Customer Spending Analysis And Mall Segmentation
Raj Prashantkumar Patel
03/09/2022

1. Problem Statement

Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits. It is amazing to realize the fact that how machine learning can aid in such ambitions. And I'm trying a build the model that will be helpful for shopping complexes and small shops.

2. Market/Customer/Business need Assessment

The covid-19 pandemic and the lockdown has negatively affected businesses like Individual shops and small shopping complexes. They have been forced to shut down their shops and stalls early, which has resulted in significant and rapid decrease in sales. Also, the customer buying preferences have been significantly changed due to the pandemic. Therefore, by using this technique, we aim to provide small shopping complexes and shops with useful insights from the available data on customer spending score and their monthly income also ways to generate more revenue. Apart from that , how to increase the customer satisfaction and retention rate.

3. Target Specification

The proposed system/service will provide the management and shopkeeper with some

techniques so that their sales soar up and they can increase their customer retention rate and satisfaction . It will suggest them to give vouchers or offer to customers to low spending score customer and organize their stores according their customer need , based on the analysis performed by the algorithm, so that the customer tempt to buy the products and satisfy their needs. It also enhances the spending score of the high earning customers, so the small complexes and shops can increase their profits. It also allows the Shop owners to understand the customer purchase trend.
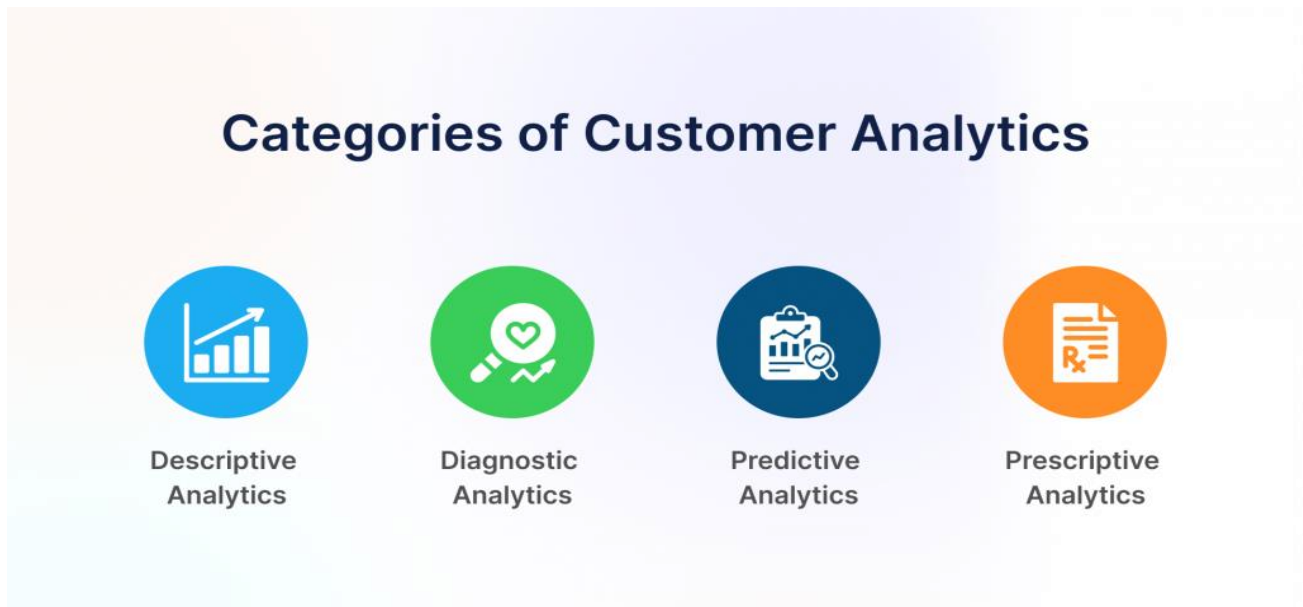
4. External Search

I have taken Dataset from the Kaggle:

Dataset Link : https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

Related links:
1) https://www.procuredesk.com/spend-analysis/
2) https://towardsdatascience.com/mall-customers-segmentation-using-machine-learning-274ddf5575d5
3) https://www.neenopal.com/CustomerSpendDashboard.html
4) https://www.revechat.com/blog/customer-analytics/

5. Benchmarking

Categories of Customer Analytics

Descriptive Analytics · Diagnostic Analytics · Predictive Analytics · Prescriptive Analytics

Most of the big companies uses Descriptive Analysis, Diagnostic Analysis, Predictive Analysis and Prescriptive Analytics to analyse the customer behaviour and their spendings.

But these type of analysis skills also be applicable to small complexes and shops to analyse their customers. And most of companies uses DBSCAN algorithms to analyse the data, but

For small Complexes and individual shops k-means algorithm gives better insight of data.

6. Applicable Constraints
● Data Collection from Shopping Complexes and Individual Shops
● Continuous data collection and maintenance
● Lack of technical knowledge for the user(small shopkeepers)
● Data Storage Spaces
● Convincing the shopkeepers and Shopping Complexes to implement the system in their   shops.

7. Applicable Regulations
● Data protection and privacy regulations(Customers)
● Govt Regulations for small businesses

8. Business Opportunity

Every small complexes and shops that depends on sales can and would want to go for
using this service in order to always know what their customers want. The emergence of
every small business is thus a fairly great business opportunity for the service provided
by us. Therefore, there is a prominent chance of this service being a great business opportunity.

9. Final Product Prototype And Code Implementation

In this Prototype , I have used Jupyter Notebook and K-means Clustering Algorithm.

Data Set:

```
#Loading the data from csv file

df = pd.read_csv("Mall_Customers.csv")
```

Attributes :

1) **CustomerID**
2) **Gender**
3) **Age**
4) **Annual Income (k$)**
5) **Spending Score (1-100)**

```
df.head()
```

	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

Dependencies and Libraries:
   1) **Sklearn**
   2) **Numpy**
   3) **Pandas**
   4) **Matplotlib**
   5) **Seaborn**

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.cluster import KMeans

import warnings
warnings.filterwarnings('ignore')

from sklearn.metrics import silhouette_score
```

Data Pre-processing and EDA:

```python
df.shape
```

```
(200, 5)
```

```python
df.info()
```

```
<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
```
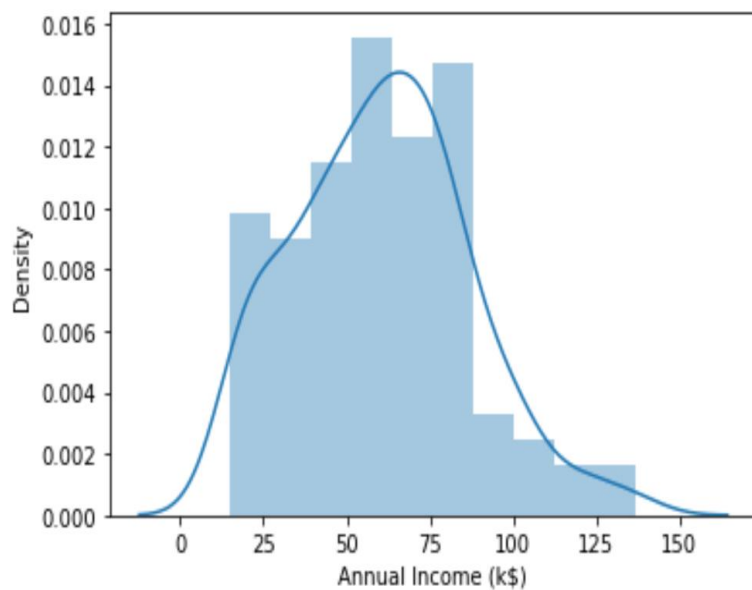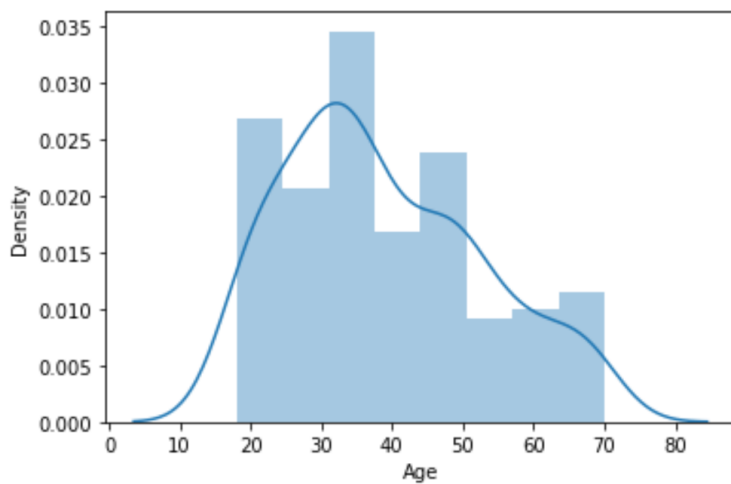
Handling Missing Values:

```
#Checking for the missing data

df.isnull().sum()
```
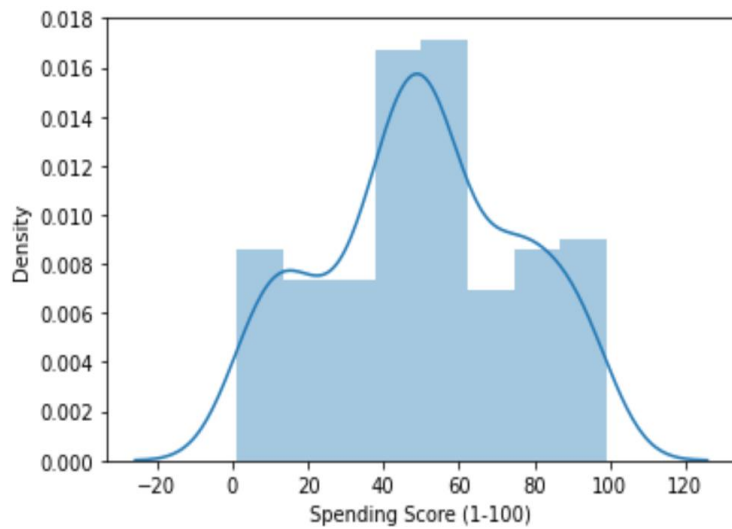
```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

Visualisation of Data:

```
columns = ['Age', 'Annual Income (k$)','Spending Score (1-100)']
for i in columns:
    plt.figure()
    sns.distplot(df[i])
```

Correlation Heat Map:

```
dataplot = sns.heatmap(df.corr(), cmap="YlGnBu", annot=True)
plt.show()
```


Choosing Columns for Forming clusters:

In this Prototype, I have chosen Annal Income and Spending score of the customer to form the clusters because this columns are closely coupled with this problem.

```
#choosing annual income and spending score to form clusters

clm = df.iloc[:,[3,4]].values
print(clm)
```

```
[[ 15  39]
 [ 15  81]
 [ 16   6]
 [ 16  77]
 [ 17  40]
 [ 17  76]
 [ 18   6]
 [ 18  94]
 [ 19   3]
 [ 19  72]
```

Model Building:

I have used K-means Clustering Algorithm to form the clusters.

Elbow Method :

```
#choosing the numbers of clusters using elbow method

wcss = []

for i in range(1,21):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=45)
    kmeans.fit(clm)
    wcss.append(kmeans.inertia_)
```

```
#plotting the elbow graph

sns.set()
plt.plot(range(1,21), wcss)
plt.title('Elbow Graph')
plt.xlabel('Number of Clusters')
plt.ylabel('wcss')
plt.show()
```

Based on the values of WCSS the elbow was plotted which will help to draw to draw the Elbow graph to determine the Number of clusters to be formed.

Elbow Graph
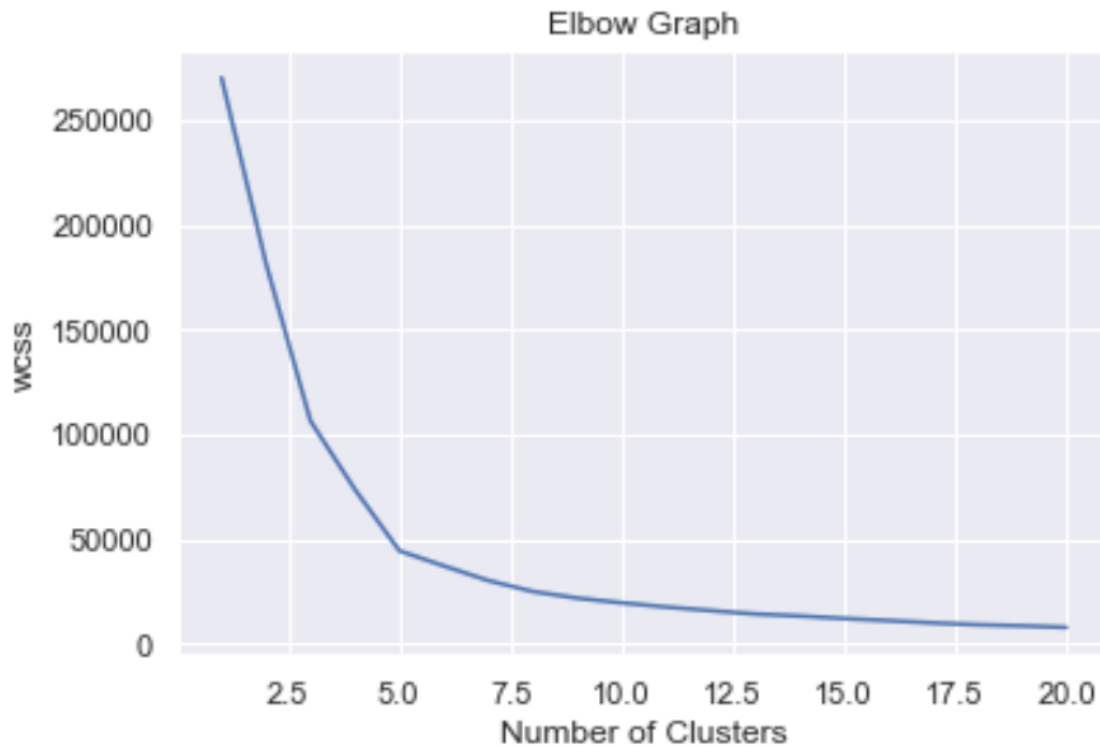
number of clusters are 5

Formation of Clusters:

```python
kmeans =KMeans(n_clusters=5, init='k-means++', random_state=0)

#labeling of clusters
y = kmeans.fit_predict(clm)
print(y)
```

```
[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 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 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]
```

Visualisation of Cluster:

```python
plt.figure(figsize=(8,8))
plt.scatter(clm[y==0,0], clm[y==0,1], s=50, c='green', label='cluster 1')
plt.scatter(clm[y==1,0], clm[y==1,1], s=50, c='red', label='cluster 2')
plt.scatter(clm[y==2,0], clm[y==2,1], s=50, c='blue', label='cluster 3')
plt.scatter(clm[y==3,0], clm[y==3,1], s=50, c='orange', label='cluster 4')
plt.scatter(clm[y==4,0], clm[y==4,1], s=50, c='yellow', label='cluster 5')

plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s=100, c='black', label='Centeroids')

plt.title('Customers Groups')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.show()
```
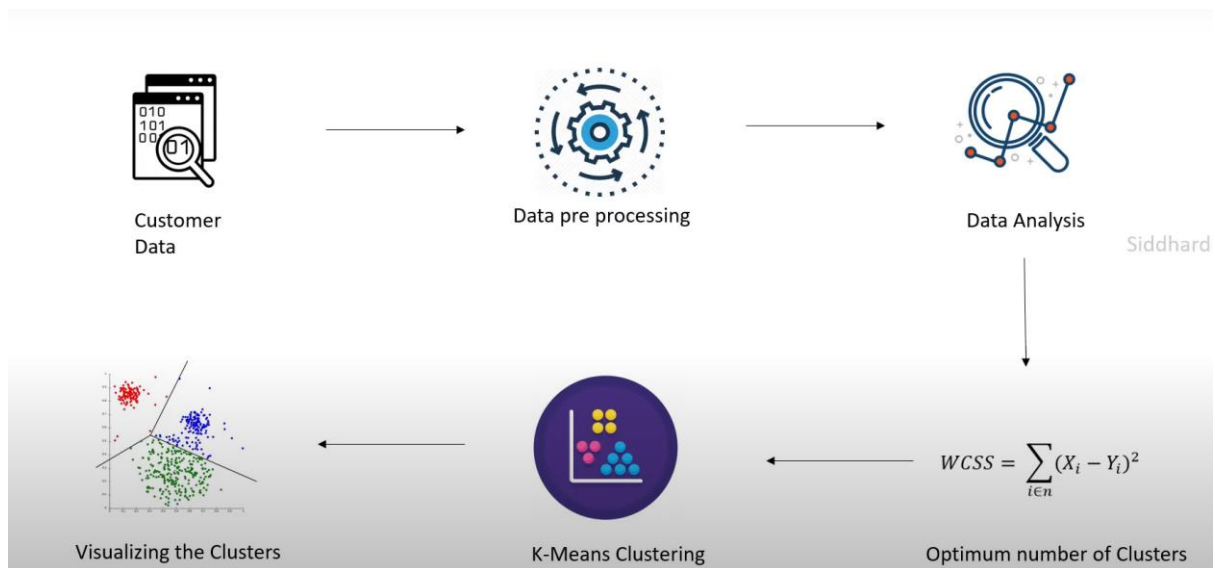
Customers Groups
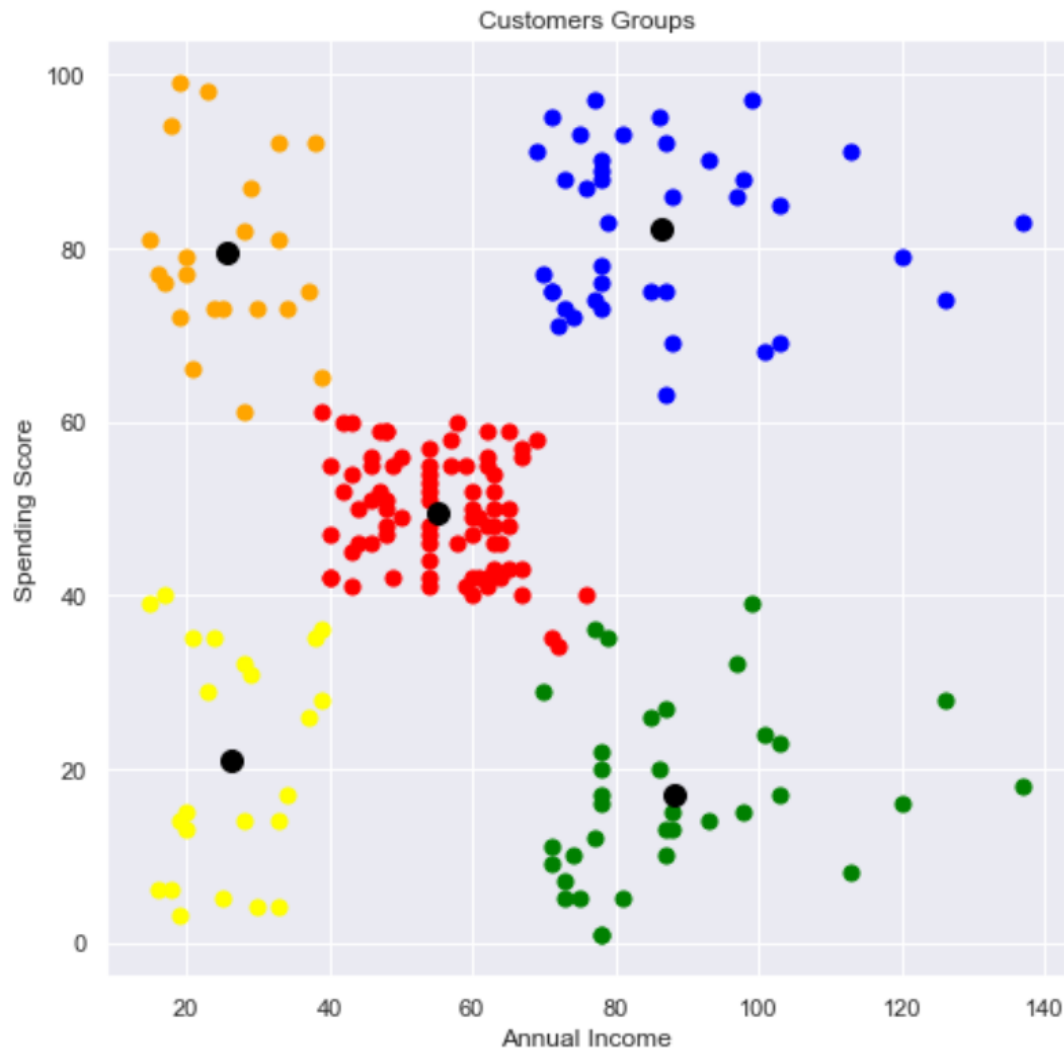
In above Graph the Different colors represents the different clusters and black dots represents the cluster centeroids.

Git-Hub Link :

https://github.com/Patelraj8694/Internship_Projects/tree/main

Schematic Diagram:

In this Prototype there are multiple steps involved. Firstly, data will be collected from different sources. Secondly, Pre-process the data. Thirdly, Analyse the cleaned data, Fourthly, forming the Optimum number of clusters. Fifthly, apply the ML algorithm K-means. Lastly, Visualise the clusters.

Customer Data → Data pre processing → Data Analysis

Siddhard

Visualizing the Clusters ← K-Means Clustering ← $WCSS = \sum_{i \in n}(X_i - Y_i)^2$ Optimum number of Clusters

10. Conclusion:

In Nutshell , I have described the problem statements and the details about this projects And rules and regulations applicable to the Project. In the later segment , I have used K-means Clustering Machine Learning model to identify the relation between the customer income and their spending and forming them in the cluster and visualize it. After that, there is a Diagram which describes the whole process.