REPORT

Major Project Synopsis

On

Customer Segmentation Using K means Clustering

In partial fulfillment of requirements for the degree

Of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Harshita Pandit (20100BTCSAII07163)

Mohammad Nadeem Khan (20100BTCSAII07177)

Ahmed Faraz Nagori (20100BTCSAII08311)

Under the guidance of

Prof: Om Kant Sharma

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE

JULY – DEC 2022

Acknowledgment

We are grateful to the number of persons for their advice and support during the time of completion of our project work. First and foremost our thanks goes to **Dr. Jigyasu Dubey**; head of the Department of Information Technology and **Prof. Om Kant Sharma** the guide of our project for providing us valuable support and necessary help whenever required and also helping us explore new technologies by the help of their technical expertise. This direction, supervision and constructive criticism were indeed the source of inspiration for us.

We would also like to express our sincere gratitude towards our director, **Dr. Anand Rajavat** for providing us valuable support.

We forward our sincere thanks to all the teaching staff and non-teaching staff o Inormation Technology Department – SVIIT, Indore or providing necessary information and there kind co-operation.

We would like to thanks our parents and family members, our classmates and our friends or their motivation and there valuable suggestion during the project. Last, but not the least, we thanks all those people, who have helped us directly or indirectly in accomplishing this work. It has been a privilege to study at Shri Vaishnav Institute of Information Technology, Indore.

Abstract

These days, we can personalize everything. There's no one-size-fits-all approach. But, for business, this is actually a great thing. It creates a lot of space for healthy competition and opportunities for companies to get creative about how they acquire and retain customers. One of the fundamental steps towards better personalization is customer segmentation. This is where personalization starts, and proper segmentation will help us make decisions regarding new features, new products, pricing, marketing strategies, even things like in-app recommendations. As the legal industry emerges from its nascent stages, there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. While many other operators view the state-mandated traceability as a necessary burden, it provides a goldmine for internal customer analysis. Traditionally, segmentation analysis focuses on demographic or RFM (recencyfrequency-monetary) segmentation. Yet, neither of these methods has the capacity to provide insight into a customer's purchasing behavior. With the help of 4Front Ventures, a battle-tested multinational coperator, this report focuses on segmenting customers using specific data and machine learning methods (K-Means and K- means Clustering) to generate new found ways to explore a dispensary's consumer base. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them. Although the results are meaningful, this report could benefit with exploring more clustering algorithms, comparing results across dispensaries within the same state, or investigating segmentations in other state markets.

Introduction

Customer segmentation is important for businesses to understand their target audience. Different advertisements can be curated and sent to different audience segments based on their demographic profile, interests, and affluence level. Customer segmentation simply means grouping your customers according to various characteristics (for example grouping customers by age). It's a way for organizations to understand their customers. Knowing the differences between customer groups, it's easier to make strategic decisions regarding product growth and marketing.

The opportunities to segment are endless and depend mainly on how much customer data you have at your use. Starting from the basic criteria, like gender, hobby, or age, it goes all the way to things like "time spent of website X" or "time since user opened our app".

There are many unsupervised machine learning algorithms that can help companies identify their user base and create consumer segments. In this project we will be using a popular unsupervised learning technique called **K-Means clustering.** This algorithm can take in unlabelled customer data and assign each data point to clusters.

The goal of K-Means is to group all the data available into non-overlapping sub-groups that are distinct from each other. That means each sub-group/cluster will consist of features that distinguish them from other clusters.

K-Means clustering is a commonly used technique by data scientists to help companies with customer segmentation. It is an important skill to have, and most data science interviews will test your understanding of this algorithm/your ability to apply it to real life scenarios.

Problem Domain

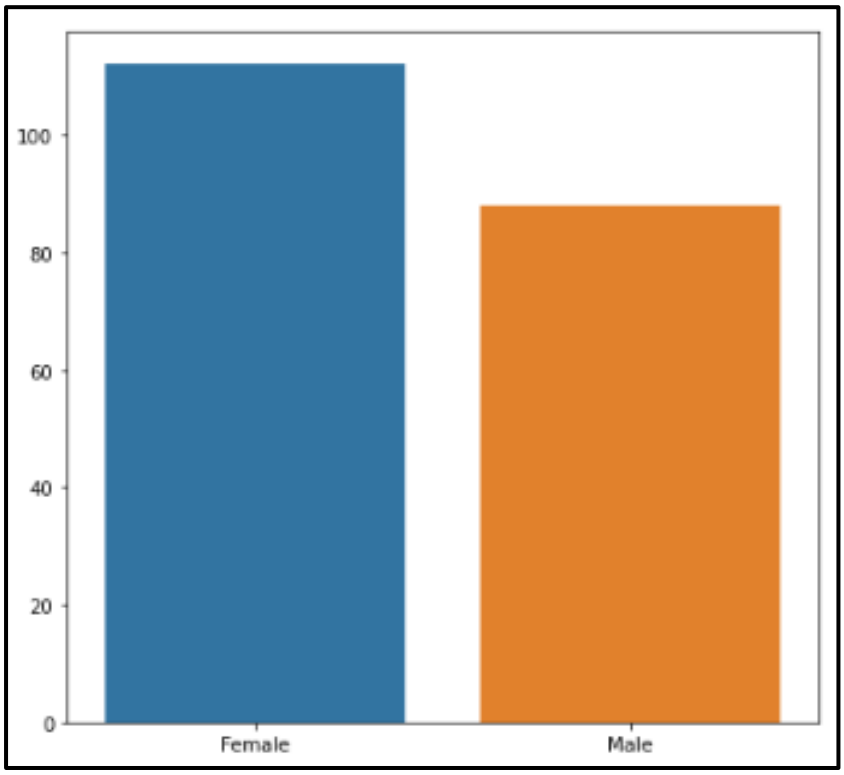
Any company in retail, no matter the industry, ends up collecting, creating, and manipulating 1 data over the course of their lifespan. These data are produced and recorded in a variety of contexts, most notably in the form of shipments, tickets, employee logs, and digital interactions. Each of these instances of data describes a small piece of how the company operates, for better or for worse. The more access to data that one has, the better the picture that the data can delineate. With a clear picture made from data, details previously unseen begin to emerge that spur new insights and innovations.

Any company in retail, no matter the industry, ends up collecting, creating, and manipulating 1 data over the course of their lifespan. These data are produced and recorded in a variety of contexts, most notably in the form of shipments, tickets, employee logs, and digital interactions. Each of these instances of data describes a small piece of how the company operates, for better or for worse. The more access to data that one has, the better the picture that the data can delineate. With a clear picture made from data, details previously unseen begin to emerge that spur new insights and innovations. The sheer size and complicated nature of data in the real world make the above task much easier said than done, though. The rise of performance metrics and interactive dashboards have ushered in a new era of looking at data.

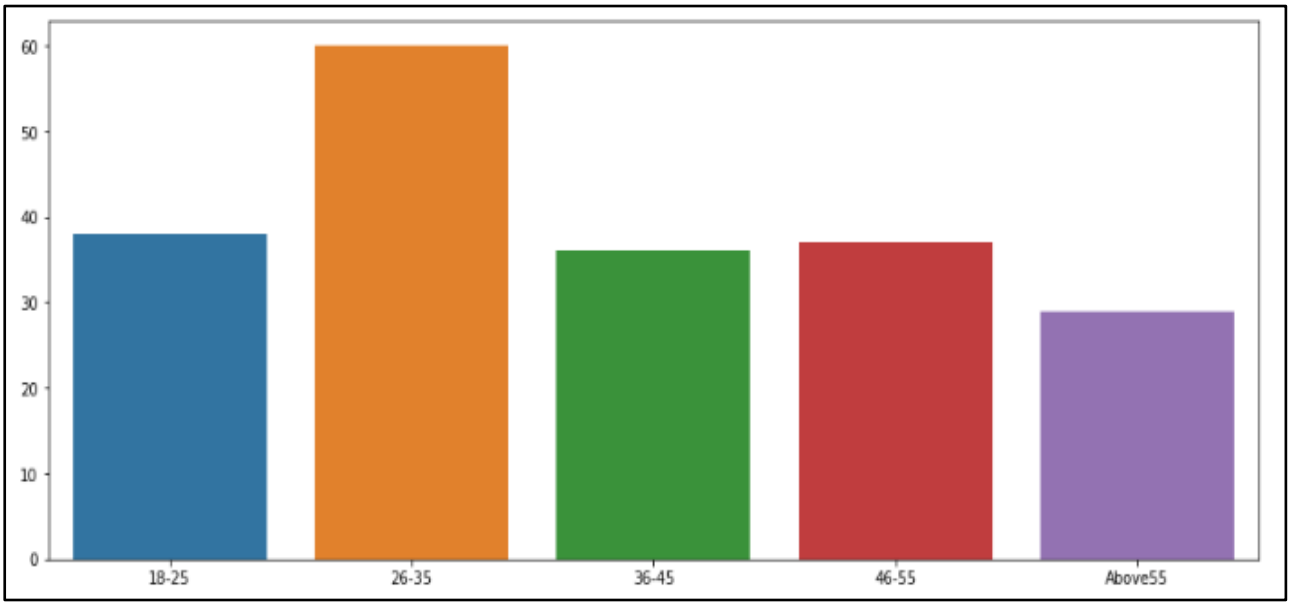
Companies that utilize proper data science and data mining practices allow themselves to dig further into their own operating strategies, which in turn allows them to optimize their commercial practices. As a result, there are increasing motivations for investigating phenomena and data that cannot be simply answered: Why is product B purchased more on the first Saturday of every month compared to other weekends?, If a customer bought product B, will they like product C?, What are the defining traits of our customers? Can we predict what customers will want to buy? It is the latter half of the last question that will be the broad focus of this project work.

Analysis of Data

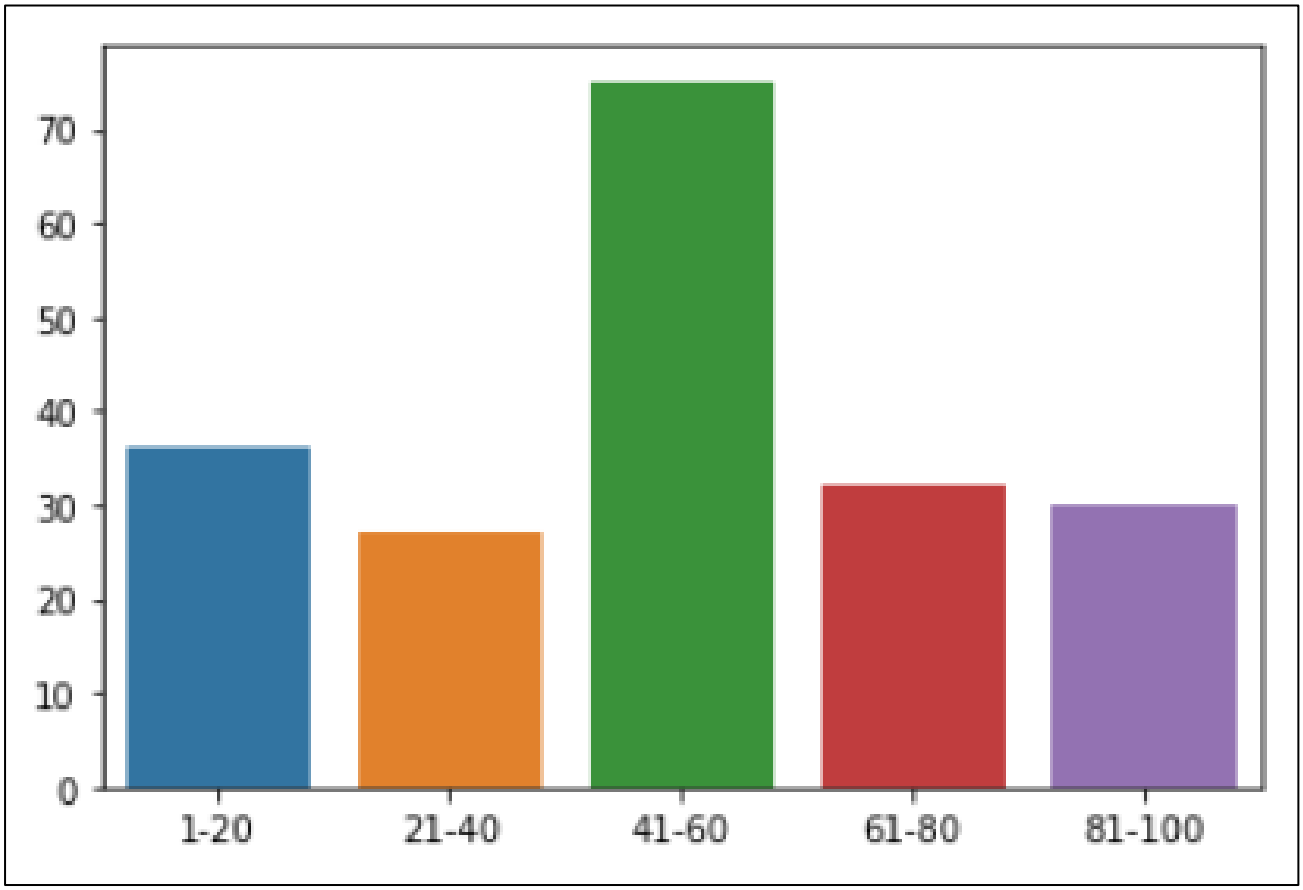
Bar Graph (Gender vs No of Customers)



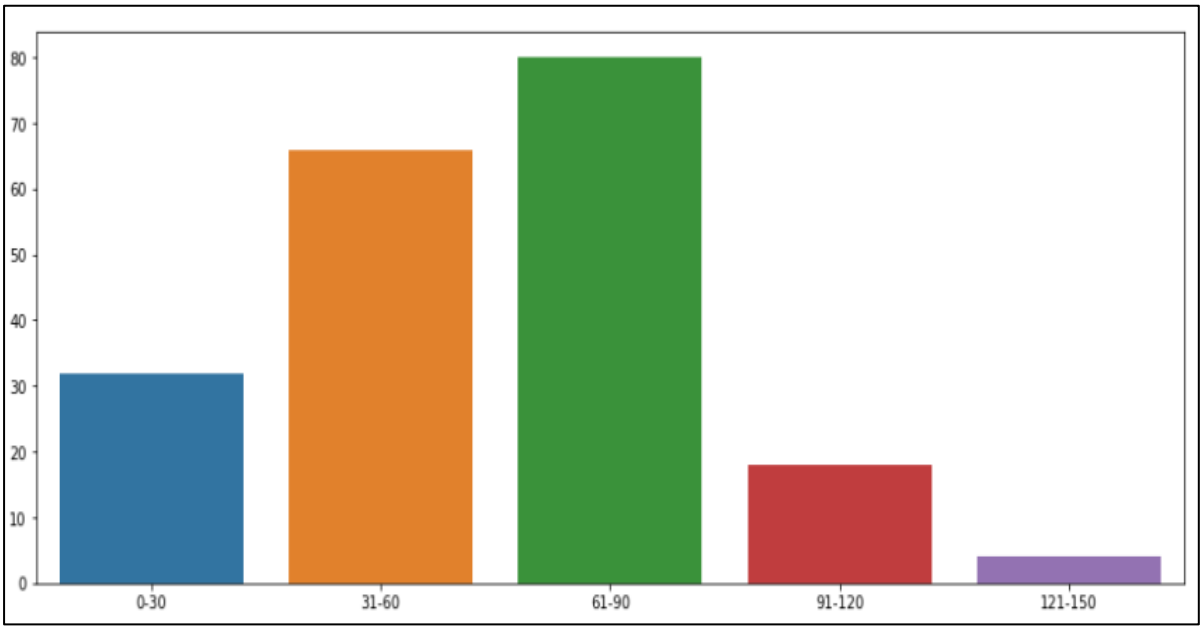
Bar Graph(Age vs No of Customer)



Bar graph(Spending Score vs No of Customers)



Bar Graph (Annual Income vs No of Customers)



Methodology

Demographic, psychographic, behavioral and geographic segmentation are considered the four main methods of customer segmentation, but there are also many other strategies we can use, including numerous variations on the four main methods. Here are several more methods that we may want to look into.

Value segmentation: Some businesses will split up a market based on the "transactional worth" of their customers — how much they're likely to spend on their products. To determine a customer's transactional worth, you can look at previous purchase data such as how many purchases they make, how often they make purchases and the value of the items they purchase.

Firmographic segmentation: Business-to-business (B2B) companies may use firmographic segmentation to divide up the businesses in a market. This is similar to demographic segmentation with individual consumers but instead looks at the characteristics of companies that may become customers. Examples of data to look at include industry, revenue, number of employees and location.

Generational segmentation: Businesses may segment consumers by generation and group them into categories that include Gen Z, Millenials, Generation X, Baby Boomers and the Silent Generation. These generations are believed to share certain preferences, behaviors, personality traits and beliefs. Of course, not every member of a generation is the same, but generational segmentation can give you some additional insight into your audience.

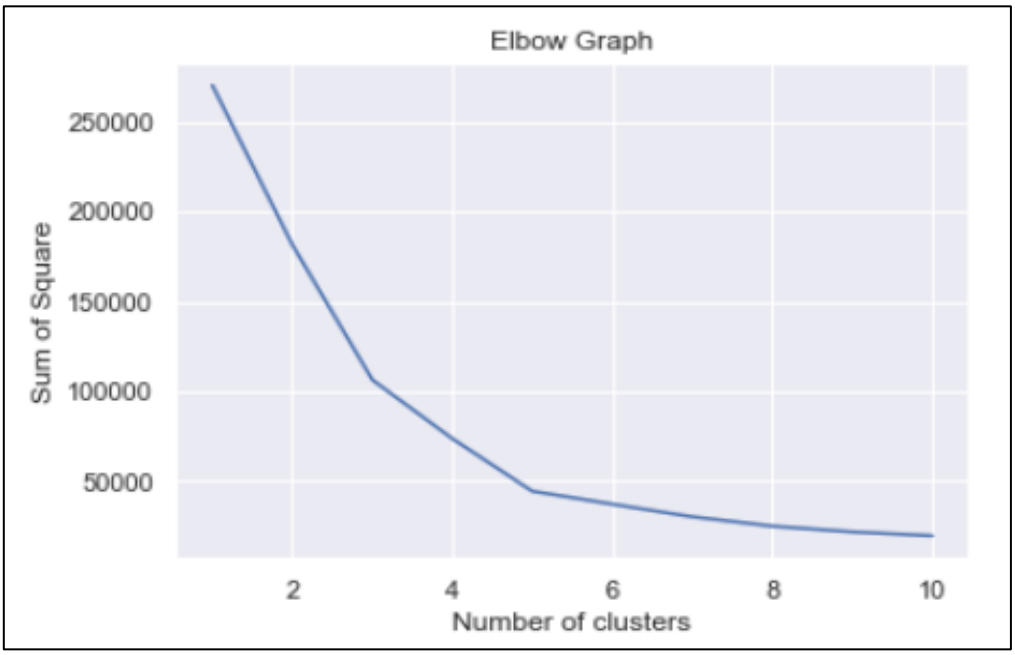
Lifestage segmentation: You can also segment your market into groups based on where they are in their lives. Going to college, getting married and having children are examples of key life events to consider. People at different stages of life need different things. For instance, soon-to-be college students may need apartment furniture. New parents will be looking to purchase baby food.

Seasonal segmentation: Similarly to how people buy different products in different periods of their lives, people also buy different items at different times of the year. Major holidays such as Christmas and Hanukkah also significantly impact purchasing behaviors.

Elbow Method: The Elbow method is used to find out the optimal value to be used in Kmeans In the line chart it resembles the arm with the elbow. The inflection in the curve resembles the underlined model that fits best at that point. So now we will use the elbow method to find out the number of clusters needed. We will first we will import means from sklearn lib and use wcss formula WCSS measures the sum of distances of observations from their cluster centroids which is given by the below formula.

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

In the below graph; when we are analysing the graph there is a cutoff point that means there is a certain drop of point here on the cluster point number 3 and point 5. In the graph we've two elbow points after these two points there is no sharp significant drop for the correct optimum number of clusters.



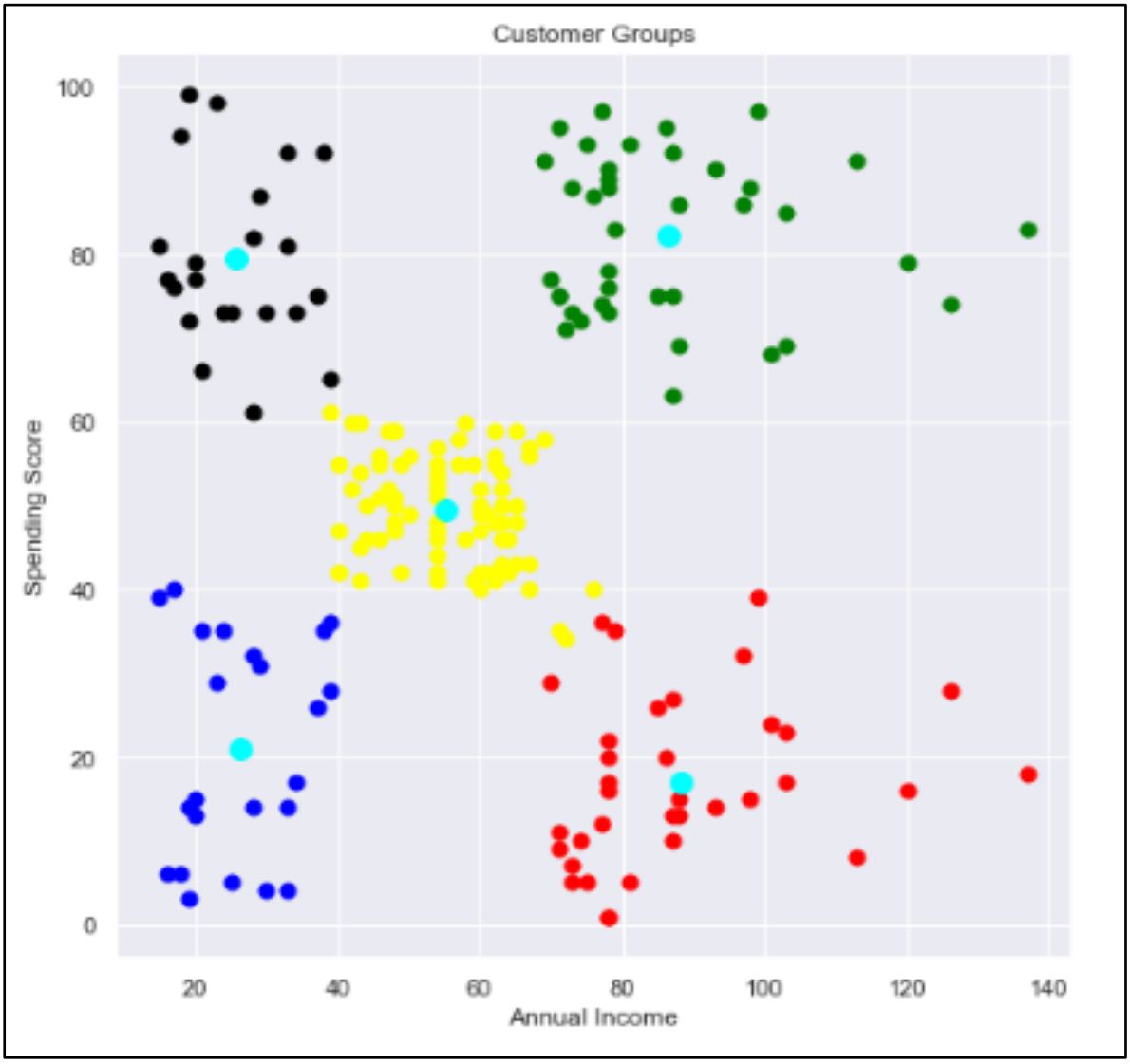
Result

Visualization of Cluster of All data

Here each customer represents one color customer1 as red, customer 2 as blue, customer3 as green, customer4 as orange, and customer5 as pink. The black dots represent the **Centroids** of the cluster.

Centroids are the cluster average for each of the variables; each cluster is defined by a single set of coordinates, the averages of coordinates of all individual observations belonging to that cluster.

We have classified the customers into 5 clusters through which we can see that customer1 is having average spending scores with the average income so this range of customers can be targeted in order to increase sales.



Conclusion

We have successfully built a K-Means clustering model for customer segmentation. We also explored cluster interpretation, and analyzed the behaviour of individuals in each cluster.

Finally, we took a look at some business recommendations that could be provided based on the attributes of each individual in the cluster.

We can use the analysis above as starter code for any clustering or segmentation project in the future.

CODE FILE

Shri Vaishnav Vidyaapeeth Vishwavidyalaya, Indore

Department of Computer Science and Engineering

Session 2021-22

3rd Year

V Semester

Subject: Introdction to Data Science

Subject Code: BTIBM505

Major Project Code File

Prepared by:-

Harshita Pandit

Ahmed Faraz Nagori

Mohammd Nadeem Khan

Submitted to:-

Om Kant Sharma

CODE:-

- **Importing the library**

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

- **Importing function from cluster model in sklearn library**

```
from sklearn.cluster import KMeans  #for forming clusters of customer
```

- Reading the Dataset

```
customerData=pd.read_csv("C:\\Users\\Harshita\\Desktop\\customers.csv")    #from csv file "customers.csv"
```

- Visulazing the Dataset

```
#Bar Graph(Gender vs No of Customers)
genders=customerData.Gender.value_counts()
plt.figure(figsize=(7,7))
sns.barplot(x=genders.index, y=genders.values)
plt.show()

#Bar Graph(Age vs  No of Customer)
df=customerData
age18_25 = df.Age[(df.Age<=25)&(df.Age>=18)]
age26_35 = df.Age[(df.Age<=35)&(df.Age>=26)]
age36_45 = df.Age[(df.Age<=45)&(df.Age>=36)]
age46_55 = df.Age[(df.Age<=55)&(df.Age>=46)]
age55above = df.Age[(df.Age>=56)]

x=["18-25","26-35","36-45","46-55","Above55"]
```

```
y=[len(age18_25.values),
len(age26_35.values),len(age36_45.values),len(age46_55.values),len(age55
above.values)]


plt.figure(figsize=(15,6))
plt.title=("Number of customers and ages")
plt.xlabel=("Ages")
plt.ylabel=("Number of customers")
sns.barplot(x=x,y=y)
plt.show()


#Bar graph(Spending Score vs No of Customers)
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)"]<=100)]

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

sns.barplot(x=x , y=y)
plt.figure(figsize=(10,20))
plt.title=("Spending scores of the customers")
plt.xlabel=("Spending Scores")
plt.ylabel=("score of customers")
plt.show()
```

```
#Bar Graph (Annual Income vs No of Customers)
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)]

x=["0-30","31-60", "61-90","91-120","121-150"]
y=[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=x,y=y,)
plt.title=("Annual Income of customers")
plt.xlabel("Annual Income in k$ ")
plt.ylabel("Number of customers")
plt.show()
```

- Analysing Dataset

```
#printing first 5 rows of teh Data Frame
customerData.head(8)
#to get the number of rows and columns from the data
customerData.shape


#Information about the dataset
customerData.info()


#Checking for the mission values in Dataset
customerData.isnull().sum()
x=customerData.iloc[ : , [3,4] ].values
```

- **Finding no. of Clusters using ELBOW method**

```
#choosing number of clusters
L = []
for i in range(1,11):
km=KMeans(n_clusters=i,init='k-means++',random_state=40 )
km.fit(x)
L.append(km.inertia_)
```

- Plotting the elbow graph.....

```
sns.set()
plt.plot(range(1,11),L)
plt.title('Elbow Graph')
plt.xlabel('Number of clusters')
plt.ylabel('Sum of Square')
plt.show()
#Optimum number of cluster will be 5
#Training the Model(K Means Clustering Model)
km=KMeans(n_clusters=5,init='k-means++' ,random_state=0)
y=km.fit_predict(x)
y
```

- **Visualization**

```
#Visualizing all the clusters

plt.figure(figsize=(8,8))

plt.scatter(x[y==0,0], x[y==0,1], s=50, c='red', label='1')
plt.scatter(x[y==1,0], x[y==1,1], s=50, c='yellow', label='2')
plt.scatter(x[y==2,0], x[y==2,1], s=50, c='green', label='3')
plt.scatter(x[y==3,0], x[y==3,1], s=50, c='black', label='4')
plt.scatter(x[y==4,0], x[y==4,1], s=50, c='blue', label='5')

plt.scatter(km.cluster_centers_[ : ,0],km.cluster_centers_[ : ,1],
s=100, c='cyan', label='Centroids'    )

plt.title('Customer Groups')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.show()
```

POWERPOINT

PRESENTATION