Clustering Methods for Mall Customer Data (June 2019)

Jason Huggy, Student, Lewis University

Abstract—The purpose of this study was to use clustering methods to determine the spending score for mall customers. Features used to determine this information include gender, age, and annual income. A spending score is assigned based on a customer's spending behavior and amount. Three clustering methods were used: K-Means, Agglomerative Clustering, and DBSCAN. Two out of the three methods were able to accurately cluster the data into five clusters. This showed that age and annual income could be used to estimate a customer's spending behavior at the mall.

I. INTRODUCTION

This study focuses on the effects of gender, age, and annual income on the spending behavior customers at an anonymous mall. Through clustering and basic statistical measures, the goal was to show that age and annual income would have an affect on a person's spending habits at the mall. Clustering was the general method utilized to conduct this study, and three specific types of clustering are used to show their effects on the results. By using clustering, the categorization of different classes of customers is made simple. The importance of this study is to show the mall's owner how to target future customers. There are many factors that go into a customer buying more products at a mall, such as the stores that are in the mall, the appearance of the mall, the area that that the mall is located, and many more. However, this study shows the roots to the problem, and may help the mall's owner bring in more customers and get them to spend more at his/her stores.

The three clustering methods utilized for this study are K-Means, Agglomerative Clustering, and DBSCAN.

The K-Means algorithm clusters data by separating samples into a number of groups of equal variance, minimizing within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. K-Means works well with large, but flat, datasets, and is one of the most widely used clustering methods. [1]

The Agglomerative Clustering method performs hierarchical clustering using a bottom up approach. In this action, each observation starts in its own cluster, and clusters are successively merged together. Using the ward linkage criteria, the algorithm minimizes the sum of squared differences within all clusters. [1]

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Clusters found by DBSCAN can be any shape, as opposed to K-Means which assumes that clusters are convex shaped. There are two parameters to the algorithm, min_samples and eps, which define formally what is said by saying dense. Higher min_samples or lower eps indicate a higher density needed to form a cluster. [1]

II. METHODOLOGY

To conduct this study, a dataset from Kaggle.com [2] was used for the information about 200 customers at an anonymous mall. The information retrieved from each customer was their gender, age, and their annual income. The mall owner then assigned a spending score to rate each customer's shopping behavior and spending habits (The owner's exact method to score spending habits is unknown). Even though the dataset is labeled, clustering, an unsupervised learning model, is utilized to conduct the analysis of this study. The methods to create each model completed is shown in the accompanying pdf file, which is a Jupyter Notebook file. This file shows the work completed in primarily the Pandas, Seaborn, Matplotlib, and scikit-learn packages in Python.

A. Import Data

To begin, Numpy, Pandas, Seaborn, and Matplotlib are imported in the Jupyter Notebook. Then the mall data is converted from a .csv file to a data frame in Pandas.

B. K-Means

To conduct K-Means on the data, KMeans is imported from sklearn.cluster. To begin, the number of clusters was set to 3, and the model was fit to the data. Clusters were assigned to each customer in the dataset, and then the clusters were graphed. After viewing the results the number of cluster's was changed to 4, and then finally to 5. Five clusters showed the most tightly fitted clusters when looking at annual income versus spending score, so it was settled at this point. Basic statistics were then recorded for each cluster.

C. Agglomerative Clustering

To conduct Agglomerative Clustering, AgglomerativeClustering is imported from sklearn.cluster. As with K-Means, 3 clusters is used in the beginning, but ended with 5 clusters due to best fit. The clusters were plotted, and basic statistics recorded.

D. DBSCAN

To utilize DBSCAN, DBSCAN is imported from sklearn.cluster. There is no need to set the number of clusters, because DBSCAN tries to determine that on its own. To begin, the model is fit to the data using the default settings of the scikit.learn algorithm: an EPS of 0.5 and min_samples set at 5. When plotted, there was mostly noise, or -1. This led to the change of EPS to 10. This EPS and the default min_samples proved to be as good as the model could get, with the least amount of noise possible. Yet, the model still had too much noise visible, so for this reason no basic statistics were recorded.

III. RESULTS & DISCUSSION

The overall clustering process proved to be a success. While gender had a minor impact on the results, age and annual income proved to be a major factor with regards to each customer's spending score. K-Means and Agglomerative Clustering established themselves be the most useful algorithms for this study. DBSCAN wasn't far off, but it had too much noise to be used properly.

A. K-Means

The results of K-Mean using five clusters had an outstanding effect on the visualization of the different types of mall customers. Figures 1 and 2 below display the clusters displayed using age and annual income. In this case, cluster 0 had a mean age of 43, mean annual income of $55,291, and mean spending score of 49.56. Cluster 1 had a mean age of 45, mean annual income of $26,304, and mean spending score of 20.91. Cluster 2 had a mean age of 25, mean annual income of $26,304, and mean spending score of 78.56. Cluster 3 had a mean age of 40, mean annual income of $87,750, and mean spending score of 17.58. Cluster 4 had a mean age of 32, mean annual income of $86,538, and mean spending score of 82.12.

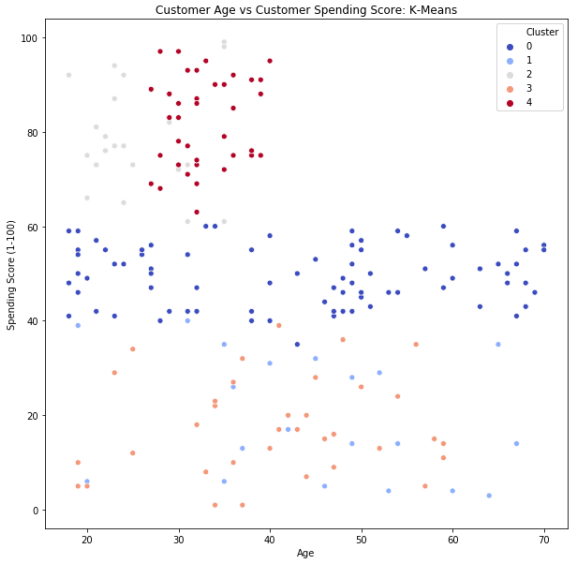

Figure 1: Customer Age vs Customer Spending Score using K-Means

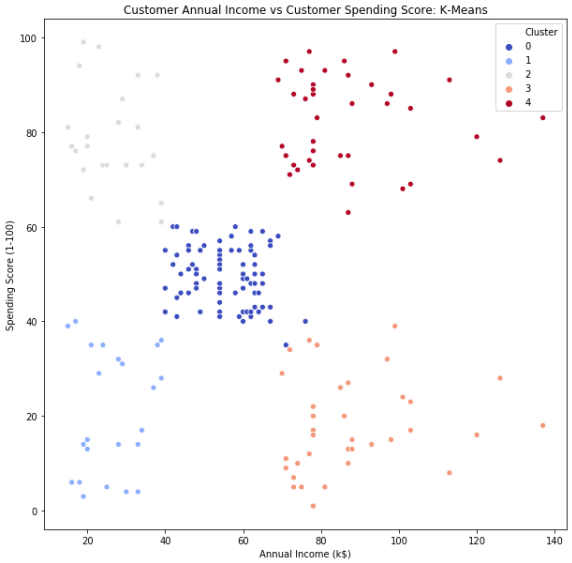

Figure 2: Customer Annual Income vs Customer Spending Score using K-Means

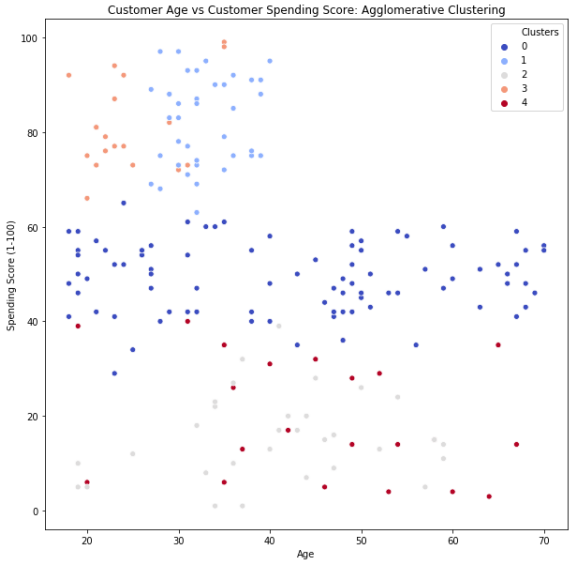

Figure 3: Customer Age vs Customer Spending Score using Agglomerative Clustering

B. Agglomerative Clustering

The results of Agglomerative Clustering using five clusters worked almost identically as K-Means. There are a few points that are different towards the middle of the graph when looking at annual income versus spending score. Figures 3 and 4 below display the clusters displayed using age and annual income. In this portion the legend for the clusters are assigned to different clusters, so be aware. Cluster 0 had a mean age of 42, mean annual income of $55,448, and mean spending score of 49.26. Cluster 1 had a mean age of 32, mean annual income of $86,538, and mean spending score of 82.12. Cluster 2 had a mean age of 41, mean annual income of $89,406, and mean spending score of 15.59. Cluster 3 had a mean age of 24, mean annual income of $24,950, and mean spending score of 81.00. Cluster 4 had a mean age of 45, mean annual income of $26,304, and mean spending score of 20.91.

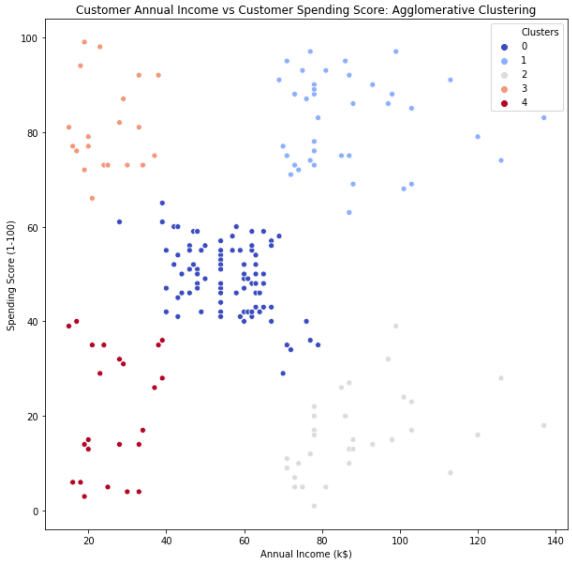

Figure 4: Customer Annual Income vs Customer Spending Score using Agglomerative Clustering

C. DBSCAN

DBSCAN worked well in trying to identify the clusters, but the amount of noise it encountered ended up being too much for the information to be of any use. Figure 5 shows the effects of the noise on the graph for annual income versus spending score. You can see that the middle and top two clusters are almost completed, but there is still too much noise. The results got worse when raising or lowering the EPS or min_samples.

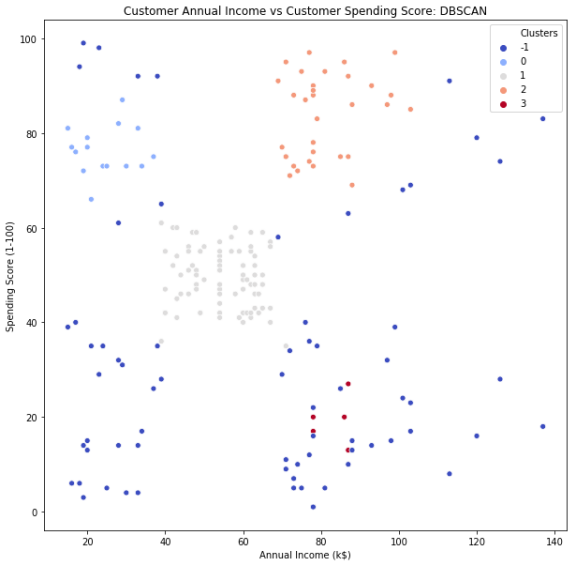

Figure 5: Customer Annual Income vs Customer Spending Score using DBSCAN

D. Clustering Evaluation

Overall, K-Means and Agglomerative Clustering were the most accurate in defining the classes of each customer. The reason is for the appearance of the clusters. In both cases you can see a clear separation and nice grouping between all five clusters. DBSCAN was not able to get the same results due to the noise it received. This may have been because DBSCAN works better on a larger dataset, not for the 200 lines of data in this study. One could say K-Means was more accurate than Agglomerative Clustering, but the process was done to easily analyze our data in a more generalized way.

E. Resulting Clusters

In the end, K-Means and Agglomerative Clustering painted a picture that is very simple to see. At this anonymous mall, there are five clusters of customers. Very generally they are depicted as such. There are two clusters with high spending scores (above 60). One of these clusters includes a low-income group making less than $40,000 a year, and the other group is a high-income group that makes more than $60,000 a year. In total, these two groups are aged 40 or less, with the high-income group being almost 7 years older on average. The higher income group seems to a more mature group; having entered the point in their life where they are finally working at serious jobs in a well-paid position, and their income shows it. The lower income group looks to be mostly college age citizens, so it is unlikely that they will have high paying jobs at this point. This low-income group is most likely people that are still stuck in a materialistic setting. They don't have the money to buy things, but they still do so people think they are cool. This group probably also has major debt problems, especially if they are college students.

Next there are two clusters that almost exactly match the annual income of the high-spenders above, but these two groups are low spenders (below 40 spending scores). The two groups have ages averaging in the 40's. The two groups may most likely be low spenders due to two things: older people tend go out less (which includes malls) and have most likely grown out of their materialistic days. The higher income, low-spenders are also most likely big investors.

The final cluster is in the middle of it all. They are a group that ranges from ages 18 to 70, and they have an average spending score of about 50. They seem to be neither poor or rich, remaining in the $40,000 to $60,000 range for annual income. This is the group that knows they have enough money to purchase things at the mall often, but they also know they aren't rich enough to be heavy spenders.

IV. CONCLUSION

In closing, the purpose of this study was to determine the spending habits of an anonymous mall's customers using clustering methods. The models utilized customer gender, age, and annual income, and used it to help determine how customer's spending scores are rated. Three clustering methods were used: K-Means, Agglomerative Clustering, and DBSCAN. K-Means and Agglomerative Clustering were most accurate in clustering the data. From this, it was easy to see that there were 5 primary classes of customers. This showed that age and annual income could be used to estimate a customer's spending behavior at the mall. To further this study, it would be best to look at the additional factors behind each group's spending habits, such as where they live, what they do for a living, and the size and frequency of purchases.

V. References

[1] "2.3. Clustering," *scikit* [Online. Available: https://scikit-learn.org/stable/modules/clustering.html#dbscan. [Accessed: 01-Jun-2019].

[2] V. Choudhary, "Mall Customer Segmentation Data," *Kaggle*, 11-Aug-2018. [Online]. Available:

https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python. [Accessed: 01-Jun-2019].