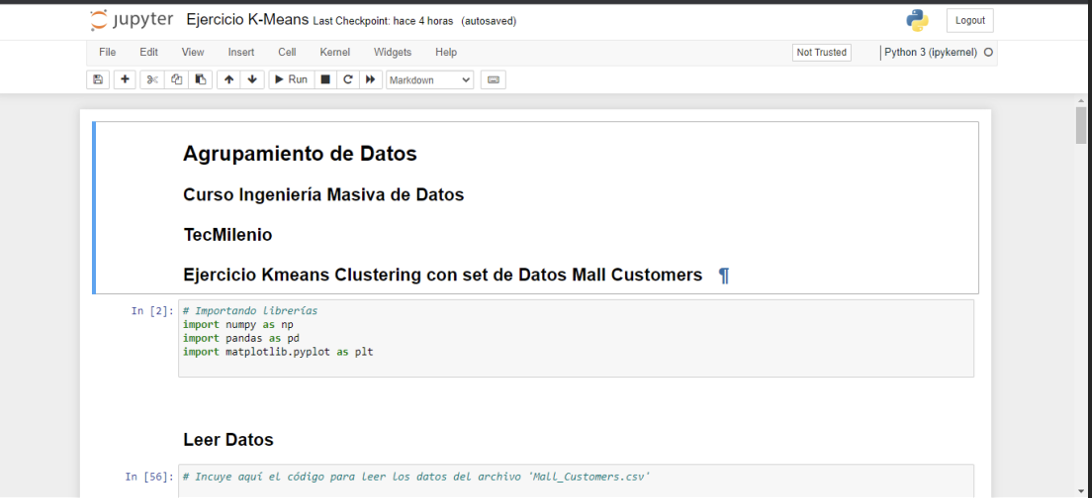
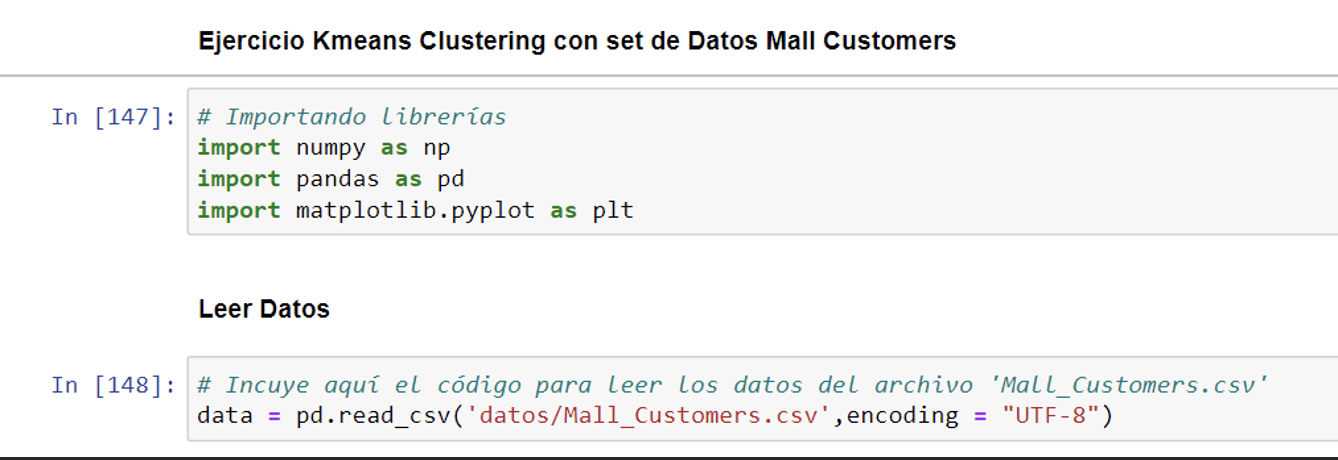
|  |  |  |
| --- | --- | --- |
| **Nombre**: Kathya, Gerardo, Guillermo, Emiliano | | **Matrícula**: |
| **Nombre del curso:** Ingeniería de datos masivos | **Nombre del profesor**:  **Miguel de Jesús Martínez Felipe** | |
| **Fecha:** | **Act7** | |

1. Prepara el ambiente de Python en el equipo de cómputo donde se realizará esta actividad.
2. Crea un nuevo Jupyter Notebook con el título Actividad 7.
3. Copia a tu computadora los siguientes archivos

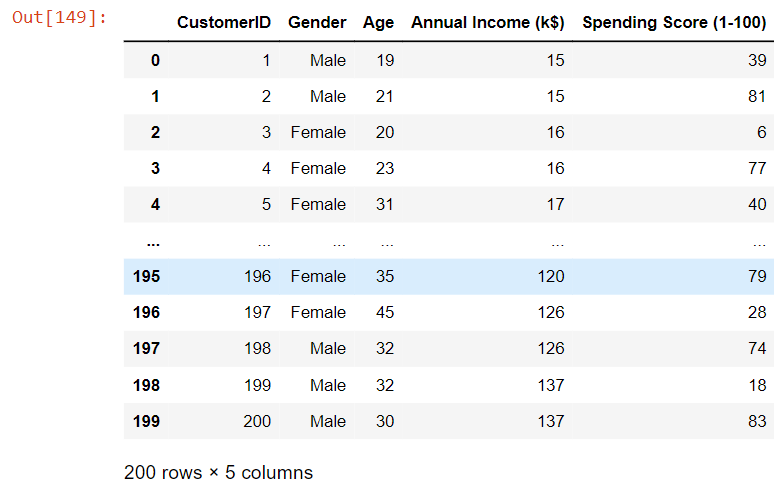


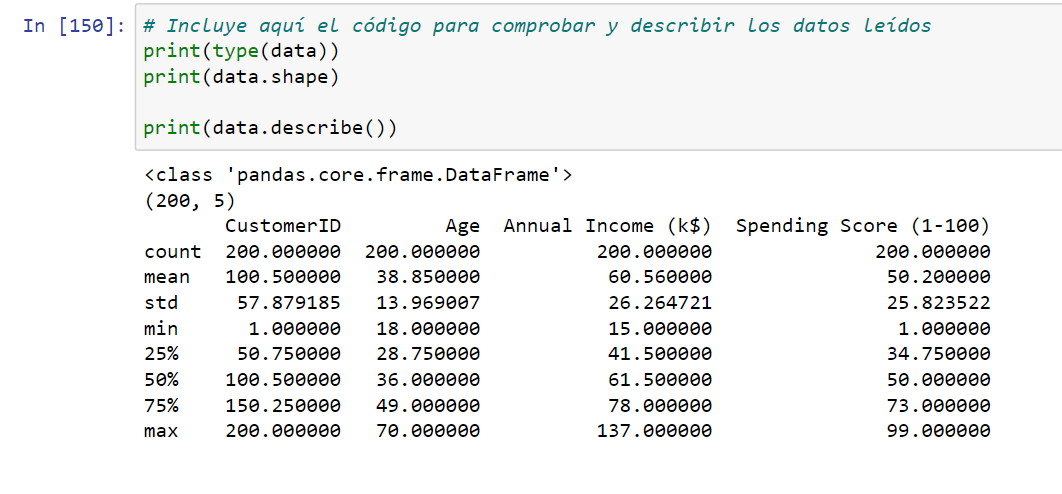


1. Haz los cambios necesarios al notebook de Jupyter para hacer una segmentación de clientes usando K-Means con los datos de *Mall\_Customers.csv,* incluyendo histogramas de exploración de datos, el número y descripción de clusters, además de la gráfica del método Elbow.
2. Incluye en tu archivo los comentarios con la explicación de tu código.

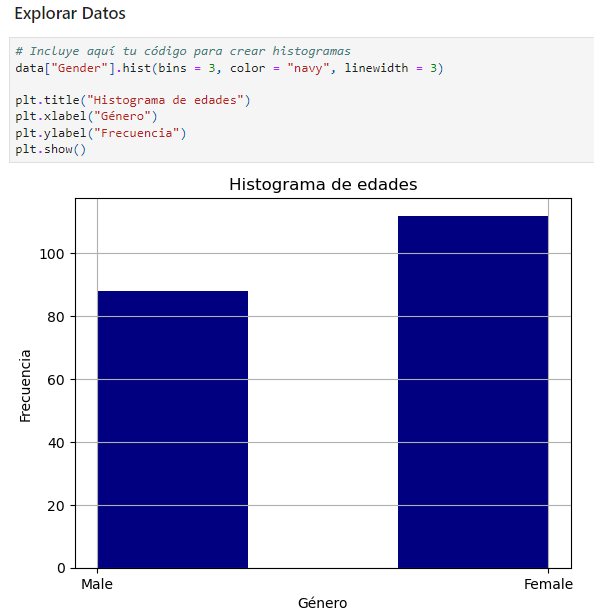
Se importan las librerías que se emplearán la lectura de datos del CSV y para generar las gráficas de datos.

Se hace un ***read\_csv*** para poder leer el archivo que contiene la información de los clientes.



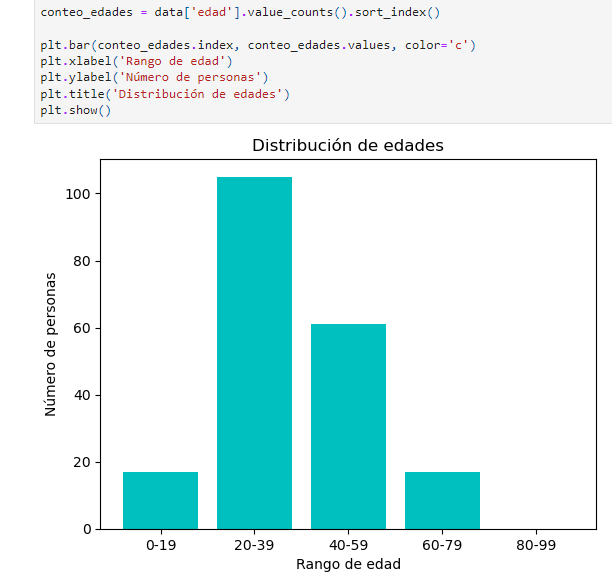


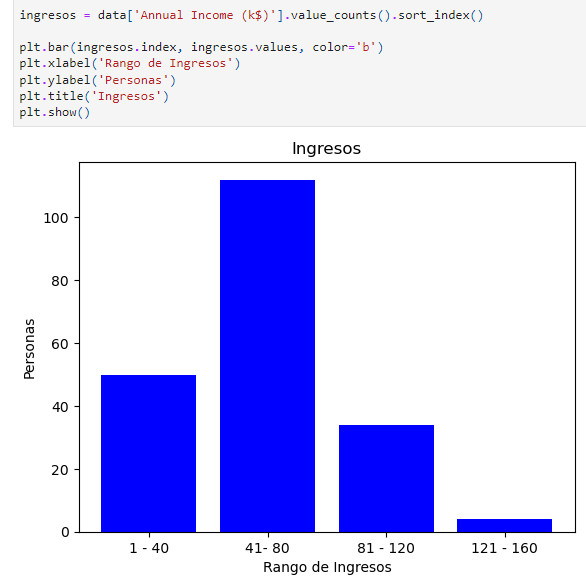
Se hace una descripción automática de los datos del dataframe mediante el método ***describe*** para generar ciertos cálculos matemáticos, los cuales no influyen, pero podrían ser útiles si alguien desea consultarlos.



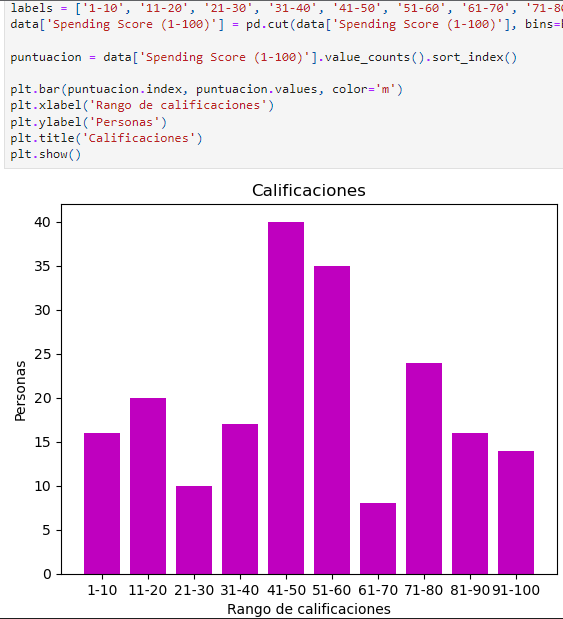
Se genera un histograma basado en la cantidad de veces que se repite cada uno de los valores de la columna, en este caso, solo son dos valores, para hombres y mujeres, por lo que el histograma solo son dos barras que expresan la cantidad de hombres y mujeres en la tabla.

Se crea un histograma con los valores de la columna de **Age,** el cual incluye valores entre 1 a 100, el código permite generar las barras basados en rangos de edad que van de 20 en 20.

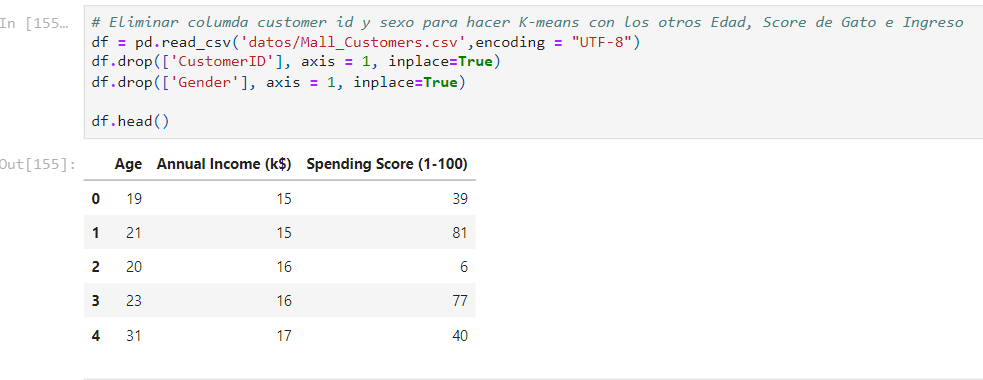




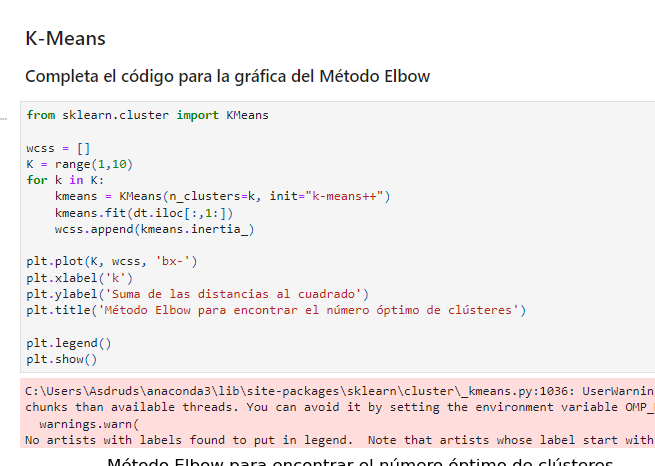
Se genera un histograma con los valores de la columna de ***Annual Income*** donde se generaron rangos para poder representar de manera más sencilla los valores registrados, ya que el rango abarca después de 100, por lo que podría hacer menos atractivo a la vista si se generar diversas barras de datos. De cualquier forma, se puede modificar el código para tener diferente cantidad de rangos.

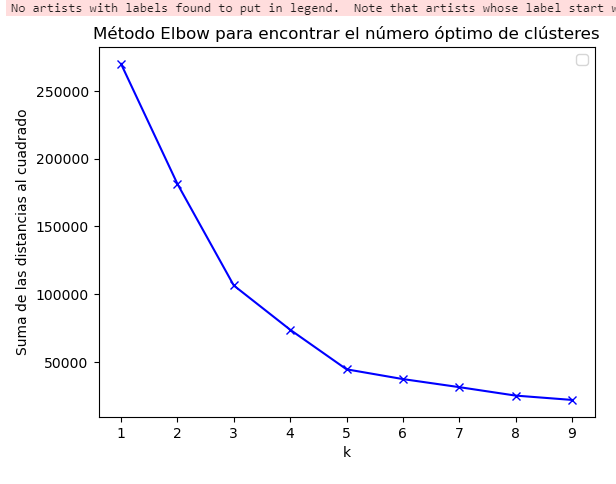


Se genera un histograma con los valores de la columna de calificaciones, en este caso se generaron mayor cantidad de rangos para poder tener una clasificación más específica de las calificaciones, lo que permite una mejor toma de decisiones para cada uno de los rangos propuestos en la tabla.



Se genera un nuevo dataframe para evitar accidentes con loas demás gráficos que emplean la información del dataframa, al nuevo dataframe se le remueven las columnas de Género y el Id, ya que no representan una utilidad para poder realizar las gráficas visuales.



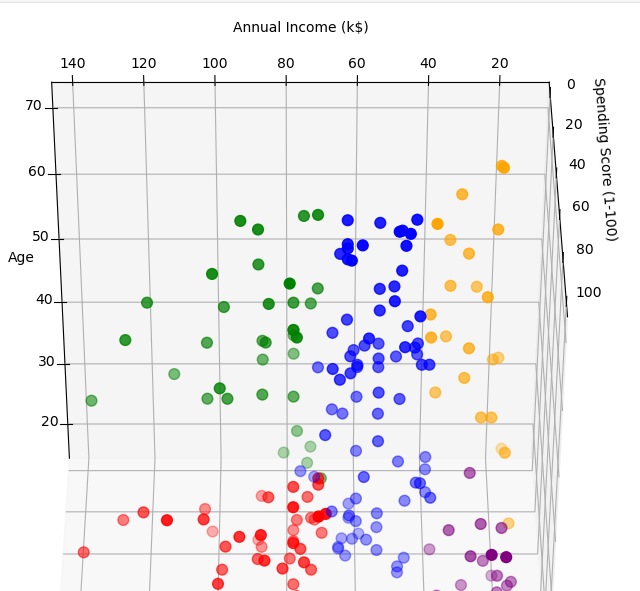
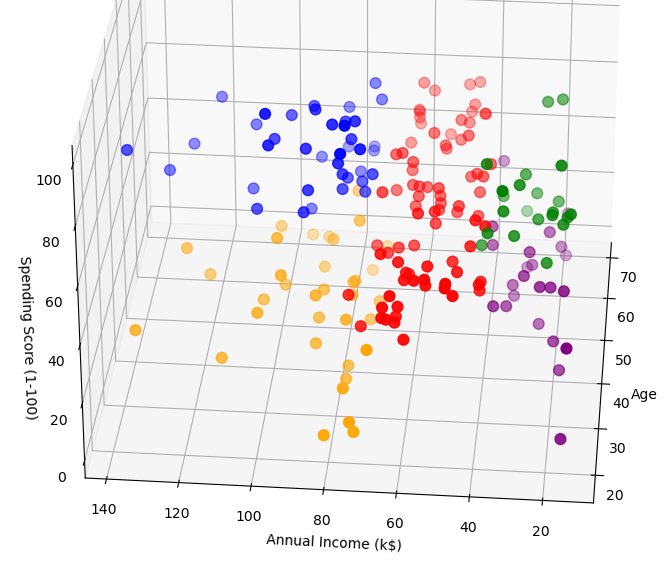
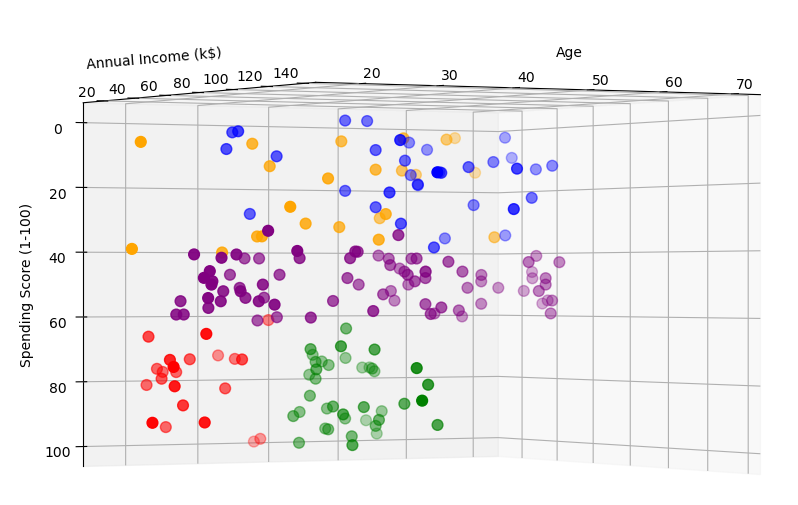


Se realiza la gráfica que nos permite saber cual es la cantidad adecuada de clústeres donde se debe observar que cantidad de clústeres representa una cantidad aceptable de errores.

Se genera el código para representar de forma visual los datos recopilados divididos en los clústeres que se decidieron, y se le asigna una orientación a la gráfica para ser mostrada.

Se utilizan diferentes orientaciones para poder observar los datos desde diferentes perspectivas.





Definición de los clústeres:

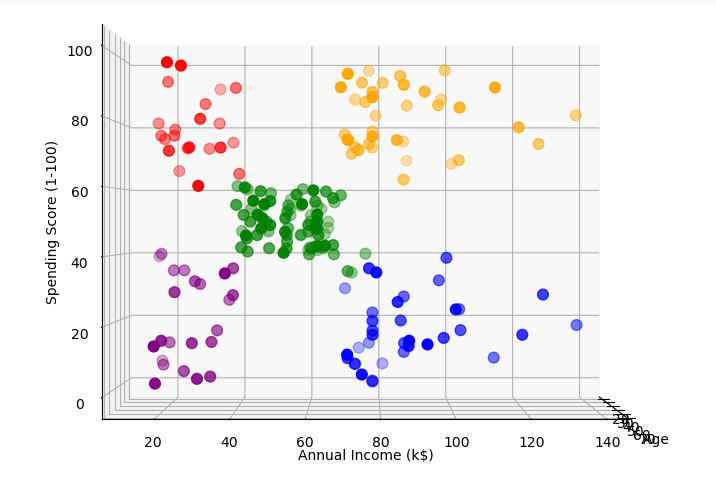
Clúster 1 (Verde): Este clúster concentra la mayor cantidad de persona en un punto medio tanto en el puntaje de gatos como en los ingresos, por lo que este clúster representa las personas que se encuentran en el centro donde los ingresos son medianos y el puntaje igualmente es mediano, igual se puede observar que este cluster tiene a las personas más cercanas entre sí.

Clúster 2 (Rojo): Este clúster representa personas con un alto porcentaje de Gasto (Spending Score) pero un ingreso muy bajo ,siendo casi de los ingresos más bajos de la gráfica, Las personas aquí generan varían compras pero registran bajos ingresos.

Clúster 3 (Morado): Este clúster representa personas con un bajo porcentaje de gastos y un ingreso anual bajo ,por lo que se entiende es el grupo de personas que menos consumen productos o servicios. Y se puede observar que estas personas son las de menor edad.

Clúster 4 (Amarillo): Este clúster representa a las personas con un alto porcentaje de gastos y de igual forma son las personas que mayores ingresos anuales registra, por lo que pueden ser considerados como el mejor mercado ya que son los que más pueden consumir los productos, por sus ingresos, lo que se ve reflejado en su porcentaje de Gasto.

Clúster 5 (Azules): Este clúster representa a las personas que registran altos ingresos anuales pero tienen un bajo porcentaje de consumo o gasto dentro de la tienda, por lo que podrían ser considera como un publico que puede ser atraído si se genera una buena oferta, ya que estas personas si tienen ingresos que podrían permitir acceder a diversos productos.



<https://drive.google.com/file/d/1KAhvaL2vB2Mco1HW23BywU_KdDsehSK6/view?usp=sharing>