# REPORT ON CUSTOMER SEGMENTATION PROJECT

Abstract

Customer segmentation is a very popular problem in the business world. We can target our products to the right customers only when we know which of the customers are/can be potential buyer of our product. By knowing the right customer base, we can make strategies and the right products to target the right customer. We can also target them with the right type of ads only if we know the likes and dislikes of our customers.

Segmentation problems are basically clustering problems and this comes under the domain of unsupervised learning problems. Unsupervised learning is often much more challenging in comparison to the supervised ones i.e., classification problems etc. The exercise tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response. Unsupervised learning is often performed as part of an exploratory data analysis.

Here, our main objective was to analyse the different features of the data using various visualisation techniques like distribution plots, scatter plots, kde plots and heat map etc. Then we performed K Means clustering using silhouette scores to divide the data into different clusters based on the features we analysed. We were able to extract very useful information from the data using these techniques.

# INDEX

1. INTRODUCTION………………………………………………………………………………….2
2. DATA VISUALISATION………………………………………………………………………...4
3. PROPOSED METHOD WITH ARCHITECTURE……………………………………….12
4. METHODOLOGY………………………………………………………………………………...13
5. IMPLEMENTATION……………………………………………………………………………..15
6. CONCLUSION………………………………………………………………………………………16

# Introduction

As the name itself suggests, that we have to group the customers in different clusters based on their characteristics. For this, we have a data of 200 mall customers. For each customer, we have the following features:

Customer ID, Age, Gender, Annual Income (k$), Spending Score (1-100)

We used the pandas library to store the data in a dataframe. Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.

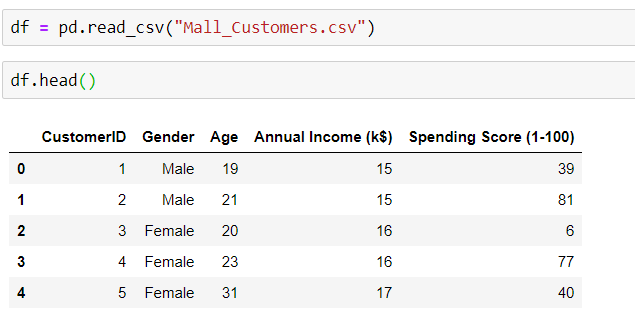


Figure 1 Storing the data in a pandas data frame and showing some of the starting rows

Then we checked whether there were any null values in the data i.e., missing values.

# Data Visualisation

A picture speaks a thousand words.

You give a raw data to someone he might not a get single thing from thing by analysing it but if you provide him with the different visualisations of the same data, he might extract useful information out of it by merely looking at it.

With the same purpose we tried to visualise the given data using seaborn library. Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps us explore and understand our data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets us focus on what the different elements of our plots mean, rather than on the details of how to draw them.

First of all, we plot the distribution of males and females in the dataset. As we can clearly see from the data that there are more females than males in our data.

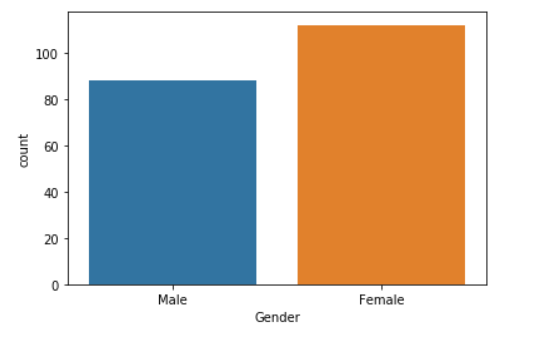


Figure 2 Distribution of male and female

Distribution plot of Age, Annual Income(k$), Spending Score (1-100):

In the below plot, along with the histograms we also have the kde curve showing distribution of different features.

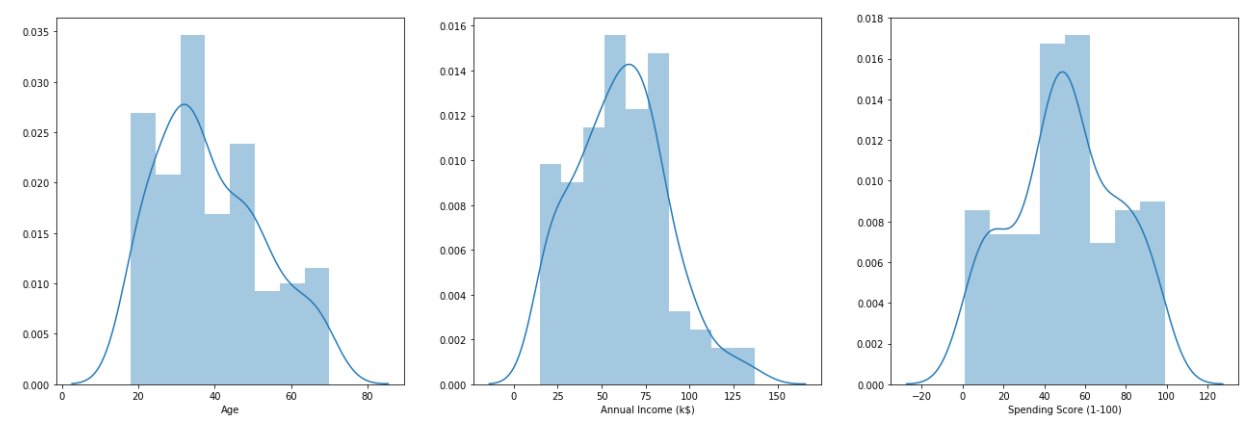
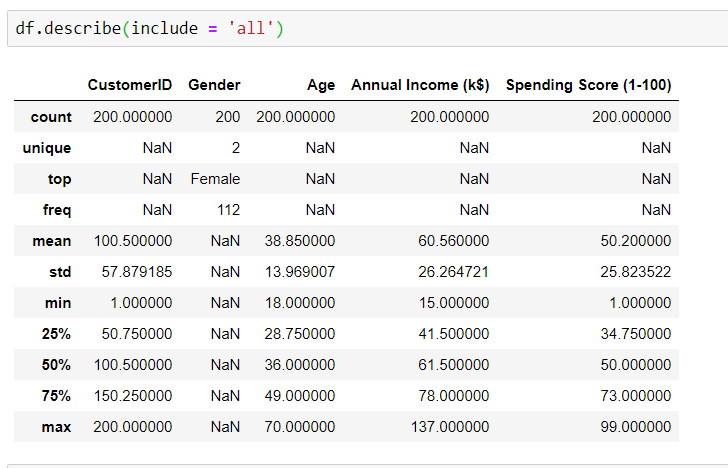


Figure 3 Distribution of age, spending score and annual income

Five-point summary and other useful parameters of our data can be simply shown using the code show in the diagram below. The five-point summary includes: min, max, Q1(25%), median and Q3(75%). The five-number summary gives you a rough idea about what our data set looks like.



We can show the bivariate relationship between any of the two features of the data using the pairplot() function. Gender feature is shown as a third variable in the plot using the hue argument. If we look at the scatter plot of Spending Score vs Annual Income in figure 4, we can clearly see different clusters in the data. A look at the age vs spending score scatter plot (figure 5) reveals that after the age of approx. 40 spending score kind of decreases drastically. This point is also proved from the correlation matrix as the value of correlation for spending score and age is negative (-0.33). This may be a useful information from business perspective that people below 40 are more likely to buy a product than the older ones. This information can also be interpreted as people after 40 are most likely married and have children whose needs they have to fulfil. And for this they are saving more from their income for their future generation. A closer look at age vs annual income scatter plot in figure 6 shows that there are few males in early 30s who are earning more than 110k $ whereas there is no woman at that age earning that much. Only at late 30s we see woman earning above 100k $. This is showing the income disparity between the two genders.

Mean spending score of females are more than that of males so females are more spendthrift.

Here, In Annual Income vs Spending Score, there is one area where density is very high which is in the middle. And other four areas where user shows different patterns. It might be because of the different purchasing power of the customer and their different spending habits. Five groups from the observations which can be deciphered are:

Low Income & High spending habits

Low Income & Low Spending habits

Moderate Income & Moderate Spending habits

High Income & High Spending habits

High Income & low Spending habits

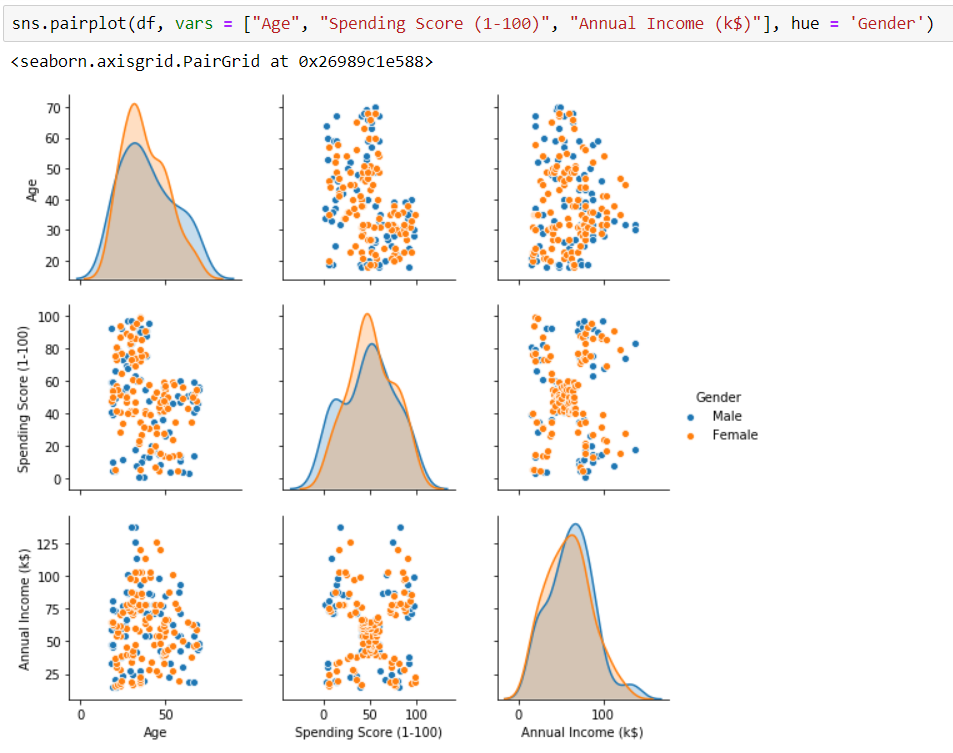
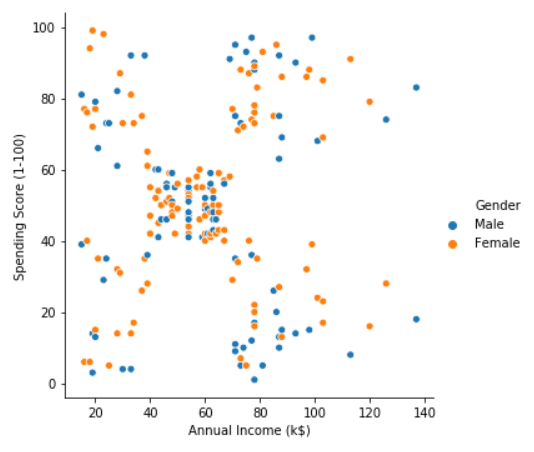


Figure 4 Pair plot of all features



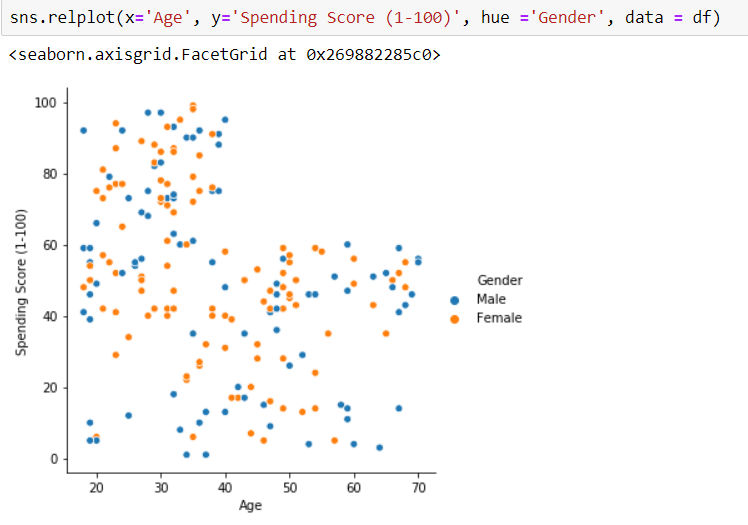


Figure 5 Scatter Plot of Age vs Spending Score

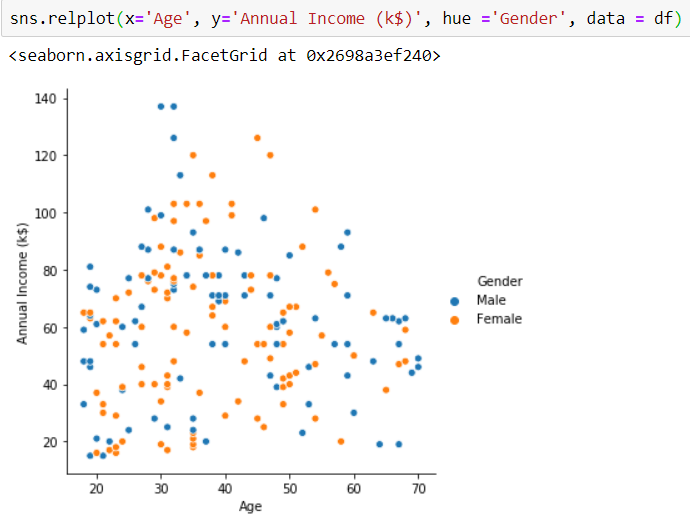


Figure 6 Scatter Plot of Age vs Annual Income

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It’s a common tool for describing simple relationships without making a statement about cause and effect.

Figure 7 Heatmap of correlation matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

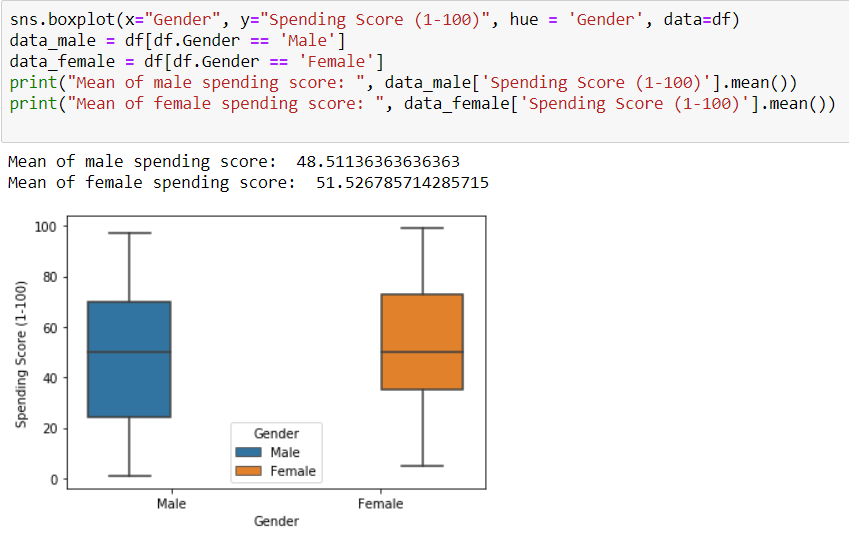


Figure 8 Box plot of spending scores of male and female

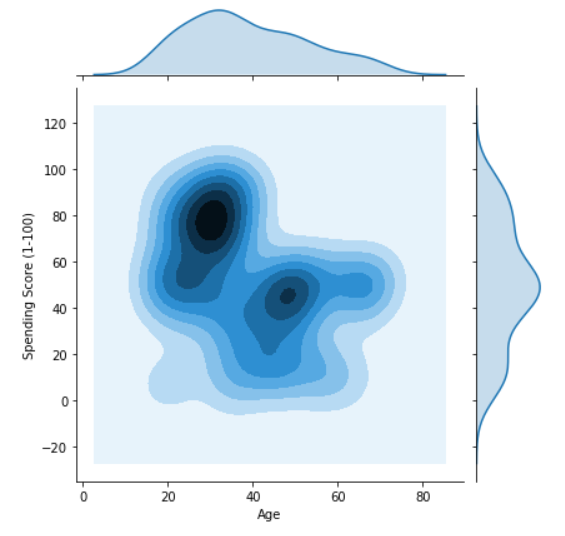
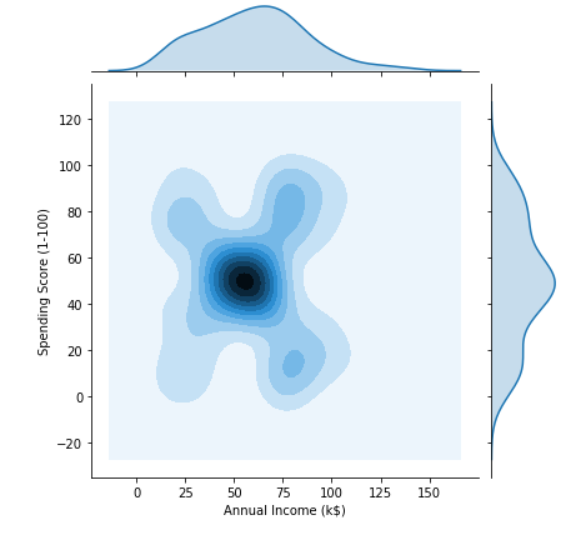
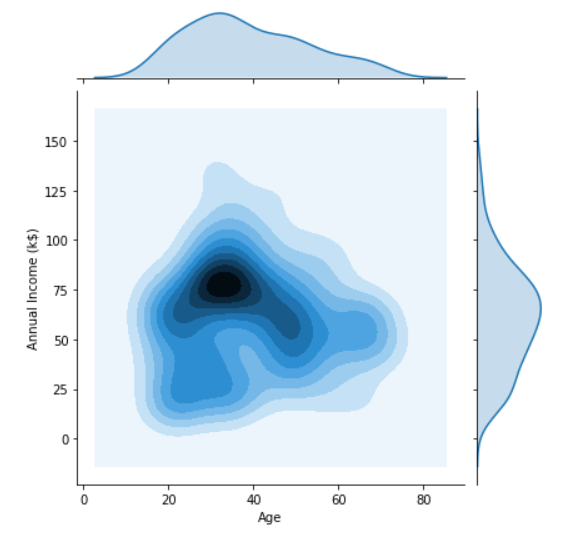


Figure 9 Joint plot of Age vs Spending Score (1-100), kind = kde



# Proposed Method with Architecture

As is evident from the above pair plot in figure 4, that Annual income vs Spending score plot clearly has distinct cluster. So, we will try to cluster our data based on these features only i.e., annual income and spending score. For this we will use K-means clustering algorithm with silhouette score. In other words, annual income and spending score are the features which captures most of the variability of the dataset.

DATA

DIFFERENT CLUSTERS PLOTTED

K MEANS CLUSTERING ALGORITHM

FEATURES SELECTED FROM THE DATA

ANALYSIS USING VISUALISATION TOOLS

This flow diagram gives a complete picture of how our data get through different stages and finally got clustered and plotted.

# METHODOLOGY

## CLUSTERING

Clustering refers to a very broad set of techniques used for finding subgroups or clusters, in a data set. When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other. Of course, to make this concrete, we must define what it means for two or more observations to be similar or different. Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Since clustering is popular in many fields, there exist a great number of clustering methods. Here, we will mainly focus on perhaps the two best-known clustering approaches: K-means clustering and hierarchical clustering. In K-means clustering, we seek to partition the observations into a pre-specified number of clusters. On the other hand, in hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

Here, we have used the K-Means clustering approach to cluster our data.

Algorithm K-Means Clustering

1. First of all, set the no. of clusters. For instance, say k=3.
2. Randomly select 3 distinct data points from the sample. It will work as centroid of the cluster for the time being.
3. Measure the Euclidean distance of first data point from these 3 centroids.
4. Assign the first point to the nearest cluster.

Repeat 3 and 4 one by one for all other data points.

1. Calculate mean of the cluster. This will act as new centroids of our cluster

Repeat steps 3,4 and 5 until the data points stay in the same cluster.

Choosing the right no. clusters i.e., K=?

Silhouette Score: This is a better measure to decide the number of clusters to be formulated from the data. It is calculated for each instance and the formula goes like this:

Silhouette Coefficient = (x-y)/ max (x, y)

where, y is the mean intra cluster distance: mean distance to the other instances in the same cluster. x depicts mean nearest cluster distance i.e., mean distance to the instances of the next closest cluster.

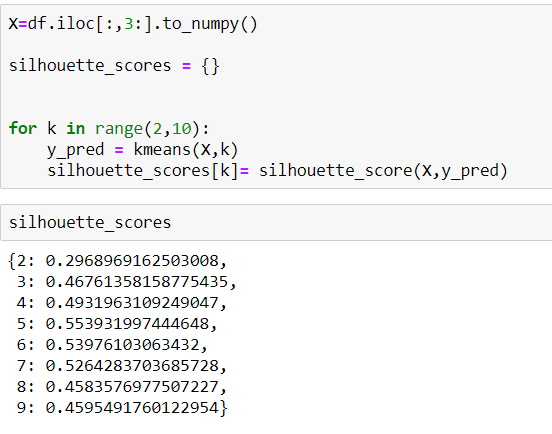
The coefficient varies between -1 and 1. A value close to 1 implies that the instance is close to its cluster is a part of the right cluster. Whereas, a value close to -1 means that the value is assigned to the wrong cluster. In our K=5 has the highest Silhouette score. So, we chose K=5

Figure 10 Silhouette Scores

Figure 11 Silhouette Calculation for different values of k

# 5. IMPLEMENTATION

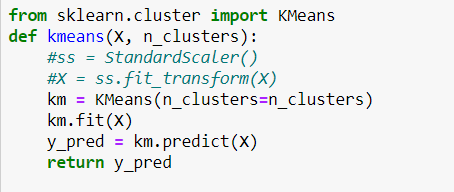
To use the K-Means clustering algorithm we use the sklearn library. ‘cluster’ module of the sklearn library provides the K-Means algorithm model. We can train this on our dataset using this.

Figure 12 Importing KMeans and instantiating it

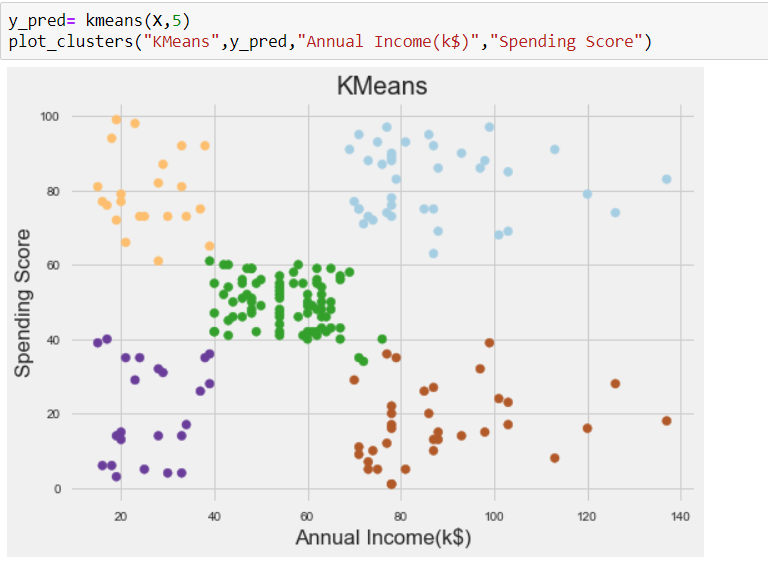
Here we defined a function named kmeans which takes the data and the no. of clusters as argument. Inside the definition KMeans model is instantiated by the name of km and then trained on the data using fit method. Then silhouette score was calculated using values of k ranging from 2 to 9 as shown in figure 8. And finally, K=5 was as it had the highest silhouette score. Then these points were plotted using matplotlib library using the plot clusters function defined in our code.

Figure 13 Different cluster of the data identified and plotted

# CONCLUSION

Visualising the data using different visualisation tools and then performing the clustering gives us the following insights of the data:

1. There are five types of people in our dataset based on the annual income and spending score.
   1. People in the sky-blue cluster top right in figure 13. These are rich people and whose spending score is also high. These customers should be our top priority as they are reliable and have almost no chance of defaulting.
   2. People who are rich but miser (bottom right). These set of people have money but are miser so we have to put extra effort on these people to make them spend. These people should be targeted with advertisements and other convincing to make them buy.
   3. People who have low income but have high spending score (orange cluster). These are too much dependent on the credit card and can become defaulters.
   4. The other two types of people are not that much important but the mall should keep them happy and satisfied too.
2. A look on the scatter plot of Age vs Spending Score (figure 5) tells us that after 40 years of age spending score decreases drastically that means. Younger people are more likely to buy a product than the older ones.
3. A closer look at age vs annual income scatter plot in figure 6 shows that there are few males in early 30s who are earning more than 110k $ whereas there is no woman at that age earning that much. Only at late 30s we see woman earning above 100k $. This is showing the income disparity between the two genders.