Anish Ahuja
501107430

Assignment 4

Background

My dataset describes basic details about my customers such as customer id, age, gender, annual income, and spending score. This data was acquired by the owners of a mall through membership cards (can also think of how Costco works). The spending score is something they assign to their customers based on defined parameters like consumer behavior and purchasing data. This value ranges from 1-99. Age ranges from 18-70. Annual income ranges from 15 to 137 (k$). Genders are male and female – 56% female and 44% male. Customer id is a unique id that ranges from 1 to 200, and there are 200 instances or customers in total.

Methodology

The first thing I did was run various print statements to understand my data better. The result of this is shown below:

```
... print(df.head())
... print(df.tail())
... print(df.shape)
...
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
     CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
195         196  Female   35                 120                      79
196         197  Female   45                 126                      28
197         198    Male   32                 126                      74
198         199    Male   32                 137                      18
199         200    Male   30                 137                      83
(200, 5)
```

This helped me understand how many features, number of instances, and the type of data (int, string, etc.) I was working with.

Upon going through the data, I discovered that this data required very little pre-processing. This is because all the features have the same type of data which are integers, and their values are about the same range as one another. My data also did not have any null values or any extreme outliers (like someone with a one million dollar annual salary). The only non-integer form of data is gender, and its values can be encoded into 1s and 0s. Since I was going to eliminate the gender feature for my clustering (see Results for explanation), I decided to not do any pre-processing on this data for the purposes of clustering with K-means. When clustering with the other methods, I did some scaling of my feature set to show the differences between the clustering in terms of when I scale the data vs unscaled.

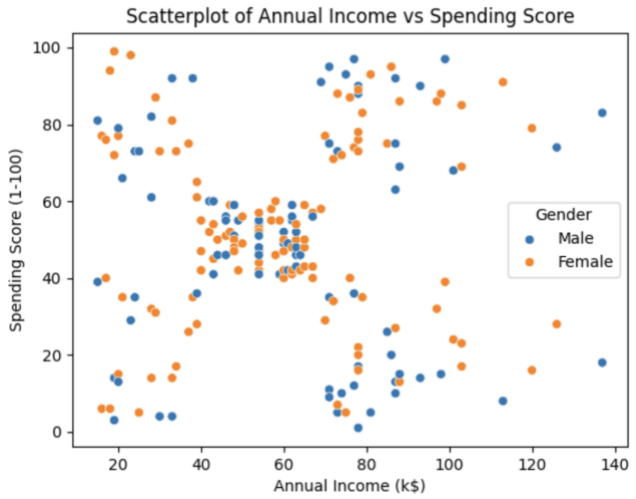
Before I can begin clustering my data, I needed to eliminate features that had no valuable impact on my data. Since I am working from the viewpoint of the mall owners that collected this data, I need to come up with results that are valuable to them. For example, any mall owner would be interested to know who spends the most money at their malls (basically has the highest spending score) and what features play a role in this. Logically speaking, I can eliminate customer id as a feature that plays no role in how much a customer spends at the mall. I, however, need to include the annual income feature as a person's income is a limitation to how much money they are able to spend at the mall. The two questionable features remaining are gender and age. I need to understand the impact of gender on mall spending, specifically whether women tend to spend more than men or vice versa. Simultaneously, an examination of the age factor is crucial, exploring whether younger individuals exhibit higher spending patterns compared to their older counterparts or if the trend is reversed.

Results

The first graph below shows a scatterplot of Annual Income vs Spending Score.

From looking at this graph, I can start to see that there are about 5 individual clusters. They are located at the top level corner, bottom left corner, center, bottom right corner, and top right corner.

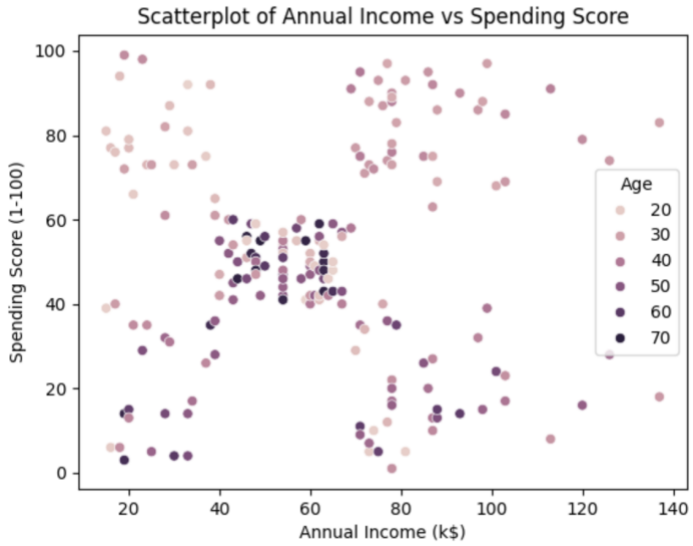
The second graph below shows the same scatterplot except it now uses a hue to separate genders. This is to see if gender plays a factor in both annual income and spending score.



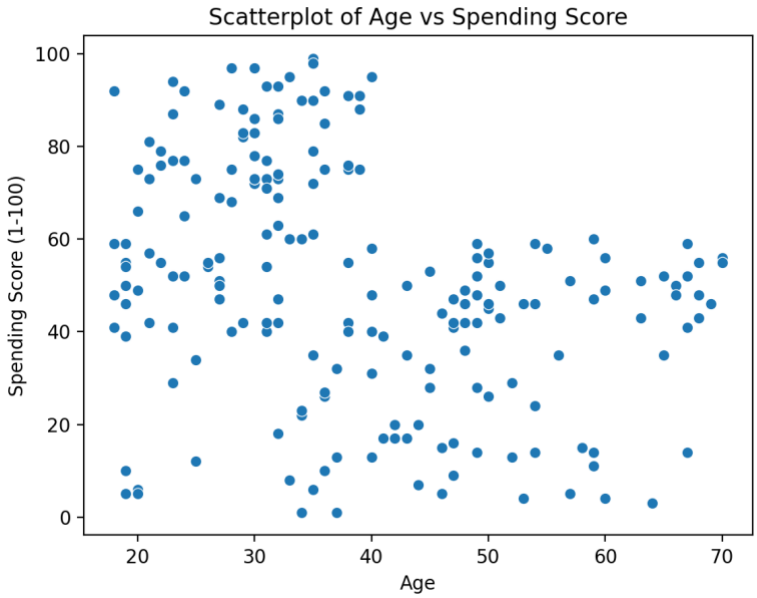
Looking at this graph, I can see that there are males and females in every cluster mentioned. This signifies an approximate equal distribution of males and females which means its not obvious if a certain gender earns more or spends more than the other.

When graphing Gender vs Spending Score, I can see that there is no clear correlation or clusters formed between gender and spending score. Both females and males have spending scores that range from the single digits to the 90s.

The next graph below shows the same scatter plot, except it now uses a hue to separate age categories. This is to see if age plays a factor in both annual income and spending score.



From this graph, I can see that the age groups are somewhat evenly spread out among the different clusters. Notably, the top left cluster seems to only have younger aged customers and this cluster also represents the lowest earning customers that have high spending scores. The top right cluster also only has 40 and under aged customers and this cluster represents customers with high incomes and high spending scores.

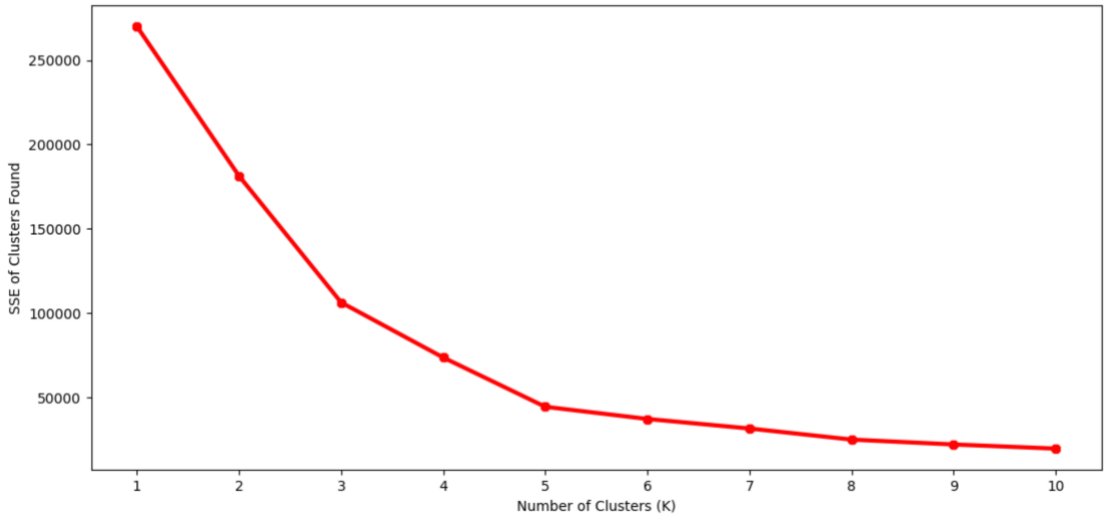


When I graph a Scatterplot of Age vs Spending Score, I don't quite see any clusters form, but I do notice a trend where older customers from the ages of over 40 to 70 have a lower spending score than those of the age of 40 or less. Nevertheless, in the absence of evident clusters, it is logical to eliminate Age as a feature.

Now that I have established that Gender and Age are features that can be eliminated, I can start clustering the original Scatterplot of Annual Income vs Age, which is restated below.

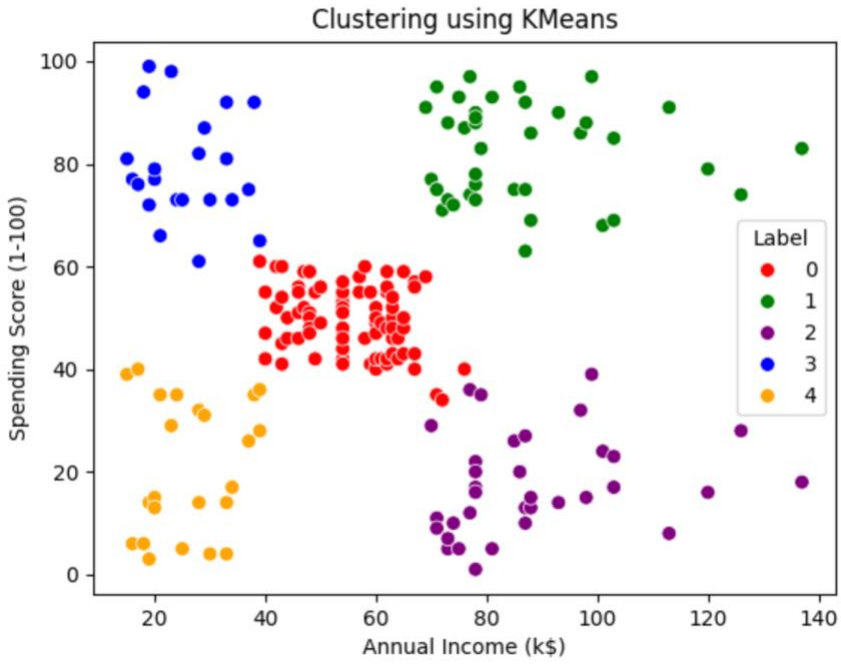
From a logical standpoint, I can see that 5 clusters are formed (top level corner, bottom left corner, center, bottom right corner, and top right corner), but I need some sort of mathematical reasoning to justify my use of 5 clusters.

The best way to do this is by building an elbow diagram that shows the SSE of Clusters found vs Number of Clusters (K). Shown below is the elbow diagram I came up with using the algorithm discussed in class:

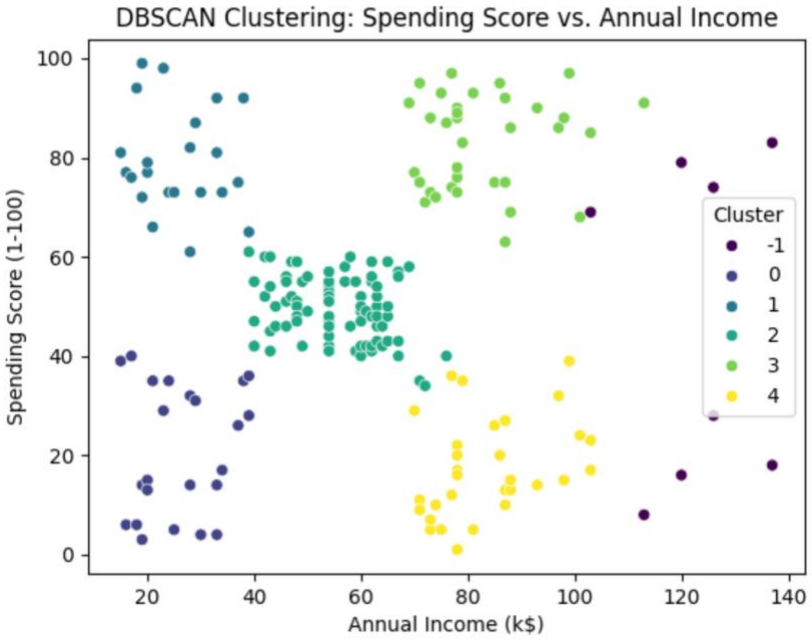


From this elbow diagram, I can see that when K is equal to 5, the rate at which SSE decreases is significantly less, which means that 5 clusters is the ideal number of clusters needed to cluster this dataset.

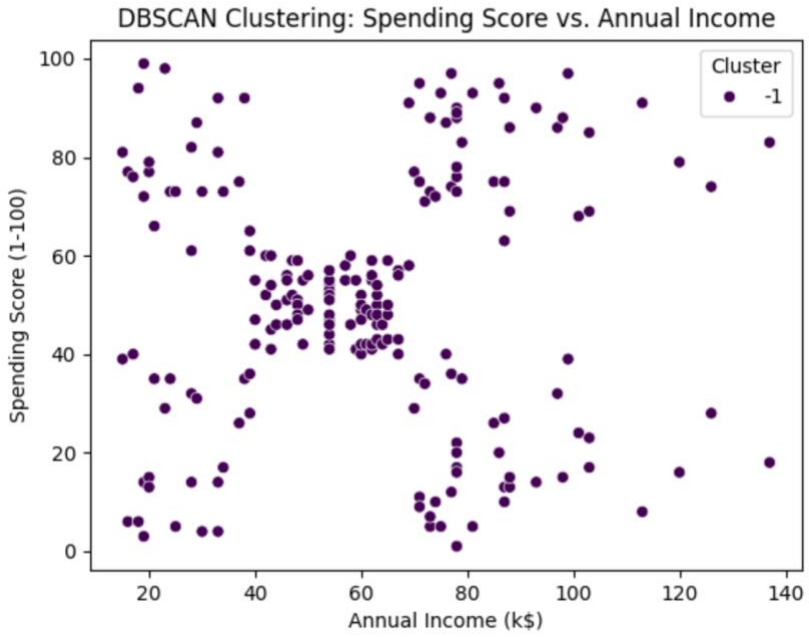
Now that I have established the number of clusters needed to cluster this dataset, I can start to cluster this dataset using the different methods I learned in class.



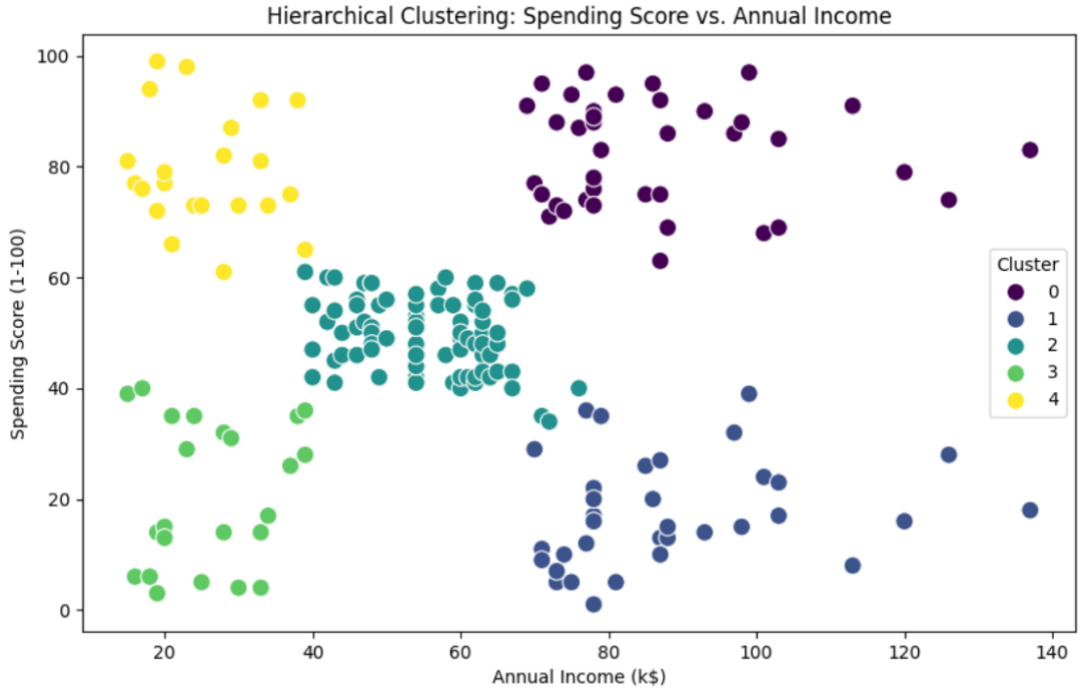
In this graph, I clustered the dataset using K-means and this gave us 5 clusters, which is the ideal number of clusters for this data set. To come up with this graph, I used the algorithm shown in class and iterated through it 10 times.



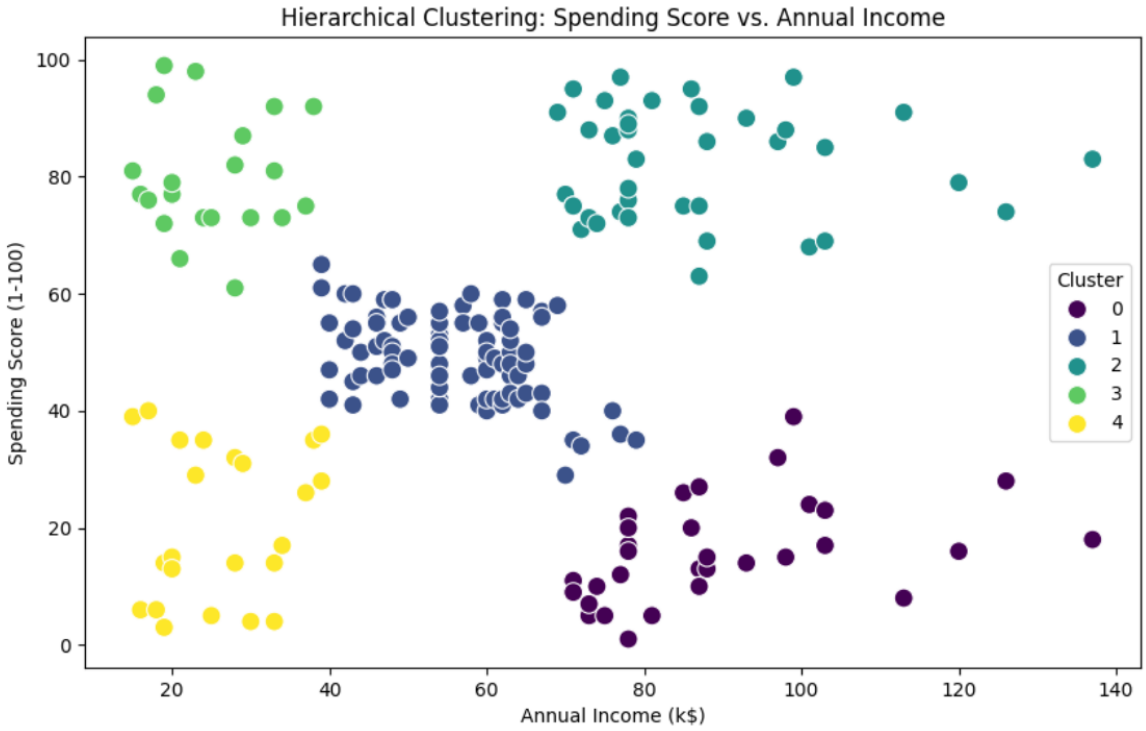
In this graph, I clustered the dataset using DBSCAN. This also gave me 5 clusters, but it also eliminated some points as noise from the other clusters. These points are labelled as -1. To come up with graph, I needed to scale the features as that is required for a density-based clustering algorithm.



If I hadn't scaled the features, my graph would have looked like this.



In this graph, I clustered using the Hierarchical clustering method. I specified the number of clusters to 5 because that was the ideal number of clusters for this dataset based on the elbow diagram shown earlier. In this graph, I scaled the features, as there were some differences in the clusters based on whether or not I scaled the dataset.



For example, this is the graph of Hierarchical clustering where the features are not scaled. The difference between both graphs is minimal, but I would argue the graph with the scaled features is slightly more accurately clustered and is the exact same as the K-means clustered graph.

Conclusion

Overall, no matter what type of clustering algorithm I used, the results were very similar. This is partly because I used K = 5 for all 3 algorithms as that was the ideal number of clusters needed to cluster this dataset based on both mathematical (elbow diagram) and logical reasoning. With unsupervised machine learning, there is always this degree of variance that exists between models. I noticed this when I clustered with both the scaled and unscaled feature set for Hierarchical clustering. The difference between both graphs was very minimal, but it nevertheless existed. I also noticed that when I used the DBSCAN method to cluster my dataset, it separated some instances and labelled them as -1. Since these instances were outliers, I assumed the modelled labelled them as noise to keep the data as accurate and precise as possible.

References

V. Choudhary. (2018). Mall Customer Segmentation Data. Kaggle. [https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python]