K-Means and K-Mediods Clustering

1st Omar Farooq
omarfarooq101@gmail.com
15-12-2020

***Abstract*—This paper aims to illustrate the use of clustering algorithms in python to analyze data. Two use cases are discussed. The first is customer segmentation at a shopping mall, and the second is analysis of a materials database.**

Keywords—k-means, k-mediods, python, clustering.

I. Introduction

Clustering allows us to group together data points that are similar. It falls among unsupervised machine learning methods, as we do not need labeled data. To execute the clustering method, at first (1) all the data points are assigned to k clusters. (2) The centroid is calculated for each of the clusters. (3) A distance measure is then utilized to evaluate the proximity of the data points to the center (known as the centroid). Once the distance has been calculated form all points to the centroid, (4) points that are not properly assigned to clusters are reassigned correctly. (4) The algorithm is terminated if no reassignment takes place. If there is a reassignment, all the above steps starting at step 2 are redone. To implement clustering, k-means and k-mediods algorithms are used. Here the data is clustered into k groups or clusters, where k is a predetermined number. In k-means, centroid is the mean of the points in the cluster. In k-mediods, there are two possibilities. If there are an odd number of points, the centroid is the central element of the sorted list of points. If there are even number of elements in a cluster, the centroid is the mean of the two central points in the sorted list.

The number of clusters are determined using the elbow method. In this method, the distortion score (or the explained variation in data) is plotted as it decreases with an increase in number of clusters. The number of clusters at the elbow (the point where gradient suddenly increases) is taken as the number of clusters for further applying clustering methods such as k-means or k-mediods. The clusters were then plotted and their silhouette score calculated for performance evaluation. The Silhouette score determines how well the data has been clustered. It is a measure of how similar the data is in one cluster, compared to other clusters. Its value ranges from +1 to -1, where the former is for optimal clustering, and the latter for worst case.

There are a number of issues when using k-means and its derivatives. Firstly, when the dataset is not huge, the initial cluster formation at step 1 of the algorithm determines the cluster formation to a large extent. Secondly, the number of clusters k must be determined and input to the algorithm. Thirdly, due to the initial random seeding of clusters, the algorithm may give different results each time its run, especially if the number of points is small. Fourthly, initial assignment to clusters is also random, which affects result reproducibility. If data is input in a different order, it would yield different clusters.

In terms of computational complexity, the algorithm is fast and efficient. Its complexity is O(tkn) where n is the number of points, k is number of clusters, and t is number of iterations.

Clustering has applications in image segmentation or vector quantization, undirected knowledge discovery, unsupervised learning of neural networks, pattern recognition and classification, optical character recognition, medical image analysis, video tracking, speech recognition, biometric identification, biological classification, statistical natural language processing, document classification, credit scoring, micro-array classification, geo-statistics, market research, shopping items grouping, social network analysis, search result grouping, software evolution, evolutionary algorithms, Markov chain Monte-Carlo methods, anomaly detection, crime analysis, field robotics, finance, climateology and recommender systems.

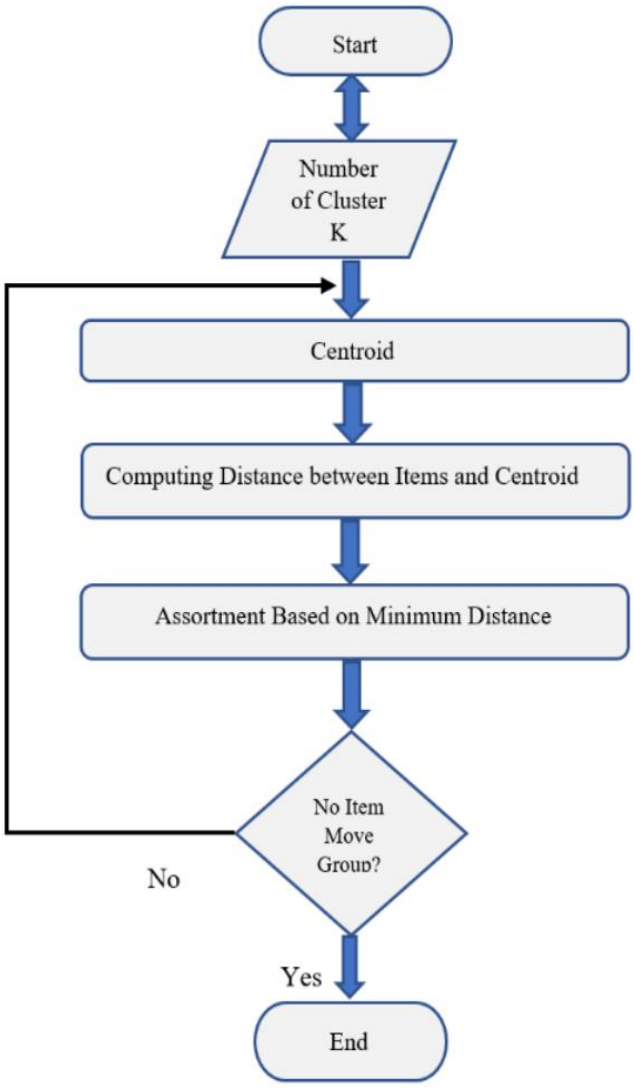

Fig.1: The clustering algorithm.

II. Experiment 1: Customer Segmentation

A. Purpose

Customer segmentation can be performed as part of market research and analysis to identify various groups among customers, so that they can be individually targeted. This allows marketers to better tailor their marketing programs for the various groups of customers, consequently increasing effectiveness and decreasing costs.

B. Dataset

Data of customers at a mall was obtained from Kaggle. The link for the dataset is given at the bottom of this page. The attributes include customer ID, age, gender, annual income, spending score.

C. Exploratory Data Analysis

After loading data onto a pandas data frame, it was encoded to remove object type variables in the gender column. Mate was encoded as '1' and female as '0'. The customer ID column was dropped. An exploratory data analysis was then performed.

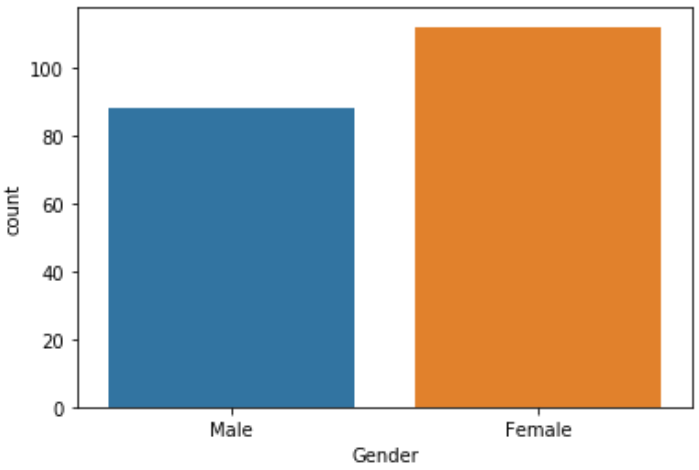

Fig 2: Count plot of genders among customers. There were slightly more female customers than male customers.

	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823522
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000	137.000000	99.000000

Fig 3: A basic description of the data.

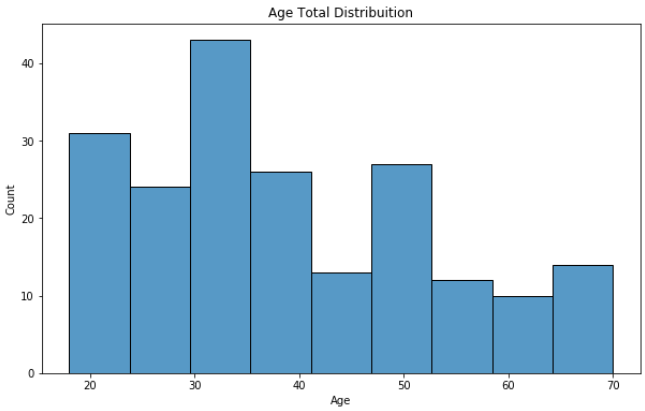

Fig 3: The distribution of ages among the members. 30-35 was the largest group

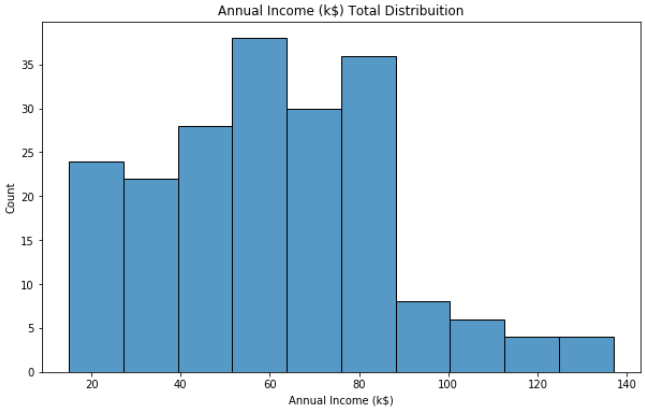

Fig 4: A histogram plot of annual income in 1000 dollars

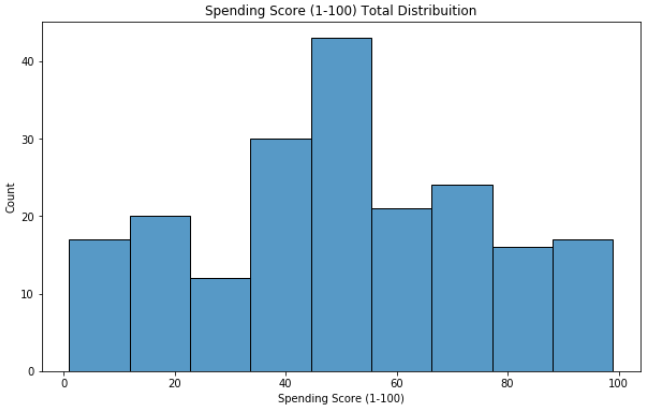

Fig 5: The spending scores of the customer sample, plotted as a histogram

Next we wanted to present the data such that all clusters were visible. For that we plotted the data as a pair plot as seen in figure 6. It was seen that the plot of spending scores with annual income is of particular interest as it describes the population demographics of interest to us. Therefore, the plot was plotted again as found in figure 7. It can be seen that 5 clusters are visible in this plot.

Dataset: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

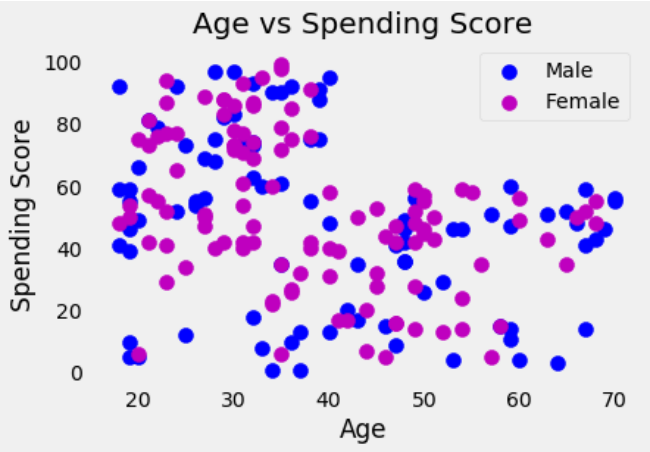

Fig 6: Age Plotted against spending score. Men and women are plotted in different colors

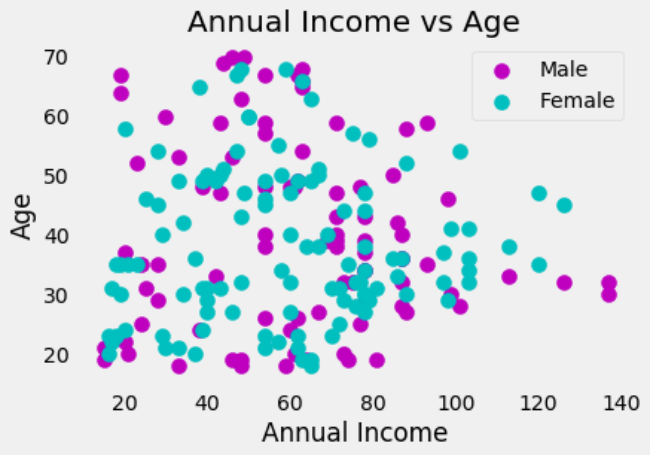

Fig 7 Plot of Annual Income with Age. Men and women are plotted in different colors

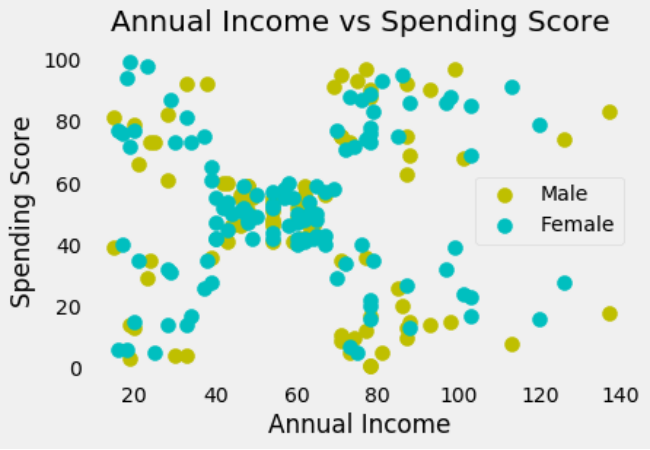

Figure 7: Plot of annual income with spending score. This plot is of particular interest to us. 5 Clusters can be seen in this plot. Men and women are plotted in different colors

D. *Elbow Method*

For finding the number of clusters in the data, the elbow method was utilized from the yellowbricks python package. The number of iterations were set to 15, and the k-means clustering package from the sklearn library was used. The Elbow Method gave the number of clusters k=5, which corresponds to our observation in the plot of Annual income with spending score in figure 7. The plot of distortion score with the number of clusters k is given in figure 8.

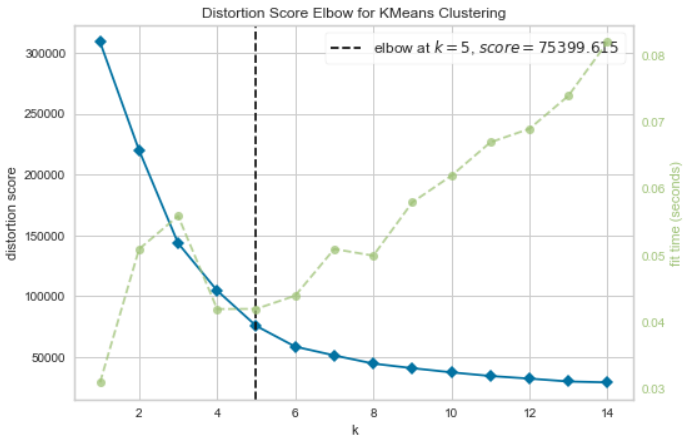

Fig 8: Plot of distortion score with the number of clusters. The result of k = 5 corresponds to our observation in fig 7 that there are five clusters.

E. *K-Means Clustering*

Next, k-means clustering was performed with k set to 5. Init was set to k-means++, for quicker convergence. A scatter plot was then plotted with hue for different clusters set to different colors.

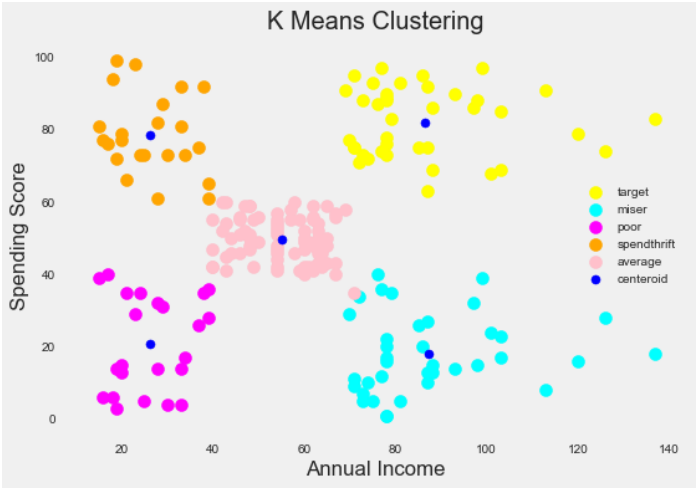

Fig 9: Scatter plot after k-means clustering with different colors for different clusters.

One problem that was observed was that every time the clustering algorithm was run, the clusters were labeled with different numbers, so the colors had to be re adjusted.

To evaluate the performance of the clustering algorithm, the silhouette score was calculated, and was given as 0.4438841244266416.

K-Mediods Clustering

The data was then clustered using the k-mediods algorithm. A scatter plot was again plotted to illustrate the new clusters.

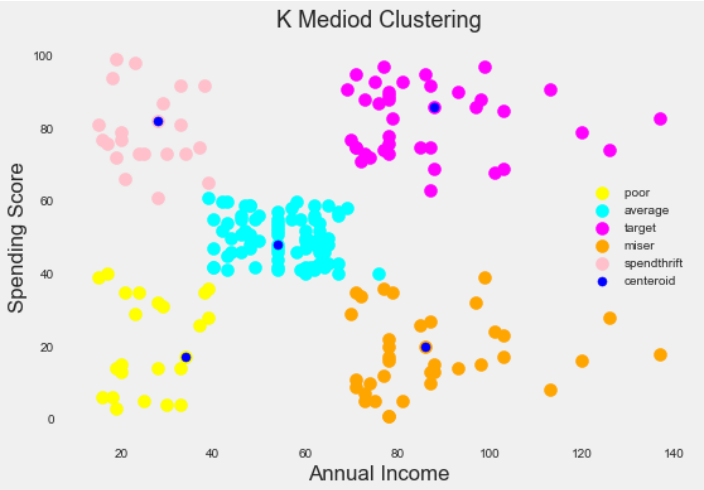

Fig 10: Scatter plot after k-Median clustering

To evaluate performance of the clustering algorithm, the silhouette score was calculated as 0.4438841244266416.

III. Experiment 2: Materials Project Database

A. Purpose

The Materials Project is an online database hosted by the US department of commerce, that contains various engineering properties of materials. This dataset illustrates the current sate of the art in Materials Science. It reflects the current state of human progress, as an age has been historically defined in terms of the materials that had been made available for engineering purpose. The current data analysis aims to categorize various materials based on their mechanical properties namely the bulk elastic modulus, elastic shear modulus and materials anisotropy.

B. Background Knowledge

The Shear modulus describes a material's tendency to shear when acted upon by opposing forces, and is given as the ratio of shear stress to shear strain

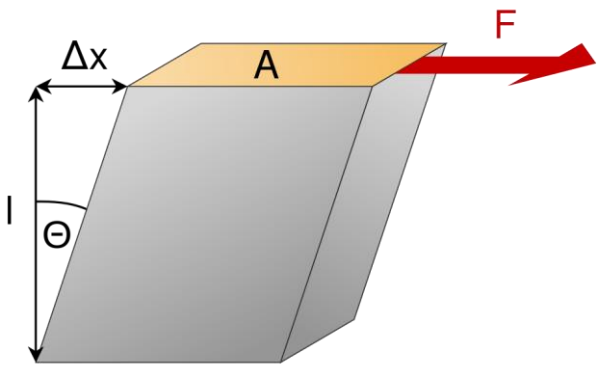

Fig 11: Shear Modulus (G) = $\frac{F/A}{\Delta x/l} = \frac{Fl}{A\Delta x}$

The Bulk Modulus is a measure of how resistant to compression a material is. It is given as the ratio of infinitesimal pressure increase to the resulting relative decrease in volume.

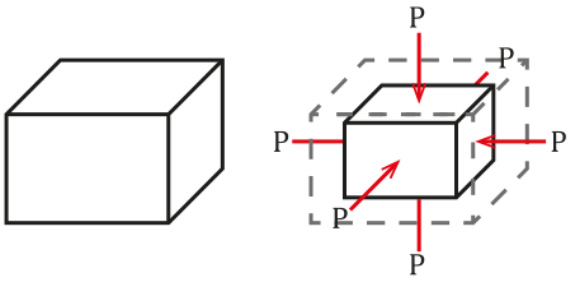

Fig 12: Bulk modulus K is for isostatic pressure deformation

Elastic anisotropy is a measure of the degree of anisotropy in the elastic modulus of the material. In an anisotropic material, material properties change with direction.

C. The Dataset

The data used was a 840mb database dump of materialsproject.org, available at the link given at the bottom of the page. It includes data of various properties of materials.

D. Exploratory data analysis

After data had been loaded into a dataframe, all columns were dropped except for materials id, formula, elastic-anisotropy, bulk elastic modulus and elastic shear modulus.

First a pair plot was plotted, using the latter three columns.

A Scatter plot was then plotted and can be found in figure 14 on the next page. The data comprised of a many outliers, and a large number of points near the origin. In order to see more clearly the centrally located points, the outliers were cropped and the data was plotted again. The plot can be found in figure 15.

E. Elbow Method

For estimating the number of clusters in the data, the elbow method was used. The plot can be seen in figure 13.

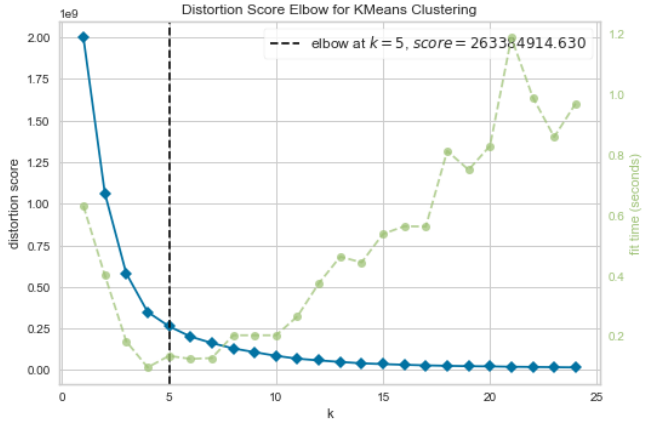

Fig 13: The elbow method indicates that there are a total five clusters in the data without outliers removed

Dataset: https://figshare.com/articles/Materials_Project_Data/7227749

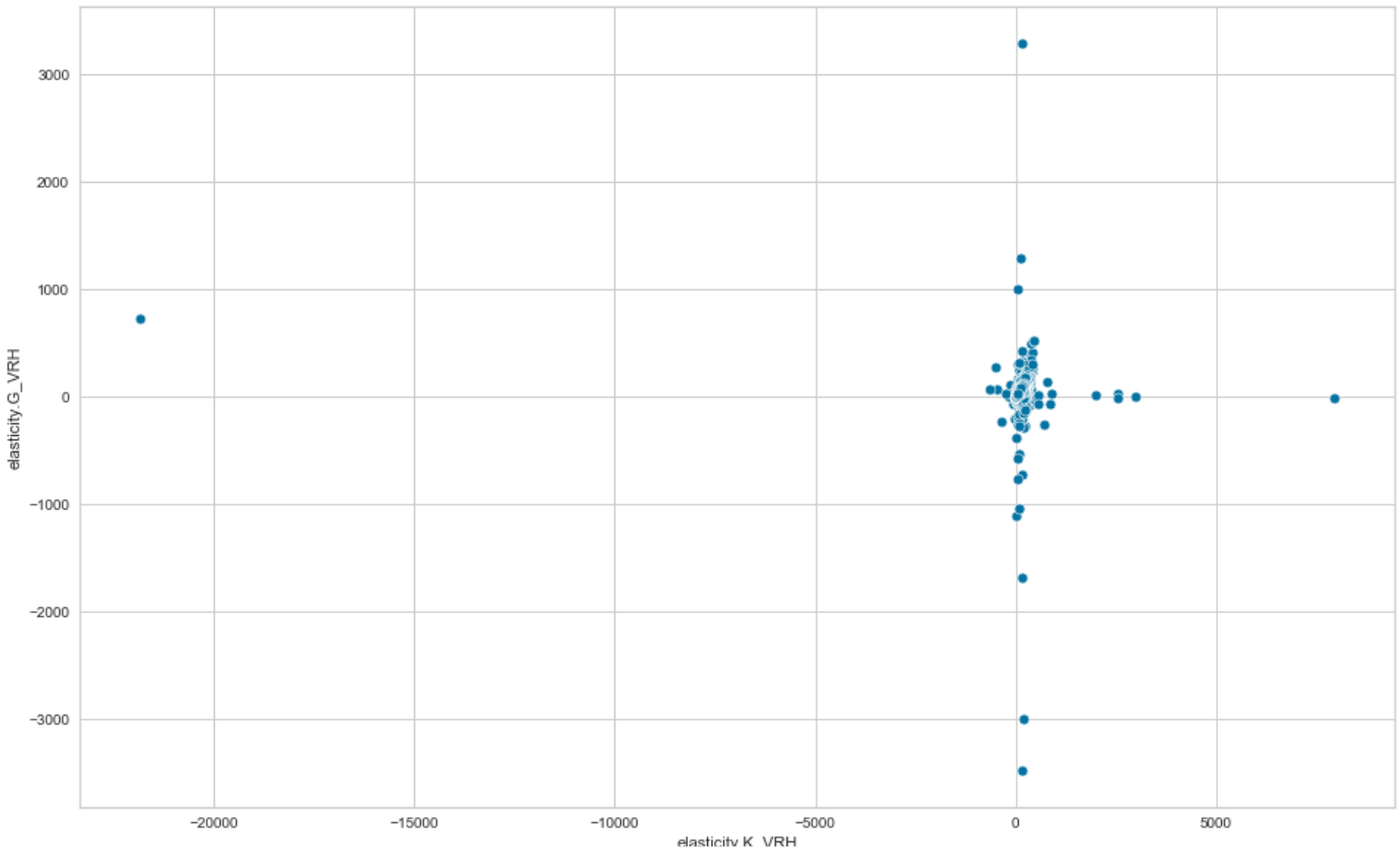

Fig 14: The values of bulk modulus (K) plotted against the values of the shear modulus (G) with no outliers removed

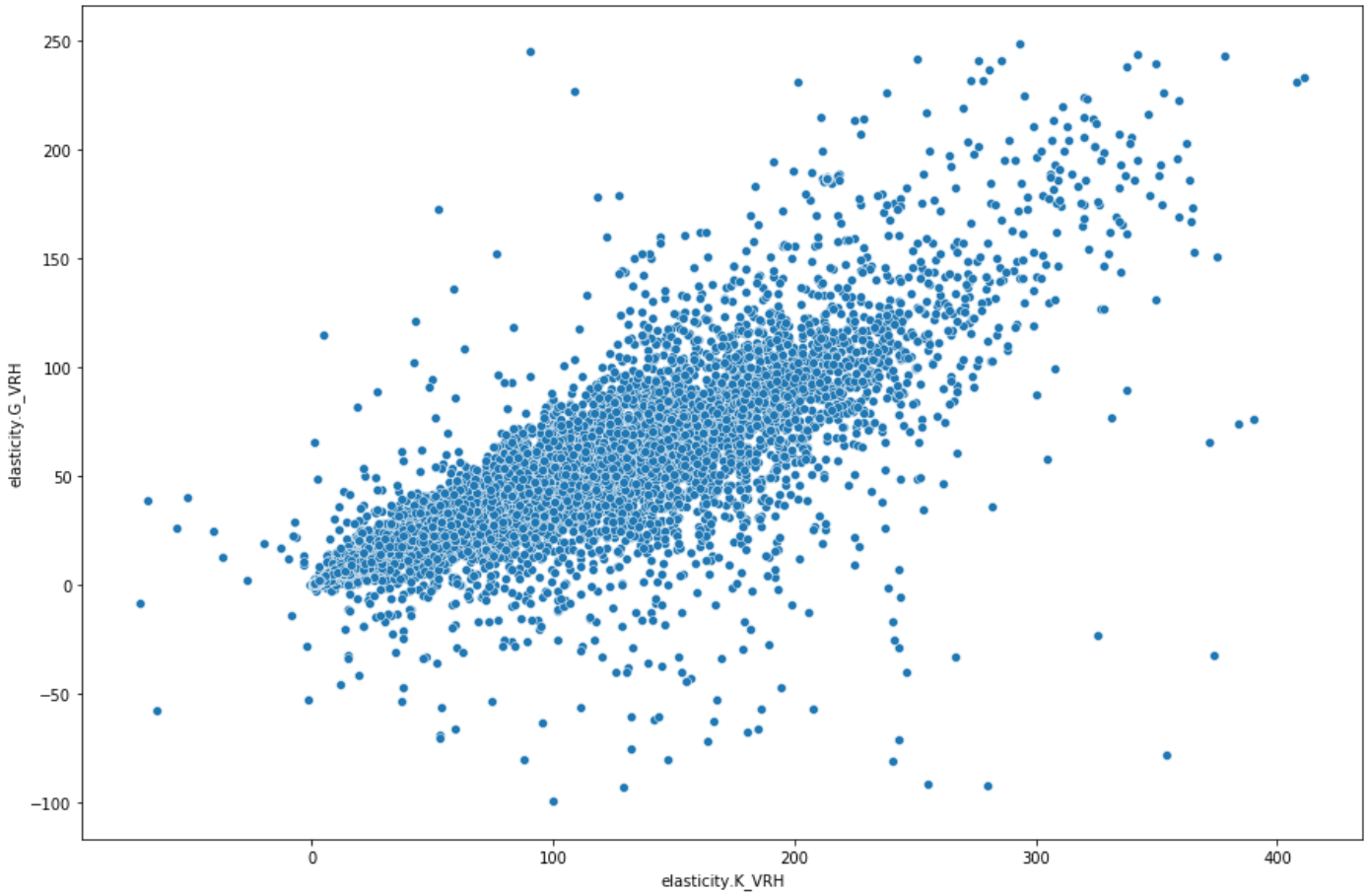

Fig 15: The values of bulk modulus (K) plotted against the values of the shear modulus (G) with outliers removed

K-means Clustering

K-means clustering was then performed with k=6. The silhouette coefficient for the clustering was 0.9839786822602455.

K-means clustering was then performed with k=5. The silhouette coefficient for the clustering was 0.9855935394543212.

K-means clustering was again performed with k=4. The silhouette coefficient for the clustering was 0.5159173839090453.

K-mediods Clustering

K- mediods clustering was then performed with k=6. The silhouette coefficient for the clustering was 0.9938422655752766.

K- mediods clustering was again performed with k=5. The silhouette coefficient for the clustering was 0.9852400804968715.

K- mediods clustering was again performed with k=4. The silhouette coefficient for the clustering was 0.5159173839090453.