Customer Segmentation using Machine Learning.

A study of customer segmentation and targeting using data science methods.

By

Deependra Dhakal
Bellevue University
MS Data Science
DSC680 T302-2225 Spring 2022
Applied Data Science
Portfolio Project II
Milestone 3: Final Paper
Date: 05/22/2022

Abstract.

The objective of this project is to perform a segmentation analysis and classify target segment members with the available data. The objective is to understand the data-driven approach in business.

Data science methods are applicable on business problems. A company is more likely to gain more profit with same resources for the marketing and sales by using data science method. Competition between businesses has created tension in finding potential customers. The ability of any business to understand its customer helps in developing customized products/services. This is possible through structured customer service. Big data and machine learning have been able to foster more acceptable customer segmentation.

Executive Summary.

The objective of this project is to perform a segmentation analysis and classify target segment members with the available data. The objective is to understand the data-driven approach in business.

After EDA and clustering on clustering, 8 clusters were picked and analyzed. Each cluster represented average customer in the dataset. Based on the analysis, there is potential to exploit by the marketing team and cluster-specific targeting of customer needs. Following is the list of cluster that was picked and analyzed:

- Cluster 0: Middled Aged, Upper-middle Income, Low Score.
- Cluster 1: Old Aged, Middle Income, Middle Score.
- Cluster 2 - Middled Aged, Upper-middle Income, High Score.
- Cluster 3 - Young Aged, Low Income, High Score.
- Cluster 4 – All Age, Low Income, Low Score.
- Cluster 5 – Young Aged, Middle Income, Middle Score.
- Cluster 6 – Middle Aged, High Income, Low Score.
- Cluster 7 – Middle Aged, High Income, High Score.

Cluster 7 seems one of the most profitable clusters which consists of high spending score, high income, middles aged 60% females. Offering high end products might be a strategy to keep them spend.

Contents

List of Figures.

1.Introduction.

1.1Background and Problem Statement

Competition between businesses has created tension in finding potential customers. The ability of any business to understand its customer helps in developing customized products/services. This is possible through structured customer service. Big data and machine learning have been able to foster more acceptable customer segmentation.

Increase in competition and availability of historical data has helped in extensive use of data mining techniques to discover important and strategic information. Growth in competitions has prompted businesses to discover ways to maximize profit. And this has been achieved by been able to foster more acceptable customer segment and focusing on them through customized products/ services.

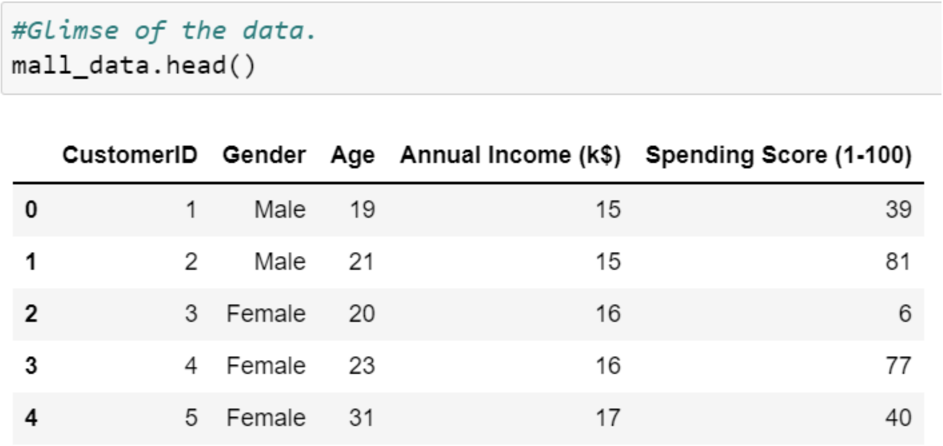
Customer segmentation is the division of customers in different segments based on their characteristics that can directly or indirectly influence the business. The importance of customer segmentation includes the ability of a business to customize their market plan and make risky decisions.

The purpose of this study is to be able to make one of the applications of machine learning and find the best segment of customer to focus on.

1.2 Data Source

The data for the project "Mall Customer Segmentation Data" will be obtained from Kaggle, https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python. This dataset was created to learn the customer segmentation concept. The scenario for the data is you own a supermarket mall and have amassed information through membership cards. Now you want to understand the customers to plan strategically. The dataset has following features:

1. CustomerID. Unique Id assigned to the customer.
2. Age. Age of the customer.
3. Gender. Gender of the customer.
4. Annual Income. Annual income of the customer.
5. Spending Score. Score assigned to customers based on behavior and purchasing data.



```
#Glimse of the data.
mall_data.head()
```

	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

Fig1: Glimpse of Data.

2.Methodology

I am trying to group/cluster customers with similar purchasing habits so that they can be targeted accordingly. This is an unsupervised machine learning problem. I would be using clustering (K-means and Hierarchical) for the purpose of this study.

2.1 K-means Clustering.

K-mean clustering is one of the unsupervised machine-learning algorithms which is used when you have data without defined categories or groups (i.e., unlabeled data). The objective of this algorithm is to find groups in the data and discover the underlying patterns, with the number of groups represented by variable K.

The k-means algorithm uses iterative process to produce result i.e., it iterates between two steps until a criterion to stop is met:

1. Data assignment step: Based on the squared Euclidean distance, each data points gets assigned to its nearest centroid.
2. Centroid update step: Centroid is recalculated by taking the mean of all data points assigned to that centroid's cluster.

The two metrics that are used for evaluating the performance of models based on different K clusters are:

1. Elbow method: Elbow method gives an idea on what a good K number of clusters would be based on the sum of squared distance between data points and their assigned centroids.
2. Silhouette Analysis: Silhouette analysis can be used to determine the degree of separation between clusters. For each sample:
   a. Calculate the average distance from all data points in the same cluster.
   b. Calculate the average distance from all data points in the closest cluster.
   c. Compute the coefficient. The coefficient can take values in the interval (-1, 1).

i. If it is 0, the sample is very close to neighboring clusters.
ii. If it is 1, the sample is far away.
iii. If it is -1, the sample is assigned to wrong clusters.

3. Exploratory Data Analysis.

Some of the initial insights from the data are as follows:

- Age, Income, and Score features are multimodal distributions.
- Income and Score plot seems to have formed 5 dense regions, 4 on the sides and 1 in the center.
- Female customers are larger in number than male by 12%.
- 20-30, 30-40 are the most common age groups.
- The biggest customer age group is 30 years old comprising mostly of women.
- Most of the customers earn between $50k and $80k annually.
- Women spend more than men.
- Five visually recognizable clusters seem to represent customer.

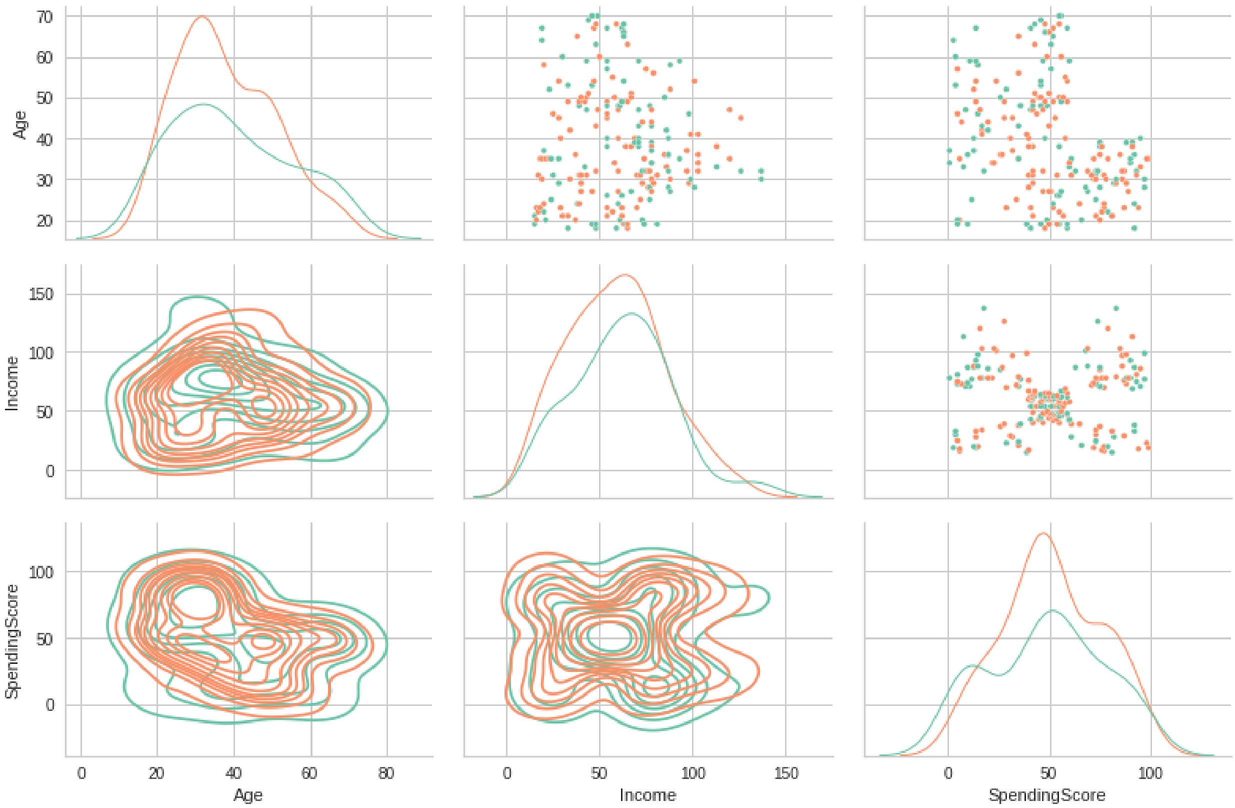


Fig2: Distribution of the data.

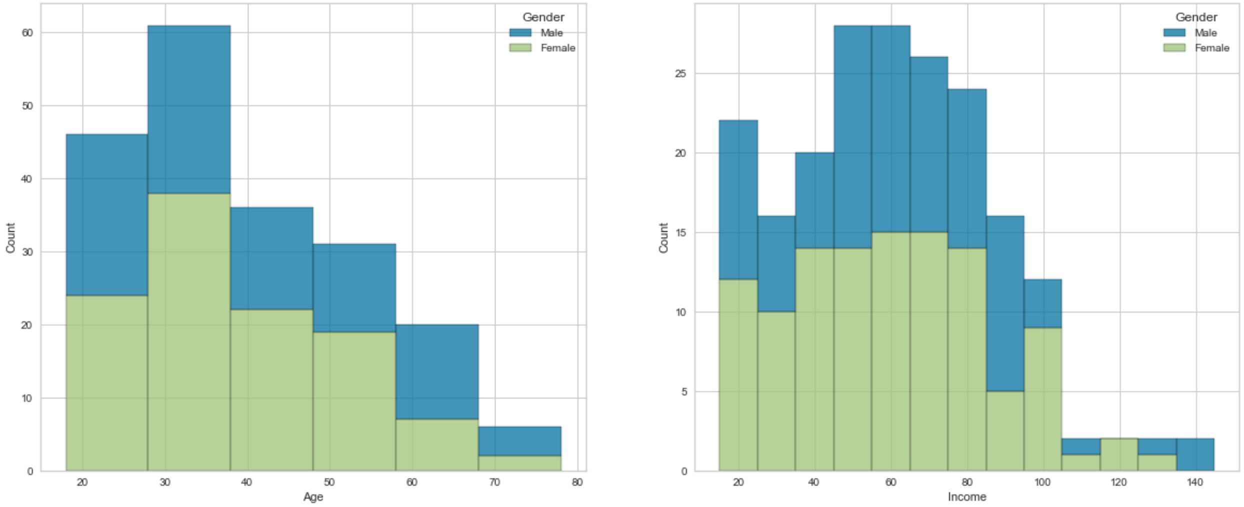


Fig3: Customer Profile (gender).

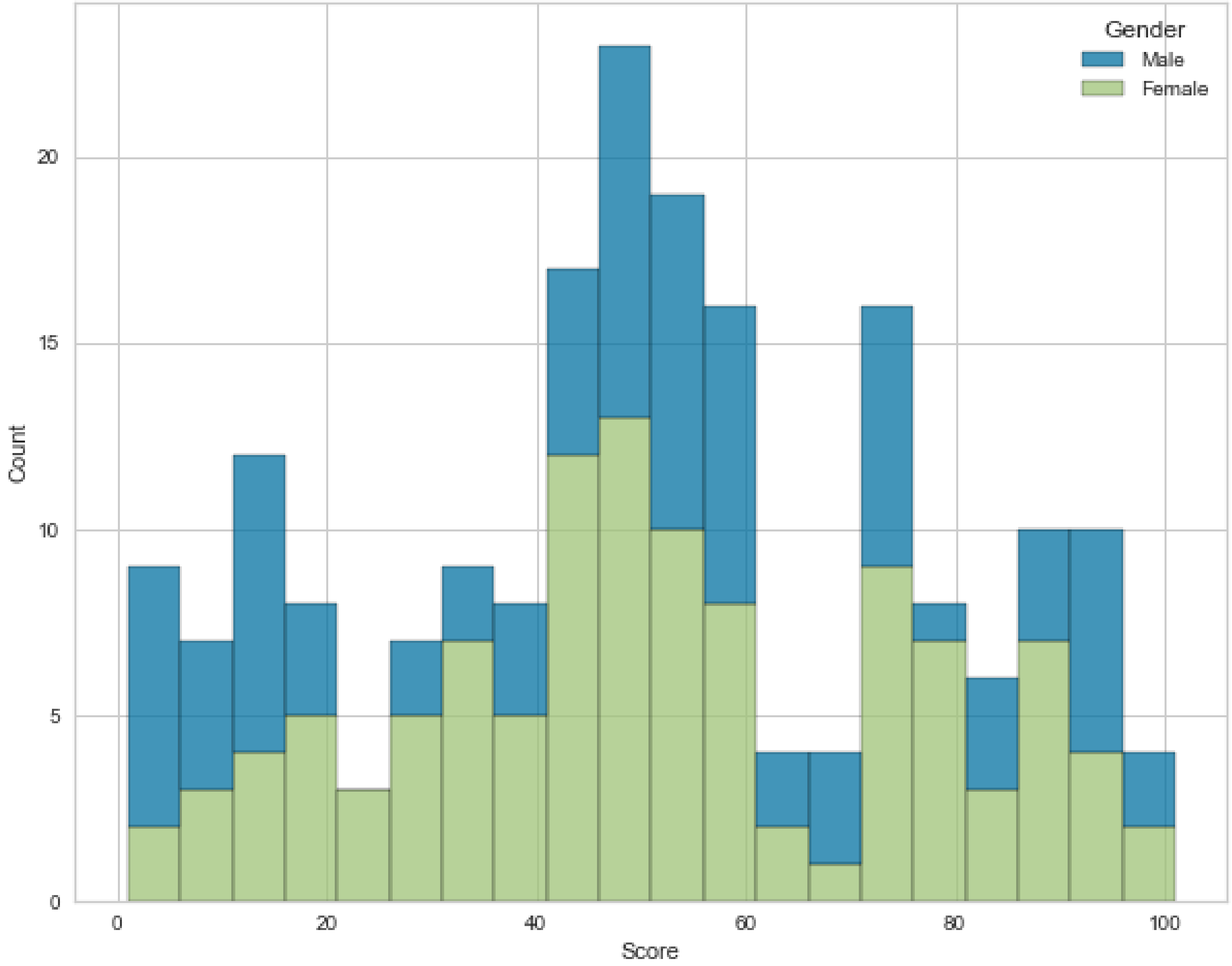


Fig4: Gender and Score.

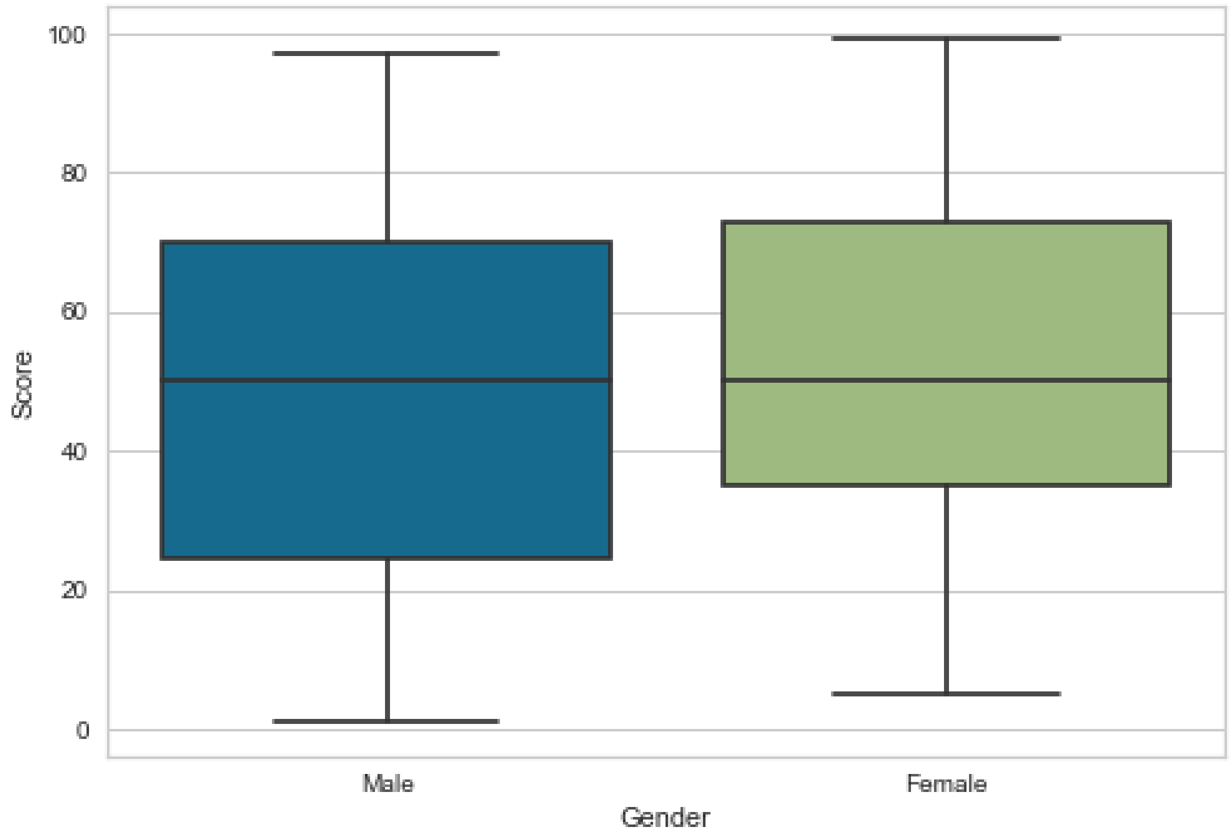


Fig5: Spending Score based on gender.

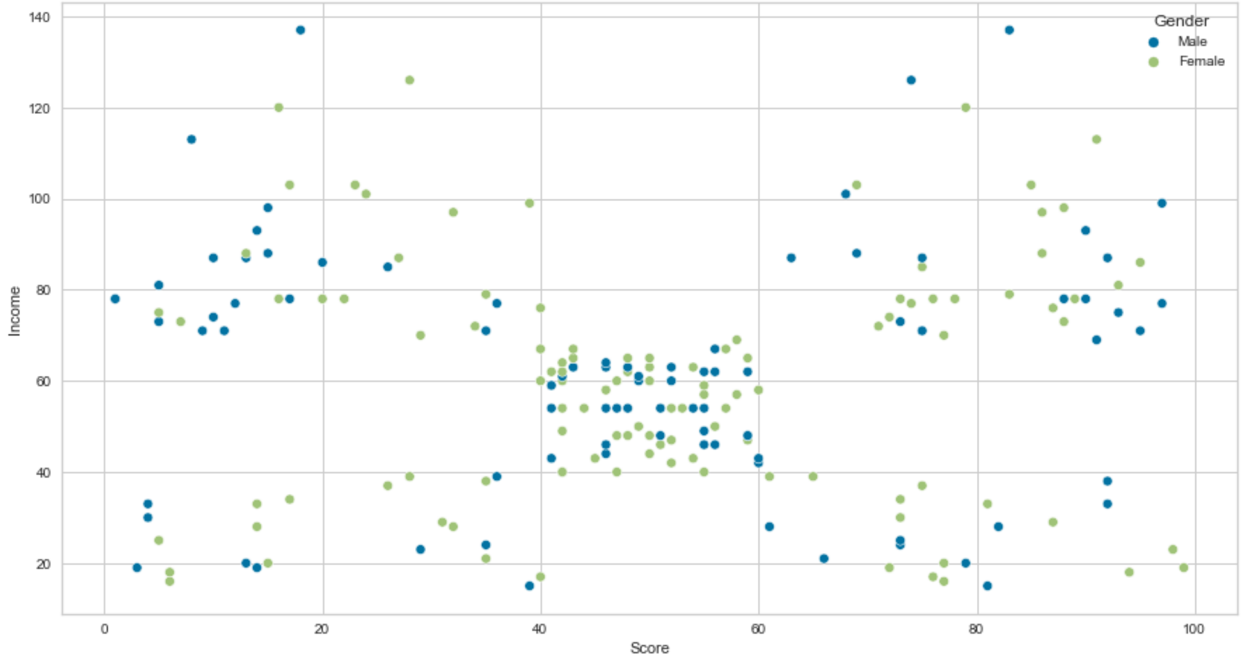


Fig6: Scatterplot.

4. Analysis.

K-means clustering works best when there is similar distance from centroids. To use k-means we need to know the number of clusters in the dataset which is 5. But let's not content with only this and use the elbow method and silhouette diagram for additional clues.

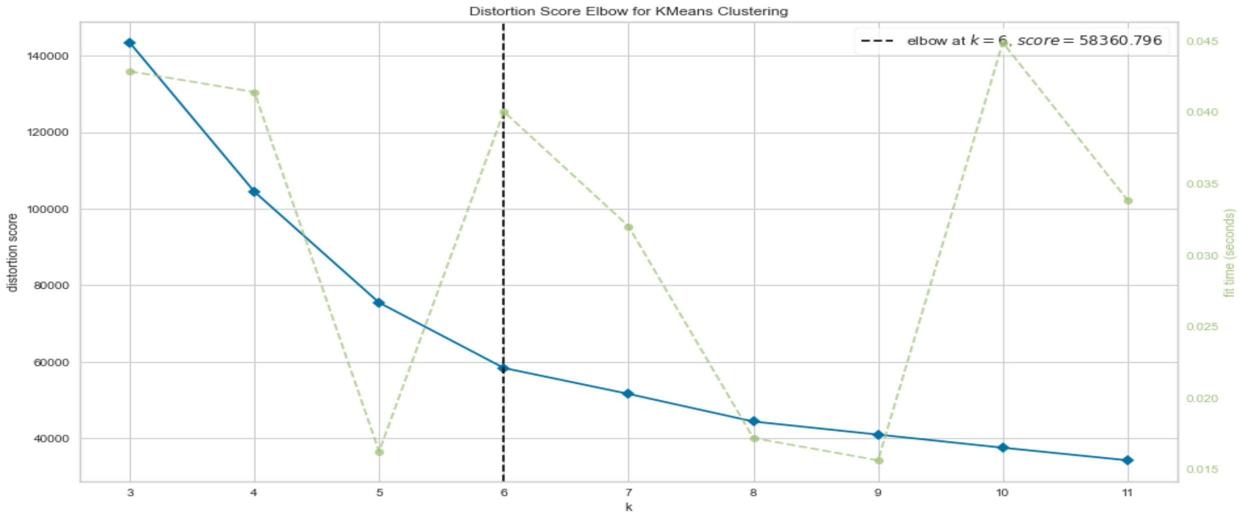
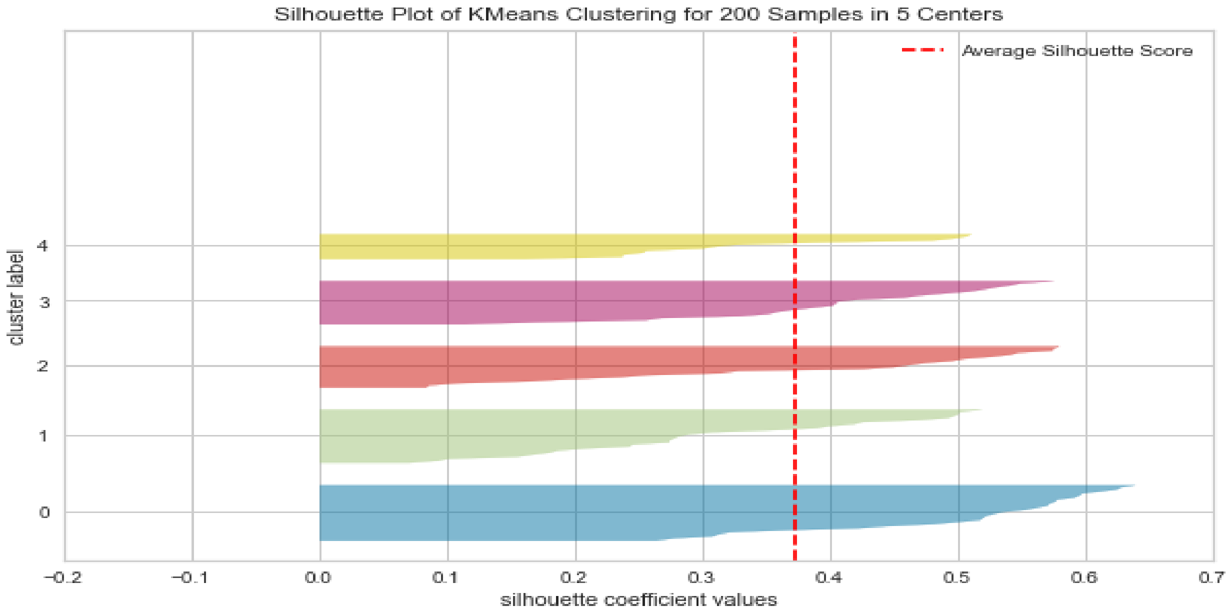


Fig7: Clustering by elbow method.

Elbow method suggest 6 clusters rather than 5 which might be correct as we only have at the data two dimensionally.



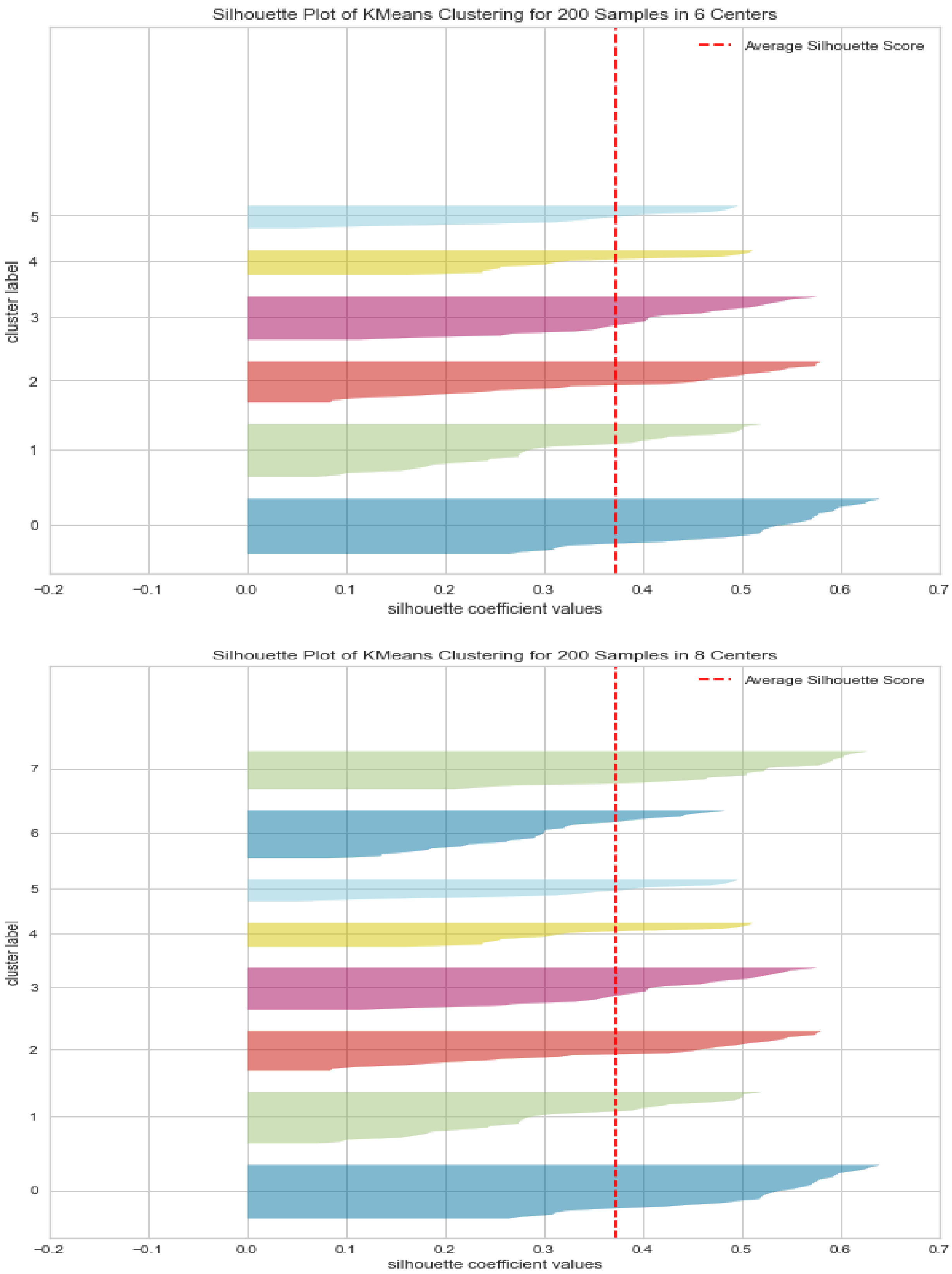


Fig8: Silhouette Plot.

We can see that plot for 5,6,8 clusters are similar in terms of inter-cluster connectivity and size. But 2d space is not enough to show all 8 clusters thus, exploring in the higher dimension will give us better idea.

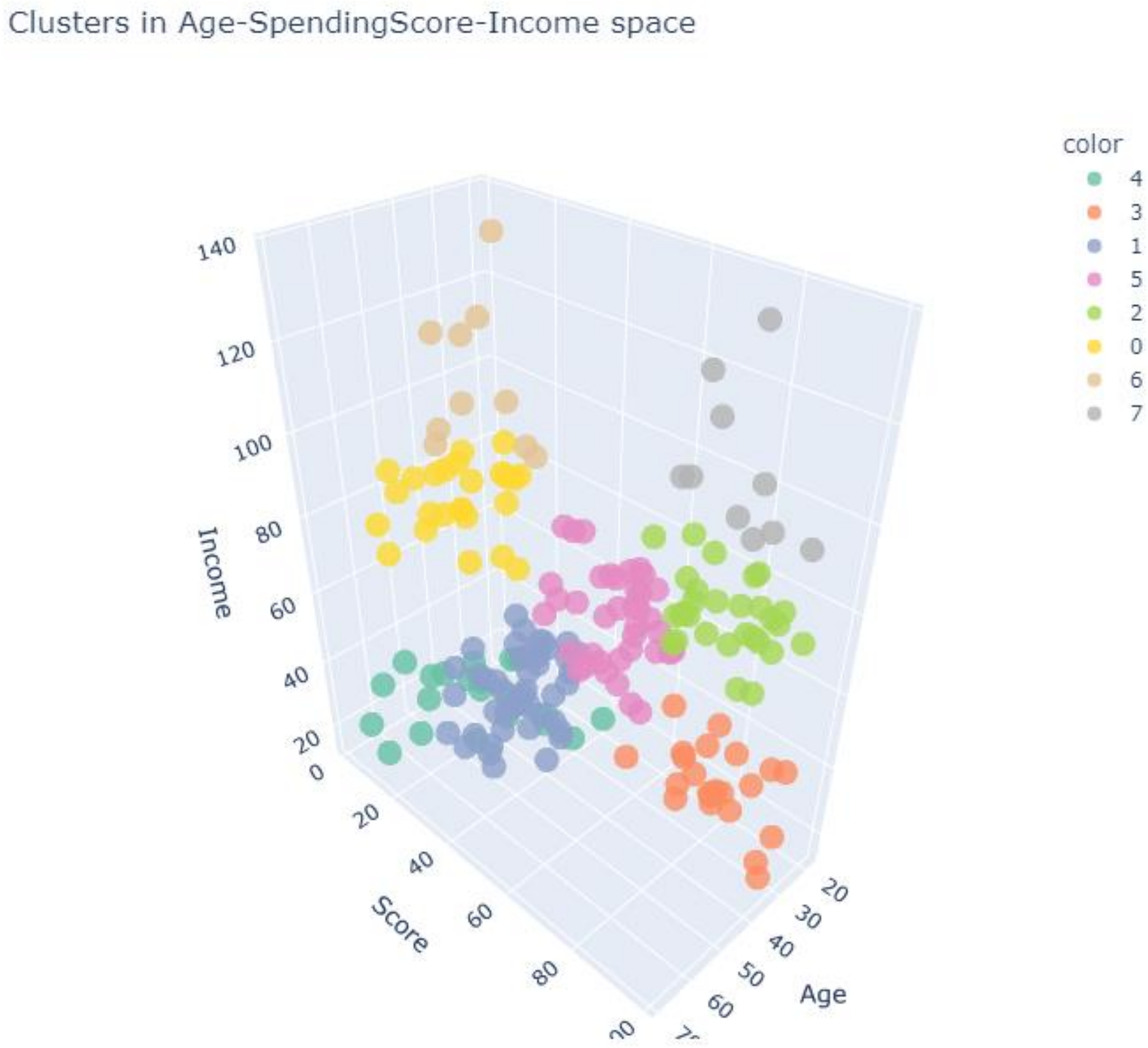
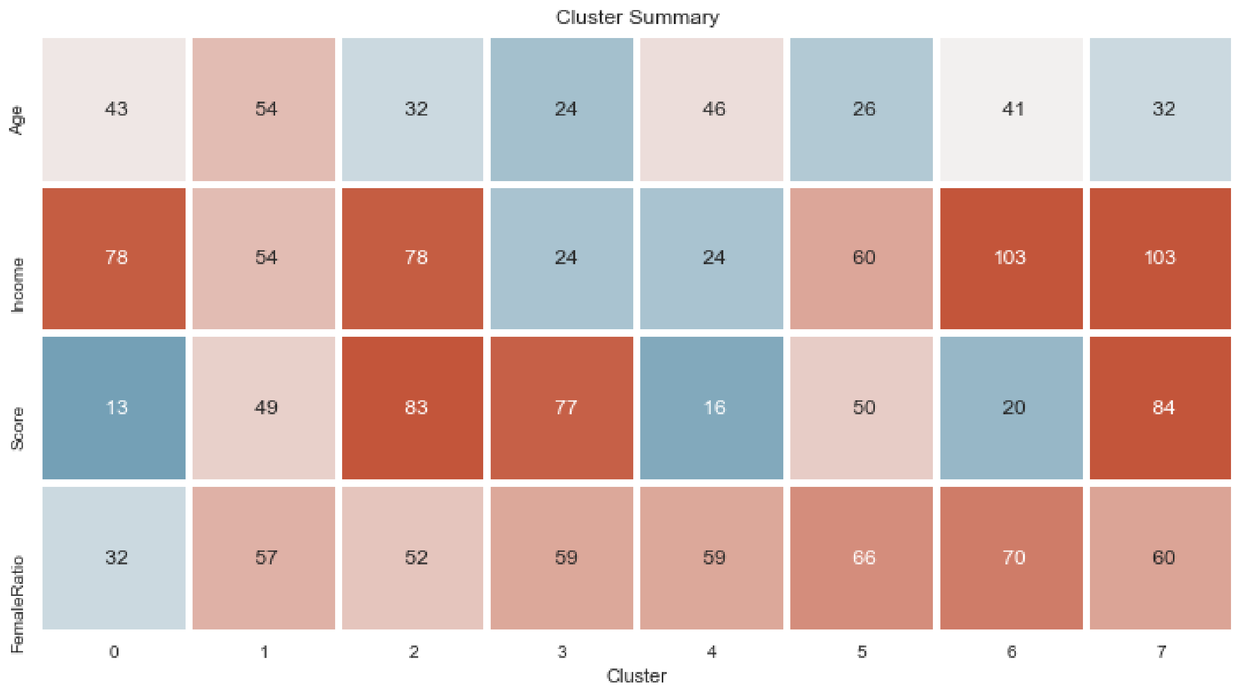


Fig9: Clusters in higher dimension.

5.Conclusion.

- ***Cluster 0: Middled Aged, Upper-middle Income, Low Score.*** The cluster has great potential with upper middle income. The cluster consists of most men, strategy would be to go for product lines that males buy.

- ***Cluster 1: Old Aged, Middle Income, Middle Score.*** The cluster consists of middle-income group and the female dominates the cluster group by 7%. Analyzing the regular products and focusing on them would be a good strategy.
- ***Cluster 2 - Middled Aged, Upper-middle Income, High Score.*** One of the most profitable cluster with upper middle income with high spending score. High end product might be a good option.
- ***Cluster 3 - Young Aged, Low Income, High Score.*** This cluster consists of younger generation with low income but high spending score.
- ***Cluster 4 – All Age, Low Income, Low Score.*** Cluster is made of all age group with low income and low spending score. People from this cluster like to save, so any programs to help them save might be helpful.
- ***Cluster 5 – Young Aged, Middle Income, Middle Score.*** This cluster has younger generation with middle income and mid spending score. This cluster present a good opportunity.
- ***Cluster 6 – Middle Aged, High Income, Low Score.*** Middled aged people with high income but low spending score with mostly female.
- ***Cluster 7 – Middle Aged, High Income, High Score.*** One of the most profitable segments with middled age people, high income, and high spending score. Offering high end products focusing female would be great.



	0	1	2	3	4	5	6	7
Age	43	54	32	24	46	26	41	32
Income	78	54	78	24	24	60	103	103
Score	13	49	83	77	16	50	20	84
FemaleRatio	32	57	52	59	59	66	70	60

Fig10: Heatmap.

Heatmap contains Female Ratio Column. We have slightly higher number of females (56%) in the dataset, however same distribution is not preserved in all clusters.

6.Ethical Considerations

Ethical problem arises from conflicts and disagreement. Especially in marketing where each party have the expectations on way of doing business and existing relationship, conflicts and disagreement are bound to occur on each facet. Some of the issues to handle are:

- Invasion of privacy and stereotyping.
- Unethical market exclusion.
- Deceptive advertising.

7.Challenges and Issues.

Data quality is one of the major issues with customer segmentation. Inaccurate data means inaccurate segmentation. For example, the attributes of the customer like age, gender, income, marital status plays a major role while segmentation but if the data is not maintained properly, the information will be less useful.

8. Future Uses/ Additional Application.

Customer segmentation can be used to analyze the customer as per the business, identify business opportunities, and strategize accordingly.

9. Recommendations / Implementations.

Customer segmentation can be used to collect feedback for the product. Knowing the customer base will be of great value to offer customized product/services and communicating for satisfied customer. It will help in maximizing the profit by focusing on higher value sales opportunity.

10. Acknowledgment

I would like to express my sincere gratitude towards the Bellevue University, my professors and fellow students for valuable guidance and helpful comments.

References:

- Allen, K. (2021, November 20). The Importance of Ethics &amp; Marketing Segmentation. Bizfluent. Retrieved April 24, 2022, from https://bizfluent.com/info-8516645-importance-ethics-marketing-segmentation.html
- Boundless. (n.d.). Boundless marketing. Lumen. Retrieved April 24, 2022, from https://courses.lumenlearning.com/boundless-marketing/chapter/ethics-in-marketing/
- Customer segmentation using K means clustering. KDnuggets. (n.d.). Retrieved April 24, 2022, from https://www.kdnuggets.com/2019/11/customer-segmentation-using-k-means-clustering.html
- Kumar, A. D., Dhiraj Kumar A Data Scientist &amp; Machine Learning Evangelist. Follow me on, Kumar, D., A Data Scientist &amp; Machine Learning Evangelist., &amp; on, F. me. (2021, December 13). Implementing customer segmentation using machine learning [beginners guide]. neptune.ai. Retrieved April 24, 2022, from https://neptune.ai/blog/customer-segmentation-using-machine-learning
- Lemos, A. R. (2021, September 22). Customer segmentation project. Medium. Retrieved April 24, 2022, from https://medium.com/geekculture/customer-segmentation-project-116c47d7a4df
- (Yueh-Han), J. C. (2021, September 12). Mall Customer Segmentation and forming Growth Strategies. Medium. Retrieved May 22, 2022, from https://medium.com/geekculture/mall-customer-segmentation-and-forming-growth-strategies-cc4130a0f4d7
- Baker, K. (2021, July 28). The Ultimate Guide to Customer Segmentation: How to organize your customers to grow better. HubSpot Blog. Retrieved May 22, 2022, from https://blog.hubspot.com/service/customer-segmentation
- Course hero. Boundless Marketing | Course Hero. (n.d.). Retrieved May 22, 2022, from https://www.coursehero.com/study-guides/boundless-marketing/ethics-in-marketing/
- Rawat, S. (2019, December 9). Mall customers segmentation-using machine learning. Medium. Retrieved May 22, 2022, from https://towardsdatascience.com/mall-customers-segmentation-using-machine-learning-274ddf5575d5
- Why is customer segmentation important? Brand Interaction Platform: Conversational AI. (n.d.). Retrieved May 22, 2022, from https://www.ada.cx/customer-segmentation#:~:text=Customer%20segmentation%20is%20not%20only,improve%20customer%20experience%20and%20satisfaction.

Questions and Answers.

What is mall customer segmentation?

It is a dataset created to work on customer segmentation using machine learning.

How will you check if the segmenting is right?

Using the strategy to focus on the cluster accordingly and keeping record of each transaction.

How do you determine customer segments?

Based on different factors such as age, gender, income, spending habit.

What are the growth opportunities as per your segmentation?

Segmentation helps in focusing on cluster accordingly and being strategically ready for each and every cluster.

How customer segmentation can be effectively used in marketing?

Customer segmentation helps in focus on customer segment you are trying to reach. Target customer as per your product.

What are the ethical implications of the project?

Some of the issues to handle are:

- Invasion of privacy and stereotyping.
- Unethical market exclusion.
- Deceptive advertising.

How do you collect data?

Through survey.

How can you identify different market segments?

Based on the demographics, we can study the market and make suggestions.

How to do customer segmentation?

Define your target customer. Do some research and determine which section of the market to target.

How market segmentation helps?

It helps in creating stronger marketing messages, create most effect marketing strategies and advertisement.