1. Introduction of the project:

Customer Segmentation is also known as **Market Basket Analysis**. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. This technique can be used by companies to outperform the competition by developing uniquely appealing products and services.

Customer segmentation is the action of breaking the customer base into groups depending on demographic, psychographic, etc. Using customer segmentation in marketing means that we can target the right people with the right messaging about our products. This will increase the success of our marketing campaigns.

➢ Why Customer Segmentation? :

The owner of mall/shopping complex wants to understand the customers like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly.

In supermarket malls through membership cards, we can have some basic data about our customers like customer-id, age, gender, annual income and spending score.
Spending Score is something we can assign to the customer based on our defined parameters like customer behaviour and purchasing data.

➢ Advantages of Customer Segmentation:

Customer segmentation is great for many different types of businesses who are looking to get smart about how they do marketing and sales activities. Some of these benefits include:

1. **Ability to Personalize Communication**: Personalizing marketing communication for customers leads to a better relationship between the customer and the business. This can greatly improve customer loyalty. Acknowledging the customer as more than another member of the email database can go a long way for the brand equity.
2. **Upselling / Cross-selling Opportunities**: Knowing the customer’s buying behavior or income level can help identify customer segments who can buy further products. Alternatively providing special offers to those who have just bought from us can increase sales of related products also.
3. **Higher ROI and CRO**: Sending out targeted campaigns to customers gets better ROI as they have already shown interest in buying from us. When we send out generic campaigns customers are less likely to engage with the content. This is because it is not relevant to them, or not of interest to them. Taking the time to segment the customers will greatly improve the CRO.
4. Determine appropriate product pricing.
5. Develop customized marketing campaigns.
6. Design an optimal distribution strategy.
7. Choose specific product features for deployment.
8. Prioritize new product development efforts.

- By using this solution we would be able to target the customers with whom we can start marketing strategy [easy to converse].
- **Segmenting** allows to more precisely reach a customer or prospect based on their specific needs and wants. **Customer Segmentation** will allow to: Better identify the most valuable customer segments. Improve the return on marketing investment by only targeting those likely to be the best customers.

2. Project Problem:

You are owing a market place and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. You 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.

Resolving problem with Customer Segmentation:

1- Determine appropriate product pricing.

2- Develop customized marketing campaigns.
3- Design an optimal distribution strategy.
4- Choose specific product features for deployment.
5- Prioritize new product development efforts.

Customer segmentation enables a company to customize its relationships with the customers, as we do in our daily lives.

When you perform customer segmentation, you find similar characteristics in each customer’s behaviour and needs. Then, those are generalized into groups to satisfy demands with various strategies. Moreover, those strategies can be an input of the

- **Targeted marketing activities to specific groups**
- **Launch of features aligning with the customer demand**
- **Development of the product roadmap**

It is crucial to keep in mind that machine learning can only be used to memorize patterns that are present in the training data, so we can only recognize what we have seen before. When using Machine Learning we are making the assumption that the future will behave like the past, and this isn’t always true.

3. Key Contributions:

Contribution of all team members are equal in all phases of Final Project development.

The following people have contributed to the Customer segmentation Machine Learning final project:-

Gaurav Srivastava:-

Contributed to all phases of the Final Project and lead to Model Deployment and Model implementation.

Sumit Sharma:-

The key contribution in final projects shown in Data pre-processing, Model implementation, and also in documentation.

Ujjval Bhardwaj:-

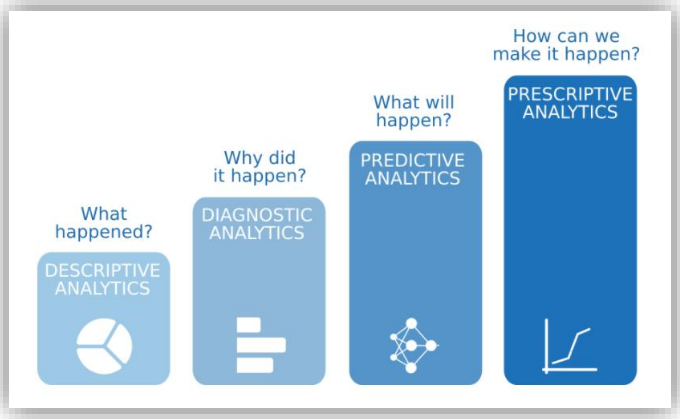
Start with research for all required phases and provided his contribution to Data pre-processing and model implementation.

4. Machine Learning:

Customer Segmentation with Machine Learning is the process of analysing customer features (such as age, education, interests, and spending habits) to select those customers who are more prone to a target product or service.

Machine learning can be used to memorize patterns, visualize cluster and give prediction of result to segment customer that are present in the training data.

Companies from many different industries have realized that Machine Learning can select potential clients much better than traditional methods, which allows the design of more effective marketing.



The volume, variety, and velocity of information stored in organizations are increasing significantly. The intelligent analysis of these data can be a differentiating factor for companies that adopt this technique.

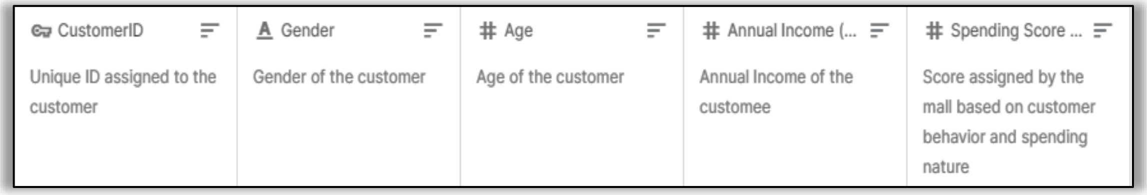
However, most companies only perform descriptive analyses of their data, which serve to know what has happened in the past.

These organizations can get much more value from their data using more sophisticated techniques. This process is known as Machine Learning.

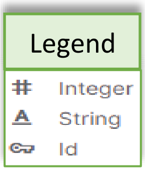
4.1. Dataset Recourse:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

Some basic data about customers like customer-id, age, gender, annual income and spending score can easily be captured through membership card and the information they provide in the visit feedback. These data are very critical to work for Customer Segmentation as once having an accurate model it is very helpful to take confidence of the unsatisfied customers.



🔑 CustomerID	A Gender	# Age	# Annual Income (...	# Spending Score ...
Unique ID assigned to the customer	Gender of the customer	Age of the customer	Annual Income of the customee	Score assigned by the mall based on customer behavior and spending nature



Legend	
#	Integer
A	String
🔑	Id

- ➢ Data column description:

  There are 5 columns that are considered important to Customer Segmentation:

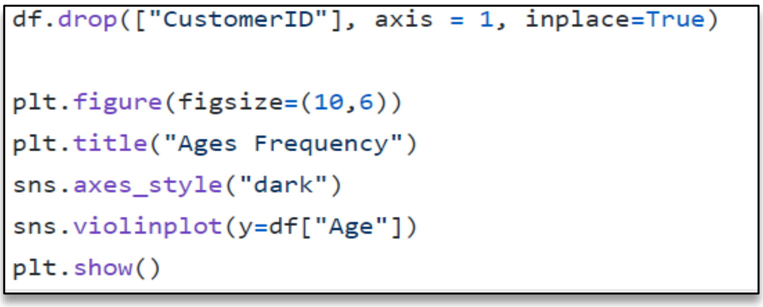
S. No.	Columns	Data type	Description
1.	Customer-id	ID (Integer)	Unique ID assigned to the customers
2.	Gender	String	Gender of the customer
3.	Age	Integer	Age of the customer
4.	Annual Income	Integer	Annual Income of the customer
5.	Spending Score	Integer	Score assigned by the mall based on customer behaviour and spending nature

4.2. Data Set Pre-processing:

Unsupervised Machine Learning for Customer Segmentation: we will train unsupervised machine learning algorithms to perform customer market segmentation. Market segmentation is crucial for marketers since it enables them to launch targeted ad marketing campaigns that are tailored to customer's specific needs.

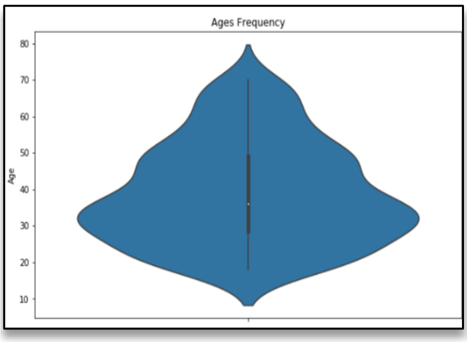
- **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.
- **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer’s city, state, or even country of residence.
- **Psychographics**, such as social class, lifestyle, and personality traits.
- **Behavioural data**, such as spending and consumption habits, product/service usage, and desired benefits.

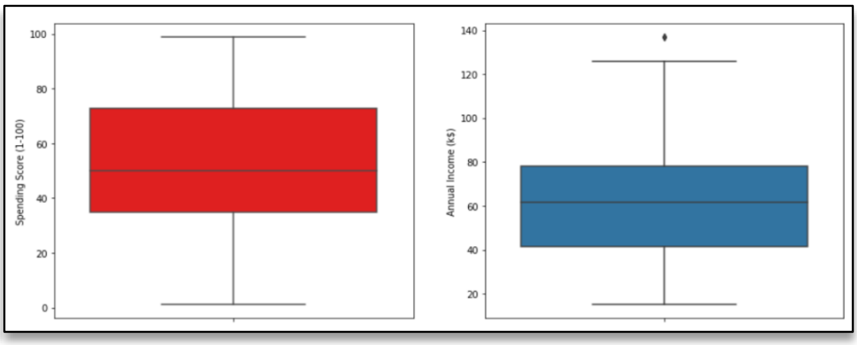
First we loading all the libraries and dependencies. The columns in the dataset are customer id, gender, age, income and spending score.



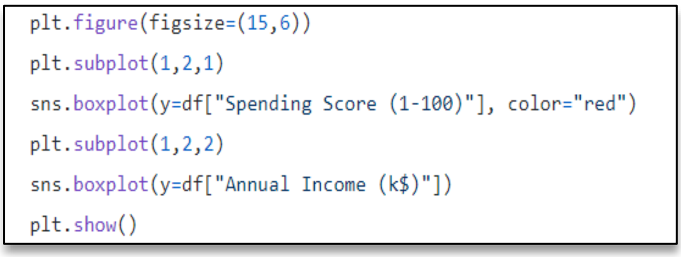
```
df.drop(["CustomerID"], axis = 1, inplace=True)

plt.figure(figsize=(10,6))
plt.title("Ages Frequency")
sns.axes_style("dark")
sns.violinplot(y=df["Age"])
plt.show()
```

Dropped the id column as that does not seem relevant to the context. Also we have plotted the age frequency of customers.



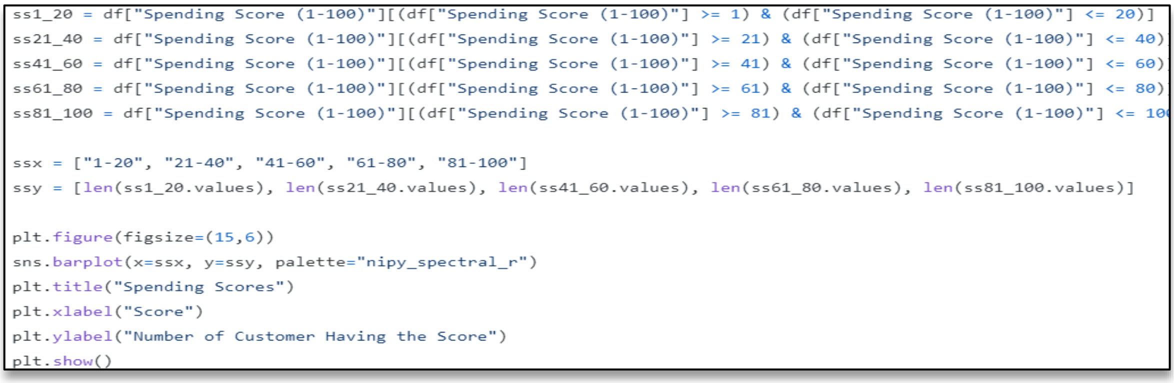


And box plot of spending score and annual income to better visualize the distribution range.

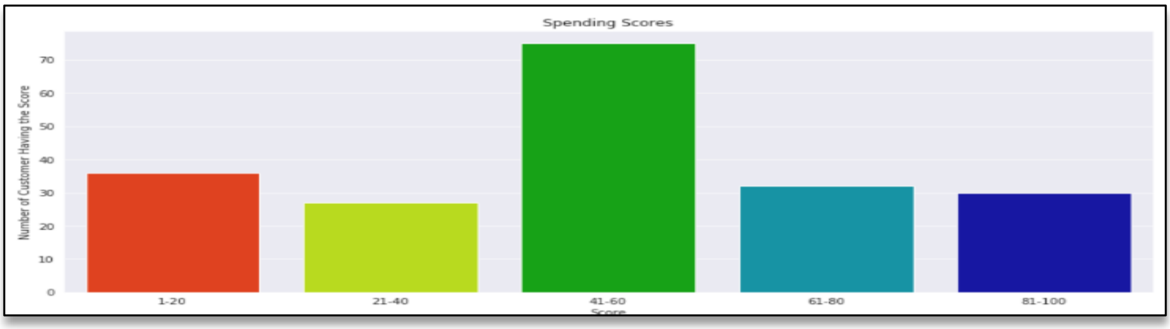


```
plt.figure(figsize=(15,6))
plt.subplot(1,2,1)
sns.boxplot(y=df["Spending Score (1-100)"], color="red")
plt.subplot(1,2,2)
sns.boxplot(y=df["Annual Income (k$)"])
plt.show()
```

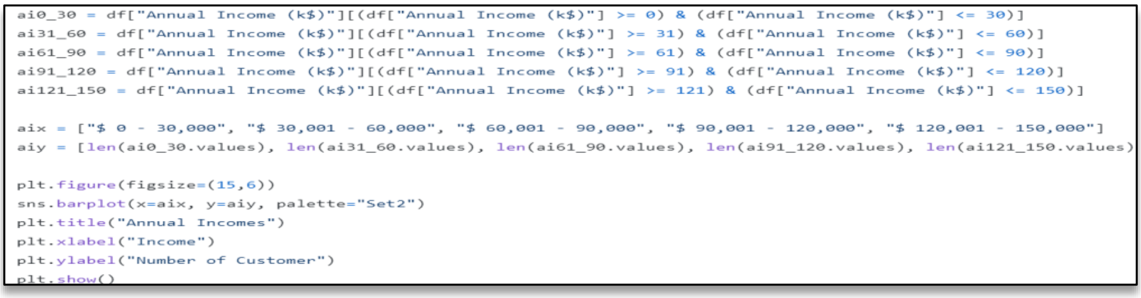
Continued with making a bar plot to visualize the number of customers according to their spending scores.



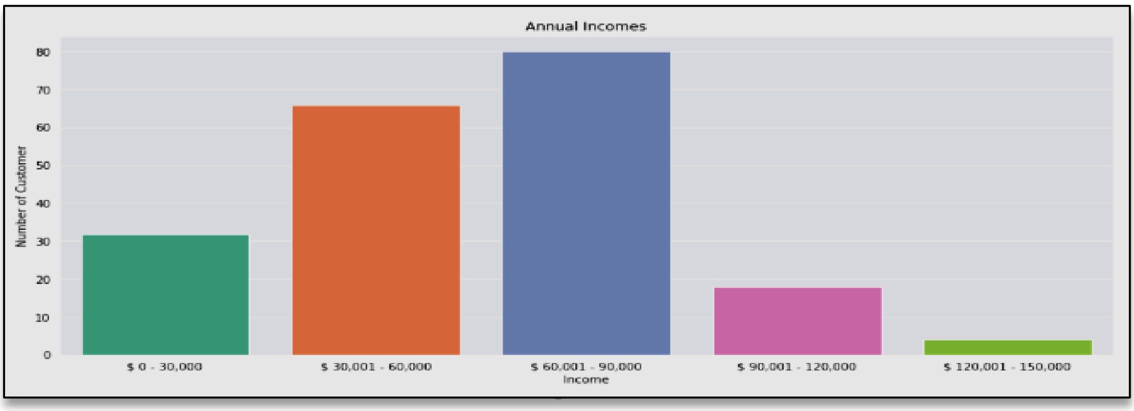
```
ss1_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) & (df["Spending Score (1-100)"] <= 20)]
ss21_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21) & (df["Spending Score (1-100)"] <= 40)
ss41_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41) & (df["Spending Score (1-100)"] <= 60)
ss61_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) & (df["Spending Score (1-100)"] <= 80)
ss81_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 81) & (df["Spending Score (1-100)"] <= 10

ssx = ["1-20", "21-40", "41-60", "61-80", "81-100"]
ssy = [len(ss1_20.values), len(ss21_40.values), len(ss41_60.values), len(ss61_80.values), len(ss81_100.values)]

plt.figure(figsize=(15,6))
sns.barplot(x=ssx, y=ssy, palette="nipy_spectral_r")
plt.title("Spending Scores")
plt.xlabel("Score")
plt.ylabel("Number of Customer Having the Score")
plt.show()
```



Made a bar plot to visualize the number of customers according to their annual income. The majority of the customers have annual income.



```
ai0_30 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 0) & (df["Annual Income (k$)"] <= 30)]
ai31_60 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 31) & (df["Annual Income (k$)"] <= 60)]
ai61_90 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 61) & (df["Annual Income (k$)"] <= 90)]
ai91_120 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 91) & (df["Annual Income (k$)"] <= 120)]
ai121_150 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 121) & (df["Annual Income (k$)"] <= 150)]

aix = ["$ 0 - 30,000", "$ 30,001 - 60,000", "$ 60,001 - 90,000", "$ 90,001 - 120,000", "$ 120,001 - 150,000"]
aiy = [len(ai0_30.values), len(ai31_60.values), len(ai61_90.values), len(ai91_120.values), len(ai121_150.values)

plt.figure(figsize=(15,6))
sns.barplot(x=aix, y=aiy, palette="Set2")
plt.title("Annual Incomes")
plt.xlabel("Income")
plt.ylabel("Number of Customer")
plt.show()
```



4.3. Finding the best Machine Learning Model for the Problem Objective:

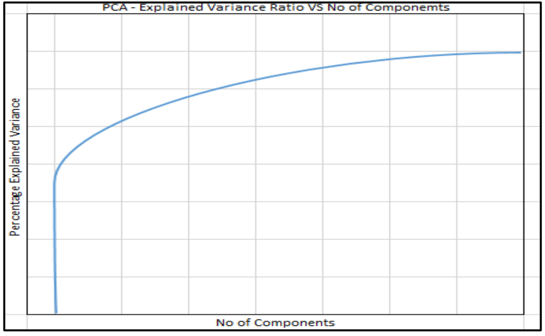
- Customer Segmentation using Unsupervised Machine Learning:

For cluster segmentation, there are two steps to be performed.

1. Dimensionality Reduction
2. Clustering

- **Dimensionality Reduction**:

Out of 50+ feature we are using most 5 important features for it, i.e. some features might be the same for all the people.

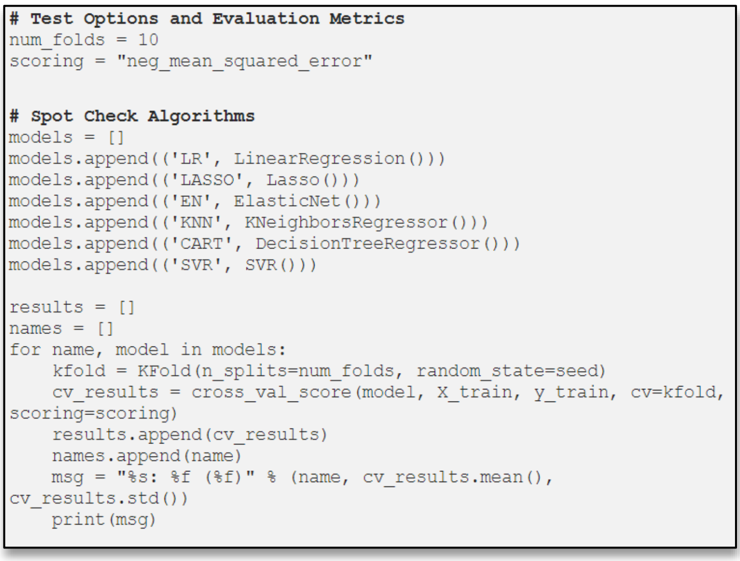


Principal Component Analysis (PCA) has been performed to analyse the explained variance of the PCA components. PCA applies a linear transformation on the data to form a new coordinate system such that the components in the new coordinate system represent the variation in the data.

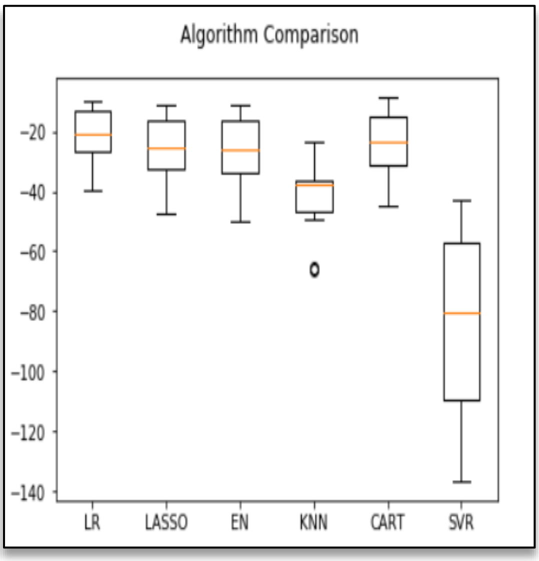
- **Clustering:**

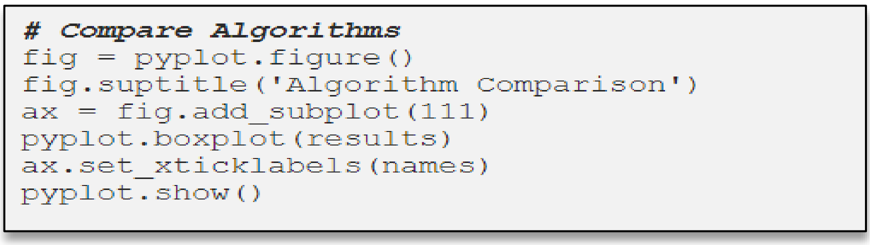
After the dimensionality reduction, the next step is to divide the general population and customer population into different segments. **K-Means clustering algorithm** has been chosen for this task. Since it is simple and is apt for this task since it measures the distance between two observations to assign a cluster. This algorithm will help us in separating the general population with the help of the reduced features into a specified number of clusters and use this cluster information to understand the similarities in the general population and customer data.

The number of clusters is a hyper parameter when working with clustering algorithms. The basic idea behind the clustering algorithms is to select the number of clusters to minimize the intra-cluster variation. Which means the points in one cluster are as close as possible to each other.



```
# Test Options and Evaluation Metrics
num_folds = 10
scoring = "neg_mean_squared_error"


# Spot Check Algorithms
models = []
models.append(('LR', LinearRegression()))
models.append(('LASSO', Lasso()))
models.append(('EN', ElasticNet()))
models.append(('KNN', KNeighborsRegressor()))
models.append(('CART', DecisionTreeRegressor()))
models.append(('SVR', SVR()))

results = []
names = []
for name, model in models:
    kfold = KFold(n_splits=num_folds, random_state=seed)
    cv_results = cross_val_score(model, X_train, y_train, cv=kfold,
scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(),
cv_results.std())
    print(msg)
```





```
# Compare Algorithms
fig = pyplot.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
pyplot.boxplot(results)
ax.set_xticklabels(names)
pyplot.show()
```

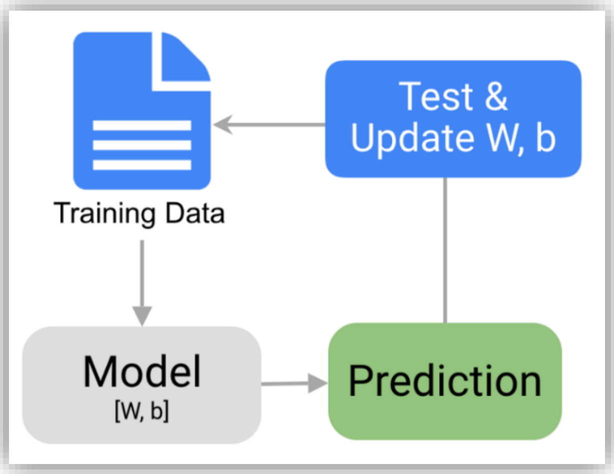
The Elbow Method—

Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

The optimal K value is found to be 5 using the elbow method.

4.4. Training and Testing the Model:

One of the most common methods for finding a good model is cross validation. In cross validation we will set:



- A number of folds in which we will split our data.
- A scoring method of **K-Means clustering algorithm.**
- Some appropriate algorithms that we want to check.

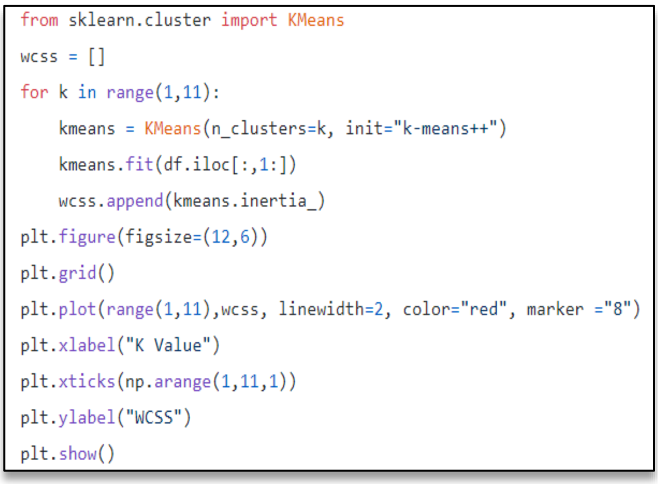
We'll pass our dataset to our cross validation score function and get the model that yielded the best score. That will be the one that we will optimize, tuning its hyper parameters accordingly.

Sum of Squares (WCSS)--

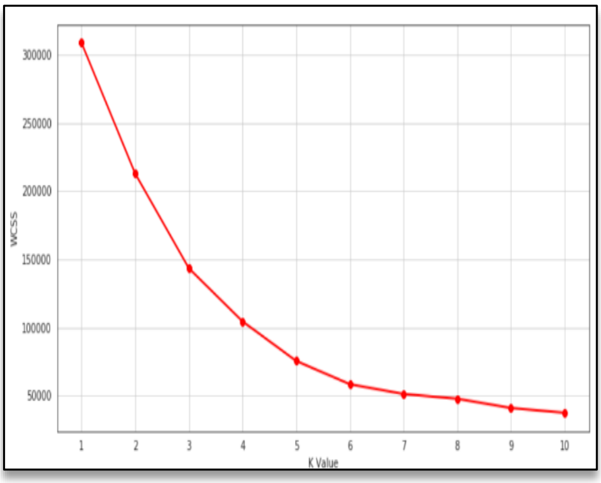
We are plotted Within Cluster Sum of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

Where *Yi* is centroid for observation *Xi*. The main goal is to maximize number of clusters and in limiting case each data



```
from sklearn.cluster import KMeans
wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(df.iloc[:,1:])
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show()
```



4.4. Model Evaluation:

Once the goal is clear, it should be decided how is going to be measured the progress towards achieving the goal. The most common evaluation protocols are:

Hold Out Validation Set:

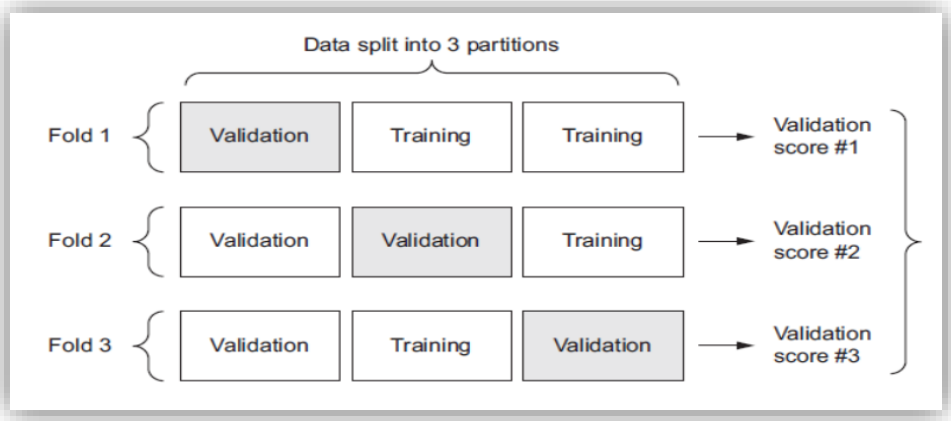
This method consists on setting apart some portion of the data as the test set.

The process would be to train the model with the remaining fraction of the data, tuning its parameters with the validation set and finally evaluating its performance on the test set.

K-Fold Validation

K-Fold consists in splitting the data into K partitions of equal size. For each partition i, the model is trained with the remaining K-1 partitions and it is evaluated on partition i.

The final score is the average of the K scored obtained. This technique is especially helpful when the performance of the model is significantly different from the train-test split.



Iterated K-Fold Validation with Shuffling

This technique is especially relevant when having little data available and it is needed to evaluate the model as precisely as possible.

It consist on applying K-Fold validation several times and shuffling the data every time before splitting it into K partitions. The Final score is the average of the scores obtained at the end of each run of K-Fold validation.

This method can be very computationally expensive, as the number of trained and evaluating models would be I x K times. Being I the number of iterations and K the number of partitions.

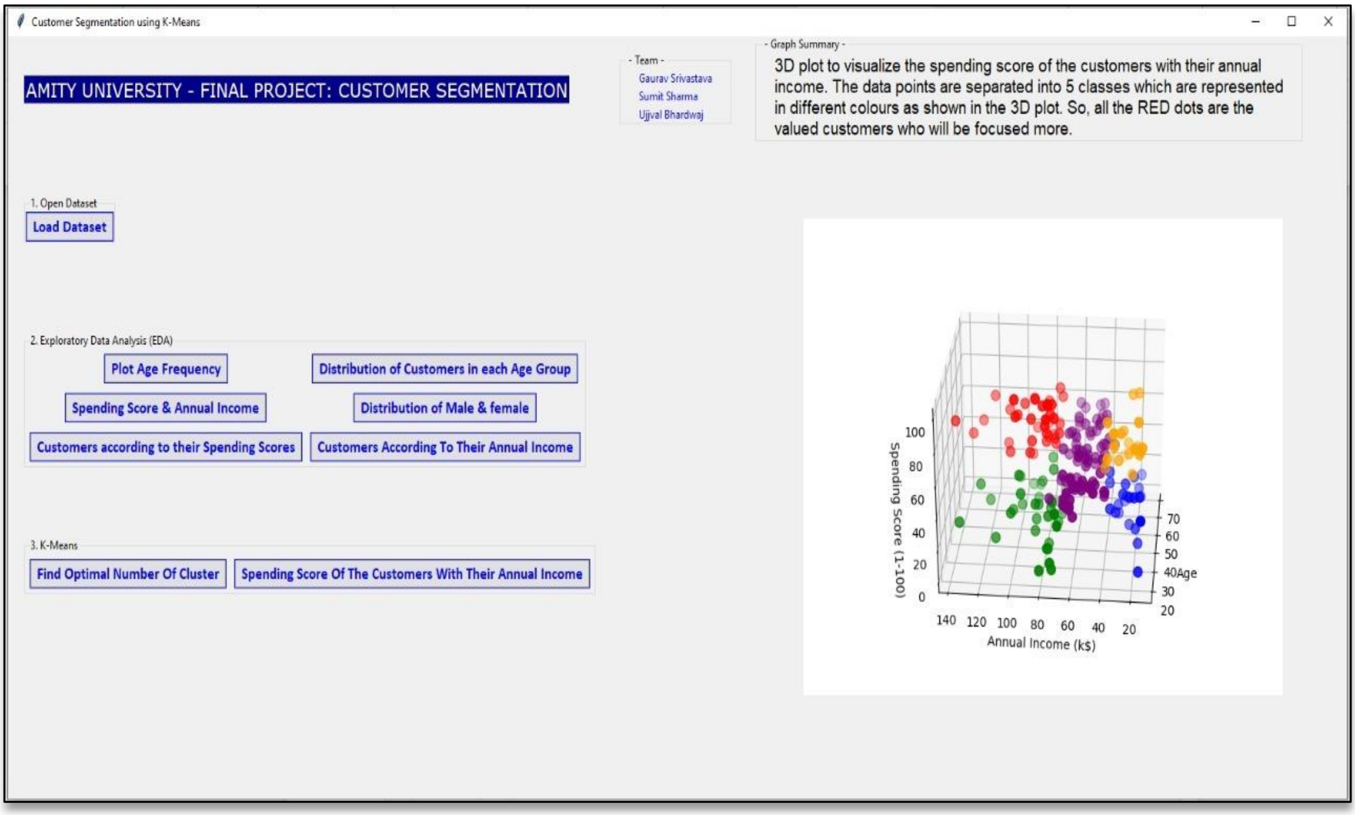
5. Analysis of results:

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 potential clusters to have probable customers are.

By using this model we can target the customers with whom we can start marketing strategy [easy to converse].

This segmentation enables marketers to create targeted marketing messages for a specific group of customers which increases the chances of the person buying a product. It allows them to create and use specific communication channels to communicate with different segments to attract them.

Application for Customer Segmentation using K Means-



Customer Segmentation using K-Means

AMITY UNIVERSITY - FINAL PROJECT: CUSTOMER SEGMENTATION

- Team -
Gaurav Srivastava
Sumit Sharma
Ujjval Bhardwaj

- Graph Summary -
3D plot to visualize the spending score of the customers with their annual income. The data points are separated into 5 classes which are represented in different colours as shown in the 3D plot. So, all the RED dots are the valued customers who will be focused more.

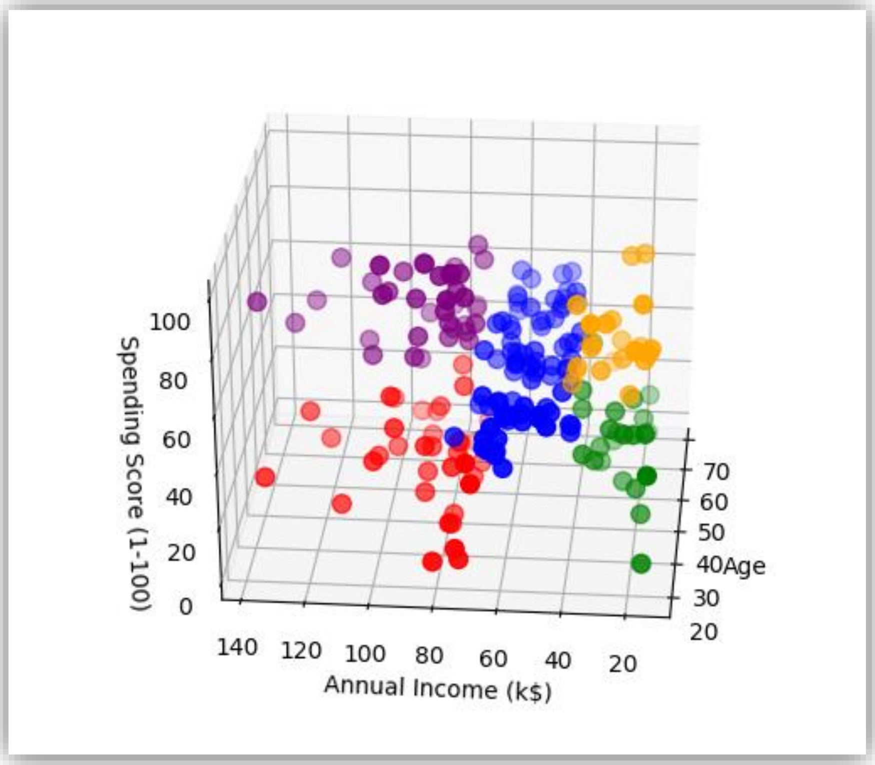
1. Open Dataset
Load Dataset
2. Exploratory Data Analysis (EDA)
Plot Age Frequency
Distribution of Customers in each Age Group
Spending Score & Annual Income
Distribution of Male & female
Customers according to their Spending Scores
Customers According To Their Annual Income
3. K-Means
Find Optimal Number Of Cluster
Spending Score Of The Customers With Their Annual Income
Spending Score (1-100)
100
80
60
40
20
0
Annual Income (k$)
140
120
100
80
60
40
20
Age
70
60
50
40
30
20

1- **Load Data Set –**
   We have to load data through **Load Data Set** utility in **Customer Segmentation Application** in .csv file, uploaded data set should have all require column data.
2- **Exploratory Data Analysis –**

   We can explore data through different graph, plot and many visualization EDA functionality.

3- **K- Mean Customer Segmentation as per spending score of the customers with annual income –**
   As per the functionality of customer segmentation application, we can target the customer as per provide in graph and **can find the targeted customer detail in result** .we can **plan our market strategy for targeted customer marked in red**.

Conclusion - 3D plot to visualize the spending score of the customers with their annual income.



3D plot to visualize the spending score of the customers with their annual income. The data points are separated into 5 classes which are represented in different colours as shown in the 3D plot. So, all the RED dots are the valued customers who will be focused more.