**Introducción**

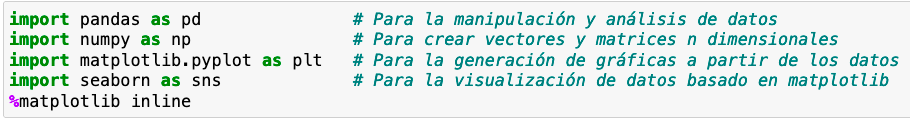
En esta práctica se utilizará el algoritmo K-medias para agrupar un conjunto de datos que se trata de un crédito hipotecario. Esta práctica es similar a la anterior pero se utiliza un algoritmo diferente para generar los clústeres. Se analizará cada uno de los clústeres y se utilizará el método del codo para definir el número de grupos en los que se agrupará. Por medio de otros conceptos como la selección de datos, visto en prácticas pasadas, se obtendrá el mejor modelo posible.

**Objetivos**

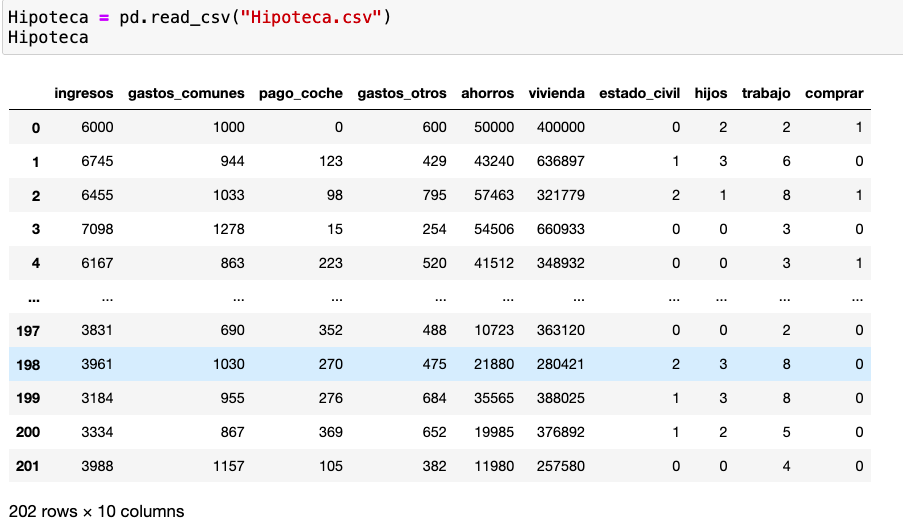
Obtener clústeres de casos de usuarios, con características similares, evaluados para la adquisición de una casa a través de un crédito hipotecario con tasa fija a 30 años.

**Desarrollo**

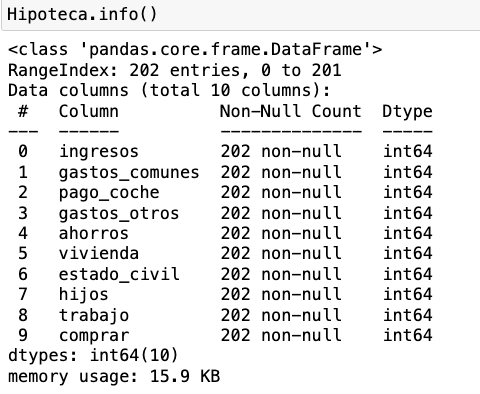
Se importan las bibliotecas necesarias para la realización de la práctica.



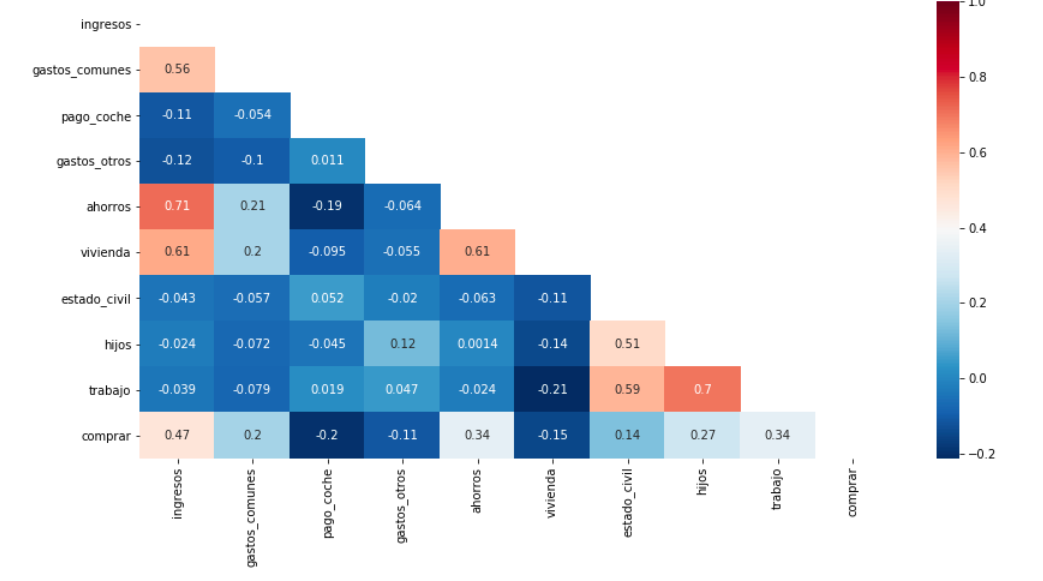
Se genera una vista de los datos que se van a utilizar durante la práctica.



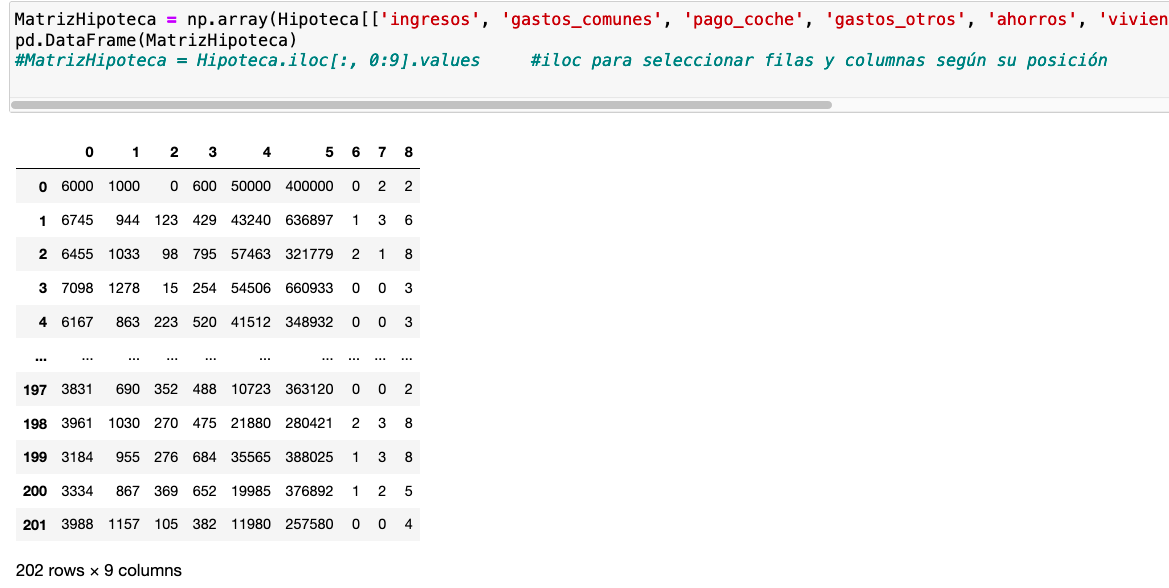
Se obtuvo la información del conjunto de datos para corroborar que no existan valores nulos dentro de los datos.



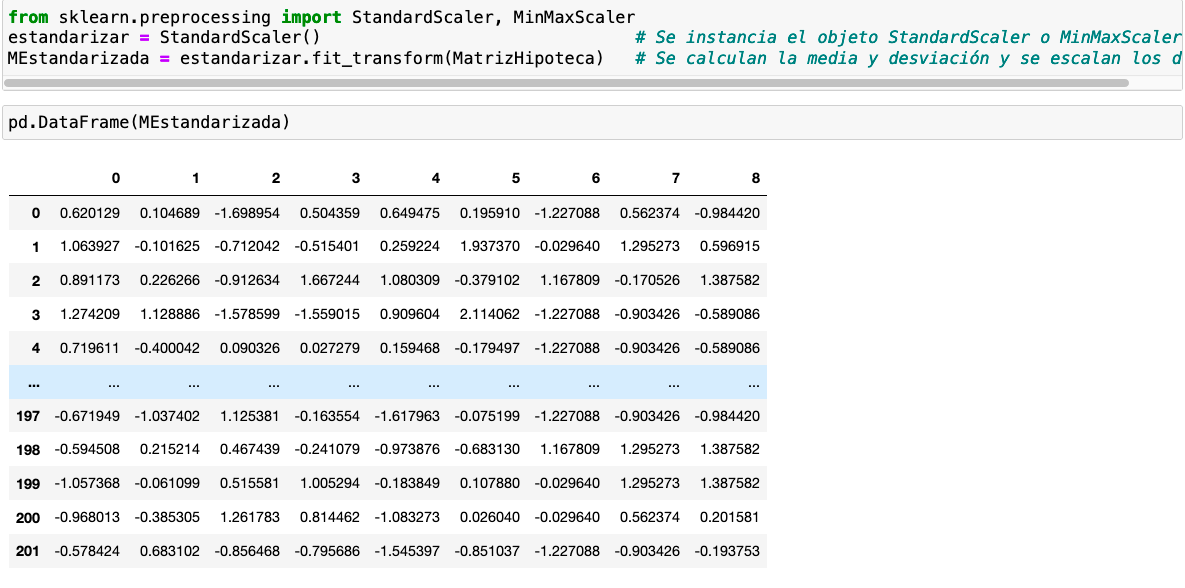
Se generó gráficas de las relaciones que existen entre todas las variables para generar una evaluación visual de qué variables están altamente correlacionadas. Además, se obtuvo la matriz de correlaciones utilizando el método de Pearson. A partir de un mapa de calor de esta matriz de correlaciones se puede ver dónde generar una reducción de dimensionalidad. Existen cuatro variables altamente correlacionadas: ingresos con ahorros y trabajos con hijos; sin embargo, estos datos son relevantes para el modelo y la correlación no es tan alta. Por esto solo se elimina la variable de comprar ya que esta variable representa un agrupamiento.



Se generó una variable a partir de las variables que se seleccionaron.



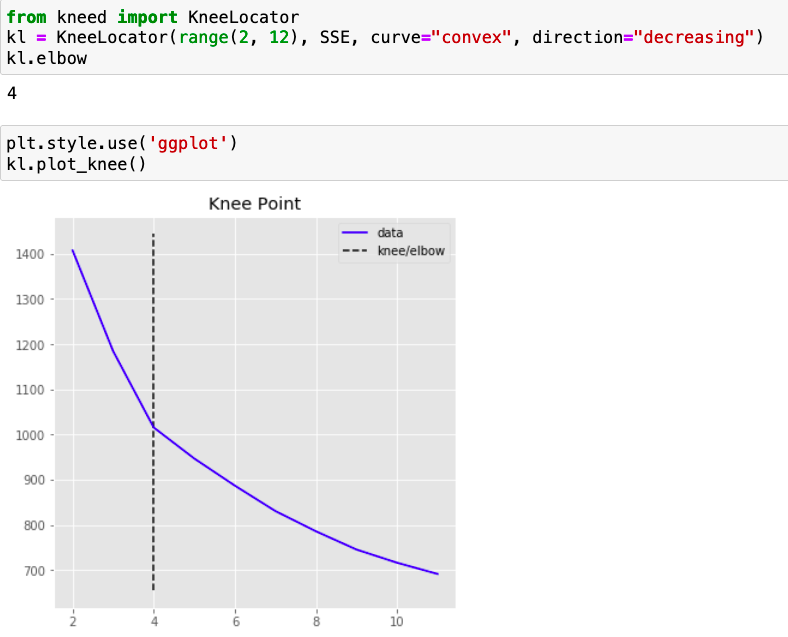
Para este tipo de algoritmos se necesita que los datos contribuyan de igual manera y para esto se genera una estandarización de los datos. La estandarización se genera como en prácticas pasadas.



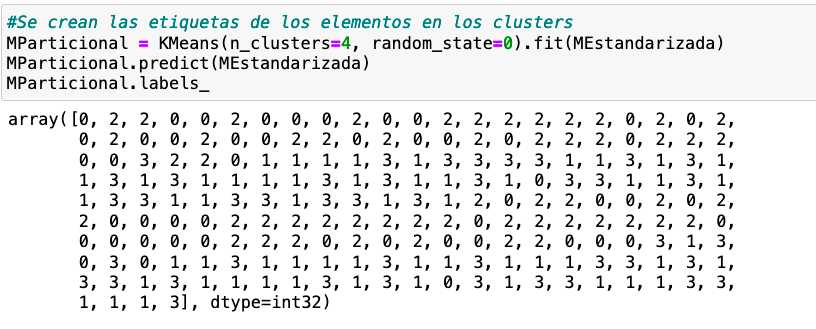
Lo primero que se realiza es el método del codo para determinar el número de agrupaciones que se deberá seleccionar para el algoritmo. Para esto se utiliza el método de *inertia* para generar las mediciones de SSE y a partir de ellas hacer la gráfica del codo donde el punto de inflexión representa el número de K.



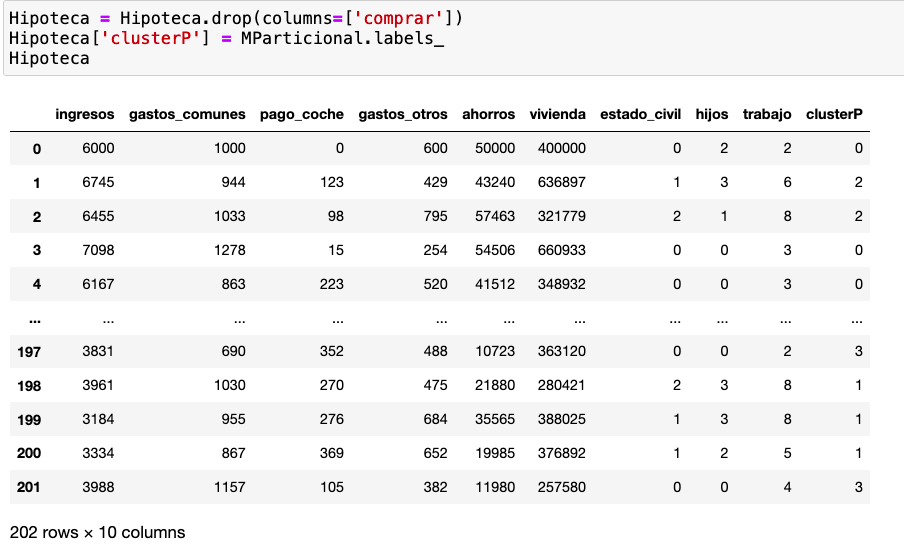
La gráfica que se genera no es totalmente clara de donde se da el punto de inflexión por lo que se utiliza una biblioteca para calcular este número. Utilizando Kneed se determinó que el número de clústeres debe de ser de 4.



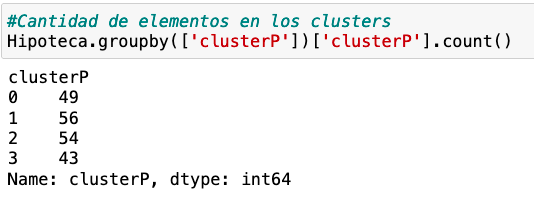
Después, se generaron las etiquetas para un K-medias de 4 clústeres a partir de la matriz estandarizada. Se utilizó un método que fue importado antes para generar dichas etiquetas.



Se reemplazó la columna de comprar que tenía las etiquetas originales de los datos por las etiquetas que generó el algoritmo.



Se agruparon los diferentes clústeres generados para ver cuántos datos contiene cada una de las agrupaciones.



Se genera una tabla de los promedios de cada una de las agrupaciones que se generaron. A partir de esta tabla se puede ver las características de cada una de las agrupaciones. Se generó un análisis de cada uno de los clústeres.



Clúster 0: Es un segmento de clientes conformado 49 usuarios, con un ingreso promedio mensual de 3502 USD, con gastos comunes de 857 USD, otros gastos de 533 USD y un pago mensual de coche de 245 USD. Estos gastos en promedio representan casi la mitad del salario mensual (1635 USD). Por otro lado, este grupo de usuarios tienen un ahorro promedio de 24129 USD, y un valor promedio de vivienda (a comprar o hipotecar) de 291900 USD. Además, en su mayoría son solteros (0-soltero), sin hijos y tienen un tipo de trabajo asalariado (2-asalariado).

Clúster 1: Es un segmento de clientes conformado 56 usuarios, con un ingreso promedio mensual de 3472 USD, con gastos comunes de 905 USD, otros gastos de 536 USD y un pago mensual de coche de 224 USD. Por otro lado, este grupo de usuarios tienen un ahorro promedio de 23957 USD, y un valor promedio de vivienda (a comprar o hipotecar) de 272010 USD. Además, en su mayoría son solteros (2-divorciado), con dos hijos y tienen un tipo de trabajo asalariado (7-autónomo y empresario).

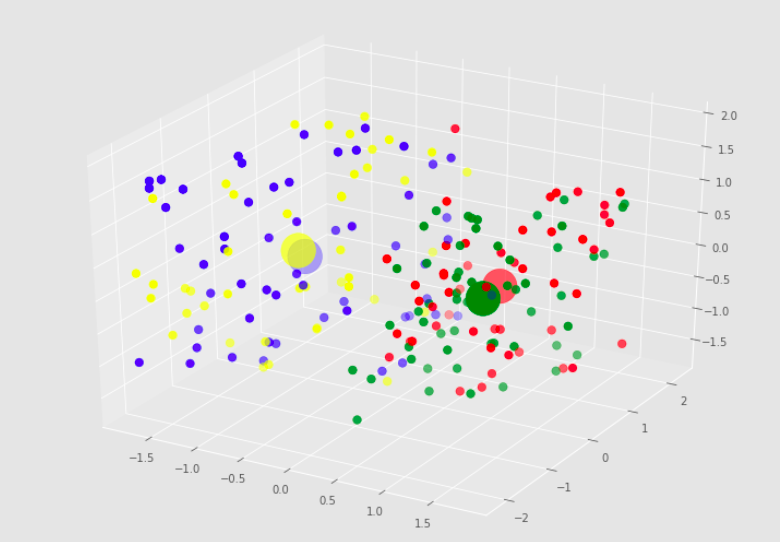
Clúster 2: Es un segmento de clientes conformado 54 usuarios, con un ingreso promedio mensual de 6389 USD, con gastos comunes de 998 USD, otros gastos de 524 USD y un pago mensual de coche de 190 USD. Por otro lado, este grupo de usuarios tienen un ahorro promedio de 54899 USD, y un valor promedio de vivienda (a comprar o hipotecar) de 430860 USD. Además, en su mayoría son solteros (1-casado), con dos hijos y tienen un tipo de trabajo asalariado (6-autónomo y asalariado).

Clúster 3: Conformado por 43 casos de una evaluación hipotecaría, con un ingreso promedio mensual de 6358 USD, con gastos comunes de 1117 USD, otros gastos de 465 USD y un pago mensual de coche de 190 USD. Estos gastos en promedio representan menos de la tercera parte del salario mensual (1772 USD). Por otro lado, este grupo de usuarios tienen un ahorro promedio de 50687 USD, y un valor promedio de vivienda (a comprar o hipotecar) de 497262 USD. Además, en su mayoría son solteros (0-soltero), casi sin hijos menores y tienen un tipo de trabajo, en su mayoría, asalariado (2-asalariado).

Por último, se generó una gráfica con tres ejes de las primeras tres variables de la tabla donde se colorea a cada uno de los diferentes grupos. En esta gráfica se puede observar la manera en que se generó la agrupación en una forma gráfica. El código para generar la gráfica fue el siguiente.



Los puntos más grandes representan a los centroides que se generaron.



**Conclusiones**

A partir de un conjunto de datos de usuarios que fueron evaluados a través de un crédito hipotecario se generaron clústeres de agrupaciones utilizando el algoritmo de K-medias. Se utilizó la matriz de correlaciones y el método de Pearson para realizar la selección de variables. También, se utilizó el método del codo para determinar el número de agrupaciones que se iban a generar. A partir de estos métodos y de una matriz con los datos estandarizados se pudo obtener un conjunto de cuatro clústeres que fueron analizados para saber las características de cada una de las agrupaciones. Este algoritmo sirve para generar agrupaciones de a quién se le puede otorgar el crédito y a quién no. Con esta práctica se pudo aplicar lo que se vio en teoría sobre el algoritmo de K-medias y se complementó con los conocimientos vistos en prácticas anteriores.