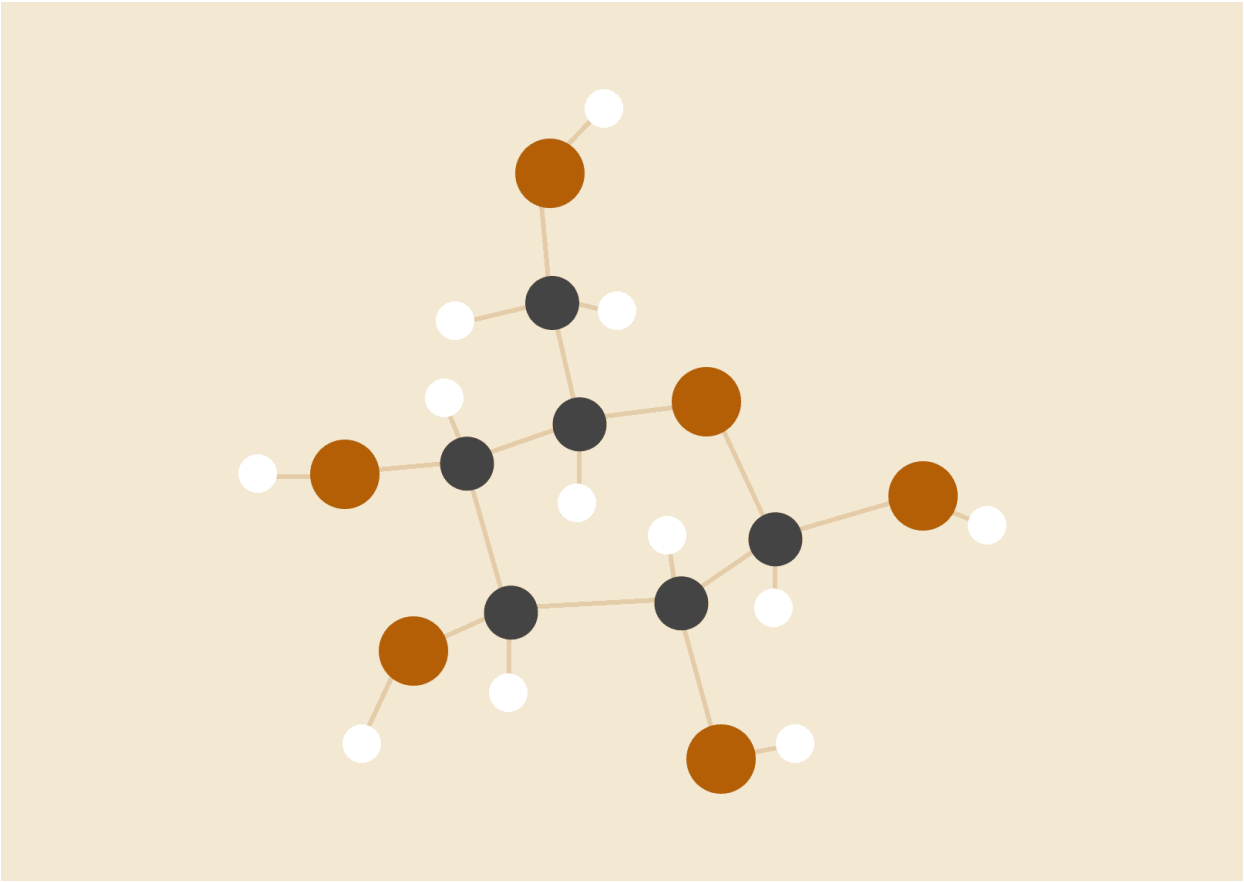
MARKET SEGMENTATION

FEYNN LABS - STUDY TASK



TEAM VISHAKHA

VISHAKHA WARUDKAR

UJWAL KOTHAPALLY

HARSH LANGADE

MINTU MEDHI

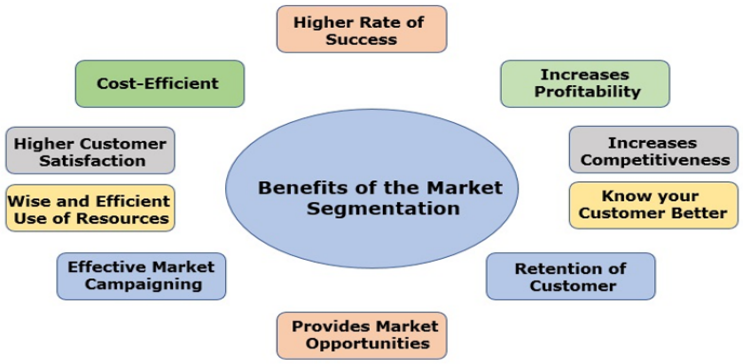
INTRODUCTION

The following report contains a brief summary of market segmentation. The report is a study project by Feynn Labs. It provides an overview of market segmentation with analysis on various steps required to complete the process of market segmentation.

What Is Market Segmentation?

Market segmentation is a marketing term that refers to aggregating prospective buyers into groups or segments with common needs and who respond similarly to a marketing action. Market segmentation enables companies to target different categories of consumers who perceive the full value of certain products and services differently from one another.

The Benefits of Market Segmentation



PROCEDURE

1. STEP 1: DECIDING (NOT) TO SEGMENT

The first step is to decide whether or not the dataset and market we choose can be enhanced with segmentation. In order to do so, we need to fathom the 'need' of the customers and how we can group the customers based on their preference. The team needs to assess the long-term organizational commitment, the resources available and potential changes/developments.

Implementation barriers should be further identified and solutions must be proactively deployed. According to McDonald and Dunbar (1995), a resolute

absolute sense of purpose and dedication is required above all, along with a willingness to appreciate the inevitable problems which will be encountered in implementing the conclusions.

2. STEP 2: SPECIFYING THE IDEAL TARGET SEGMENT

Next, we move onto selecting the ideal target segment. Out of the various segments we have identified through different categories, we are required to choose an ideal target segment that is most attractive to us. To proceed with the selection, evaluation of the segments must be done.

For example, as proposed by Solomon et al. (2011) the criteria used to evaluate market segments should be - Differentiable, Measurable, Substantial, Accessible and Actionable. There are generally two sets of segment evaluation criteria, knock-out criteria and attractiveness criteria.

The knock-out criteria includes essential and non-negotiable features of the segment whereas the attractiveness criteria involves relative attractiveness of the remaining market segments apart from the knock-out segments. A structured process should be implemented resulting in each criteria having an appropriate weight attached to it signifying the level of importance. Different perspectives must be considered for optimal results.

3. STEP 3: COLLECTING DATA

To begin with collecting data we should signify that empirical data forms the basis of both commonsense and data-driven market segmentation as it provides a clear base to work on. This data enables us to create market segments as well as give a grand picture and explanation of these segments.

The commonsense segmentation uses variables to represent one characteristic of the consumer whereas data-driven is based on multiple segmentation variables. Variables that describe personal characteristics are known as descriptor variables and are critical to creating an effective marketing mix targeting the segment.

With prior knowledge of the market, the organization can decide to choose the best segmentation criteria. The most common of the segmentation criteria are geographic, socio-demographic, psychographic and behavioral. Each criteria provides various benefits and barriers to the segmentation.

- *Geographic - Geographic segmentation is a component that competently*

complements a marketing strategy to target products or services on the basis of where their consumers reside. Division in terms of countries, states, regions, cities, colleges or Areas is done to understand the audience and market a product/service accordingly.

- *Socio-demographic - Socio-demographics are nothing more than characteristics of a population. Generally, characteristics such as age, gender, ethnicity, education level, income, type of client, years of experience, location, etc. are being considered as socio-demographics and are being asked in all kinds of surveys.*

- *Psychographic - Psychographic segmentation is the process of creating clusters of customers who share similar characteristics and then grouping them together. These clusters will form subsets (or segments) of customers based on their: Social class, Lifestyle, Interests, Values etc.*

- *Behavioral - Behavioral segmentation is a form of marketing segmentation that divides people into different groups who have a specific behavioral pattern in common. Users may share the same lifecycle stage, previously purchased particular products, or have similar reactions to your messages.*

A huge proportion of market segmentation data comes from survey data due to its feasibility and accessibility. Although the data can be disrupted due to the biases introduced as opposed to actual behavior. To tackle this we need to introduce various response options and styles such as agree, strongly agree, neither agree nor disagree so as to capture responses closest to the actual behavior.

Another important aspect is the sample size of the data that needs to be collected. The sample size should be sufficient to enable the algorithm to extract the correct segments and optimize the quality of the results. It can be concluded from the body of work studying the effects of survey data quality on the quality of market segmentation results based on such data that, optimally, data used in market segmentation analyses should -

- Contain all necessary items;
- Contain no unnecessary items;
- Contain no correlated items;
- Contain high-quality responses;

- be binary or metric;
- be free of response styles;
- Include responses from a suitable sample given the aim of the segmentation study
- Include a sufficient sample size given the number of segmentation variables (100 times the number of segmentation variables).

4. STEP 4: EXPLORING DATA

An EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent. You can't draw reliable conclusions from a massive quantity of data by just gleaning over it—instead, you have to look at it carefully and methodically through an analytical lens.

Getting a "feel" for this critical information can help you detect mistakes, debunk assumptions, and understand the relationships between different key variables. Such insights may eventually lead to the selection of an appropriate predictive model.

Data Cleaning is a critical aspect of the domain of data management. The data cleansing process involves reviewing all the data present within a database to either remove or update information that is incomplete, incorrect or duplicated and irrelevant. Data cleansing is just not simply about erasing the old information to make space for new data, but the process is about rather finding a way to maximize the dataset's accuracy without necessarily tampering with the data available.

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

5. STEP 5: EXTRACTING SEGMENTS

Many segmentation methods can be used to extract market segments using clustering algorithms, where cluster signifies the segments. As mentioned by Henning and Liao (2013) selecting a suitable clustering method is required to match the requirement of a user group like researcher. Hence it is required to

extract the segments from different clustering algorithms like Hierarchical methods, K-means method, Hybrid approach (the centroids of K-means cluster is user in Hierarchical Method to form dendrogram to decide nos of segments) etc can be used and then to decide the analytical-result of which algorithm suits as per requirement of the end user as there is no one single best algorithm for all data set. While deciding the optimum algorithm for extracting the segments we need to consider Data Set characteristics like Nos of consumers, Nos of segmentation variables, scale of segmentation variable etc and Segmentation characteristics like similarities in the segment and differences between the segments, Number and size of segments etc.

A helpful solution to decide the best clustering method is to extract the groups using all suitable clustering methods and then perform global stability and segment level stability analysis to decide the best clustering method and segments and possibly pass the segments created by the best clustering algorithm to the next step for profiling segments.

6. STEP 6: PROFILING SEGMENTS

Profiling is all about differentiating segments based on variables so that it can be observed how the groups are segmented based on segregating criteria. The goal of profiling is to differentiate between groups with proper distinction so that each group has some unique characteristics and the cluster has an ideal solution.

While profiling segments we need to count each segment which gives the number of observations present in each of the segments followed by overall average and individual segment wise average for each attribute. To find the best profiling, the above steps needs to be performed for each of the k-values(k-clusters)

Profiling is an essential tool to validate whether the groups formed through clustering techniques are business relevant or not to solve business problems.

For instance, what is the average income of each segment, and then on comparing the segments, it can be understood that one group has a high income, another group has low income. This way it is visible that the groups are getting differentiated on the basis of income i.e. income is segregating the segments.

7. STEP 7: DESCRIBING SEGMENTS

Segment describing needs to be done after segment profiling. The segments are described on the basis of variables(segment characteristics) which are not used for the purpose of segmentation. These additional variables are known as descriptor variables. Segmentation descriptions are required for the development of a customized marketing mix.

Using visualizations is the best way to describe segments because this simplifies interpretation of results for both, data analysts and the organizational officials. A cross tabulation of segment membership with descriptor variables is generally the basis for all visualizations.

- Nominal or Ordinal Descriptor variables:

  Mosaic plots are the best way to describe the segments for nominal or ordinal descriptor variables. They can be used to visualize segments for multiple descriptor variables.

- Metric variables:

  Conditional plots are generally used for this purpose. Histograms and parallel box-and-whisker plots.A modified version of segment level stability across solutions(SLSA's) plot can be used to know the true value of a descriptor variable.

Simple statistical tests can be used to test for differences between market segments for various descriptor variables. Use of chi-square tests can be done. Analysis of Variance(ANOVA) is one popular method which can be used for testing differences between means of more than two groups.

A way to learn about segments is to try to predict membership using descriptor variables. Some Linear Regression models/algorithms which can be used for this purpose are Binary Logistic Regression, Multinomial Logistic Regression and some tree-based methods.

8. STEP 8: SELECTING (THE) TARGET SEGMENT(S)

The selection of one or more target segments is a long-term decision which significantly affects the future performance of an organization. First thing to do in this step is to make sure that each of the segments complies with the knock-out criteria. Then, the attractiveness of the remaining segments and the relative organizational competitiveness for these segments needs to be evaluated.

McDonald and Dubar(1995) ; Lilien and Rangaswamy(2003), recommend the use of a decision matrix to visualize and interpret relative segment attractiveness and relative organizational competitiveness for all market segments. Whatever criteria is chosen, the decision matrix plot has two dimensions: segment attractiveness and relative organizational competitiveness/constraints.

The data on actual values that a segment has for a particular attractiveness criteria emerges from grouping,profiling and description of each market segment. The weights quantifying the impact of each of the criteria on total value of segment attractiveness is available already from step 2. Using this data, segment attractiveness is calculated. Similar data is required for calculating relative organizational competitiveness and the same procedure is deployed to calculate organizational constraints for each of the criteria i.e. weights are assigned for each criteria and corresponding values for each of the segments for each of the criteria are assigned.

The bubble size for each of the segments on the plot may represent any characteristic of the segment. Typically, profit potential is used.

The resulting plot can be used to choose one or more target marketing segments.

9. STEP 9: CUSTOMIZING THE MARKET MIX

The marketing mix is an important tool in the formulation and implementation of an effective marketing strategy. It should be used to show the possibilities why your product or service is different and better than your competitors.

Customization is the latest branding strategy as it plays a key role in increasing efficiency, ensuring maximum optimization of resources at hand, and reducing unfair wastage.

- ### Product

  Product is related to developing the right product or service for your target market. Your product or service must meet a specific consumer need. When setting a price for your product, you should consider the cost of

competition and total marketing mix in your target market location.

- ### Price

  Price develops the habit of constantly checking and rechecking the prices of the products and services you sell to make sure they are still suitable for the realities of the current market. Sometimes you need to lower your prices. At other times, it may be appropriate to increase your value.

- ### Promotion

  Marketing and sales means thinking about promotions all the time. Promotions include all the ways you tell your customers about your products or services and then how you sell and sell them in the market.

  Small changes in the way you promote and sell your products can lead to different changes in your results. Even small changes in your ad can immediately lead to higher sales.

- ### Place

  Place in the Marketing Mix is the place where your product or service is actually sold. Develop the habit of reviewing and reflecting on the exact location where the customer meets the seller. Sometimes changes in place can lead to a rapid increase in sales.

  You can sell your product in many different places. Some companies use direct selling, sending their salespeople to meet in person and talk with the prospect. Some sell by telemarketing. Some sell via catalog or mail order.

Some trades sell at shows or retail establishments. Some others sell in joint ventures with similar products or services. Some companies use representatives of manufacturers or distributors. Many companies use a combination of one or more of these methods.

10. STEP 10: EVALUATION AND MONITORING

The Objective of Evaluating the effectiveness of the Market Segmentation Strategy is to determine whether to develop a customized marketing mix for one or more segments that yields the expected benefits for the Organization.

Evaluation of Market Segmentation

We have to consider three factors while Evaluating different Market Segments.

- ### Segment Size & Growth

  Size of the Market segment must be right for that market segment to be large and it must include characteristics like growth. The Size of the Market Segment should be Consistent with the Production Capacity and Development Characteristics of the Organization. The Market Segment must meet the growth potential and the Development Potential should be able to be handled by the Organization.

- ### Segment Structural Attractiveness

  Every company needs to evaluate the attractiveness of the market segment for the long run. Market segment can be more attractive or less attractive. If the market segment is large it will be considered as more attractive when not much larger it will be less attractive.

  For Example, what if market segment is more attractive, more number of companies will come to cater i.e. the company will come to serve the segment. Therefore, that segment will be more competitive.

  For every company it is not necessary that if the market segment is large, this will be more attractive. If a small company wants to cater to a larger company, this will not happen due to lack of skills and resources. In that scenario, small firms will not be able to cater large segments. If any firm

can cater any segmentation, this will be less attractive.

Single suppliers must be less attractive because suppliers can control the price and reduce the quality or quantity of the goods.

- Company Objective & Resources

  Companies should select this market segment on the basis of its objectives and resources.

  For example, BMW is a luxury car Maker Company. Its objective is to make luxury cars with the best performance. If they manufacture cars in the luxury sector then It will not make any sense. While evaluating market segments we need to consider things like segment size and growth, attractiveness of segment, objective and resources of the company.

  After completion of Market Segmentation analysis and all strategic and tactical marketing activity have been initiated, the success of Market Segmentation strategy should be evaluated and the Market should be carefully monitored on a continuous basis.

CASE STUDY

Problem Statement :

This Mall wants to get insights on their customers. We have to build a system that can cluster customers into different groups. So one group of customers may represent those who tend to purchase more from the mall, some other groups may represent groups of customers that don't purchase that much in amount. So, this group of customers will give us insights and details about that mall to make better decisions to make better marketing strategies.

COLLECTING DATA

Importing Python Libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from scipy.spatial.distance import cdist
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
# Loading the data
data = pd.read_csv("D:\\ML Project\\Market_Mall_segmentation\\Mall_Customers.csv")
data.head(2) # Shows First two rows
```

Out[2]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81

In [3]:
```
data.shape, data.columns
```

Out[3]:
```
((200, 5),
 Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
        'Spending Score (1-100)'],
       dtype='object'))
```

The Mall Customer data is taken from Kaggle which includes features like Customer ID, Gender, Age, Annual Income(k$) and Spending Score(1-100). The data includes 200 data

points. We load the csv dataset with the help of pandas.

EXPLORING DATA

```
In [4]: data.isna().sum() # There is no missing values

Out[4]: CustomerID              0
        Gender                  0
        Age                     0
        Annual Income (k$)      0
        Spending Score (1-100)  0
        dtype: int64
```

```
In [6]: data.info() # This shows null values and data types

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 200 entries, 0 to 199
        Data columns (total 5 columns):
         #   Column                  Non-Null Count  Dtype
        ---  ------                  --------------  -----
         0   CustomerID              200 non-null    int64
         1   Gender                  200 non-null    object
         2   Age                     200 non-null    int64
         3   Annual Income (k$)      200 non-null    int64
         4   Spending Score (1-100)  200 non-null    int64
        dtypes: int64(4), object(1)
        memory usage: 7.9+ KB
```

info function will show the data type of the data points as well as missing values, if present. In this case, there are no null values in the dataset. It basically allows us to learn the shape of the object type of our data.

```
In [7]: data.describe() #This is to know descriptive statistics summary
```

Out[7]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

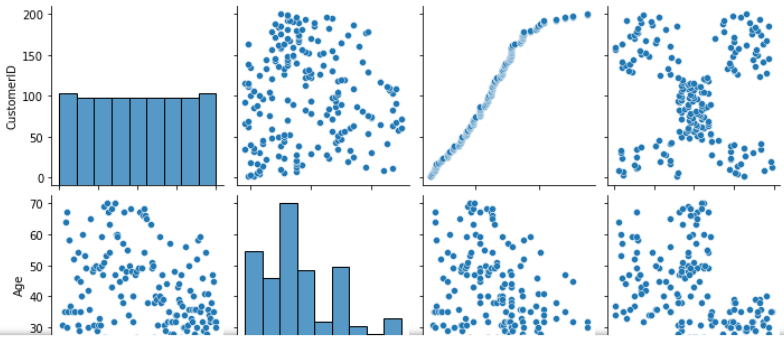
The describe() method computes and displays summary statistics for a python dataframe. It is used to generate descriptive statistics that summarize the central tendency, dispersion and shape of the dataset's distribution excluding NaN values.

```
In [8]: correlation = data.corr()

In [9]: sns.heatmap(correlation, annot = True)
        plt.show()
```

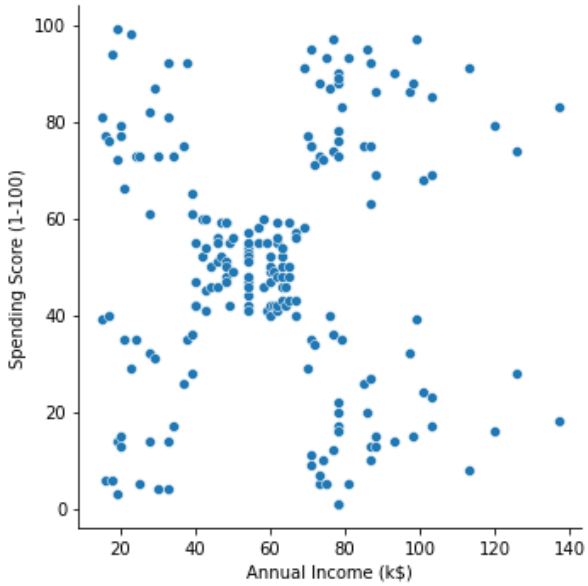
The Heatmap consists of values representing different shades of the same color for each value to be plotted. Usually the darker color in the chart represents higher values than the lighter shades. A completely different color can also be used for a very different value. It shows graphical representation of correlation of the variables of the data using color to visualize the values.

```
In [10]: sns.pairplot(data)
Out[10]: <seaborn.axisgrid.PairGrid at 0x19d17762640>
```



Seaborn pairplot used to get relations between each and every variable present in pandas DataFrame. The pairplot function creates a grid of axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. That creates plots as shown above.

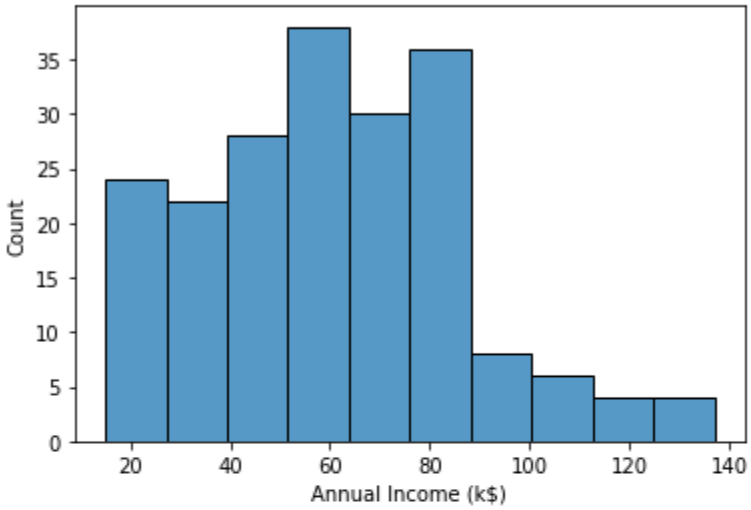
```
In [11]: sns.relplot(x = 'Annual Income (k$)', y = 'Spending Score (1-100)', data = data)
Out[11]: <seaborn.axisgrid.FacetGrid at 0x19d17739cd0>
```



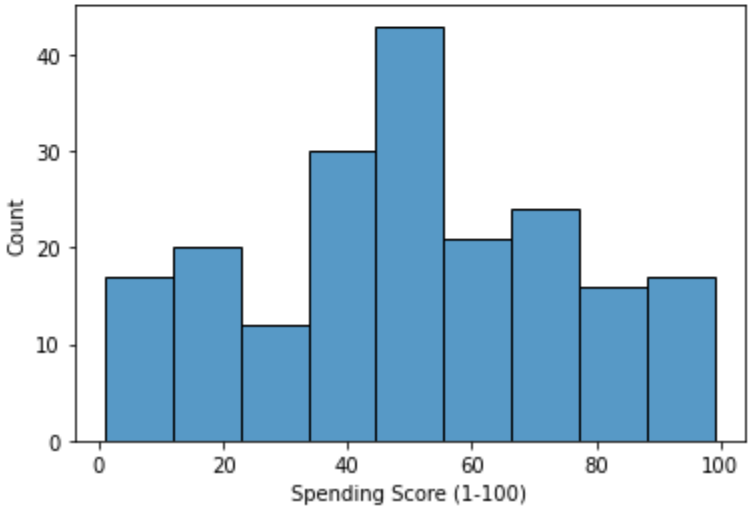
This relplot function provides us access to some other axis level function which shows

the relationship between two variables with semantic mappings of subsets. It is nothing but visualizing the statistical relationship between data points. In this case we have determined the relationship between Annual Income and Spending Score.

```
In [12]: sns.histplot(data['Annual Income (k$)'])
Out[12]: <AxesSubplot:xlabel='Annual Income (k$)', ylabel='Count'>
```




```
In [13]: sns.histplot(data['Spending Score (1-100)'])
Out[13]: <AxesSubplot:xlabel='Spending Score (1-100)', ylabel='Count'>
```

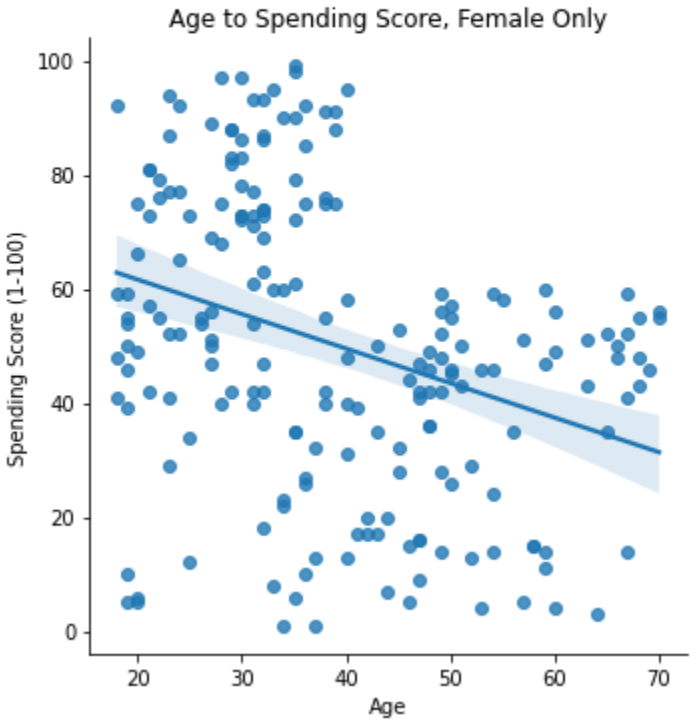


A histogram is a visualization tool that represents the distribution of one or more variables by counting the number of observations that fall within discrete bins. Increased Annual Income and Spending Score shows the symmetric relationship

between each other.

```
In [14]: sns.lmplot('Age', 'Spending Score (1-100)', data)
         plt.title('Age to Spending Score, Female Only')
```

```
Out[14]: Text(0.5, 1.0, 'Age to Spending Score, Female Only')
```



lmplot is a line plot, one of the most basic plots which shows the line on a 2-D plane. Its function creates a basic scatter plot using given data onto a FacetGrid.

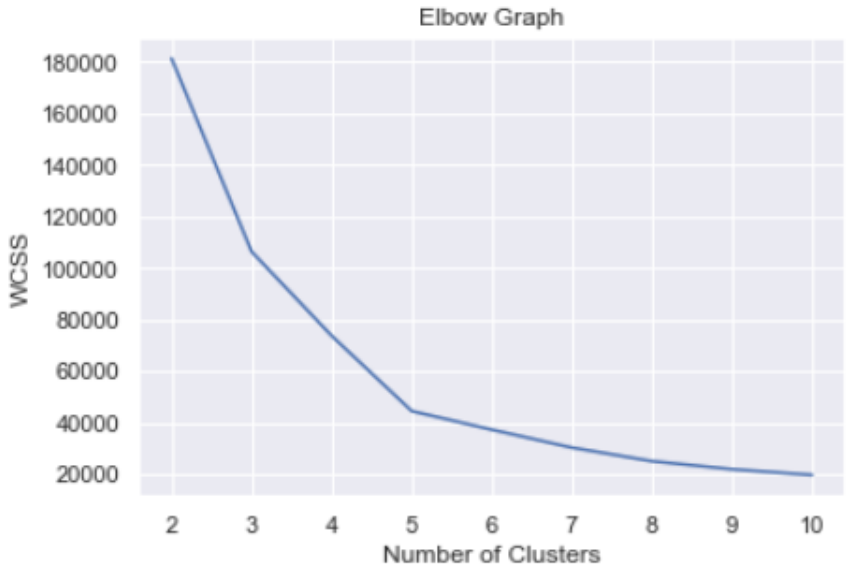
```
In [15]: x = data[['Annual Income (k$)', 'Spending Score (1-100)']].values
         x
```

```
Out[15]: array([[ 15,  39],
                [ 15,  81],
                [ 16,   6],
                [ 16,  77],
                [ 17,  40],
                [ 17,  76],
                [ 18,   6],
                [ 18,  94],
                [ 19,   3],
                [ 19,  72],
                [ 19,  14],
                [ 19,  99],
                [ 20,  15],
                [ 20,  77],
                [ 20,  13],
                [ 20,  79],
                [ 21,  35],
                [ 21,  66],
                [ 23,  29],
```

```
In [16]: # WCSS - Within Cluster Sum of Squares
         WCSS = []
         for i in range(2, 11):
             kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
             kmeans.fit(x)
             WCSS.append(kmeans.inertia_)
```

WCSS is the sum of squared distance between each point and centroid in a cluster. When we plot WCSS with the K value, the plot will look like an elbow. As the number of clusters increases, the Value of WCSS starts to decrease. It should be large when k = 1.

In [17]:
```
sns.set()
plt.plot(range(2, 11), WCSS)
plt.title("Elbow Graph")
plt.xlabel("Number of Clusters")
plt.ylabel("WCSS")
plt.show()
```



As we have to determine the value of k, the Elbow method is used. This will help to choose the optimum value of k by fitting the model with the range of value of k. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters.

In [18]:
```
# Training the Model
kmeans = KMeans(n_clusters = 5, init = 'k-means++', random_state = 0)
kmeans.fit(x)
#return the label for each data point based on cluster
y = kmeans.predict(x)
y
```

Out[18]:
```
array([4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3,
       4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 1,
       4, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 0, 2, 1, 2, 0, 2, 0, 2,
       1, 2, 0, 2, 0, 2, 0, 2, 0, 2, 1, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
       0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
       0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
       0, 2])
```

K- Means clustering is one of the unsupervised learning algorithms. It is used when the data is not already divided into groups or categories(unlabelled data). The aim of this clustering algorithm is to search and find the groups in the data, where variable k represents the number of groups.

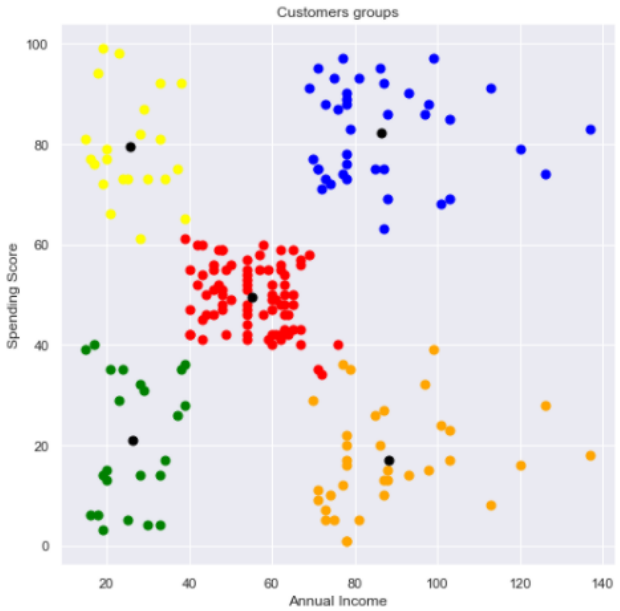
The k-means clustering algorithm mainly performs two tasks:

1. Determines the best value for k center points or centroids by an iterative process.
2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, constitute a cluster.

In [19]:

```
#Plotting all the Clusters
plt.figure(figsize = (8,8))
plt.scatter(x[y == 0, 0], x[y == 0, 1], c = 'orange', label = 'Cluster 1', s = 50)
plt.scatter(x[y == 1, 0], x[y == 1, 1], c = 'red', label = 'Cluster 2', s = 50)
plt.scatter(x[y == 2, 0], x[y == 2, 1], c = 'blue', label = 'Cluster 3', s = 50)
plt.scatter(x[y == 3, 0], x[y == 3, 1], c = 'yellow', label = 'Cluster 4', s = 50)
plt.scatter(x[y == 4, 0], x[y == 4,1], c = 'green', label = 'Cluster 5', s = 50)

# Plotting all the Centroids
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s = 50, color = 'black', label = 'Centroids')
plt.title("Customers groups")
plt.xlabel("Annual Income")
plt.ylabel("Spending Score")
plt.show()
```



The output image shows the clusters with their centroids. This one clearly shows five different clusters with different colors. Clusters are formed between two parameters of the dataset; Annual income of the customer and Spending Scores(expenses). We can change the color and label as per requirement or choice.

PROFILING SEGMENTS

From the above graph it can be observed that there are 5 distinct clusters.

For profiling the segments we first label each of the data points with the cluster number

In [100]:

```
data_groupded=data.groupby(by=["Clusters"]).mean().reset_index()
data_groupded.rename(columns={"Clusters":"Customer_Groups"},inplace=True)
data_groupded
```

Out[100]:

	Customer_Groups	Age	Annual Income (k$)	Spending Score (1-100)
0	0	41.114286	88.200000	17.114286
1	1	42.716049	55.296296	49.518519
2	2	32.692308	86.538462	82.128205
3	3	25.272727	25.727273	79.363636
4	4	45.217391	26.304348	20.913043

it belongs to.

Then we group the segments as per the clusters (Customer_groups) to find the mean of other segments like “Annual Income(k$)” and “Spending Score (1-100) for each cluster.

```
In [83]: CustomerMagnitue_df = pd.DataFrame(data.Clusters.value_counts().reset_index())
         CustomerMagnitue_df.rename(columns={"index": "Customer_Groups",
                                             "Clusters": "Customer Group Magnitude"},inplace=True)
         CustomerMagnitue_df
```

Out[83]:

	Customer_Groups	Customer Group Magnitude
0	1	81
1	2	39
2	0	35
3	4	23
4	3	22

In the final step of profiling, the Customer_magnitude data frame is merged with data_grouped dataframe with a common key as Customer_Groups (clusters).

```
In [102]: profilling=pd.merge(data_groupded,CustomerMagnitue_df, on="Customer_Groups")
          profilling
```

Out[102]:

	Customer_Groups	Age	Annual Income (k$)	Spending Score (1-100)	Customer Group Magnitude
0	0	41.114286	88.200000	17.114286	35
1	1	42.716049	55.296296	49.518519	81
2	2	32.692308	86.538462	82.128205	39
3	3	25.272727	25.727273	79.363636	22
4	4	45.217391	26.304348	20.913043	23

From profiling we can observe that the highest Customer Group Magnitude is 81 with mean spending score of 49.51 and Annual Income 55.29 k$ and the mean Age is 42. i.e Customer group number 1 having mean spending score of 49.51 and mean annual Income 55.29 k$. Similarly, other customer groups can be analyzed.

DESCRIBING SEGMENTS

Now, we proceed to the describing step.

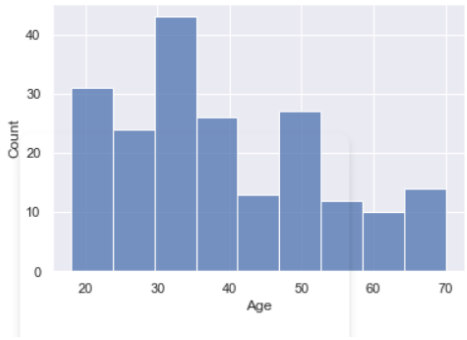
The descriptor variable used here is 'Age'. Histograms are plotted for visualizing age distribution in each of the segments. A box plot is also used to visualize it properly.

```
data_new = pd.read_csv("Mall_Customers.csv")
data_new["Clusters"]=kmeans.labels_
data_new.head()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Clusters
0	1	Male	19	15	39	4
1	2	Male	21	15	81	3
2	3	Female	20	16	6	4
3	4	Female	23	16	77	3
4	5	Female	31	17	40	4

```
sns.histplot(data['Age'])
```

```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```

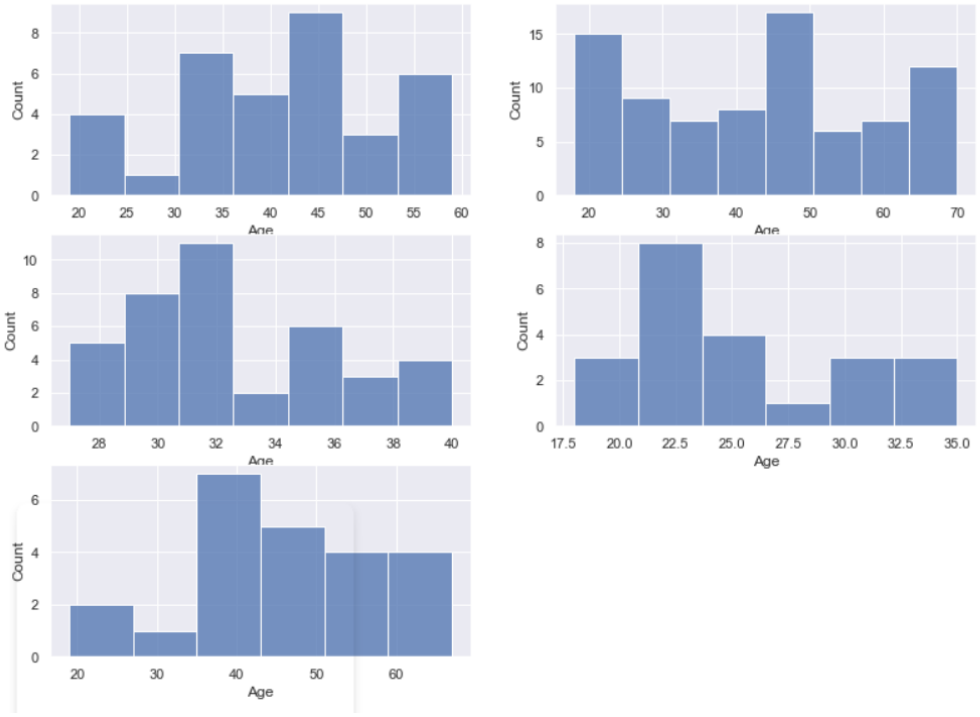


The age distribution(considering the entire dataset) shows that the age group 30-35 has more customers as compared to other age groups.

```
clust_1 = data_new.loc[data['Clusters'] == 0]
clust_2 = data_new.loc[data['Clusters'] == 1]
clust_3 = data_new.loc[data['Clusters'] == 2]
clust_4 = data_new.loc[data['Clusters'] == 3]
clust_5 = data_new.loc[data['Clusters'] == 4]
```

```
plt.figure(figsize=(12,9))
plt.subplot(3,2,1)
sns.histplot(clust_1['Age'])
plt.subplot(3,2,2)
sns.histplot(clust_2['Age'])
plt.subplot(3,2,3)
sns.histplot(clust_3['Age'])
plt.subplot(3,2,4)
sns.histplot(clust_4['Age'])
plt.subplot(3,2,5)
sns.histplot(clust_5['Age'])
```

```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```

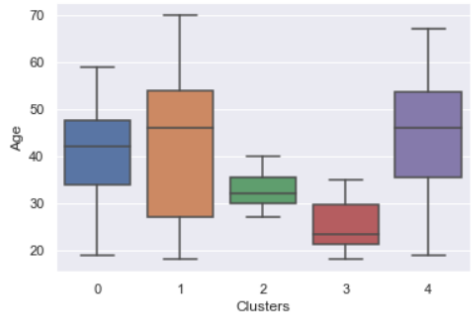


From these histograms, the number of customers in each of the segments is seen.

The 1st segment has more customers in the age group 40-45, 2nd one has more customers in the age group 40-50. The 3rd segment has more customers in the age group 28-32, whereas the 4th segment has more customers in the age group 20-25. The 5th segment has more customers in the age group 35-50.

```
sns.boxplot(data = data,x = 'Clusters',y = 'Age')
```

```
<AxesSubplot:xlabel='Clusters', ylabel='Age'>
```



Age distribution is visualized using a box plot as well.

SELECTING (THE) TARGET SEGMENT(S)

After completing profiling and describing, now we select the target segment.

```
df1 = pd.read_csv("Segment_Evaluation_Data.csv")
df1.head()
```

	Customer_Groups	Average Age	Rating(Value)_Age	Rating(Value)_Gender	Segment Attractiveness	Average Annual Income(k$)	Rating(Value)_Annual_Income(k$)	Average Spending Score(1-100)	Rating(Value)
0	0	41.114286	5	6	5.5	88.200000	9	17.114286	
1	1	42.716049	5	7	6.0	55.296296	6	49.518519	
2	2	32.692308	8	6	7.0	86.538462	8	82.128205	
3	3	25.272727	7	7	7.0	25.727273	3	79.363636	
4	4	45.217391	5	7	6.0	26.304348	3	20.913043	

The segment attractiveness criteria used here are Age and Gender, whereas for evaluating organizational competitiveness, Annual Income(k$) and Spending Score(1-100) are used. Spending score is given a weightage of 65% and annual income is given weightage of 35%. Age and Gender are given equal weightage of 50%. The rating for each of the segments for these criteria are given assuming that the mall generally sells items which have a higher cost. So, annual income and spending score are the criteria that consumers will consider when choosing between malls they will prefer.i.e. segments with higher average annual income and higher average spending score are given a higher rating because customers of those segments will find this particular mall more attractive as compared to customers of other segments. It is also assumed that female customers generally visit malls more often and hence higher the percentage of female customers in a segment, higher is its rating for gender criteria. Similarly, it is assumed that younger people visit malls more often than older people.So,segments with lower average age of customers are given a higher rating for the age criteria. This is because the mall will prefer segments with more female customers and younger people so as to maximize their sales. Using this data, a segment evaluation plot is plotted.

```
x = np.asarray(df1['Segment Attractiveness']).astype(np.float32)
y = np.asarray(df1['Organizational Competitiveness']).astype(np.float32)
plt.scatter(x,y)
plt.xlabel('Segment Attractiveness')
plt.ylabel('Organizational Competitiveness')
plt.title('Segment Evaluation Plot')
```

```
Text(0.5, 1.0, 'Segment Evaluation Plot')
```

It is clear from this plot that the segment with Organizational Competitiveness 8.65 and Segment Attractiveness 7.0 is the best for the mall to target. Hence, Segment number 3 is chosen as the target segment.

But, the segment with the maximum number of customers can be considered as well because even if other characteristics are not optimal, a higher number of customers can nullify the lack of other characteristics.

REPLICATION OF CASE STUDY IN PYTHON

https://github.com/vishakhawarudkar96/Feynn_Labs_Project.git