Data Science assignment

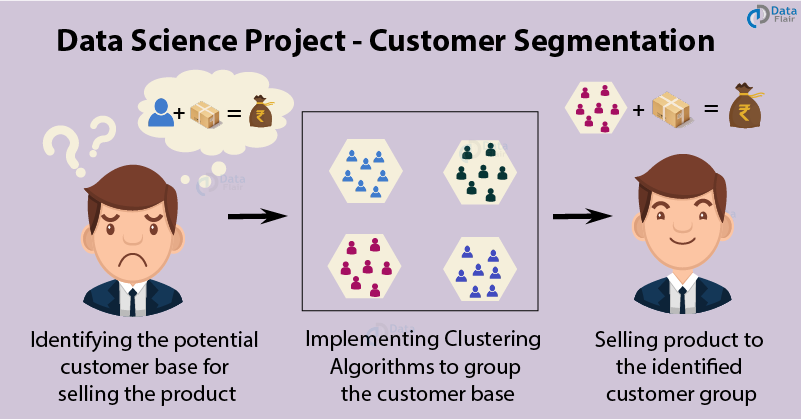
Name – Rohan Mathur

Reg.no – RA1811027010048

Sec- CSE – BD I2

Title - Customer Segmentation

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of [***K-means clustering***](https://data-flair.training/blogs/k-means-clustering-tutorial/) which is the essential algorithm for clustering unlabelled dataset.



1. Discovery

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

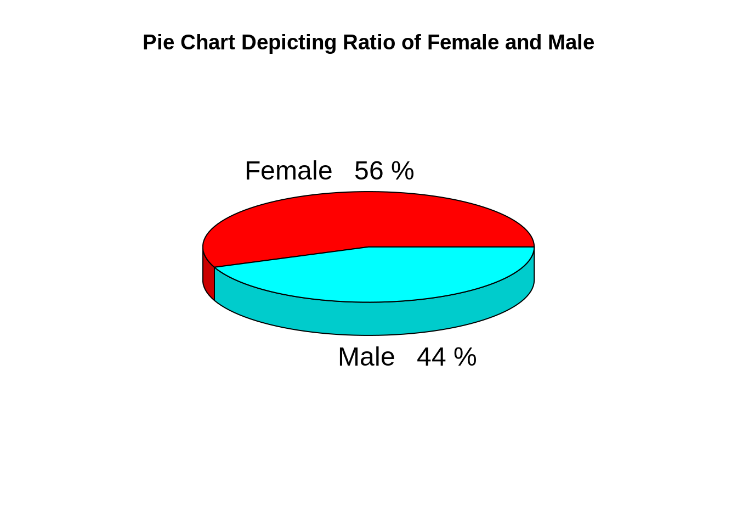
Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit.

2. Data Preparation -

This dataset has been collected as a part of survey from the various shops in a shopping mall. Now we will explore the dataset using plot visualizations.

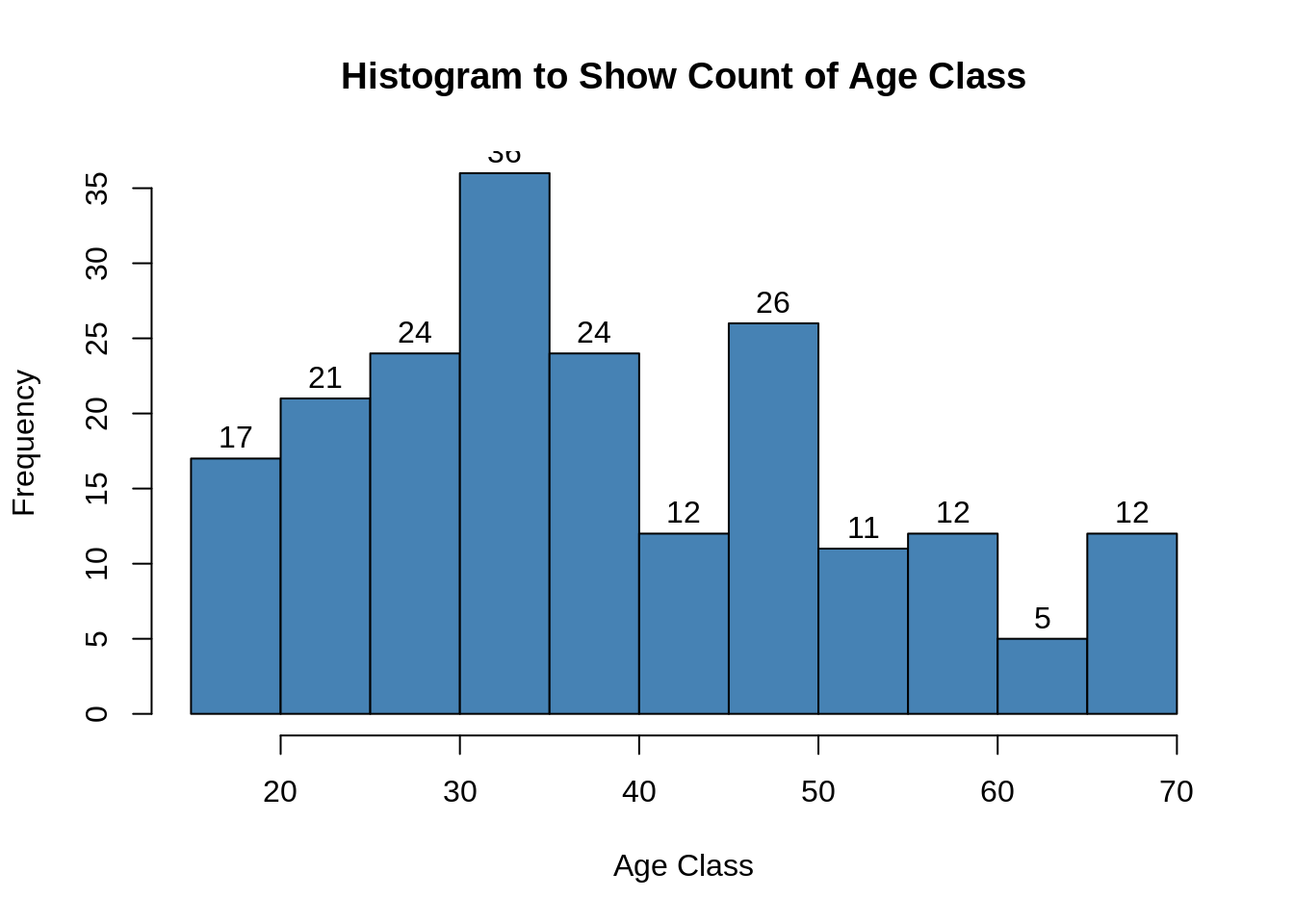
3. Model Planning – In Model planning we will visualize our dataset based on different parameters such as age, gender etc.

(i) Customer Gender Visualization – In this we visualize our customer dataset based on gender based on pie-chart.

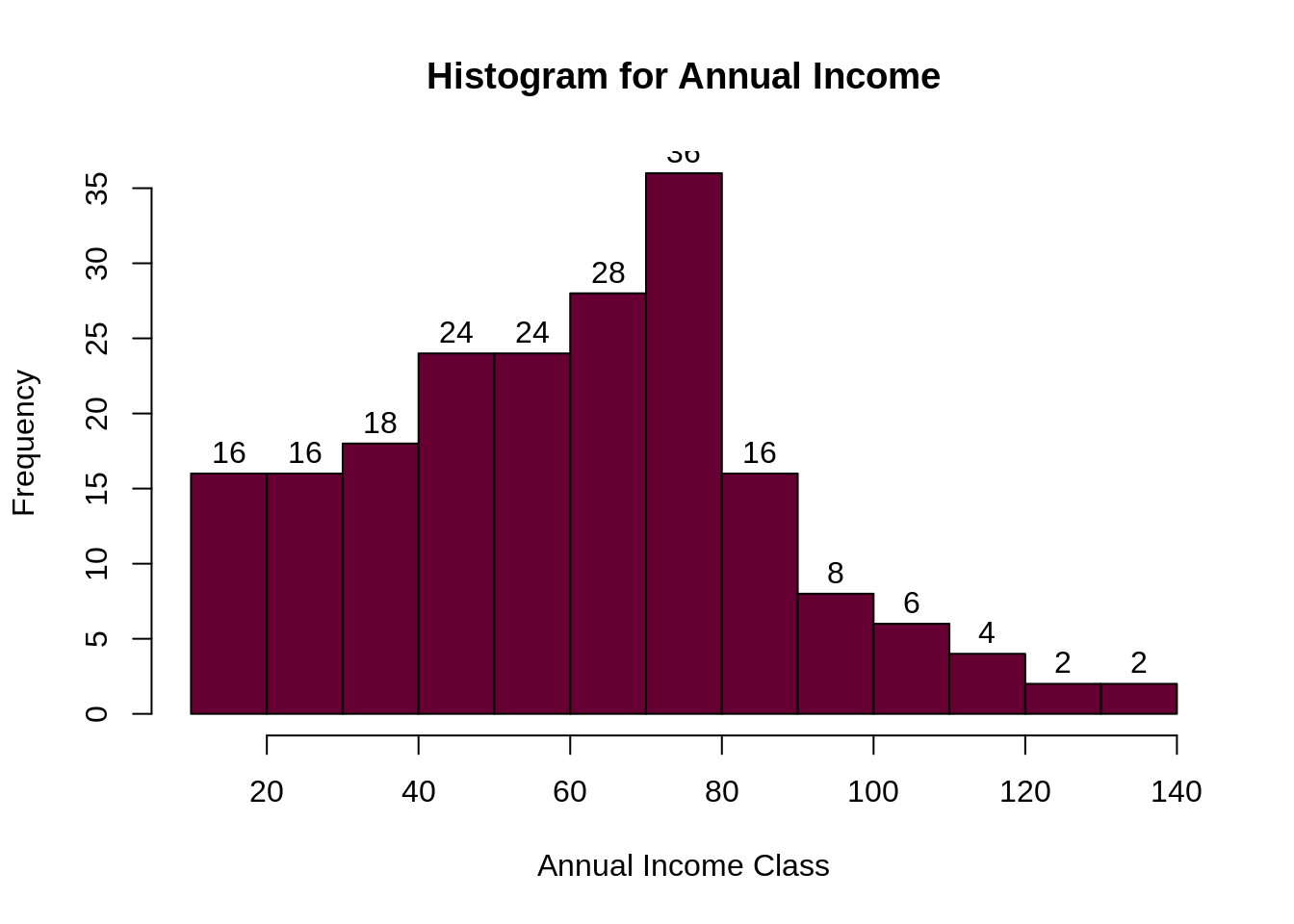


So we in this sample dataset we have 56% Females and 44% Males.

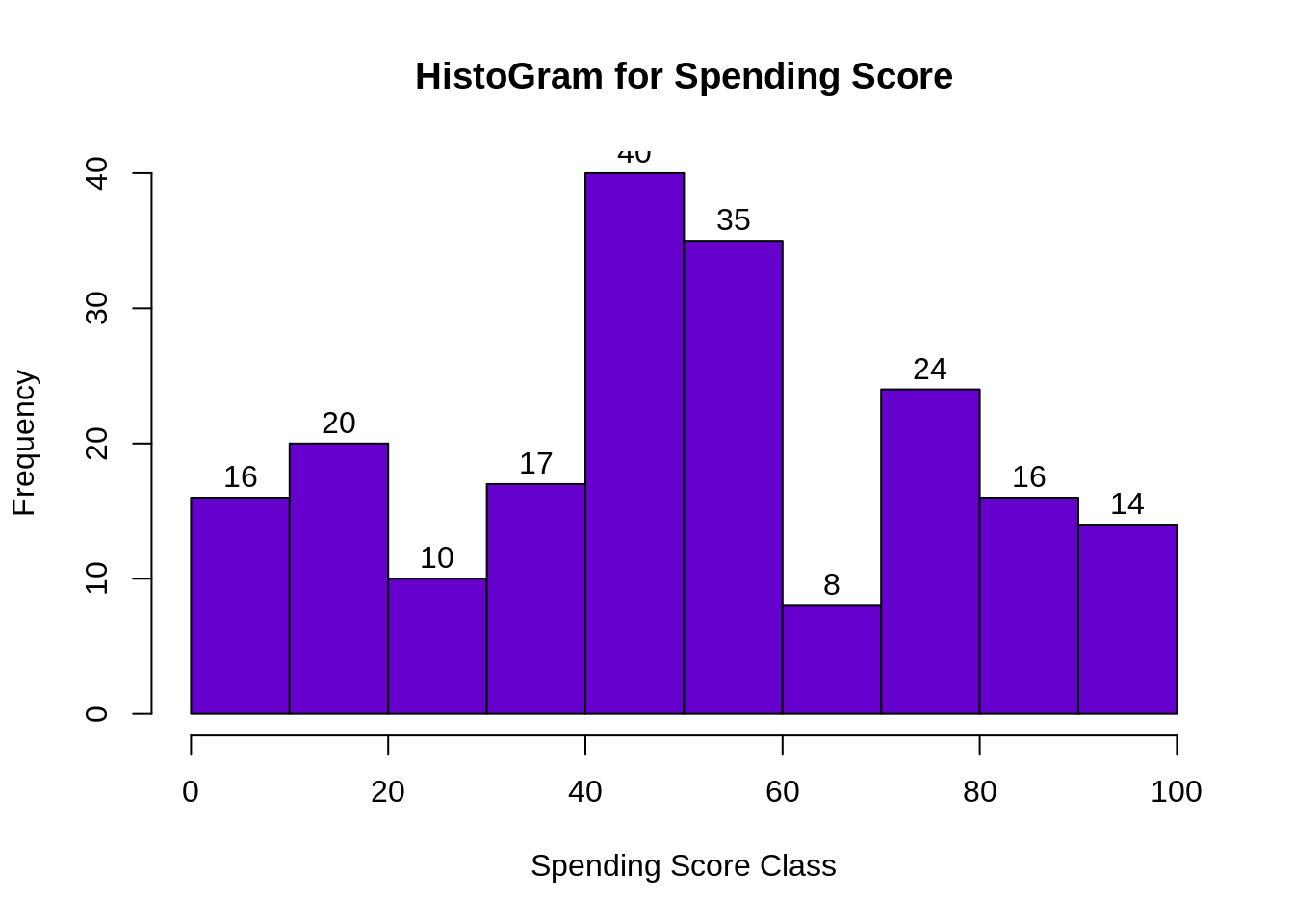
(ii) Age Visualization – Here we will plot the frequency of customer ages from sample dataset.



(iii) Customer Income Visualization – We will create visualizations to analyse the annual income of the customers. We will plot a histogram for this.



(iii) Analysing spending score – It is the score(out of 100) given to a customer by the mall authorities, based on the money spent and the behaviour of the customer.



4. Model Building – Now we build our model using K-means algorithm.

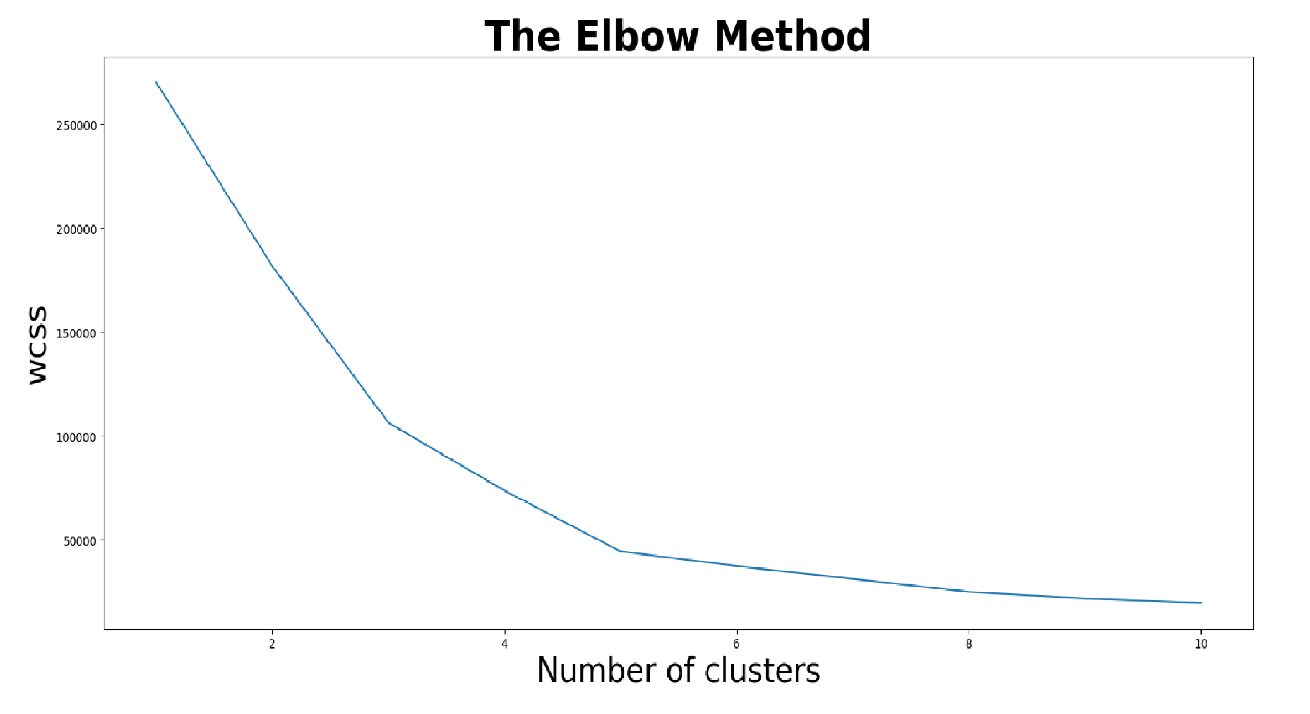
 K-means clustering –

* We specify the number of clusters that we need to create.
* The algorithm selects k objects at random from the dataset. This object is the initial cluster or mean.
* The closest centroid obtains the assignment of a new observation. We base this assignment on the Euclidean Distance between object and the centroid.
* k clusters in the data points update the centroid through calculation of the new mean values present in all the data points of the cluster. The kth cluster’s centroid has a length of p that contains means of all variables for observations in the k-th cluster. We denote the number of variables with p.
* Iterative minimization of the total within the sum of squares. Then through the iterative minimization of the total sum of the square, the assignment stop wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations.

**While working with clusters, you need to specify the number of clusters to use.** You would like to utilize the optimal number of clusters. To help you in determining the optimal clusters, there are three popular methods –

* Elbow method
* Silhouette method
* Gap statistic

5. Operations – Now supposing we take the **Elbow method** then, we calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. We then calculate the total intra-cluster sum of square . Then, we proceed to plot based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters.



6. Communicating the results – The data(clusters) are plotted on a **spending score Vs annual income curve.** Different clusters are plotted with different colours and the different cluster groups plotted are verified with the input data. The customers belonging to different clusters are classified and products are the marketed targeting different clusters.

