CUSTOMER SEGMENTATION

DATA SCIENCE

PREPARED FOR

Exposys Data Labs

PREPARED BY

Siddhesh Suresh Kadam

(Intern)

DEC 23, 2021

Table of Contents

1. Abstract

Customer segmentation is the practice of grouping your customers based on similar factors such as demographics or behaviours to better target them with advertising. These customer segmentation categories might be utilised to kick off a conversation regarding marketing personas. Because customer segmentation is often utilised to develop a brand's messaging, positioning, and sales process, marketing strategies must be properly linked with those client segments to be effective.

2. Introduction

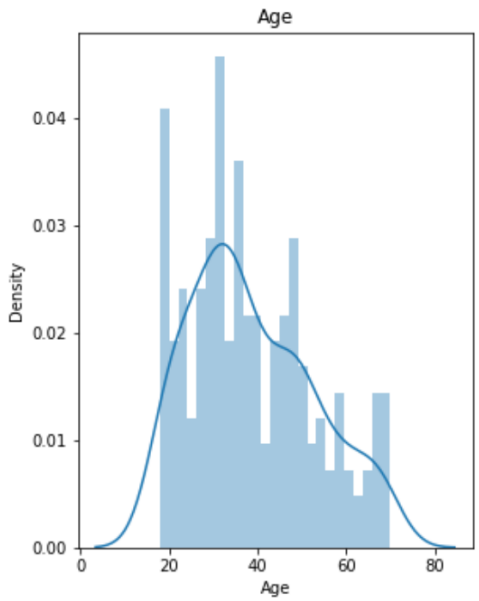
In this project, our primary goal is to bring out strategic information from the available customer data provided to us. Identifying the consumer segments using clustering and targeting the possible user population. Visualizing the gender and age distributions using K-means clustering. Customer segmentation is popular since it helps with marketing and sales. This is because you may have a better understanding of your customers' interests and needs. The financial impact is far bigger, and effective customer segmentation may help you increase client lifetime value.

3. Scope

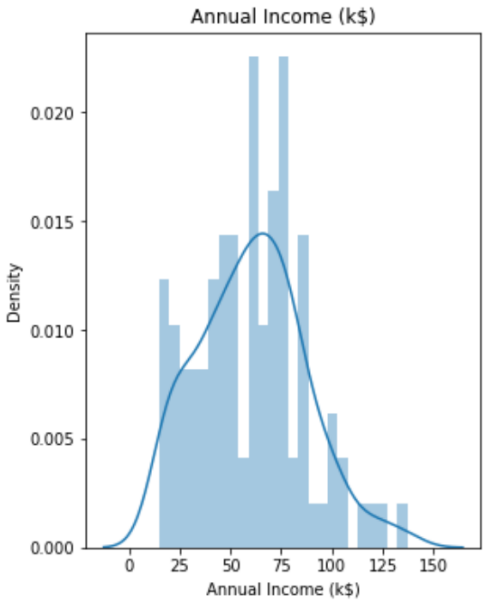
All the inferences and conclusions that would be drawn from the customer data needs to be properly visualized and presented. The algorithm is trained and tested only on the customer data provided by the Exposys Data Labs.

4. Pre-Clustering Inferences

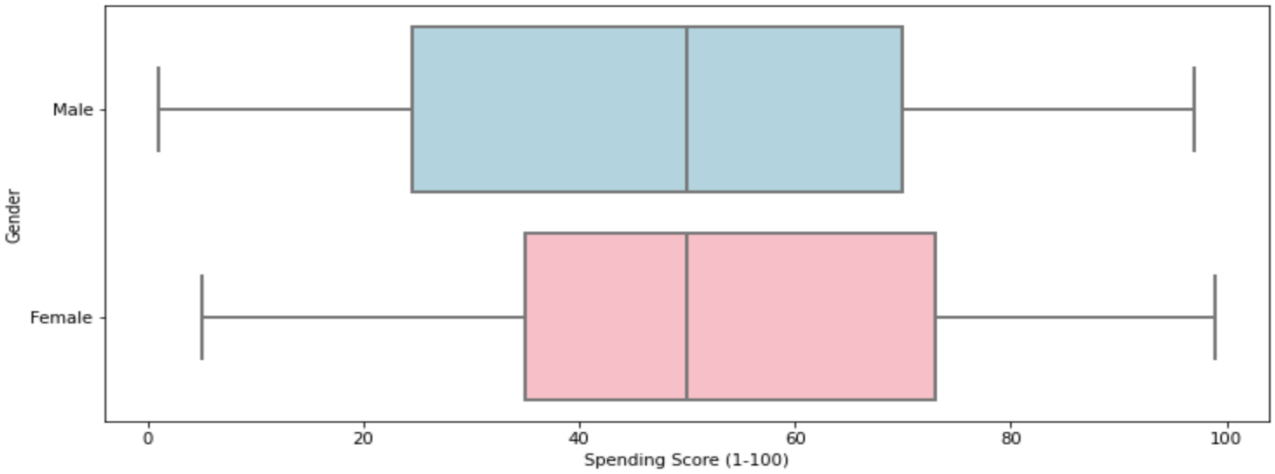
We use a histogram to visualize the density of each instance of the two attributes; Age, Annual Income (K$). From the observations from Age histogram, the customer data highly consists of people in their early 20's to mid-'30s. Moving on to the Annual Income histogram, apart from the outliers most of the customers have their annual income less than 90k$ with the maximum number of customers lying in the range of 50k$-80k$.



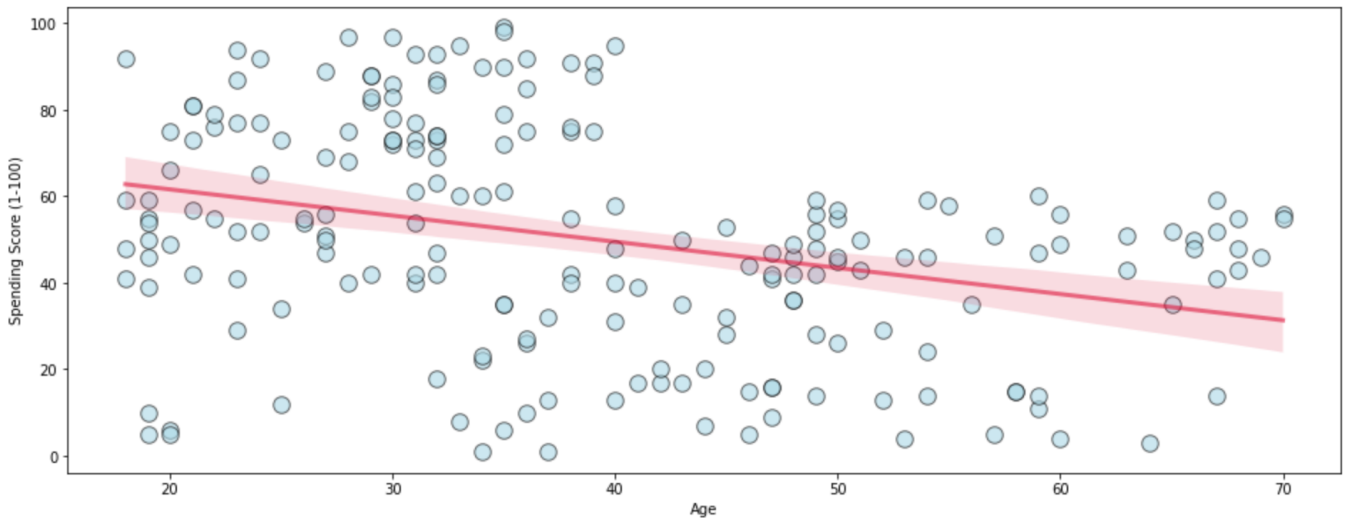
Age Density



Annual Income Density



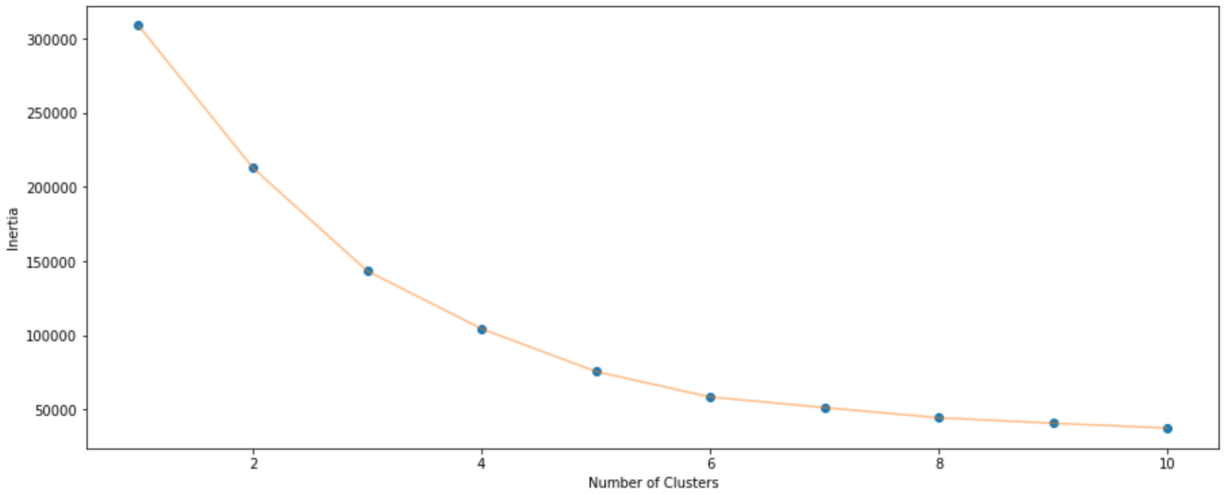
Some important inferences drawn from the initial discovery itself. Using Pearson Correlation, we observed that the attribute Age had quite an impact on the Spending Score. The two attributes were inversely related to each other. The study discovered is represented below.



Age vs Spending Score

5. KMeans Clustering Algorithm and Architecture

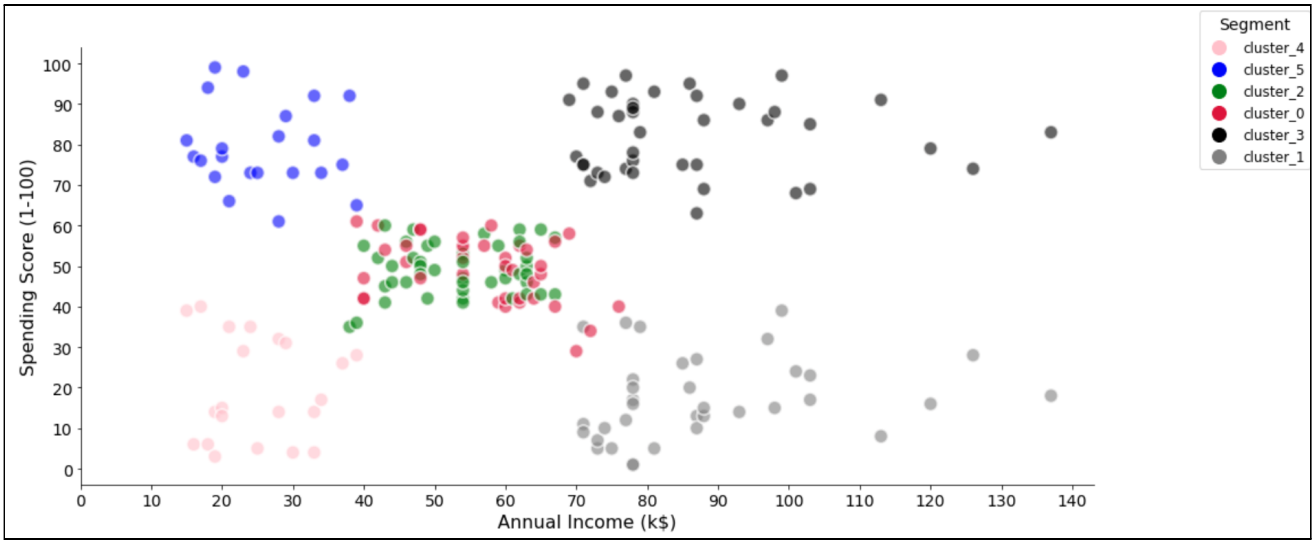
For this project we used KMeans Clustering, a type of unsupervised machine learning algorithm. It groups the unlabeled data into many clusters, where K specifies the number of predefined clusters that must be produced throughout the process; for example, if K=2, two clusters will be created. To calculate the optimal number of clusters for our model, we test our model on different numbers of clusters and evaluate each instance's inertia value. It should be noted that while selecting the elbow value, i.e. the optimal cluster value, precaution must be taken not to select a larger number of clusters as the model might overfit.



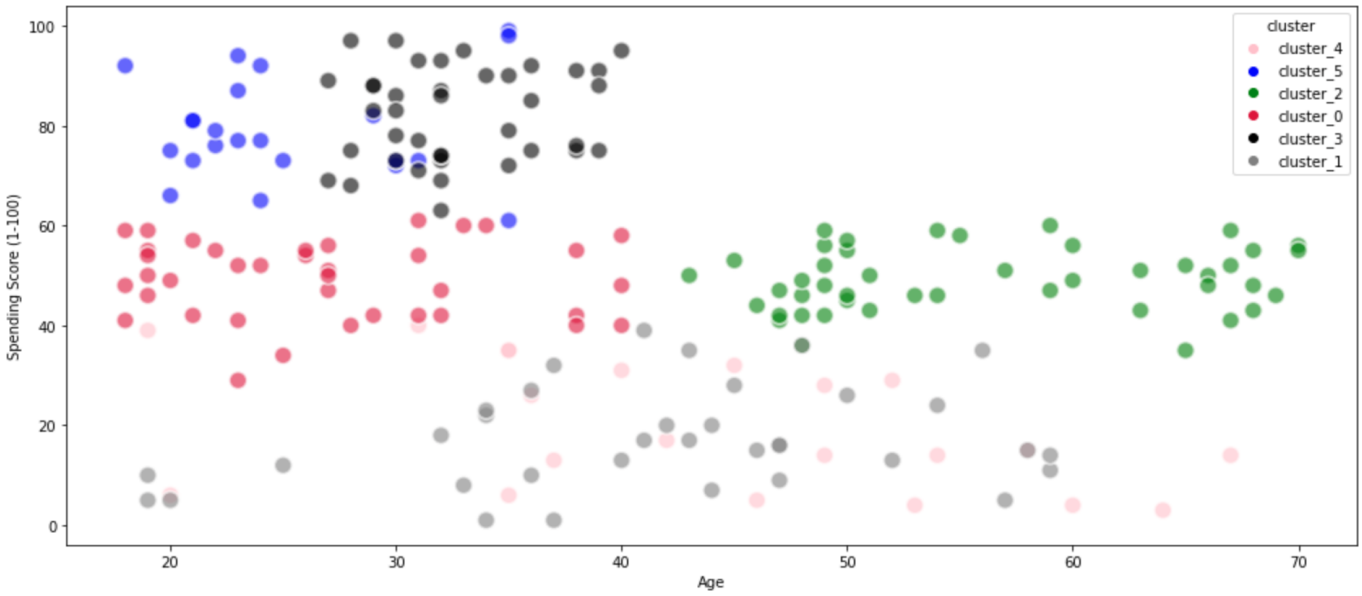
In our case the optimal value tends out to be 6. Any value higher than this will cause the model to overfit. After getting this value we train the data over the KMeans cluster algorithm with K=6.

6. Implementation and Results

The customer data is now divided into clusters and each cluster represents a group of customers sharing similar characteristics. Of the 200 customer data, all of them are divided into the 6 groups. Below is a proper visualization of the clusters.

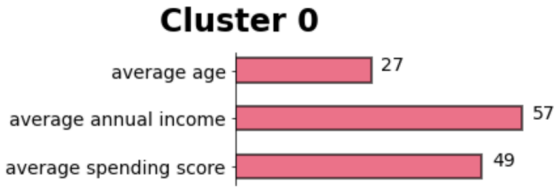


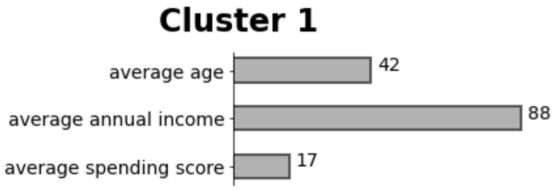
Annual Income vs Spending Score segmented by clusters

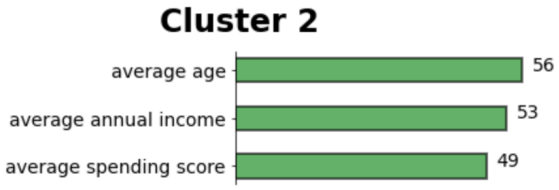


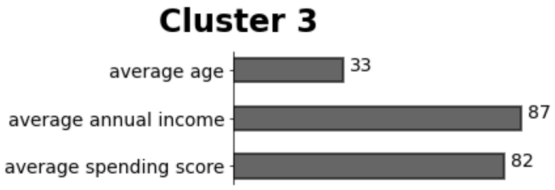
Age vs Spending Score segmented by clusters

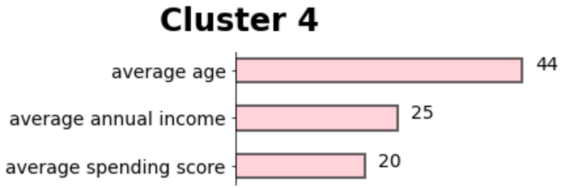
Below are the customer's average age, average annual income and average spending score of each cluster.

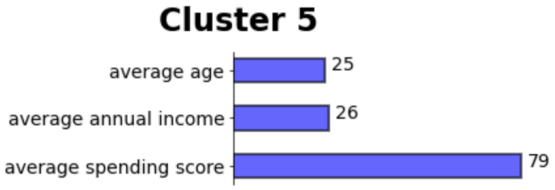












7. Software Requirements

Python 3.9.7

Python Library		
1	Numpy	1.21.2
2	Pandas	1.3.4
3	Matplotlib	3.5.0
4	Seaborn	0.11.2
5	Scikit learn	1.0.1

8. Conclusion

We have successfully performed customer segmentation on the given customer data. Major concerns with the project are with the data. The amount of data (200 instances) is very low and hence KMeans is not necessarily the optimal path for segmentation analysis. Data being incomplete (employment status, family size, etc)might be another reason why this analysis might not be up to the mark.