MEMORANDUM

DATE: Sep 22, 2022

TO: Mitch Cochran, MA, MHS, CISM, CGCIO

FROM: Thao Hau Nguyen, BA K61, NEU Vietnam

SUBJECT: Customer segmentation applied K-Means Clustering in Python

I'm writing to show the results of clustering from customer dataset, which is helpful for retail companies to decide targeted customers and run effective programs to boost sales. Clustering is significantly popular to identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively.

This report includes 2 main parts: description of dataset and results of dataset applied K-Means clustering in Python.

Firstly, I collected dataset of 200 observations related to the 5 following attributes and ensured no missing values:

- CustomerID: a unique identification number given to every customer
- Gender: Male or Female
- Age
- Annual Income: the amount of income you earn in one fiscal year, by k$
- Spending Score: a score that the mall computed for each of their clients based on several criteria. The score goes from 0 (low spends) to 100 (high spends).

```
df.info()
df.shape
```

✓ 0.6s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB

(200, 5)
```

Some general visualization and some simple graphics regarding its data were implemented to improve our comprehension about a future model.

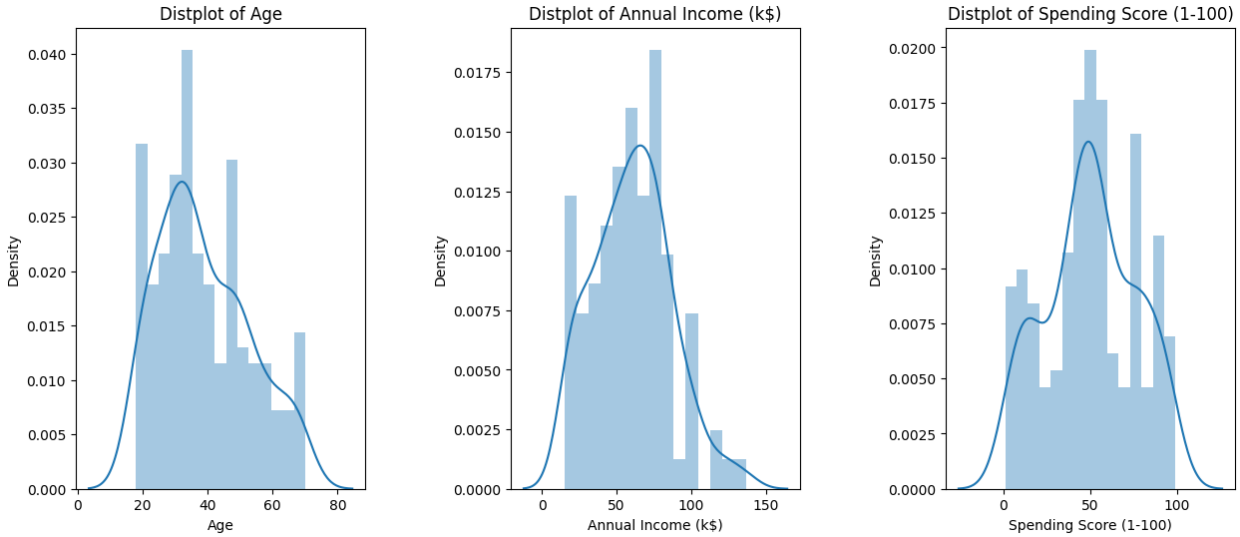

Figure 1: Distributions of Age, Annual Income and Spending Score

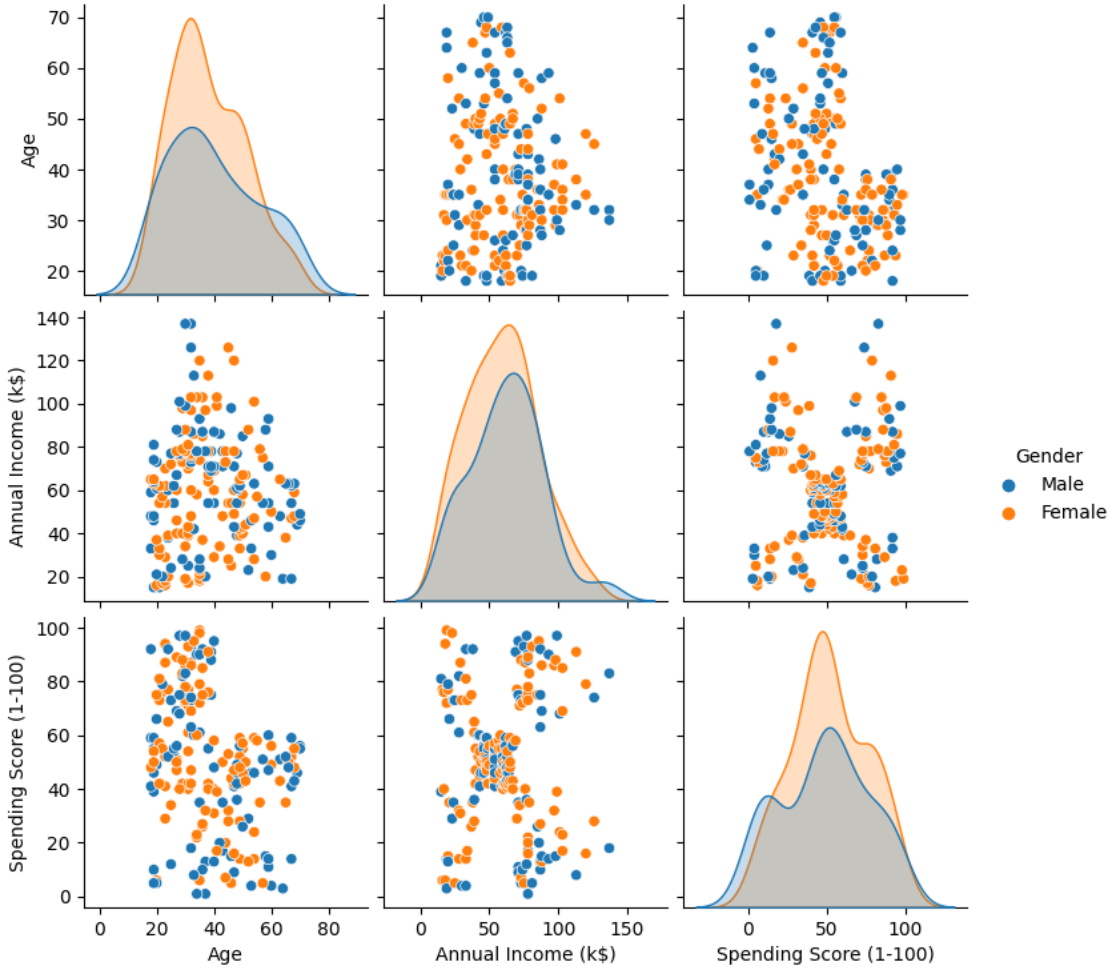

Figure 2: Scatter plots of 3 attributes by Gender

From figure 2, data analysts can cluster with 2 of 3 attributes depending on the purposes of businesses and time for projects. However, in this report, I will show the result of clustering by Age, Annual Income and Spending Score.

To satisfy the first property of clusters - minimize the distance of the points in a cluster with their centroid, I used centroid-based algorithm, which is also known K-Means. The first step of K-Means Clustering is to choose the optimum number of clusters k by using Elbow curse, where the x-axis will represent the number of clusters and the y-axis will be the inertia:

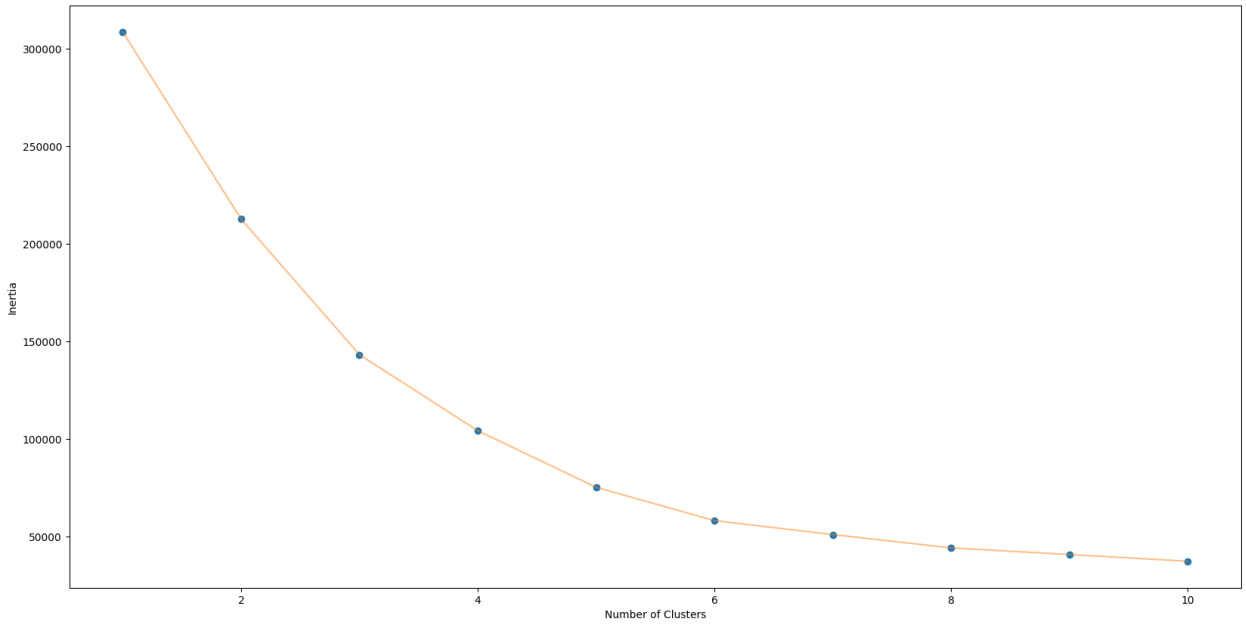

Figure 3: The Elbow method using Inertia

From a visual perspective of the graph above, I identified 6 clusters of clients, from a raw behavioral perspective:

```
print(df.groupby('cluster').mean())
✓ 0.6s
```

cluster	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
0	22.181818	44.318182	25.772727	20.272727
1	162.000000	32.692308	86.538462	82.128205
2	82.022727	56.340909	53.704545	49.386364
3	23.090909	25.272727	25.727273	79.363636
4	90.052632	27.000000	56.657895	49.131579
5	164.428571	41.685714	88.228571	17.285714

Figure 4: Mean of 3 Attributes by Clusters

To get more information about 6 clusters, not only looking the means in figure 4, I also continued to dig into the features of each cluster, which I present fully in the Python file. In conclusion, I will summarize some main characteristics of all 6 clusters below:

- Cluster 0: 39 clients of all ages (ranging from 19 to 67), with a low annual income and low spending score.
- Cluster 5: Clients with large range (from 19 to 59 years old), who have highest annual income stretching from 71k$ to 137k$, but lowest spending score.

- Cluster 1: 44 Millennials under 40 years old with 2nd in terms of annual income out of 6 groups and highest spending score.
- Cluster 2: Oldest group (Gen X and Baby Boomers) which includes 22 clients (from 43 to 70 years old) with middle annual income and middle spending score.

- Cluster 3: Youngest client's group with age from 18 to 35, with an as low annual income as cluster 0 but very high spending score.
- Cluster 4: Slightly older than cluster 3 with 38 observations, but nearly 2.3 times the annual income of that group but average spending score.

Link of project about Customer Segmentation I did apply K-Means Clustering in Python: here

Thank you for reading!

Thao Hau Nguyen