EAST WEST UNIVERSITY

Project Report

Project Title: Customer Segmentation

Course Title: Machine Learning
Course Code: CSE475
Section: 02
Semester: SPRING'21

Group ID: EWUSP2021CSE47503

SUBMITTED BY:

- **Md. Tanvir Hossain Joarddar (2018-1-60-242)**
- **Akib Adnan (2017-2-60-155)**
- **Ishrat Jaben Bushra (2018-1-60-099)**
- **Rafina Afreen (2018-1-60-119)**

1. Introduction:

Customer segmentation enables a company to customize its relationships with the customers, as we do in our daily lives. When we perform customer segmentation, we find similar characteristics in each customer's behavior and needs. Then, those are generalized into groups to satisfy demands with various strategies. Given this statement, we can conclude that we have to compare the existing customer data and the general population data in some way to deduce a relationship between them. A manual way of doing this is to compare the statistics between the customers and the general population. For example, the mean and standard deviation of age can be compared to determine which age group is more likely to be a customer or the salaries can be compared to see what group of people fall into customers, etc. But this analysis would give out many results which again have to be analyzed to come up with a final strategy. This process will require a lot of time, and by the time this analysis completes, the competitor in the market will capture most of the population, and the company will be out of business. Today with the advent of Machine Learning (ML) techniques used in every domain, this problem can also be addressed with the help of ML algorithms like using different clustering techniques.

1.1 Objectives:

Regarding this problem some objectives are:

- To divide customer in different groups
- Find our target customers with whom we can start marketing strategy
- Target marketing activities to specific groups
- Launch of features aligning with the customer demand
- Development of the product roadmap
- Understanding the relation between customer and their behavioral pattern

1.2 Motivation:

Let's imagine we have a supermarket mall and through membership cards, we have some basic data about our customers like Customer ID, age, income and spending score etc. various information which is something you assign to the customer based on your defined parameters like customer behavior and purchasing data. The main aim of this problem is learning the purpose of the customer segmentation concepts, also known as market basket analysis, trying to understand customers and separate them in different groups according to their preferences, and once the division is done, this information can be given to marketing team so they can plan the strategy accordingly.

1.3 Existing works:

- We have found customer segmentation problem solved using other machine learning technics like classification methods (random forest).

Link: https://www.kaggle.com/kaydenvan/randomforest-logisticregression-xgboost-catboost

- We have also found the Customer Segment using Naive Bayes.

Link: https://www.kaggle.com/microdegree/predict-the-customer-segment-using-naïve bayes

1.4 Necessity:

The general population and customer population have been compared and segmented using an Unsupervised learning algorithm. We were able to determine which clusters have more customers and which are potential clusters to have probable customers. The main intention of using clustering is trying to find out more customers who has higher possibility to do this.

2. Methodology:

We have used K-means and Agglomerative clustering in our project. Elbow Method is used for determining the optimal number of clusters and silhouette score is used for validation while clustering.

Elbow Method:

Elbow method is applied to calculate value of K for the dataset 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.

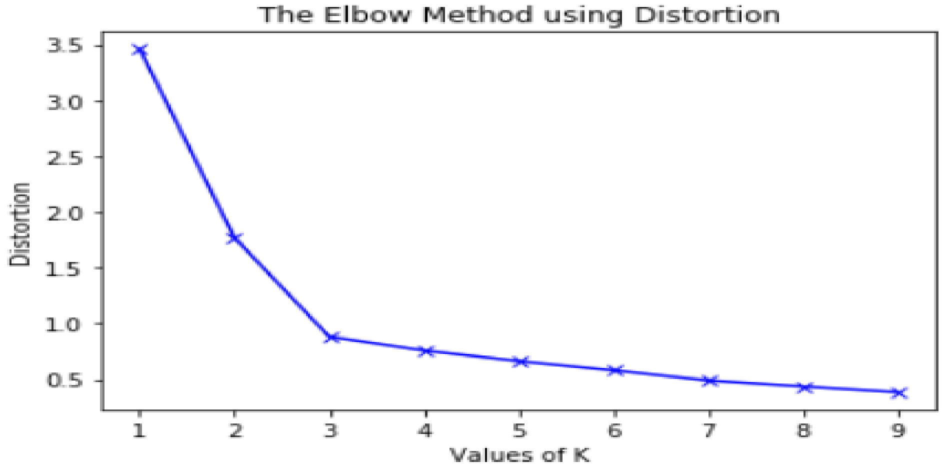
Formula:

$$\sum_{i=1}^{k} \sum_{Xj \in Si} \|Xj - \mu i\|^2$$

Xj = data point in Si cluster

μi = centroid of the cluster

After applying elbow method, a sample graph like below will show for our project.



The Elbow Method using Distortion

Silhouette Method:

With the help of the silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well within the cluster is the data object. If we obtain a high silhouette width, it means that we have good clustering. The silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the silhouette over significant values for k cluster.

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

- **1:** Means clusters are well apart from each other and clearly distinguished.
- **0:** Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- **-1:** Means clusters are assigned in the wrong way.

Silhouette Score = (b-a)/max(a,b)

Where,

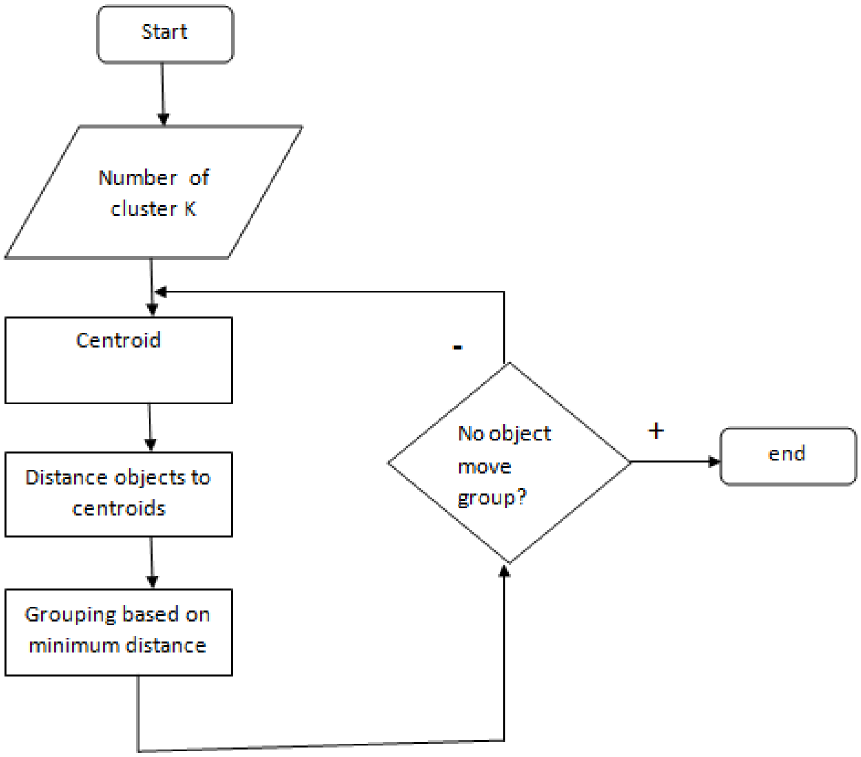
a = average intra-cluster distance i.e the average distance between each point within a cluster.

b = average inter-cluster distance i.e the average distance between all clusters.

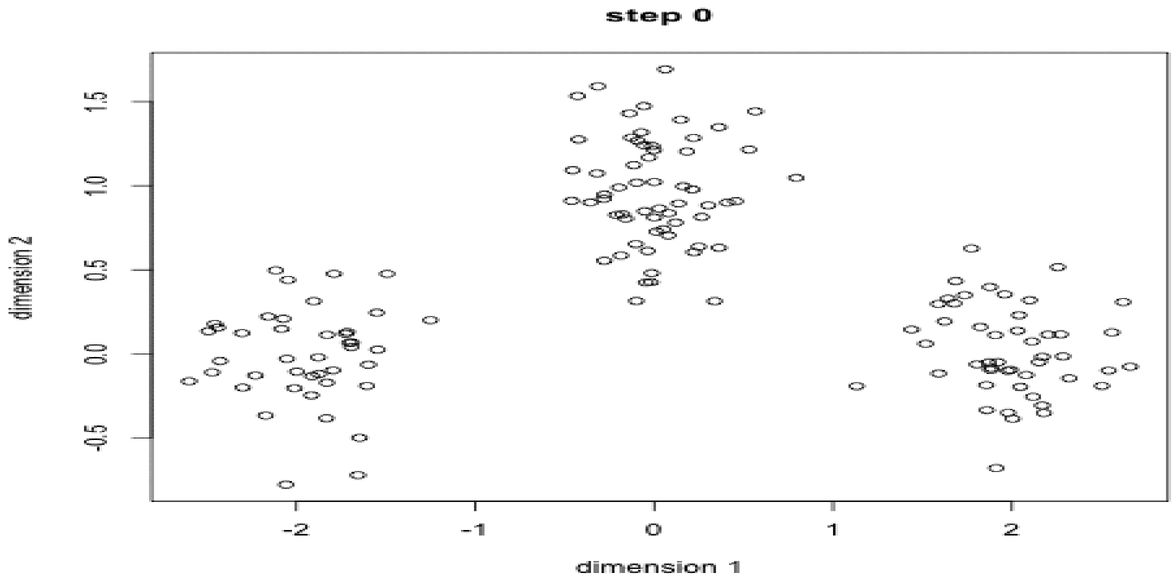
K-means Algorithm:

It is 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 (discussed in later section), 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.

Working Method:


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

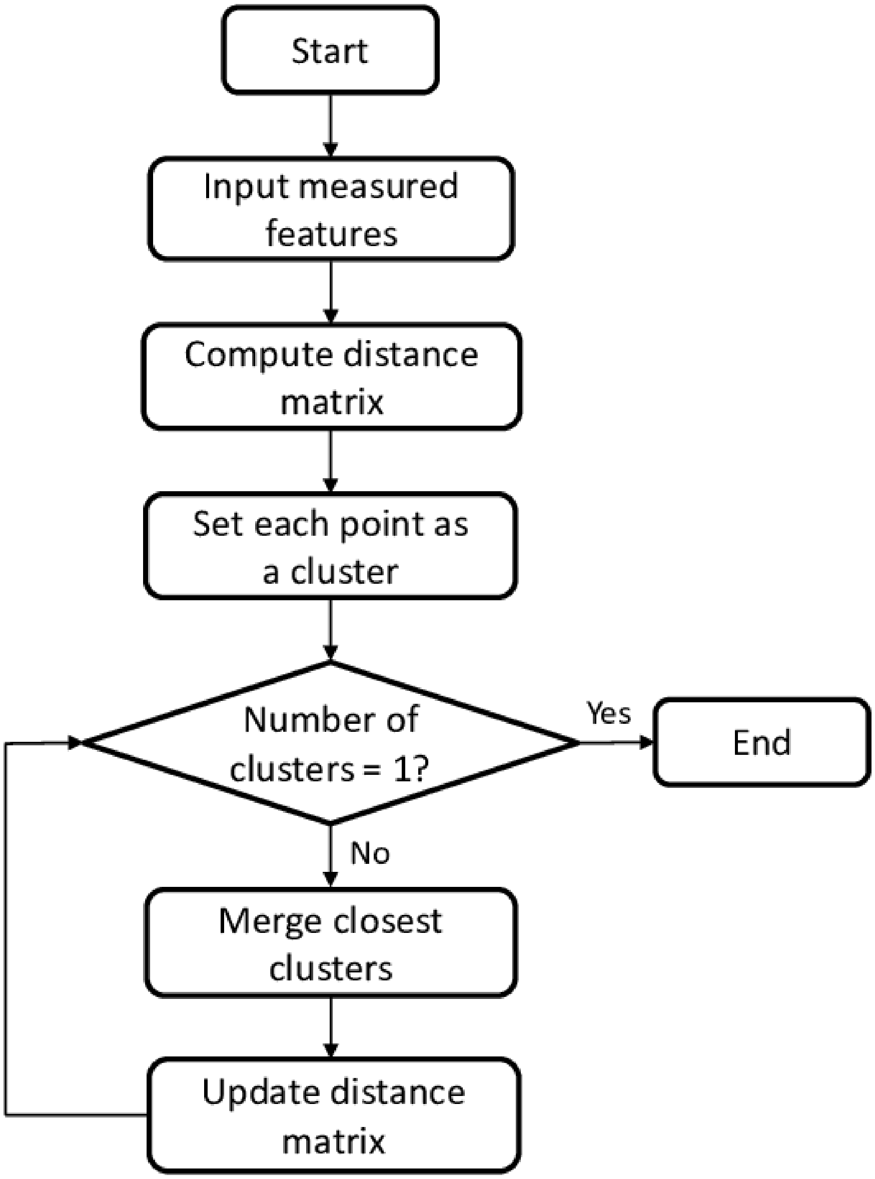
After clustering we get a result like the sample below:



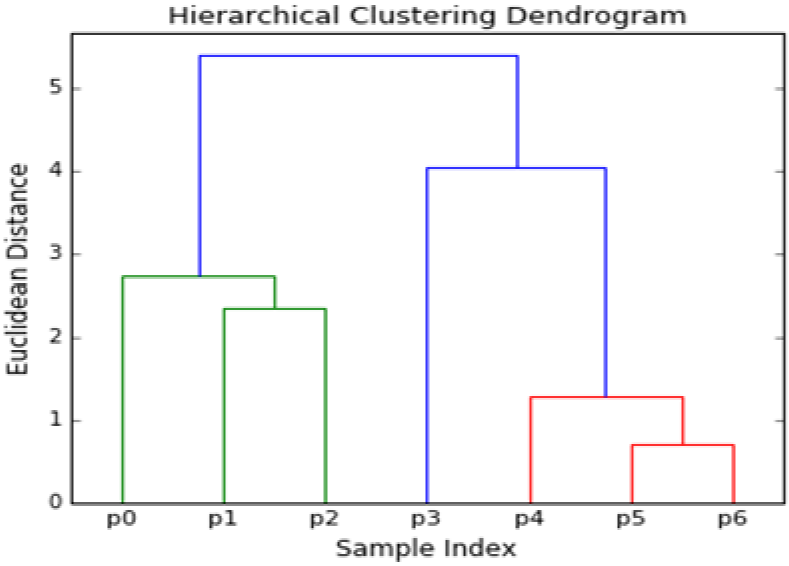
Agglomerative Clustering:

Agglomerative Clustering is based on forming a hierarchy represented by dendrograms. Dendrogram acts as memory for the algorithm to talk 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.

Working Method:



After clustering we get a result like the sample below:



3. Implementation:

At first, we have collected a dataset then we have processed and visualized the dataset, then we have performed elbow method and silhouette score finding technics to find out optimal number of clusters for kmeans algorithm. After clustering we have plotted some graph to make decision from the clusters. Then we have also made cluster using agglomerative clustering technic and plotted graph from it to check if the kmeans result and agglomerative result are same or not.

3.1 Data Collection:

Link: https://www.kaggle.com/somesh24/customer-segmentation

Data Information:

This dataset is composed by the following nine features:

- Customer id: Unique ID assigned to the customer
- Age: Age of the customer
- Edu: Customer education qualification
- Years Employed: The work experience of customer in years
- Income: Annual Income of the customer
- Defaulted: Determine whether the customer is defaulted or not
- Address: Address of the customer
- DebtIncomeRatio: Percentage of ratio between spending score and income
- Spending_score: Score assigned by the mall based on customer behavior and spending nature.

NB: We have created spending score column by adding Card Debt and Other Debt column.

In this particular dataset we have 850 samples to study.

3.2 Data Processing:

- we have checked and dropped null values
- dropped few unnecessary columns: (Customer id, Defaulted, Address, DebtIncomeRatio)
- handled categorized values that exits in dataset
- Standardizing data, so that all features have equal weight.

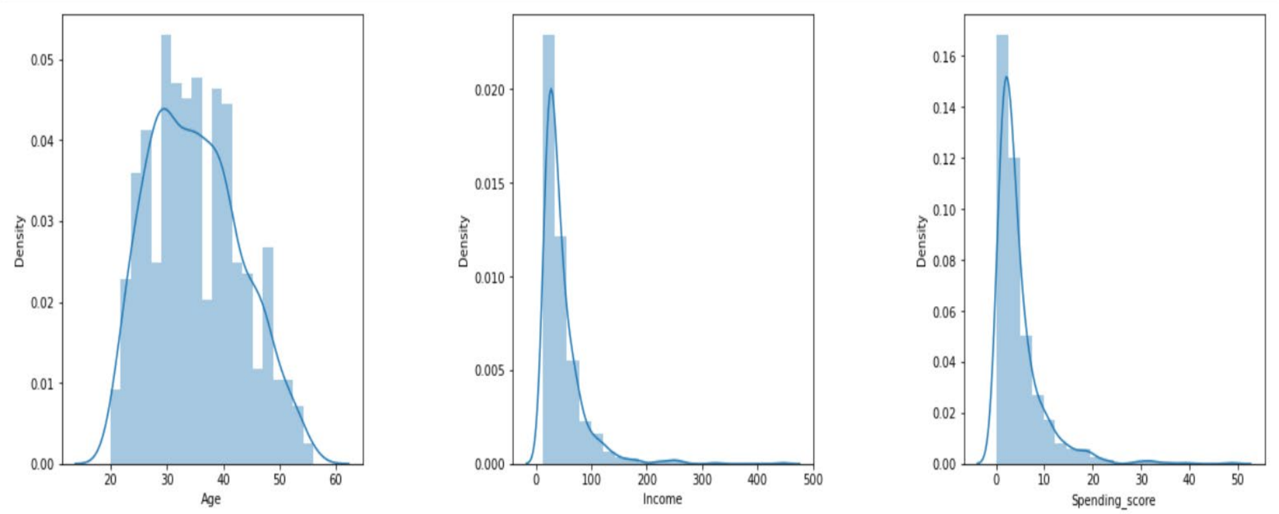
3.3 Model Development:

We have used Anaconda, Jupiter Notebook as our project environment, used Python language to develop our code.

3.4 Result:

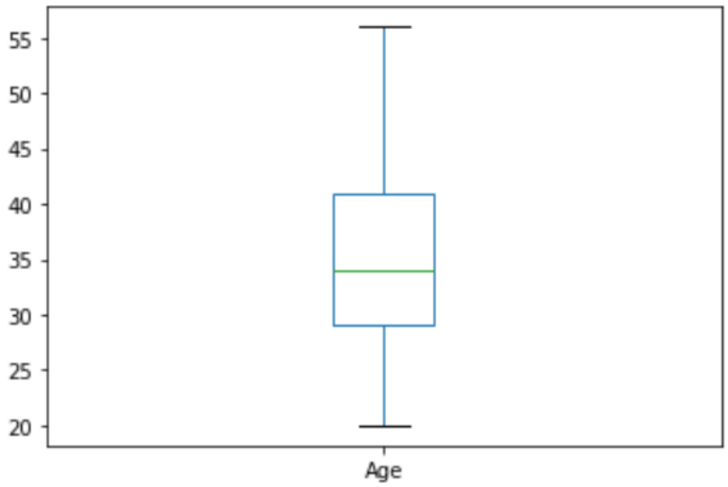
Data visualization: We have studied and analyzed our dataset by various graph and plots they are:

- We are plotting the histograms where we are only going to consider the following features: Age, Income and Spending_score.



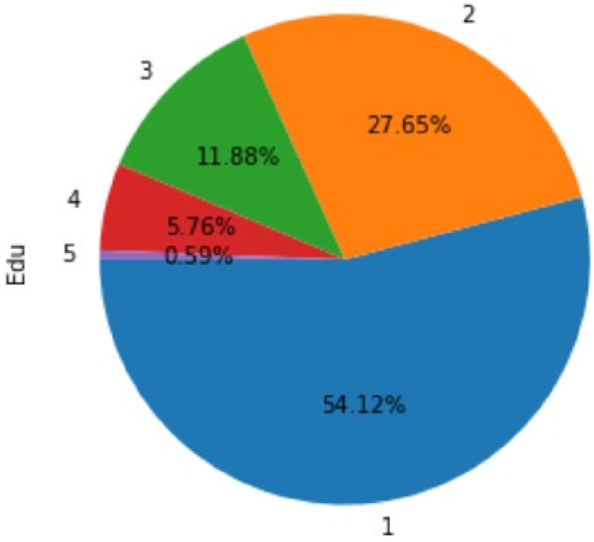
In these histograms we can observe that the Age distribution of these values resembles a Gaussian distribution, where the vast majority of the values lay in the middle with some exceptions in the extremes.

- We analyze the "Age" feature using Boxplot, it's a continuous variable.



From the above graphs, we can obviously conclude that most of the customers have an age between 29 and 41, also the minimum age of customers is 20, whereas the maximum age is 56.

- Now we will create a pie chart and a Count plot on "Edu" and "Years Employed" to show distribution across our customers dataset



From the above graph, we conclude that the percentage of

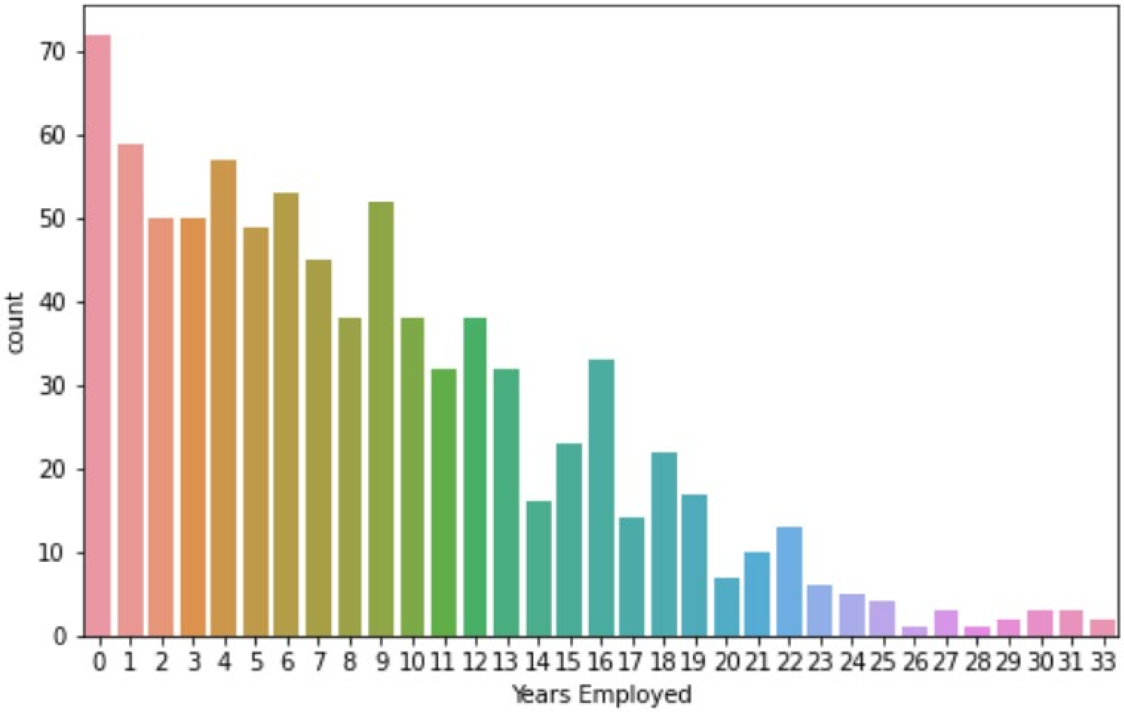
Edu 1 is 54.12%,

Edu 2 is 27.65%,

Edu 3 is 11.88%,

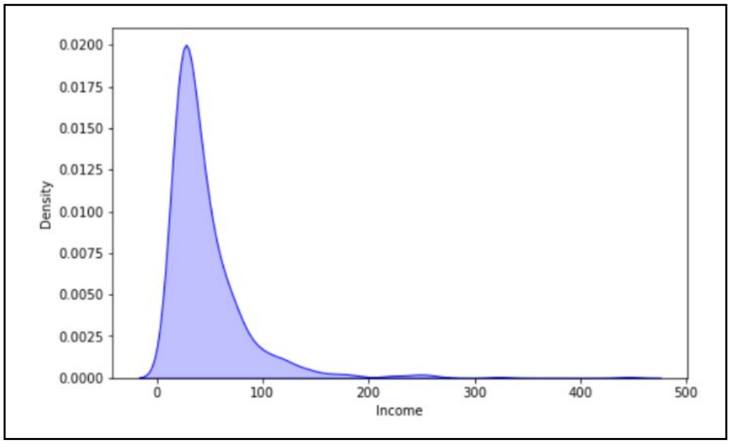
Edu 4 is 5.76%,

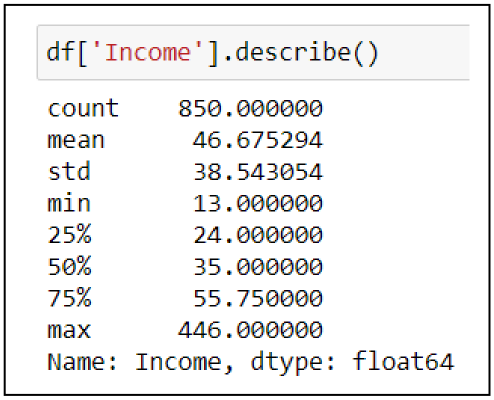
Edu 5 is 0.59%.



Here we see the number of each Year Employed of our customers.

- Now we will explore the "Income" and "Spending_score" feature using a histogram and a density plots and Boxplot to get insights from this feature

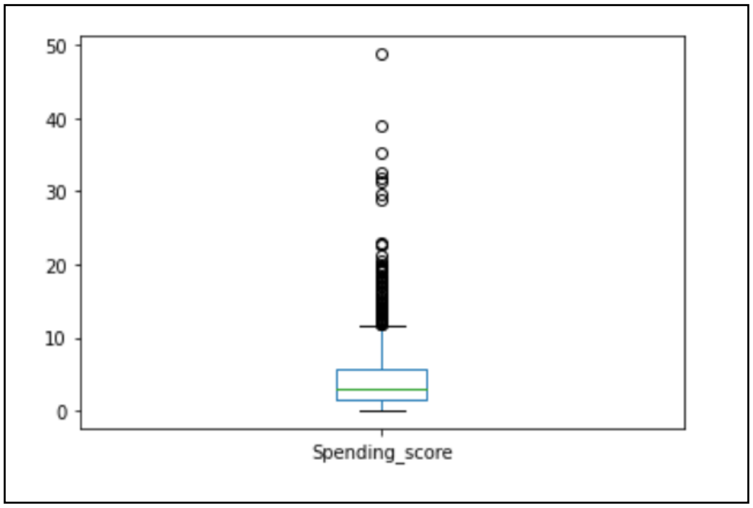


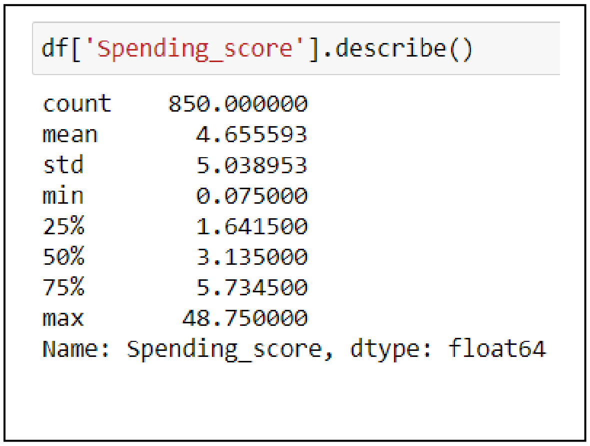


```
df['Income'].describe()
```

```
count    850.000000
mean      46.675294
std       38.543054
min       13.000000
25%       24.000000
50%       35.000000
75%       55.750000
max      446.000000
Name: Income, dtype: float64
```

From the above graphs, we can obviously see that the minimum annual income of the customers is 13 while the maximum income is 446. People earning an average income of 446 have the highest frequency count in our histogram distribution. The average income of all the customers is 46.67. In the Kernel Density Plot that we displayed above, we observe that the annual income has not a normal distribution.



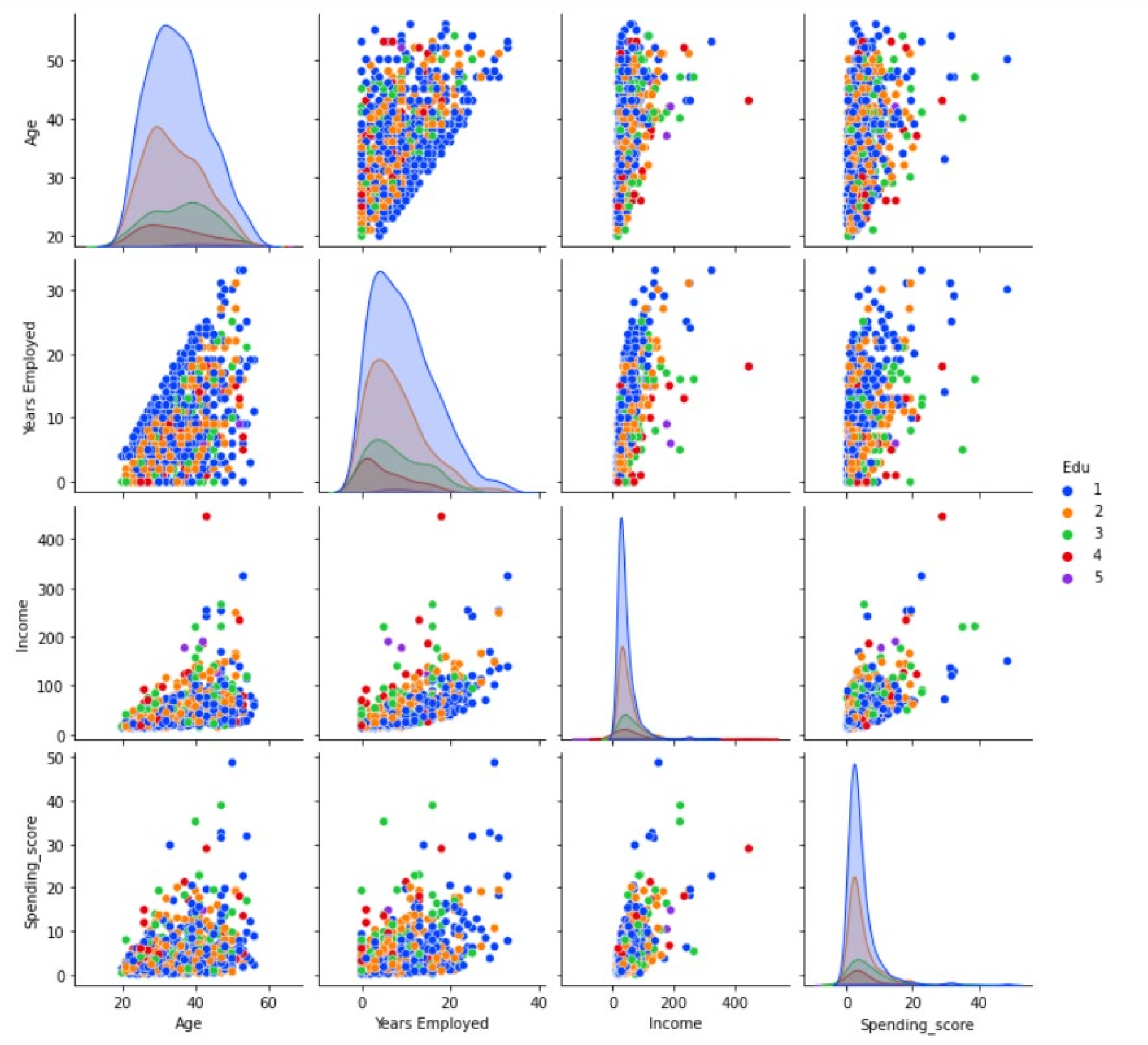


```
df['Spending_score'].describe()
```

```
count    850.000000
mean       4.655593
std        5.038953
min        0.075000
25%        1.641500
50%        3.135000
75%        5.734500
max       48.750000
Name: Spending_score, dtype: float64
```

The minimum spending score is 0.075 while the maximum is 48.75 and the average is 4.65. From the Box plot, we can conclude that class of customers having a spending score between 1 and 12 have the highest frequency among all classes.

- let's plot the relation between variables using "Edu" as a class distinction. In order to do so we are using the function pairplot and some parameters as well so we can visualize the "Edu" class separation better.

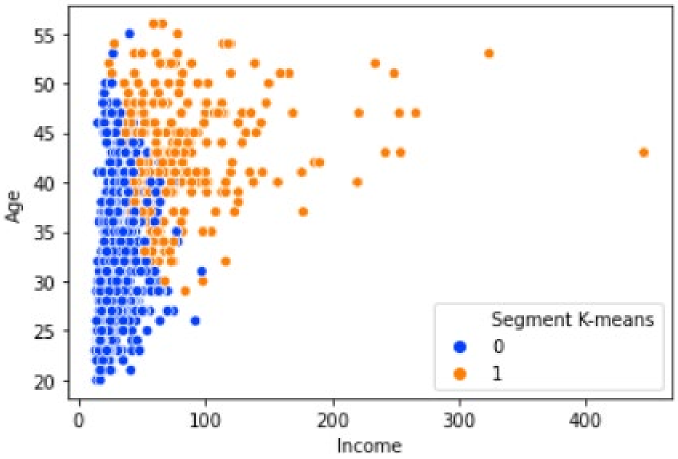


After analyzing our dataset and clustering we have made some questions from which we can say how clustering help customer segmentation

Decision after clustering:

We can understand the variables much better and take good decisions, 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.

1. **Which age group has highest income?**

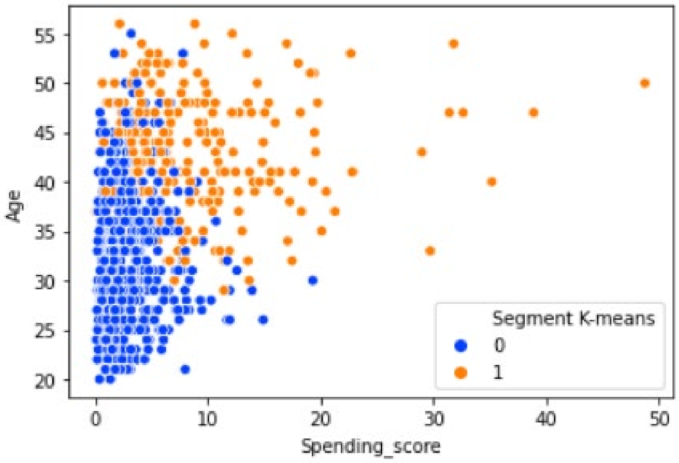


- In this graphic we can clearly see how people in their thirty threes, forties and fifties tend to earn more money annually than the ones younger than thirty threes or older than fifty years old. That is to say people whose age lays between thirty-three and fifty years old seem to get better jobs since they might be better prepared or be already more experienced than younglings or older people. In the graphic we can also see how cluster 0 people tend to earn a little bit more money than cluster 1 people, at least until fifty years old.

2. **Which age group comes into the mall most and their income?**

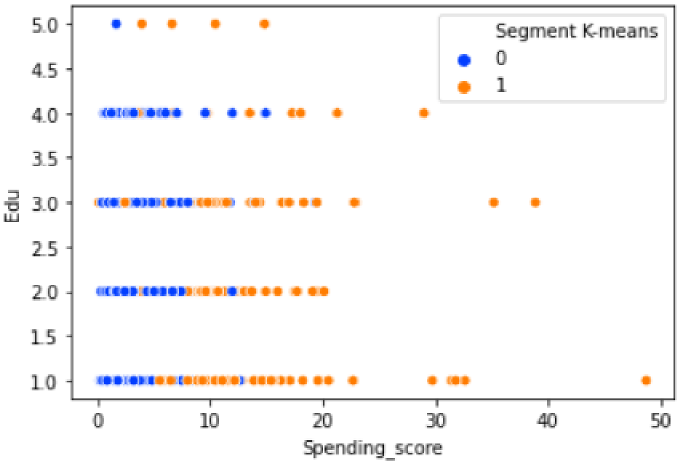
- We can see people age from 29 to 41 come to mall most and their Income is in between 10k to 80k.

3. **Which Age group comes into mall most and their spending score?**



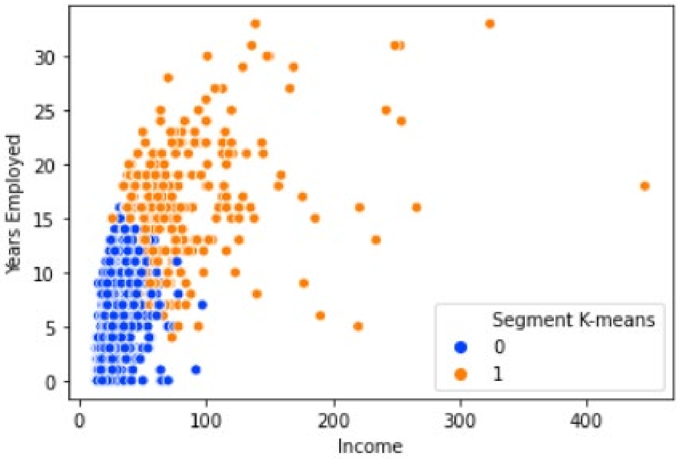
- we can see people age from 29 to 41 come to mall most and their Spending_score is in between 1k to 14k.

4. **Which Education groups have highest spending score?**



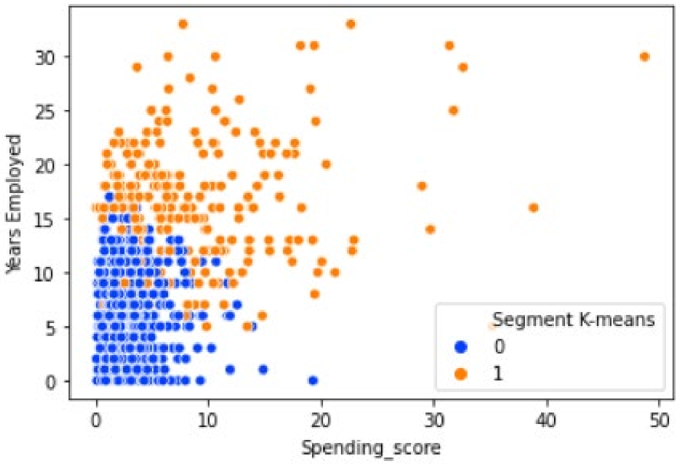
- We noticed that 54.12% people of our Mall are in Edu group-1 which have high spending score of 1 to 20 and most of them are from Cluster 0. Again about 27.65% people of our Mall are in Edu group-2 and their spending score of 1 to 15 and most of them are also from Cluster 0.

5. What years employed range have highest income?



- From the above graph we see that the person whose Years Employed in between 12 to 23 have higher income and most of them are from Cluster 0.

6. Relation between years employed and spending score.



- Here we can observe how a better Years Employed leads to having a higher spending score. With the increasing of Years Employed people have more spending score in the Mall.

From our cluster analysis,

Yellow Cluster:

- The yellow cluster groups middle-aged people with moderate to low annual income who actually spend a lot.
- As middle-aged people spend more money in malls, the main target when it comes to marketing, then we have to doing deeper studies about what they are interested in may lead to higher profits.

Blue Cluster:

- The blue cluster basically groups people of all ages whose salary isn't pretty high and their spending score is low.
- Although middle-aged seem to be the ones spending the most, we can't forget there are more people we have to consider, like people who belong to the blue cluster, they are what we would commonly name after "Younger" and it seems to be the biggest cluster.

Promoting discounts on some shops can be something of interest to those who don't actually spend a lot and they may end up spending more.

4. Conclusions:

Clustering is a powerful technique in order to achieve a decent customer segmentation. And it is a good way to understand the behavior of different customers and plan a good marketing strategy consequently. In order to cluster the customer, we have applied some unsupervised machine learning algorithms such as- K-MEANS and Agglomerative clustering method. By plotting some graphs using agglomerative clustering we see that the result of the clustering of customers is almost similar to what was done by K-Means clustering.

4.1 Challenges:

1. We had to use many kinds of graphs and plots at first to analyze our data and only selected a few plots because all plots were not good for analysis.
2. We have faced accuracy determining problem at first because it is unsupervised technique.
3. We did not find any dataset with sufficient quality.
4. We have faced a problem that is to convert categorical to numeric after dropping null values.

4.2 Limitations:

1. Our project cannot handle any dataset containing images.
2. We have to manually set which columns have to be dropped and converted from categorical to numerical.
3. We have to manually set which columns we have to choose for clustering.

4.3 Future decision:

- Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.
- This work can be extended to perform incremental way of clustering as new data is streamed in.
- Also, identifying appropriate number of clusters for a particular data set through some kind of evaluation instead of empirical analysis could also be a possible extension.