

DHA Suffa University

Department of Computer Science

Course Project - Data Science



Customer Segmentation

Project Report

By:

Danish Hudani (CS171078)
Muhammad Bilal (CS181002)

Course Teacher:
Muhammad Adeel Mannan

1. Summary:

The project contains datasets of the following attributes; age, spendings of about 200 customers. It will then analyze that depending on the age from 1-100 how much an average person spends in his/her age bracket. From this, different marketing strategies could be devised to target the age that spends the most, which would profit the organization.

2. Background

a) Problem Statement

We want to understand the similarities between different customers based upon their age, spending score and annual income to find any pattern that could be used by the marketing team to plan their strategy accordingly to easily converge target customers.

b) Factors / Parameters

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Gender                                200 non-null    object
1   Age                                   200 non-null    int64
2   Spending Score (1-100)               200 non-null    int64
3   Annual Income (k)                    200 non-null    int64
dtypes: int64(3), object(1)
memory usage: 6.4+ KB
```

From the above information we can say that we have following features in our data with respective data types:

- **Gender (str):** This attribute contains the gender of the person.
- **Age (int):** This attribute contains the age of a person.
- **Annual Income (In Thousands PKR) (int):** This attribute contains the income of an individual in PKR and in thousands.
- **Spending Score (1-100) (int):** This attribute contains the self judged estimated score that person spends on shopping indicating his/her shopping habit.

c) Research Question

Based upon our problem statement, we can derive two research questions that this study will acknowledge:

RQ1: Is there any association between the attributes of customers to analyze their spending habits?

RQ2: Can we distribute the customers in groups or clusters based upon these attributes to target them for any potential marketing strategy?

d) Objective

The major objective of this project would be to define marketing strategies that analyze the category of people that spend the most and hence target them to shop in order to boost the sales and revenue of the company.

e) Scope

The scope of this project extends to make sectors, but the context of this study is related to domain of commerce to market the particular brand to particular customers. Although the concept used in this study can be extended to different sectors like health, energy etc.

3. Methodology

a) Quantitative / Qualitative

This study contains three quantitative attributes and one qualitative attribute. The statistics associated with those attributes are given in the following tables:

Quantitative Attributes: Age, Spending Score, and Annual Income

	Age	Spending Score (1-100)	Annual Income (k)
count	200.000000	200.000000	200.000000
mean	38.850000	50.200000	846.720000
std	13.969007	25.823522	315.176654
min	18.000000	1.000000	300.000000
25%	28.750000	34.750000	618.000000
50%	36.000000	50.000000	858.000000
75%	49.000000	73.000000	1056.000000
max	70.000000	99.000000	1764.000000

Qualitative Attributes: Gender

Gender	
count	200
unique	2
top	Female
freq	112

b) Descriptive / Predictive

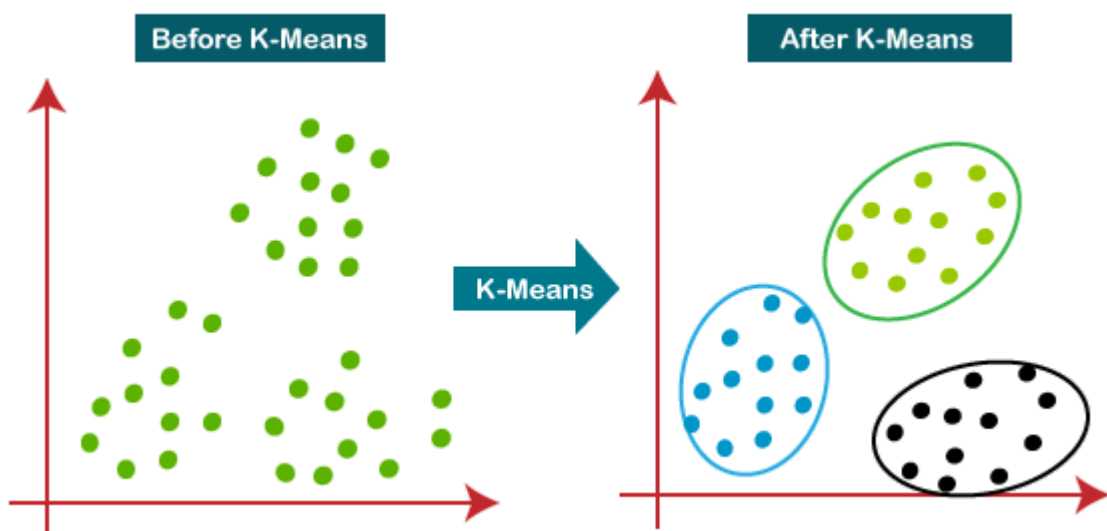
The segmentation of customers in this study comes under the descriptive approach as we are using past data and inferring it to prove our hypothesis that there exists association between Age, Spending Score, and Annual Income.

But this study can be extended to use the model in this scenario, K-Means, to predict the new customer records in the clusters based upon their attributes.

c) Types of Experiment / Approach

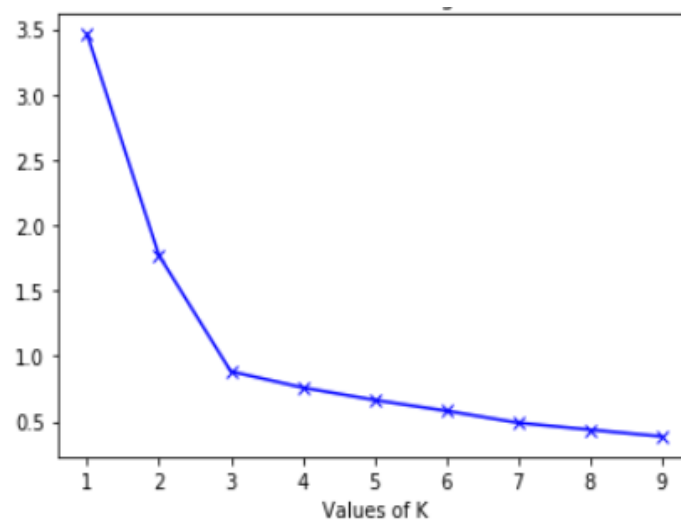
In this study we have used the Unsupervised ML approach to create segments of customers to see if the customers based on the attributes of age, spending score, and annual income can be clustered together.

K-means: When you have unlabeled data, clustering is a sort of unsupervised learning (i.e., data without defined categories or groups). The goal of this method is to find groups in the data, with K denoting the number of groups.



To pick the optimum value of K, we have used the Elbow Method.

Elbow Method: The Elbow Method can be used to determine the best K value. The elbow approach is a heuristic used in cluster analysis to determine the number of clusters in a data set. Plotting the explained variation as a function of the number of clusters and selecting the elbow of the curve as the number of clusters to utilize is the method.

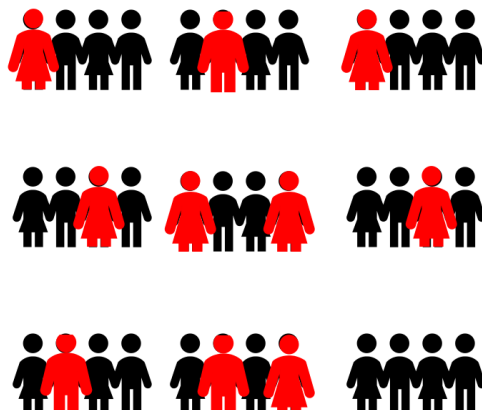


d) Sampling

In this study, we have used simple random sampling to remove any biases. We could have used cluster or stratified clustering but it would introduce biases based upon those groups and would make our clustering technique futile.

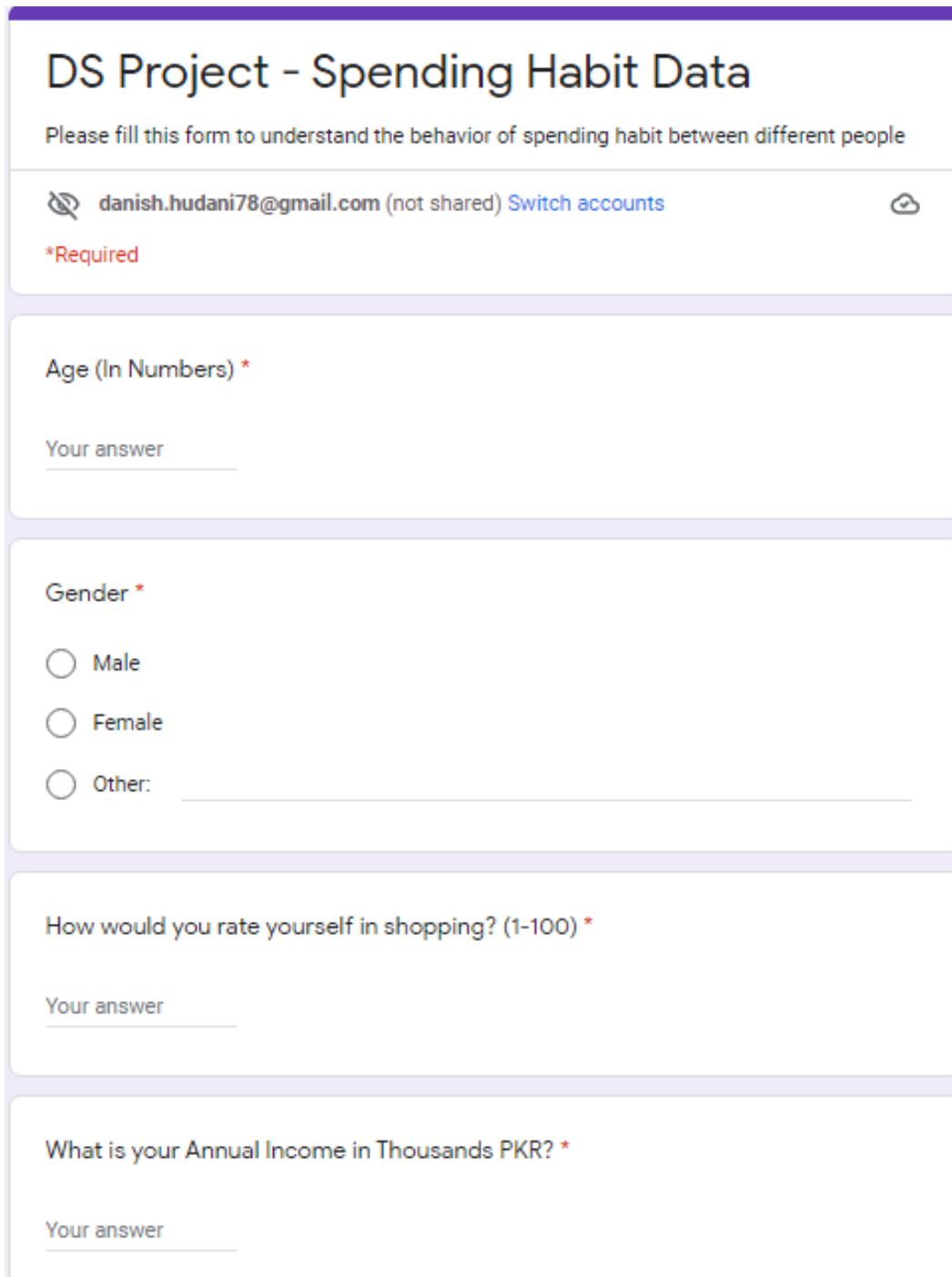
Simple Random Sampling: A simple random sample is a subset of a statistical population in which each subset member has the same chance of being chosen. A basic random sample is intended to represent a group in an unbiased manner.

Simple Random Sampling



e) Data Collection Technique & Instrument



To collect data from participants we used Google Forms to perform survey. The snap of the form is attached. Forms were distributed to different people who are either from the working class or retired, male and female. Forms were distributed on different platforms like facebook, whatsapp, and linkedin.



The image shows a Google Form titled "DS Project - Spending Habit Data". The form is designed to collect data on spending habits. It includes a header section with the title and a description: "Please fill this form to understand the behavior of spending habit between different people". Below the header, there is a section for the user's email address, which is "danish.hudani78@gmail.com (not shared)", with a "Switch accounts" link and a cloud icon. A red asterisk indicates that the email is required. The form then contains four questions, each with a text input field and a "Your answer" label. The questions are: "Age (In Numbers) *", "Gender *", "How would you rate yourself in shopping? (1-100) *", and "What is your Annual Income in Thousands PKR? *". The "Gender" question has three radio button options: "Male", "Female", and "Other: _____".

DS Project - Spending Habit Data

Please fill this form to understand the behavior of spending habit between different people

 **danish.hudani78@gmail.com** (not shared) [Switch accounts](#) 

***Required**

Age (In Numbers) *

Your answer _____

Gender *

☐ Male

☐ Female

☐ Other: _____

How would you rate yourself in shopping? (1-100) *

Your answer _____

What is your Annual Income in Thousands PKR? *

Your answer _____

8. Data Analysis

a) Demographic Analysis

We have used data from around 112 females out of which 57 are adult, 44 are middle-age and 11 are senior citizen, and 88 are male out of which 41 are adult, 29 are middle-age and 18 are senior citizen.

Age Group	Adult/Young	Middle-Age	Senior Citizen
Gender			
Female	57	44	11
Male	41	29	18

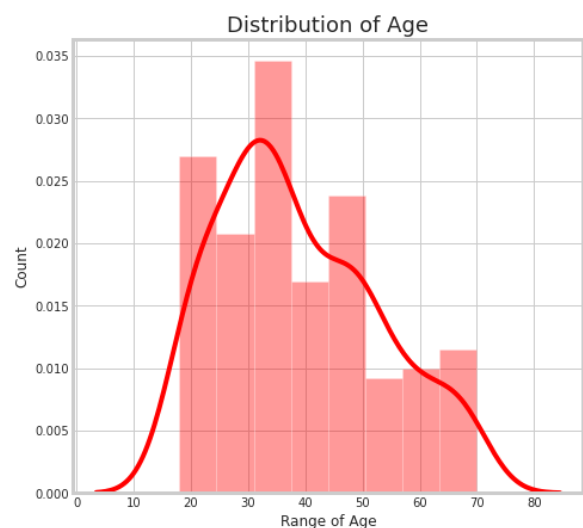
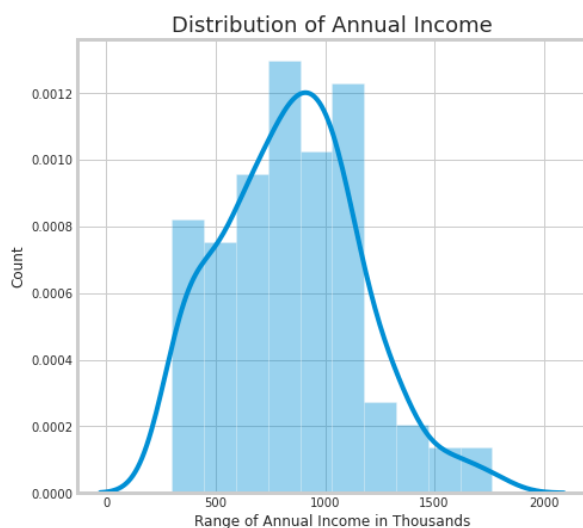
b) Exploratory Data Analysis

We have performed EDA by plotting distributions, pairplot for relation, heatmap for correlation and line plot that shows trends between those attributes.

- **Distribution of Annual Income and Age:**

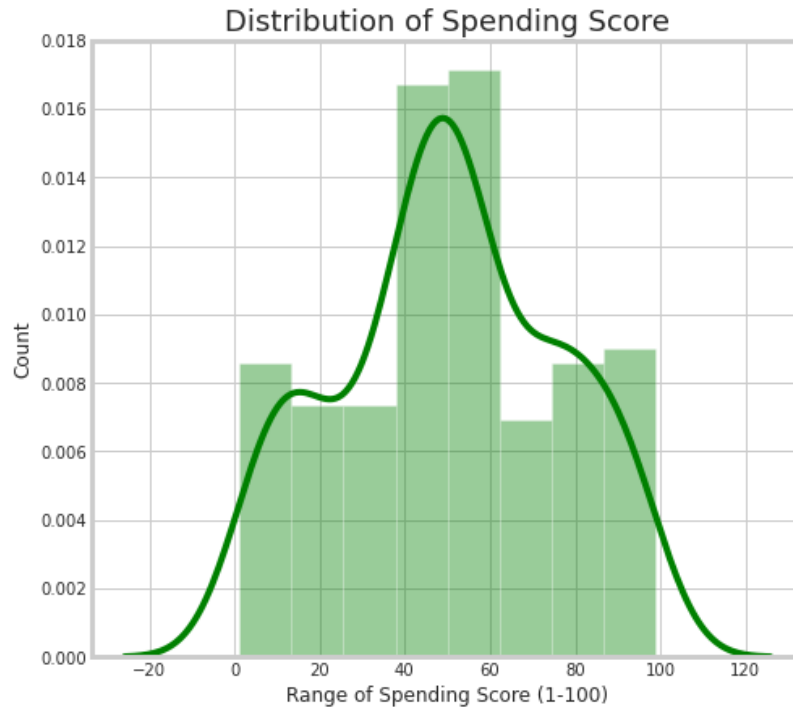
From the plot we can see that most of the customers in our data have annual income of around 300,000 PKR to 1,764,000 PKR. We can also say that the highest annual income in our data is around 1,764,000 PKR, whereas the lowest annual income in our data is around 300,000 PKR.

From the plot we can see that most of the customers in our data are of age between 18 and 50. Where senior citizens are less likely to go shopping and youngsters are also less likely to spend as compared to middle aged people.

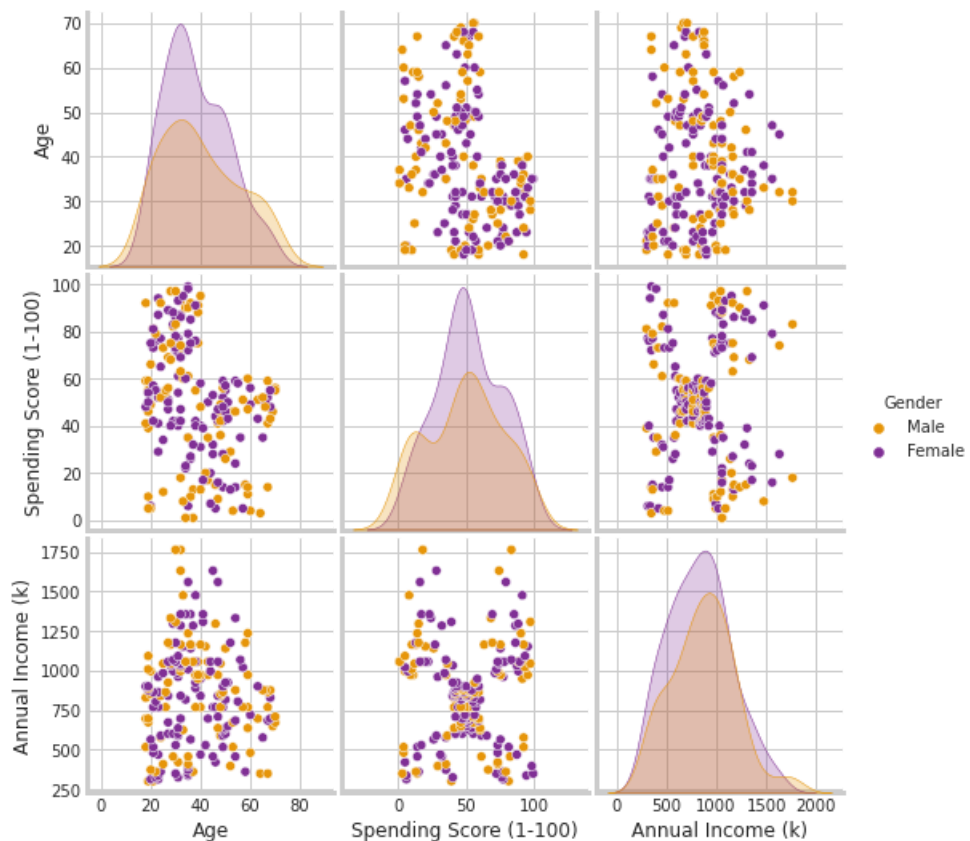


- **Distribution of Spending Score:**

From the distribution we can say that most of the spending score lies between 40 and 60. With a maximum spending score of 99 and minimum spending score of 1.

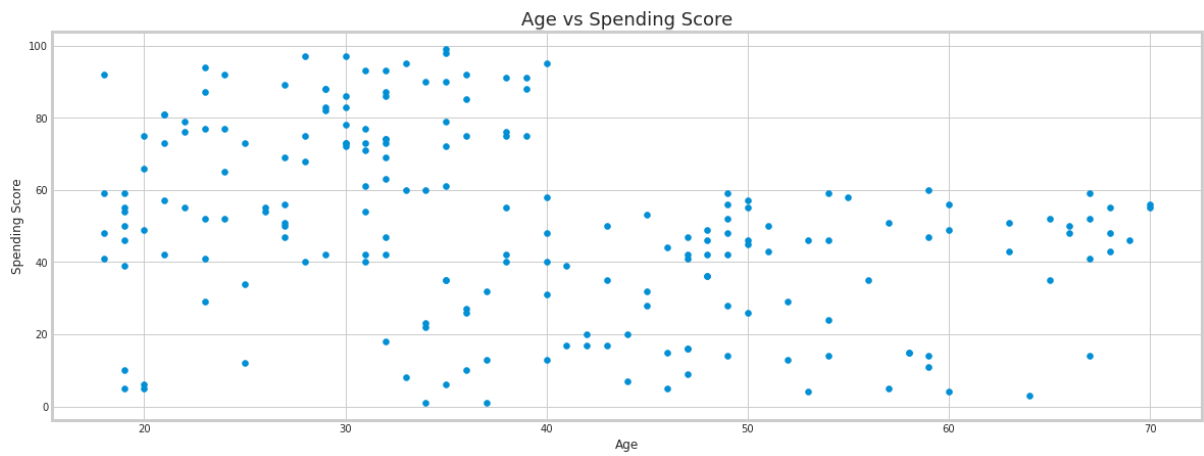


Below the pairplot (matrix of scatter plot) shows us that we do see some association or clustering between Annual Income and Spending Score, where Gender is not an attribute that we see segmented or distinguished.

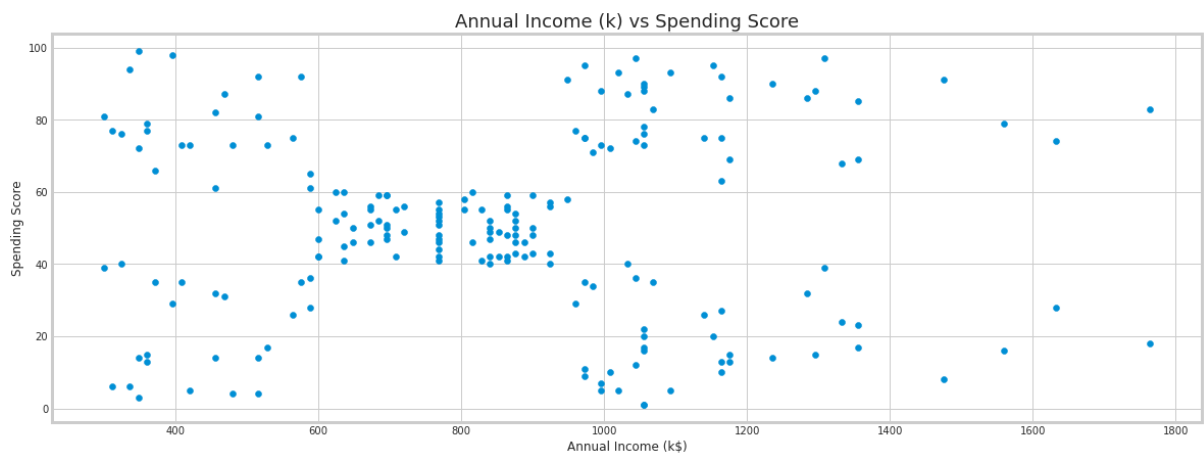


We also plotted separate scatter plots to explore the clusters.

- **Scatter Plot between Age and Spending Score:**



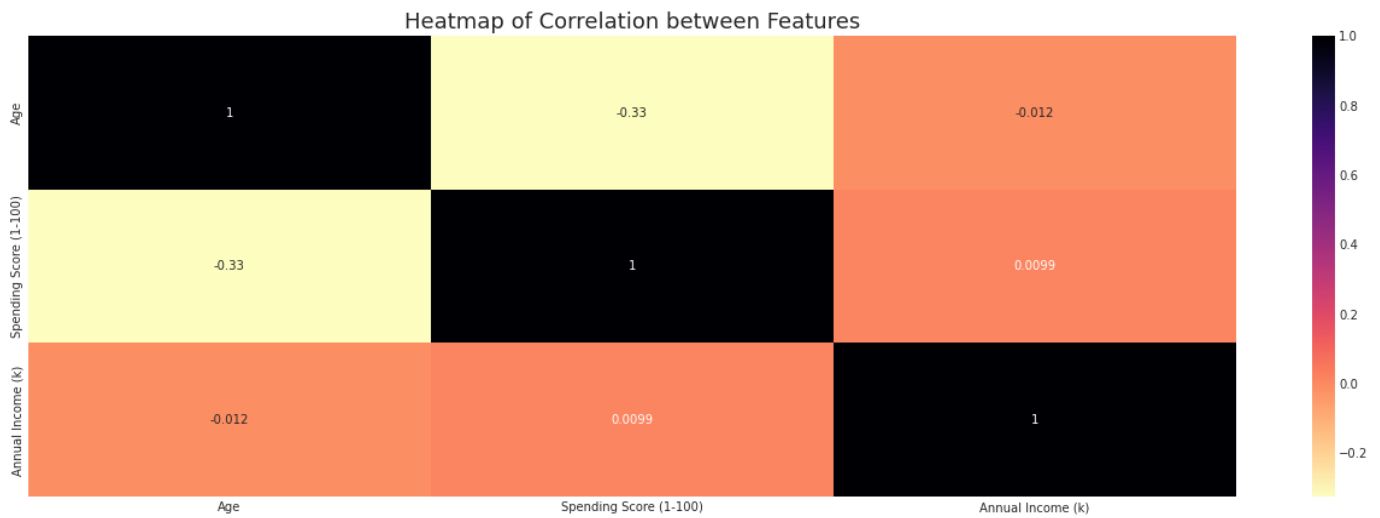
- **Scatter Plot between Annual Income and Spending Score:**



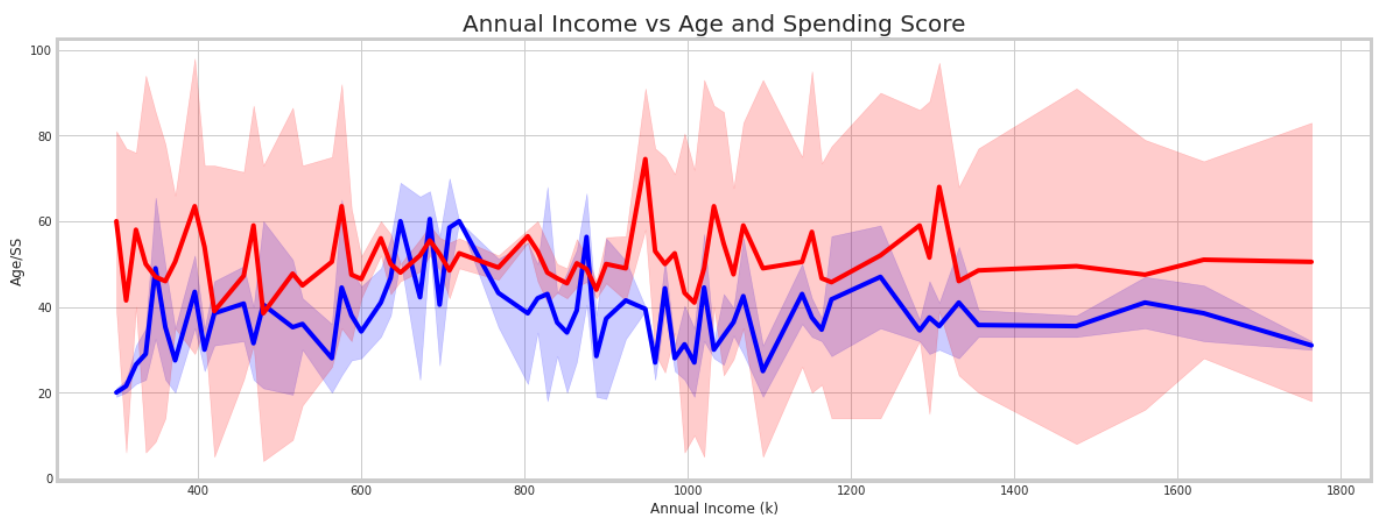
- **Scatter Plot between Gender and Spending Score:**



The below plot shows the heat map that shows the correlation annotated with its value and coded with color based on the strength of its correlation. The heatmap of correlation suggests that there aren't any strong correlations between features.



The below mentioned plot shows the line graph that explores the trend between the Age and Spending Score with respect to Annual Income and sees there is a trend between Age and Spending Score.



We can refer to the above plot which shows a clear trend between Annual Income, Age and Spending Score, using these attributes to implement our clustering algorithm.

c) Modeling

In this study we have used the K-Means clustering algorithm to create clusters of people based on the above mentioned attributes.

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The purpose of this technique is to locate groups in the data, with K representing the number of groups. Based on the attributes provided, the algorithm assigns each data point to one of K groups iteratively. Data points are grouped together based on how comparable their features are.

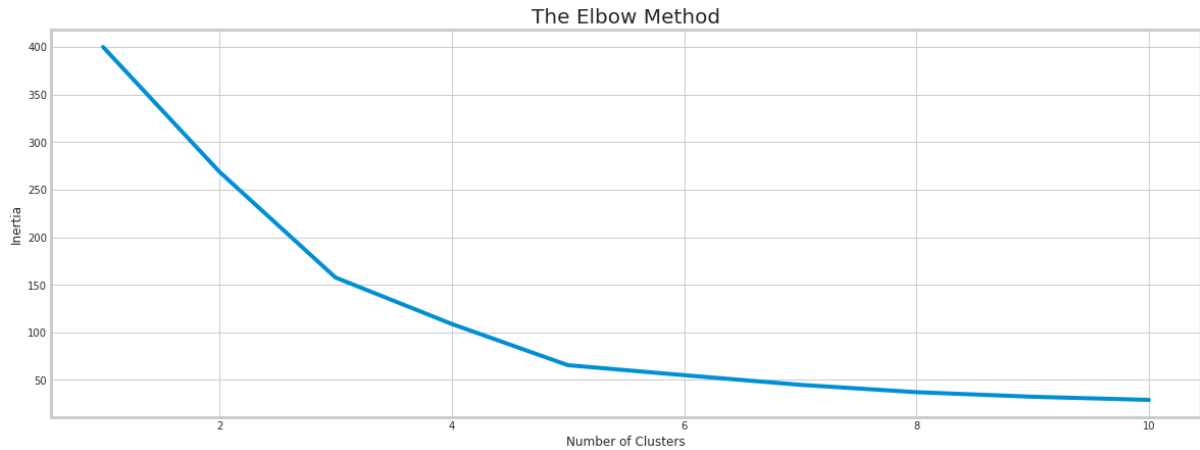
The optimum value of K can be chosen using the Elbow Method. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

As K-Means use Euclidean Distance, we need to standardize the features to avoid dominance of one feature over another, which we achieved using StandardScaler() from the sklearn module in python.

In the first scenario, we have created clusters based on Annual Income and Spending Score.



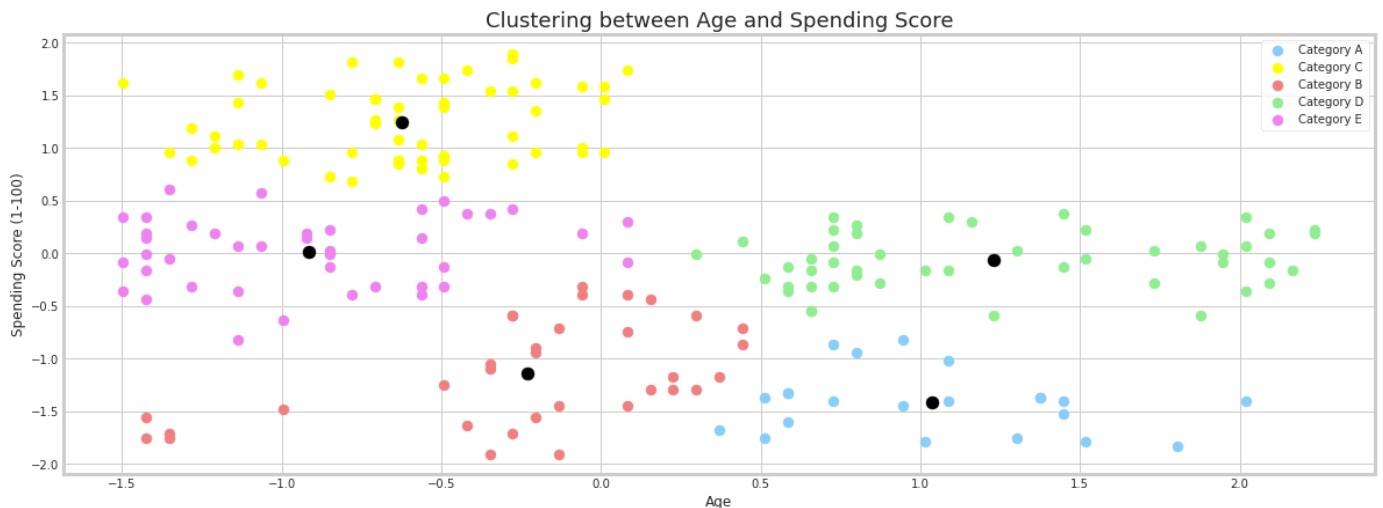
As we can see from the above plot, there are 5 centroids of clusters marked in black dots, also colored in different categories. The no of clusters have been deduced by finding the optimum point in the below mentioned graph using the elbow method.



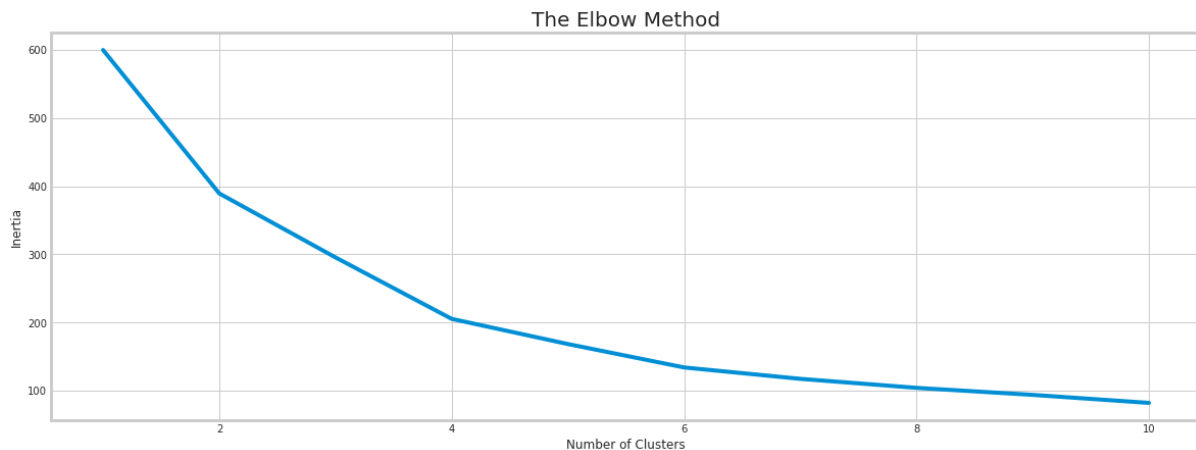
From above clustering analysis, we can say that there are 5 categories of customer in our data i.e

- A person who makes little and spends a lot is reckless: **Category A**
- A person who makes little and spends little is thrifty: **Category B**
- A person who makes the average and spends the average is normal: **Category C**
- A person who makes a lot and spends a lot is a high roller: **Category D**
- A person who makes a lot and spends little is a miser: **Category E**

In the second scenario, we have created clusters based on Age and Spending Score.



As we can see from the above plot, there are 5 centroids of clusters marked in black dots, also colored in different categories. The no of clusters have been deduced by finding the optimum point in the below mentioned graph using the elbow method.



From above clustering analysis, we can say that there are 5 categories of customer in our data i.e

- A person who is a senior citizen and spends less: **Category A**
- A person who is senior citizen and spends average: **Category B**
- A person who is young and spends the average: **Category C**
- A person is young and spends a lot: **Category D**
- A person is young and middle-aged and spends little: **Category E**

Consolidating the above separate cluster analysis into one by using all the attributes, we have got the plot, where customers are segregated into 5 clusters or 5 categories. The description and plot of these consolidated categories are given in the results and finding section.

d) Model Validation

K-Means clustering algorithm is an unsupervised machine learning algorithm such that we don't have any target variable to determine the cost of deviation from what the model should have predicted or derived. It's an inferential approach that only provides us with implicit validation in metric such as inertia.

Inertia is a metric that indicates how well a dataset was clustered.

It's determined by squaring the distance between each data point and its centroid and summing the squares throughout one cluster.

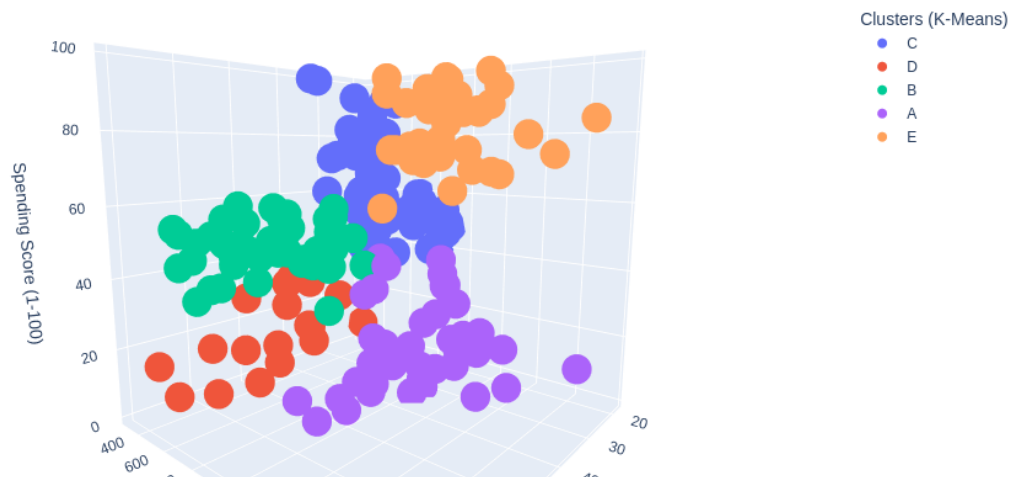
- The inertia of the 1st scenario is **65.5684**
- The inertia of the 2nd scenario is **71.093**
- The inertia of the consolidated scenario is **168.247**

The inertia in consolidated scenarios increases due to increase in features.

e) Results and Findings

The results of the study suggest that there are similarities based upon Age, Annual Income and Spending Score which could be used to design marketing strategies.

In the below attached plot, you can see 5 clusters or 5 categories which show a clear grouping of people based on those attributes.



We can conclude that there are 5 categories in our data i.e:

- **Category A:** Having an average spending score of 19.35, between 1 and 42. Age lies between 19 and 47, average around 40. Annual Income is between 64 Thousand USD and 95 Thousand USD, average between 86 Thousand USD. Preferably not to be targeted.
- **Category B:** Having an average spending score of 62.24, between 39 and 99. Age lies between 18 and 38, average around 25. Annual Income is between 15 Thousand USD and 67 Thousand USD, average between 41 Thousand USD. Preferably to be targeted.
- **Category C:** Having an average spending score of 48.85, between 1 and 27.5. Age lies between 40 and 70, average around 55. Annual Income is between 38 Thousand USD and 79 Thousand USD, average between 54 Thousand USD. Preferably to be targeted.

- **Category D:** Having an average spending score of 81, between 58 and 97. Age lies between 27 and 40, average around 32. Annual Income is between 69 Thousand USD and 137 Thousand USD, average between 86 Thousand USD. Preferably to be targeted.
- **Category E:** Having an average spending score of 18, between 3 and 36. Age lies between 20 and 67, average around 46. Annual Income is between 16 Thousand USD and 39 Thousand USD, average between 26.7 Thousand USD. Preferably not to be targeted.

The statistics of the above-mentioned groups based on the attributes Age, Annual Income, and Spending Score are attached at the end of the report in the jupyter notebook.

9. Conclusion

From the above analysis we can infer that the K-Means clustering has done a good job in creating segments of customers based upon their age, spending score, and annual income. If any brand needs to create their marketing campaigns, these attributes and model can be useful in determining their strategy.

In this study, we have reached the following conclusion:

1. There have been associations or similarities between customers of different age groups, spending scores and annual income. Which have been used to analyze their spending habits via trends between age, spending score and annual income.
2. We have successfully clustered different customers based upon 5 mentioned categories, with an idea on who to target if to sell a product from a brand.

10. Limitation

Limitations of this study includes:

- Confounding effects between attributes have not been explained. This study assumes that there isn't any confounding attributes that are affecting the attributes.
- The K-Means Clustering algorithm can be used for descriptive and predictive analysis but is highly prone to outliers, which can affect the model's ability to predict customers in new segments.
- This approach as explained before can be used for predictive analysis but this study focuses on a descriptive approach such that no concept of training and explicit validation can be explained.

11. References

- <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
- <https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84>
- <https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84>
- <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
- <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>
- <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>
- <https://www.analyticsvidhya.com/blog/2021/06/how-to-solve-customer-segmentation-problem-with-machine-learning/>