A Report

on

CUSTOMER SEGMENTATION

USING

K – MEANS CLUSTERING

*by*

Tanay Mayee

**Abstract**

As more and more business being coming up every day, it has become significantly important for the businesses to apply marketing strategies to stay in the market as the competition had been cut to throat. This is where machine learning comes into play, various algorithms are applied for unravelling the hidden patterns in the data for better decision making for the future. This elude concept of which segment to target is made unequivocal by applying segmentation. The process of segmenting the customers with similar behaviors into the same segment and with different patterns into different segments is called customer segmentation. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms which we’ll be using to perform customer segmentation. In our case we have a mall customers dataset that has 200 customers and have some basic data about these customers like the Customer ID, age, gender, annual income and spending score. Our aim is to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly.

**Contents**

|  |  |  |
| --- | --- | --- |
|  |  | **Page No.** |
|  | **Abstract** | 2 |
| **Chapter-1** | **Introduction**   1. What is Customer Segmentation? 2. Types of Customer Segmentation 3. Advantages of Customer Segmentation 4. Problem Statement 5. The Dataset | 4  4  4  4  5  5 |
| **Chapter-2** | **Existing Method**   1. What is Clustering? 2. Types of Clustering Algorithms 3. K – Means Clustering 4. Agglomerative Clustering 5. Mean Shift Clustering | 6  6  6  6  7  7 |
| **Chapter-3** | **Proposed Method with Architecture**   1. K –Means Clustering 2. The Elbow Method | 8  8  9 |
| **Chapter-4** | **Methodology**   1. Loading the Data and the Necessary Libraries 2. Label Encoding 3. Feature Scaling 4. K – Means Clustering 5. Principle Component Analysis (PCA) | 11  11  11  11  12  13 |
| **Chapter-5** | **Implementation**   1. Data Analysis 2. Data Visualization 3. K – Means Clustering and PCA | 14  14  15  23 |
| **Chapter-6** | **Conclusion** | 30 |

**Chapter 1**

**Introduction**

1. **What is Customer Segmentation?**

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

1. **Types of Customer Segmentation**

* Demographic Segmentation – based on gender, age, occupation, marital status, income, etc.
* Geographic Segmentation – based on country, state, or city of residence. Local businesses may even segment by specific towns or counties.
* Technographic Segmentation – based on preferred technologies, software, and mobile devices.
* Psychographic Segmentation – based on personal attitudes, values, interests, or personality traits.
* Behavioral Segmentation – based on actions or inactions, spending/consumption habits, feature use, session frequency, browsing history, average order value, etc.

1. **Advantages of Customer Segmentation**
2. Determine appropriate product pricing.
3. Develop customized marketing campaigns.
4. Design an optimal distribution strategy.
5. Choose specific product features for deployment.
6. Prioritize new product development efforts.
7. **Problem Statement**

Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively.

Use K – Means clustering and also visualize the gender and age distribution. Then analyze their annual incomes and spending scores.

Language: Python (or) R

1. **The Dataset**

The dataset has the data of 200 customers which includes some basic data about the customers like Customer ID, age, gender, annual income and spending score. We want to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly.

Link:(https://drive.google.com/file/d/19BOhwz52NUY3dg8XErVYglctpr5sjTy4/view)

**Chapter 2**

**Existing Methods**

1. **What is Clustering?**

It is basically a type of [unsupervised learning method](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.  
**Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

1. **Types of Clustering Algorithms**

* **K-means Clustering:**

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method, after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre’s are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.

* **Agglomerative Clustering:**

Agglomerative Clustering is based on forming a hierarchy represented by dendrograms (discussed in later section). Dendrogram acts as memory for the algorithm to tell about how the clusters are being formed. The clustering starts with forming N clusters for N data points and then merging along the closest data points together in each step such that the current step contains one cluster less than the previous one.

* **Dendrogram:**

Dendrogram is the hierarchical representation of object, it is used to determine the output of the hierarchical clustering. The way Dendrogram is interpreted is by checking the height of each clade (horizontal line), the lower the height the more associated data points are and greater the height more less associated data points.

* **Mean Shift Clustering:**

This clustering algorithm is a non-parametric iterative algorithm functions by assuming the all the data points in the feature space as empirical probability density function. The algorithm clusters each data point by allowing data point converge to a region of local maxima which is achieved by fixing a window around each data point finding the mean and then shifting the window to the mean and repeat the steps until all the data point converges forming the clusters.

**Chapter 3**

**Proposed Method with Architecture**

1. **K – Means Clustering Algorithm**

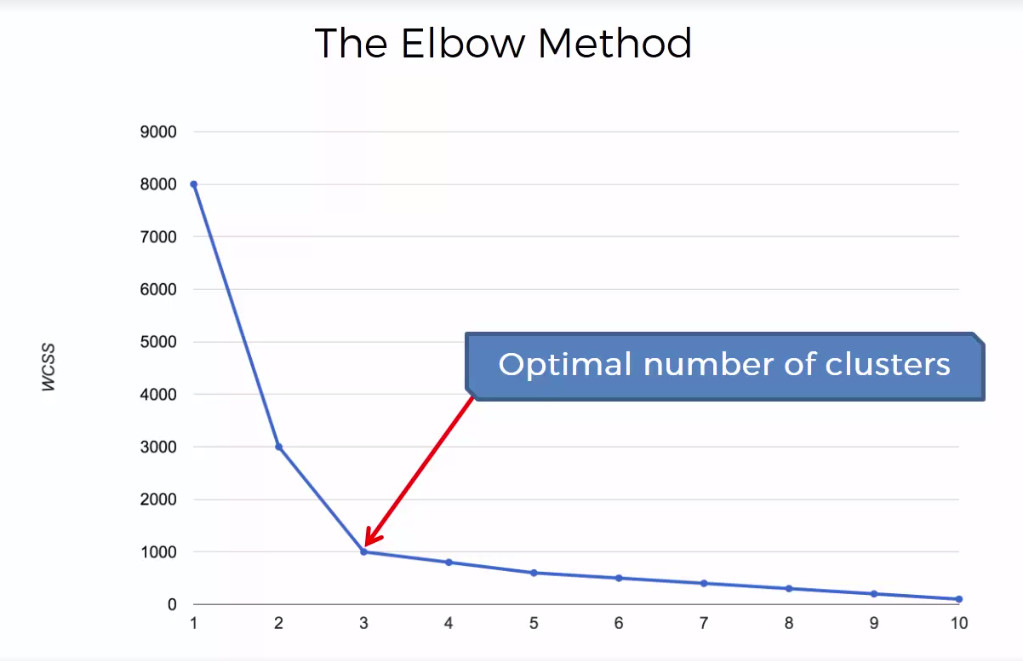
K-means algorithm in one of the most popular centroid based algorithm. Suppose data set, D, contains n objects in space. Partitioning methods distribute the objects in D into k clusters, C1,...,Ck , that is, Ci ⊂ D and Ci ∩Cj = ∅ for (1 ≤ i, j ≤ k). A centroid-based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster. Conceptually, the centroid of a cluster is its center point. The difference between an object p ∈ Ci and ci , the representative of the cluster, is measured by dist(p,ci), where dist(x,y) is the Euclidean distance between two points x and y.

**Algorithm:**

The k-means algorithm for partitioning, where each cluster’s center is represented by the mean value of the objects in the cluster. Input: k: the number of clusters, D: a data set containing n objects. Output: A set of k clusters.

1. **The Elbow Method**

Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further.

****

**Chapter 4**

**Methodology**

1. **Loading the Dataset and the Necessary Libraries**

Firstly we need to import our dataset in order to perform customer segmentation.

Following are the libraries which we’ll be using:

* Numpy
* Pandas
* Mathplotlib
* Seaborn
* Scikitlearn (LabelEncoder, StandardScaler, KMeans, PCA)

1. **Label Encoding**

We’ll need to label encode the ‘Gender’ column as so that each of the values i.e. ‘M’ and ‘F’ will be assigned a particular integer. In our case the label encoder has assigned a ‘0’ to the Female customers and a ‘1’ to the Male customers.

1. **Feature Scaling**

Now to make our analysis fair enough we’ll need to use the standard scaler in order to bring all the values in our dataset to a certain range so that it’ll become a lot easier for us to apply K – Means Clustering on the scaled values.

1. **K – Means Clustering**

**Choosing the optimal number of Clusters:**

Now let's first test out some values for the number of clusters and then we'll decide the actual number of clusters using the Elbow Method.

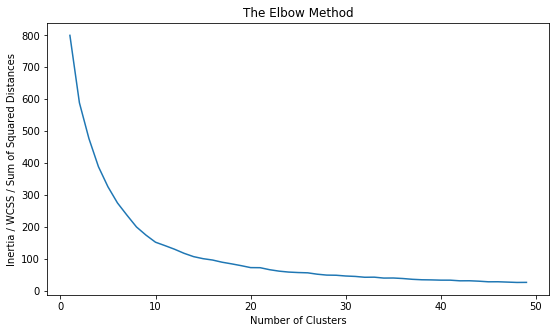
After plotting the Inertia / WCSS / Sum of Squared Errors v/s Number of Clusters graph we can now finally choose the actual number of clusters (K) that we’ll be using in our actual model.

**The Elbow Method:**

Step-1: Run the algorithm for various values of k i.e making the k vary from 1 to 50 in our case.

Step-2: Calculate the within cluster squared error.

Step-3: Plot the calculated error, where a bent elbow like structure will form (in our case it’s at k = 10 ), will give the optimal value of clusters.

****

**Algorithm:**

Step-1: Initialize the K (= 10) clusters.

Step-2: Assign the data point that is closest to any particular cluster.

Step-3: Recalculate the centroid position based on the mean of the cluster formed

Step-4: Repeat step 2 and 3 until the centroid position remains unchanged in the previous and current iteration.

1. **Principle Component Analysis (PCA)**

PCA is used in [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis) and for making [predictive models](https://en.wikipedia.org/wiki/Predictive_modeling). It is commonly used for [dimensionality reduction](https://en.wikipedia.org/wiki/Dimensionality_reduction) by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. The {\displaystyle i^{\text{th}}}principal component can be taken as a direction orthogonal to the first {\displaystyle i-1}principal components that maximizes the variance of the projected data. In our case we have reduced our dimensionality from 4 dimensions to 2 dimensions i.e. PC1 and PC2. What PCA actually does is that it reduces the dimensionality to such an extent that it becomes convenient for visualization that still contains most of the information.

**Chapter 5**

**Implementation**

1. **Data Analysis**

* Importing the Data set using the ‘read\_csv’ method of pandas.
* Following is what our dataset looks like:

| * **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 |
| ... | ... | ... | ... | ... | ... |
| 195 | 196 | Female | 35 | 120 | 79 |
| 196 | 197 | Female | 45 | 126 | 28 |
| 197 | 198 | Male | 32 | 126 | 74 |
| 198 | 199 | Male | 32 | 137 | 18 |
| 199 | 200 | Male | 30 | 137 | 83 |

200 rows × 5 columns

* Dropping the ‘CustomerID’ column as it is not necessary for our analysis.

|  | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
| --- | --- | --- | --- |
| count | 200.000000 | 200.000000 | 200.000000 |
| mean | 38.850000 | 60.560000 | 50.200000 |
| std | 13.969007 | 26.264721 | 25.823522 |
| min | 18.000000 | 15.000000 | 1.000000 |
| 25% | 28.750000 | 41.500000 | 34.750000 |
| 50% | 36.000000 | 61.500000 | 50.000000 |
| 75% | 49.000000 | 78.000000 | 73.000000 |
| max | 70.000000 | 137.000000 | 99.000000 |

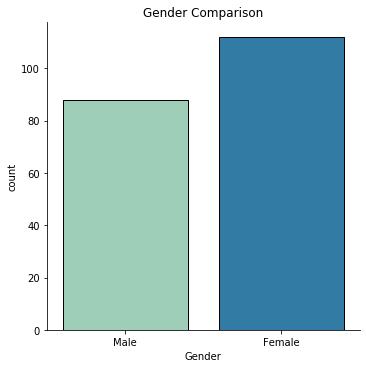
* Getting some insights on our dataset:
* Checking for any NULL values in our dataset. There were no NULL values in our dataset.

1. **Data Visualization**

Let us now perform data visualization for our dataset so that we’ll have a visual understanding of what actually our dataset represents.

Following are the graphs that were plotted to get a clear view of our data values which will ultimately help us to find some pattern between various features of the customers:

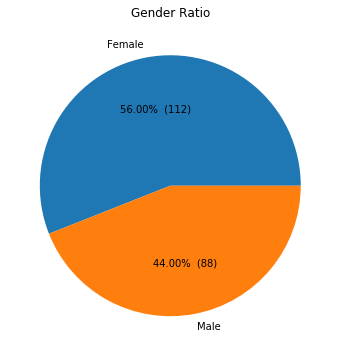
### Visualizing the Gender distributions in the dataset

****

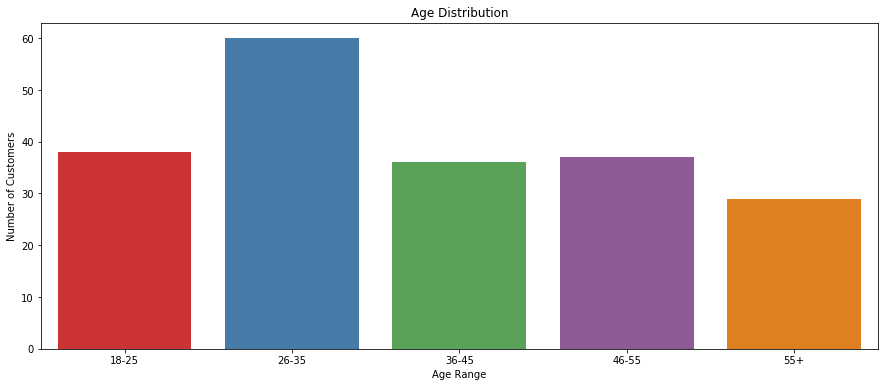
### INSIGHTS:

* Clearly there are more female customers than the male customers in our dataset.
* To be precise let's plot a pie chart for the Gender Ratio comparison.
* There are 112 female customers which is 56% of the total customers and 88 male customers which is 44% of the total customers.

### Gender Ratio



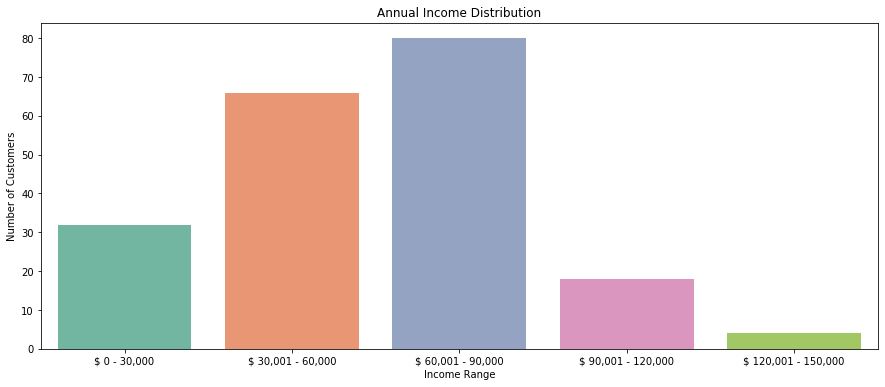
### Number of Customers according to their Age



### INSIGHTS:[¶](http://localhost:8889/notebooks/Customer%20Segmentation.ipynb#INSIGHTS:)

* The majority of customers are having their age in the range of 26 to 35,
* Then the customers having age between 18 to 25, 36 to 45, 46 to 55 are the second highest.
* Lastly, there are least number of customers having age more than 55.

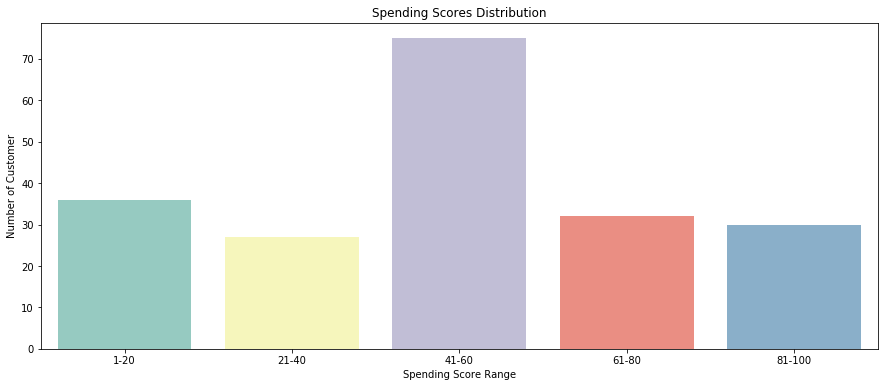
### Number of Customers according to their Annual Income (k$)



### INSIGHTS:[¶](http://localhost:8889/notebooks/Customer%20Segmentation.ipynb#INSIGHTS:)

* Most of the customers earn an amount in the range of 60k to 90k dollars.
* Then there are customers earning an amount between 30k to 60k dollars and the ones who earn less than 30k dollars.
* At last there are very few people who have huge income which is from 90k to 120k dollars and 120k to 150k dollars.

### Number of Customers according to their Spending Scores (1-100)

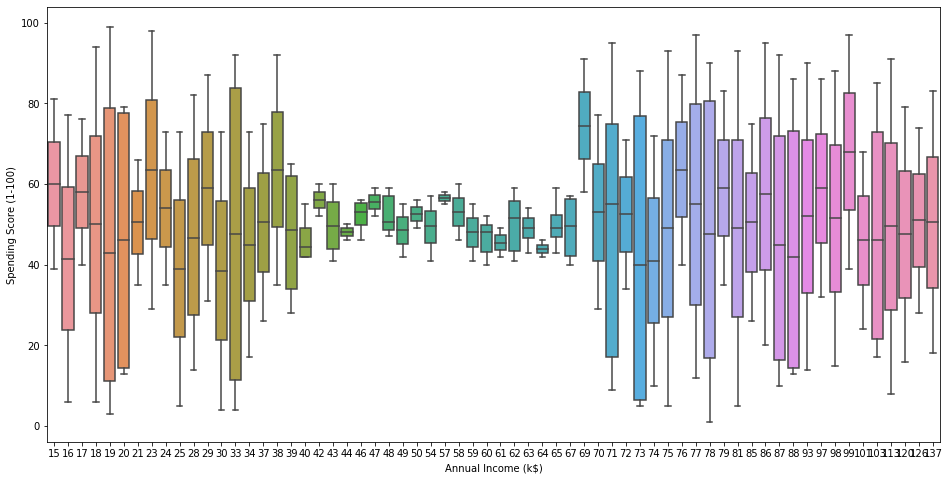


### INSIGHTS:[¶](http://localhost:8889/notebooks/Customer%20Segmentation.ipynb#INSIGHTS:)

* Most of the customers are having their Spending Scores in the range of 41 to 60.
* Then there are customers with Spending Scores in the range of 1 to 20, 61 to 80, 81 to 100 and lastly 21 to 40.

### Annual Income (k$) v/s Spending Scores(1-100)

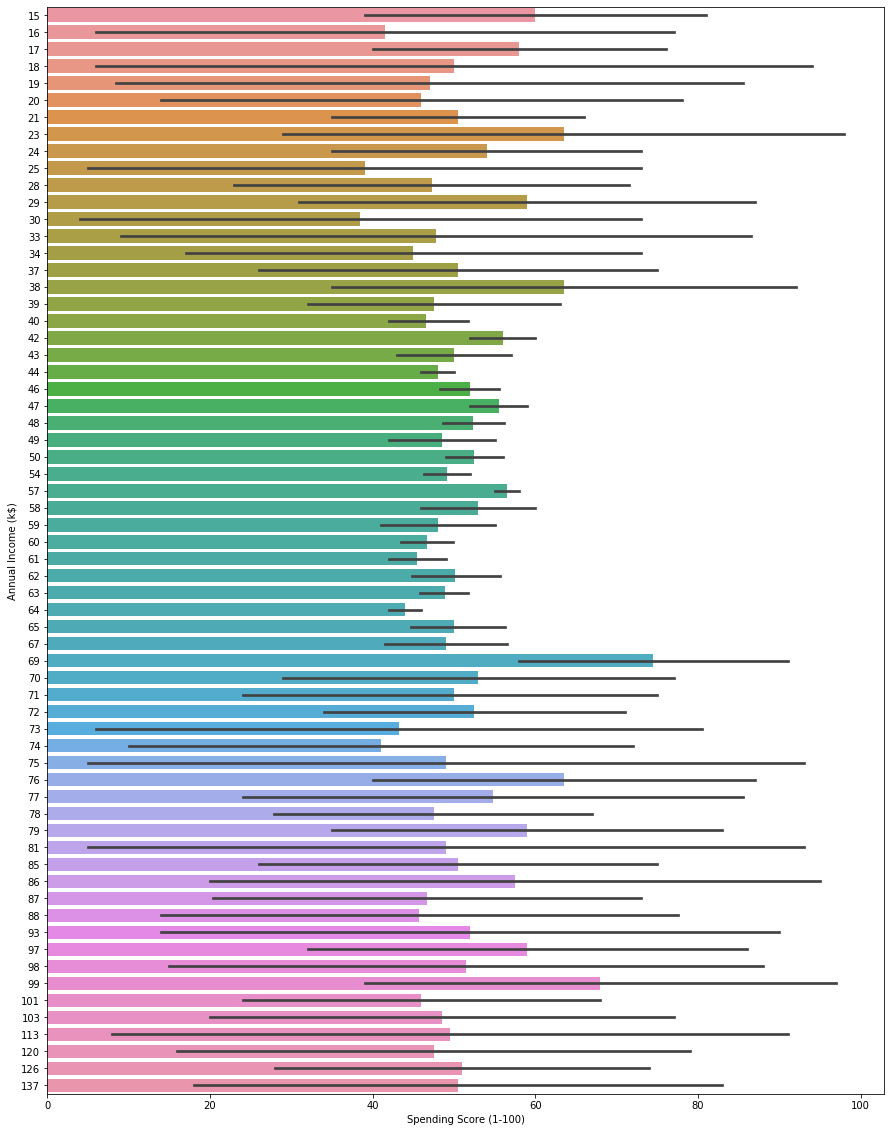
* The following boxplot will annotate the distribution of Annual Incomes of customers with respect to their Spending Scores.



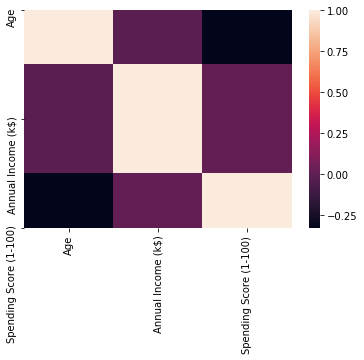
### NOTE:

* The graph on next page visualizes the customers with respect to their Annual Incomes and their Spending Scores.

### Annual Income (k$) v/s Spending Scores(1-100)



### Heatmap to find the Correlation between various Attributes of the Dataset



### INSIGHTS:

* We can clearly see that there is no correlation(~ 0.00) between any of the Attributes in our Dataset.

1. **K – Means Clustering and Principle Component Analysis (PCA)**

### Label Encoding the Gender values

This is what our dataset looks after label encoding the gender column:

| **Gender** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
| --- | --- | --- | --- |
| 0 | 1 | 19 | 15 | 39 |
| 1 | 1 | 21 | 15 | 81 |
| 2 | 0 | 20 | 16 | 6 |
| 3 | 0 | 23 | 16 | 77 |
| 4 | 0 | 31 | 17 | 40 |
| ... | ... | ... | ... | ... |
| 195 | 0 | 35 | 120 | 79 |
| 196 | 0 | 45 | 126 | 28 |
| 197 | 1 | 32 | 126 | 74 |
| 198 | 1 | 32 | 137 | 18 |
| 199 | 1 | 30 | 137 | 83 |

200 rows × 4 columns

### INSIGHTS:

* So the Gender values are now encoded, i.e. Female corresponds to '0' and Male corresponds to '1'.

### Feature Scaling on our Dataset

| **Gender** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
| --- | --- | --- | --- |
| 0 | 1.128152 | -1.424569 | -1.738999 | -0.434801 |
| 1 | 1.128152 | -1.281035 | -1.738999 | 1.195704 |
| 2 | -0.886405 | -1.352802 | -1.700830 | -1.715913 |
| 3 | -0.886405 | -1.137502 | -1.700830 | 1.040418 |
| 4 | -0.886405 | -0.563369 | -1.662660 | -0.395980 |
| ... | ... | ... | ... | ... |
| 195 | -0.886405 | -0.276302 | 2.268791 | 1.118061 |
| 196 | -0.886405 | 0.441365 | 2.497807 | -0.861839 |
| 197 | 1.128152 | -0.491602 | 2.497807 | 0.923953 |
| 198 | 1.128152 | -0.491602 | 2.917671 | -1.250054 |
| 199 | 1.128152 | -0.635135 | 2.917671 | 1.273347 |

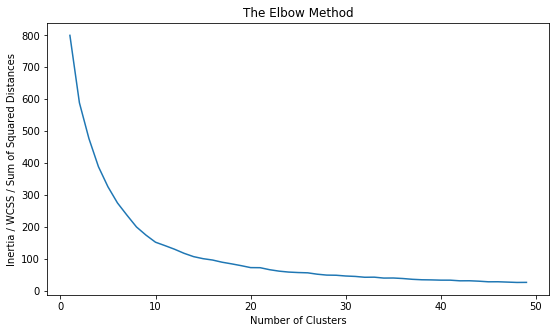
This is what our dataset looks after applying feature scaling:

200 rows × 4 columns

### Clustering using K-Means Clustering

Let's first test out some values for number of clusters and then we'll decide the actual number of clusters using the Elbow Method.

* **The Elbow Method**

****

### INSIGHTS:

* Here we can see that there is a slight bend (the elbow) when the K value i.e. the Number of Clusters are equal to 10 and hence we'll choose 10 as our Number of Clusters.
* Now let's build our actual model considering the number of clusters as 10.

### Principle Component Analysis (PCA)

We are using PCA just for the sake of simplicity because we will not be able to visualize our analysis in 4 dimensions as we have 4 different attributes, hence we'll perform dimensionality reduction and bring it down to 2 dimensions.

After applying PCA this is our dataset with new features PC1 (Principle Component 1) and PC2 (Principle Component 2).

| **PC1** | **PC2** |
| --- | --- |
| 0 | -0.406383 | -0.520714 |
| 1 | -1.427673 | -0.367310 |
| 2 | 0.050761 | -1.894068 |
| 3 | -1.694513 | -1.631908 |
| 4 | -0.313108 | -1.810483 |
| ... | ... | ... |
| 195 | -1.179572 | 1.324568 |
| 196 | 0.672751 | 1.221061 |
| 197 | -0.723719 | 2.765010 |
| 198 | 0.767096 | 2.861930 |
| 199 | -1.065015 | 3.137256 |

200 rows × 2 columns

### INSIGHTS:

* We now have the same 200 samples from our dataset, just that previously we had 4 different features i.e. the Gender, Age, Annual Income and the Spending Score which have now been reduced to 2 new features i.e. the Principle Component 1(PC1) and the Principle Component 2(PC2).

Now before we do that let's just have an overview of what clusters are assigned to the samples in our dataset.

Following are the cluster that have been assigned to every customer entry in our dataset.

| **PC1** | **PC2** | **Cluster** |
| --- | --- | --- |
| 0 | -0.406383 | -0.520714 | 2 |
| 1 | -1.427673 | -0.367310 | 2 |
| 2 | 0.050761 | -1.894068 | 7 |
| 3 | -1.694513 | -1.631908 | 8 |
| 4 | -0.313108 | -1.810483 | 7 |
| ... | ... | ... | ... |
| 195 | -1.179572 | 1.324568 | 1 |
| 196 | 0.672751 | 1.221061 | 6 |
| 197 | -0.723719 | 2.765010 | 9 |
| 198 | 0.767096 | 2.861930 | 5 |
| 199 | -1.065015 | 3.137256 | 9 |

200 rows × 3 columns

* Here is the array representing the cluster assignment to each of the 200 customers.

array([2, 2, 7, 8, 7, 8, 7, 8, 0, 8, 0, 8, 7, 8, 7, 2, 7, 2, 0, 8, 2, 2,

7, 2, 7, 2, 7, 2, 7, 8, 0, 8, 0, 2, 7, 8, 7, 8, 7, 8, 3, 2, 0, 4,

7, 8, 3, 4, 4, 4, 3, 2, 4, 0, 3, 0, 3, 0, 4, 0, 0, 2, 3, 3, 0, 2,

3, 3, 2, 4, 0, 3, 3, 3, 0, 2, 3, 2, 4, 3, 0, 2, 0, 3, 4, 0, 3, 4,

4, 3, 3, 2, 0, 4, 4, 2, 3, 4, 0, 2, 4, 3, 0, 2, 0, 4, 3, 0, 0, 0,

0, 4, 4, 2, 4, 4, 3, 3, 3, 3, 2, 4, 4, 9, 4, 1, 5, 9, 0, 9, 5, 9,

4, 1, 5, 1, 6, 9, 5, 1, 6, 9, 4, 1, 5, 9, 5, 1, 6, 9, 5, 9, 6, 1,

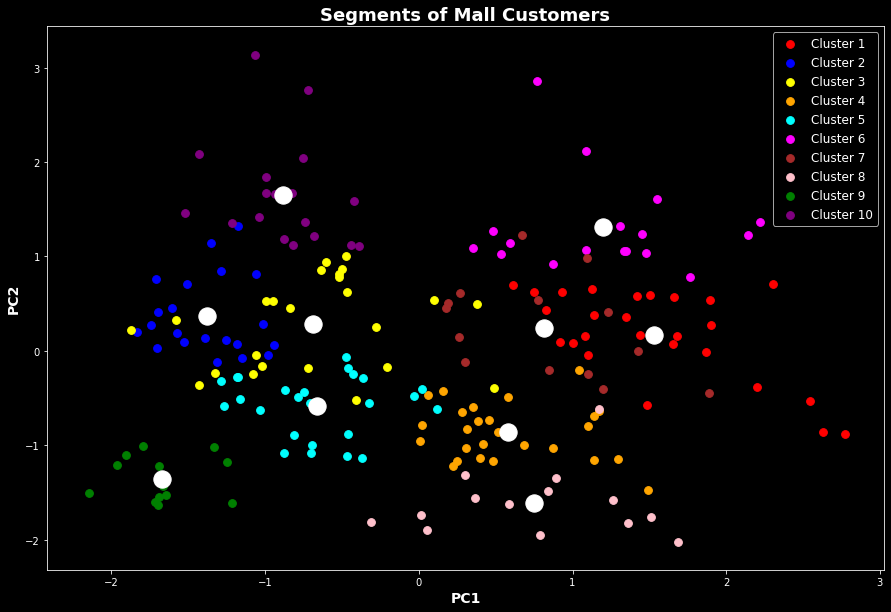
6, 1, 5, 1, 5, 1, 3, 1, 5, 1, 5, 1, 5, 1, 6, 9, 5, 9, 5, 9, 6, 1,

5, 9, 5, 9, 6, 1, 5, 1, 6, 9, 6, 9, 6, 1, 6, 1, 5, 1, 6, 1, 6, 9,

5, 9])

### Visualizing the Clusters

Following is s scatter plot which depicts the various clusters formed and their respective cluster centers.



### INSIGHTS:

* The above scatter plot has the PC1 (Principle Component 1) as it’s x axis and PC2 (Principle Component 2)as it’s Y axes.
* **Cluster 1 –** This cluster consist of customers with high PCA1 score and medium PCA2 score.
* **Cluster 2 –** This cluster represents customers having a medium PCA2 and a low PCA1.
* **Cluster 3 –** In this cluster, there are customers with a low PCA1 and a medium PCA2 score.
* **Cluster 4 –** This cluster comprises of customers with a medium PCA1 and a low PCA2.
* **Cluster 5 –** This comprises of customers with a low PCA2 as well as low PCA1 score.
* **Cluster 6 –** This comprises of customers with a high PCA2 and a high PCA1 score as well.
* **Cluster 7 –** This comprises of customers with a medium PCA2 and a medium PCA1 score.
* **Cluster 8 –** This comprises of customers with a very low PCA2 and a medium PCA1.
* **Cluster 9 –** This comprises of customers with a very low PCA2 as well as a very low PCA1 score.
* **Cluster 10 –** This comprises of customers with a high PCA2 and a low PCA1.

With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of customers, companies can release products and services that target customers based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.

**Chapter 5**

**Conclusion**

In this data science project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analyzed and visualized the data and then proceeded to implement our algorithm.

Hence, K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.