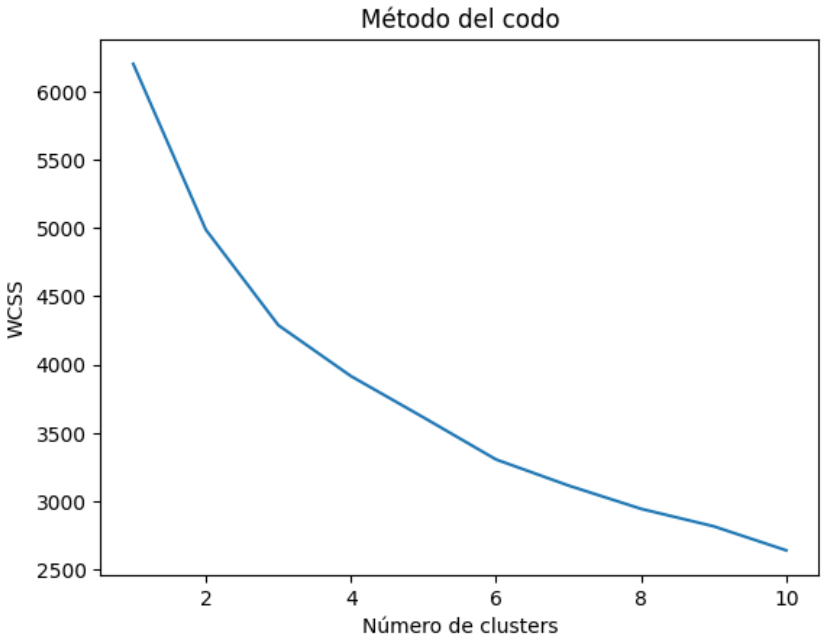
Reporte 4 Clustering

Para está practica, trabajé con un archivo obtenido de Kaggle (https://www.kaggle.com/datasets/luisenriquesguerrero/creditos-personales-actualizado), una base de datos donde muestra a clientes de un banco, ordenados por su ID, edad, experiencia laboral, ingreso y entre otros datos, la relación entre sus ingresos y gastos y si tienen algún préstamo, en resumen, un pequeño balance de su situación económica.

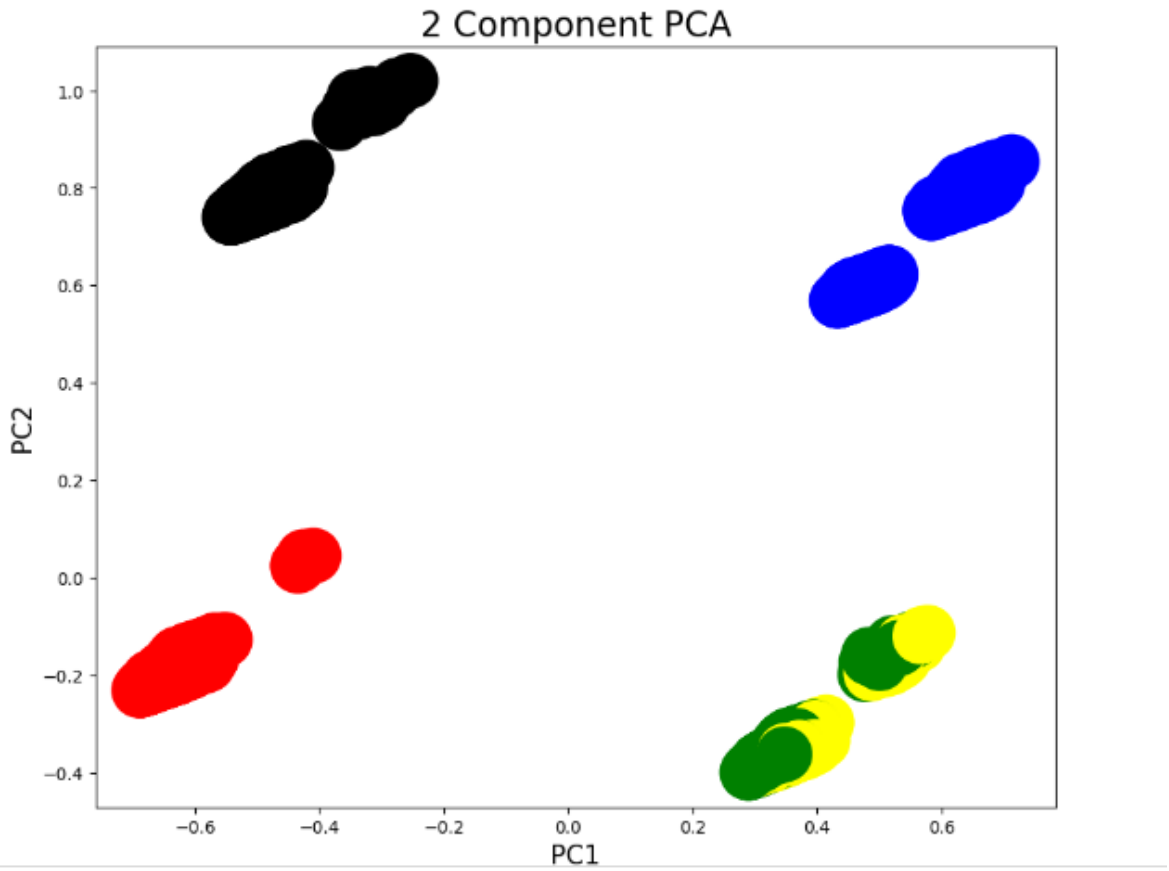
Para este trabajo, utilicé el método de K-medias, con ayuda del método del codo para saber el número de Clusters optimo para mis variables, en ese sentido tuve algunos problemas, asumo que fue por el hecho de tener algunas variables booleanas y otras mas con pocos datos ya que eran datos categóricos, estás variables y su poca variabilidad, posiblemente me causaron conflicto al utilizar el método de componentes principales, sin embargo, fue interesante re visitar estos métodos de estadística y su aplicación con un enfoque similar y a la vez diferente.

Grafica de codo




Por lo visto en la gráfica, no es muy claro donde sucede la reducción, podría ser entre 4 y 6, por lo que me decidí por 5 para el número de clusters (De igual manera si probé con 4 y vi datos muy separados que fueron parte de un mismo grupo)

Luego de realizar el método de componentes principales para mis datos y los 5 cluster, obtuve lo siguiente:




Por lo visto, tengo problemas con una serie de datos que se agruparon juntos, a diferencia de los tres restantes que si están bien segmentados, al intentar con 6 cluster el resultado no cambio en ese sentido, es decir, seguían presentándose mezclados.