Project Title: Customer Segmentation Using K-Means Clustering

Author: Yogapriya P

Internship: Infotact Solutions – Data Analytics Internship

Duration: 2 Months

Objective :

This project focuses on customer segmentation using K-Means clustering. The goal is to categorize customers into distinct groups based on their behaviour or purchasing patterns to help businesses personalize marketing strategies.

- Identifying customer segments from raw data
- Use clustering to find similar behaviour groups
- Providing actionable insights for business decisions

Tools and Technologies :

- Programming Language: Python
- Libraries: Pandas, Matplotlib, Seaborn, Sklearn, Streamlit, Joblib
- Jupyter Notebook and VS code
- Visualization: matplotlib & seaborn

Data Understanding and Data Precessing

- Dataset : https://www.kaggle.com/datasets/vishakhdapat/customer-segmentation-clustering
- Initial exploration showed missing values and outliers, which were handled during preprocessing.
- Created new columns for Age, Total Children, Total Spending, Customer Since.

Exploratory Data Analysis (EDA)

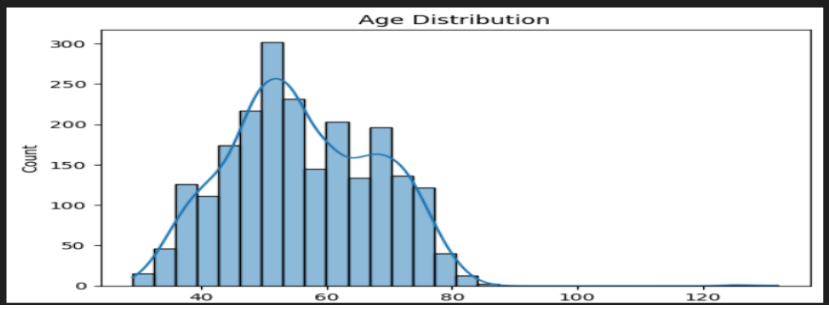
- Histograms and boxplots were used to understand distributions
- Correlation matrix showed strong relationships between some features. Visualized clusters using 2D scatter plots.

1) Histogram and Boxplots:

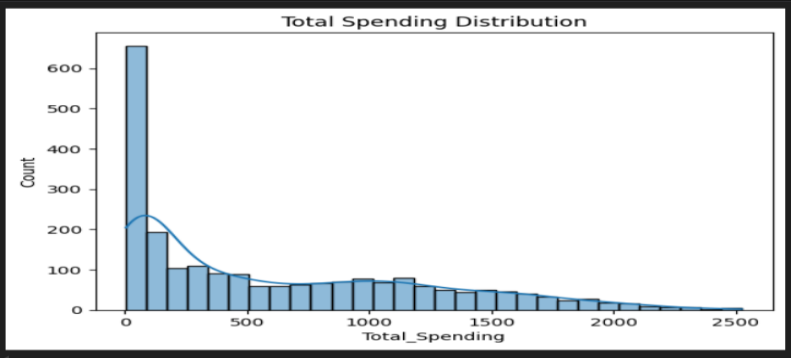
A **histogram** shows the **distribution of a single variable** — i.e., how frequently each range of values occurs. Here we used histogram for **Age Distribution, Income Distribution, Total Spending Distribution.**

```
sns.histplot(dataset["Age"], bins =30, kde = True)
plt.title("Age Distribution")
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

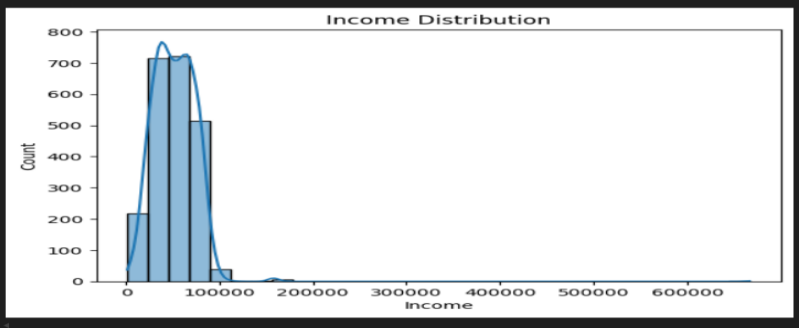



```
Click to add a breakpoint t["Total_Spending"], bins= 30, kde = True)
plt.title("Total Spending Distribution")
plt.show()
```




```
sns.histplot(dataset["Income"], bins = 30 , kde= True)
plt.title ("Income Distribution")
plt.show
```

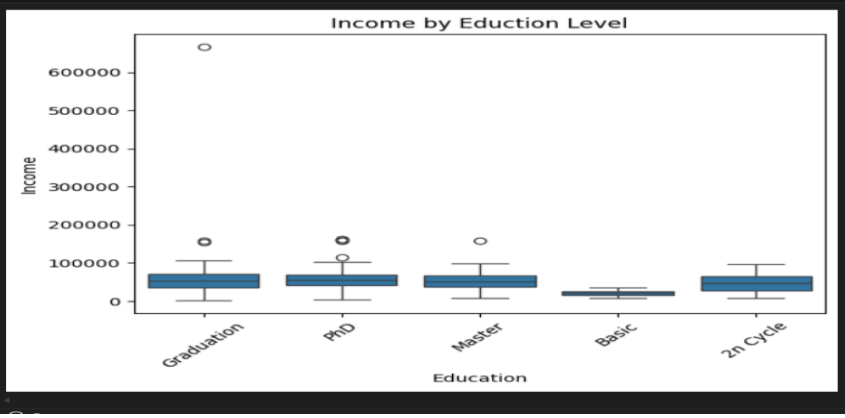
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



A **box plot** shows the **summary statistics** of a variable. Here it uses to detect **outliers** Compare **distribution between categories** like

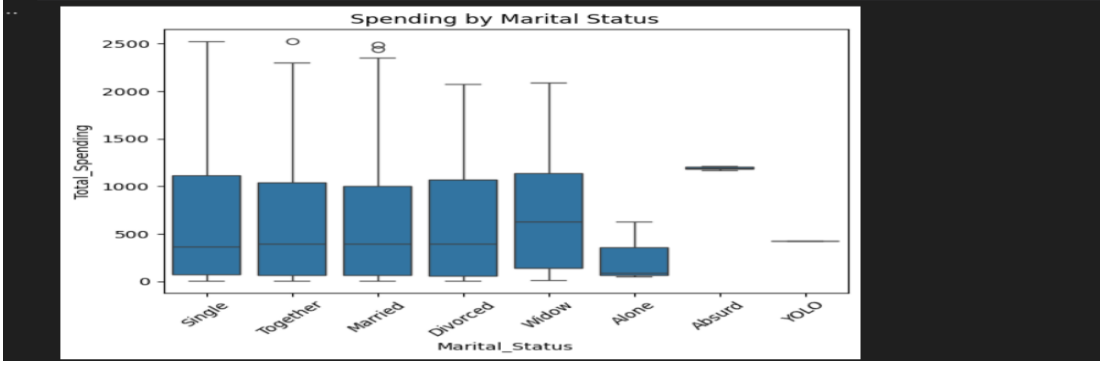
- relationship between **Education and Income**.

```
sns.boxplot(x= "Education", y= "Income", data= dataset)
plt.xticks(rotation = 45)
plt.title( "Income by Eduction Level")
plt.show()
```



- relationship **between Spending by Marital Status and Marital Status.**

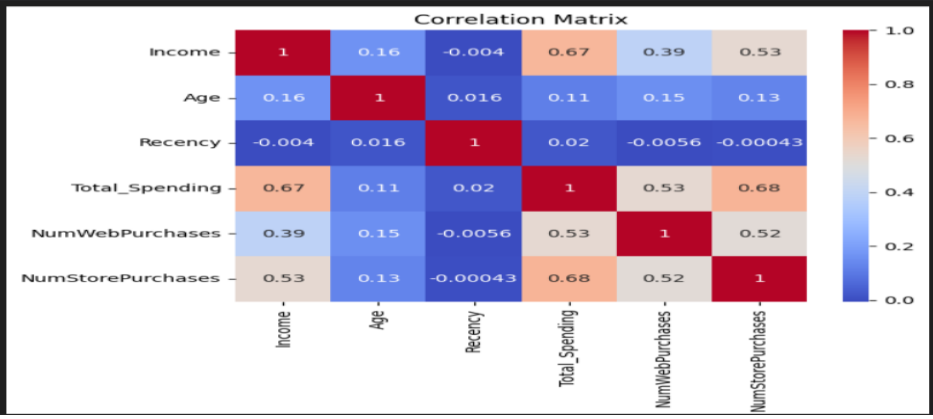
```
sns.boxplot(x="Marital_Status", y="Total_Spending", data=dataset)
plt.xticks(rotation =45)
plt.title("Spending by Marital Status")
plt.show()
```



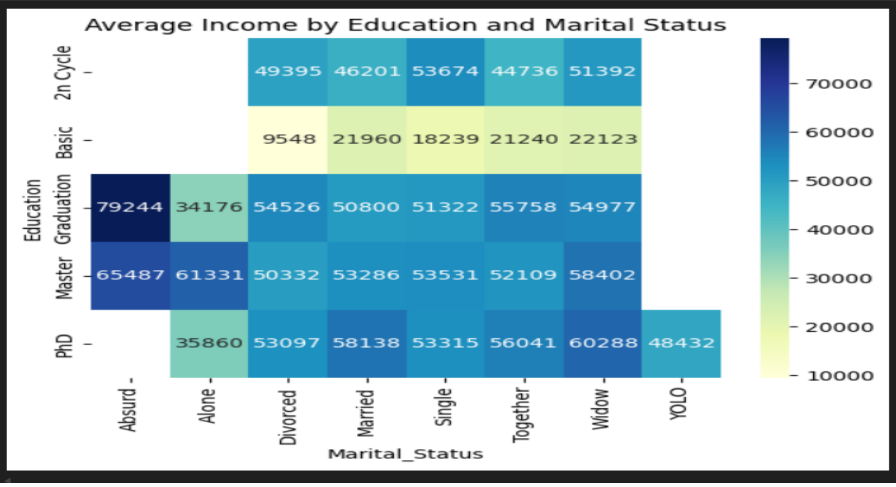
2) Correlational Matrix :

The correlation coefficients between **multiple variables** in a dataset. Each cell in the matrix shows the **relationship strength** between two variables. **Income,Age, Recency, Total Spending, No. of Purchases in Web and Store**.

```
sns.heatmap(corr, annot = True, cmap = "coolwarm")
plt.title("Correlation Matrix")
plt.show()
```

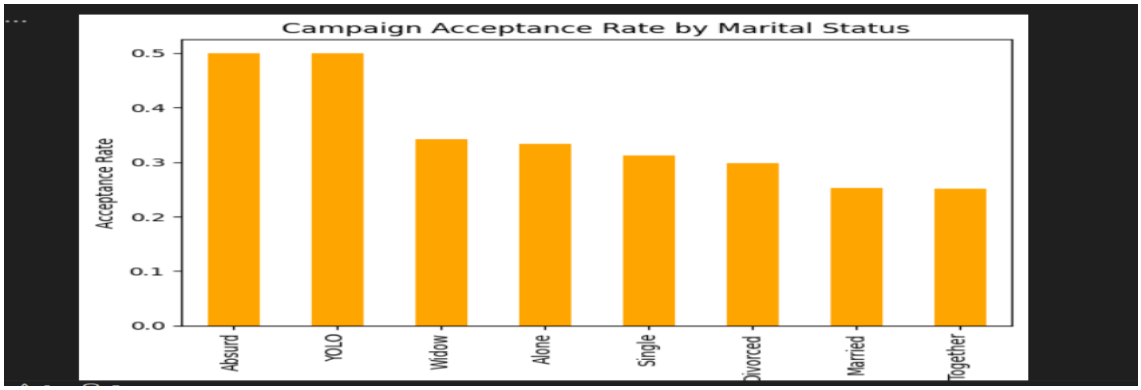


```
sns.heatmap(pivot_income, annot = True, fmt= ".0f", cmap="YlGnBu")
plt.title("Average Income by Education and Marital Status")
plt.show()
```

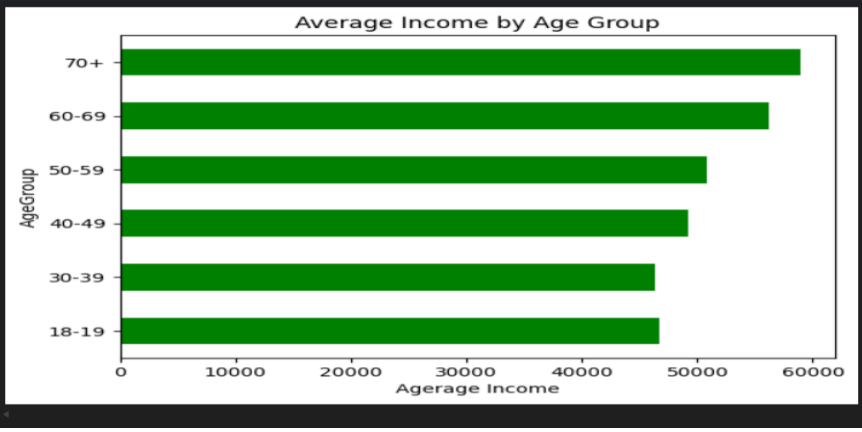


Business Analysis :

Forming several groups and their specified plot. Plots for Average Spending by Education and Campaign Acceptance Rate by Marital Status gives customer approach towards their needs. Average Income by age Group provide essential informations.




```
group3.plot(kind="barh", color ="green")
plt.title("Average Income by Age Group")
plt.xlabel("Agerage Income")
plt.show()
```



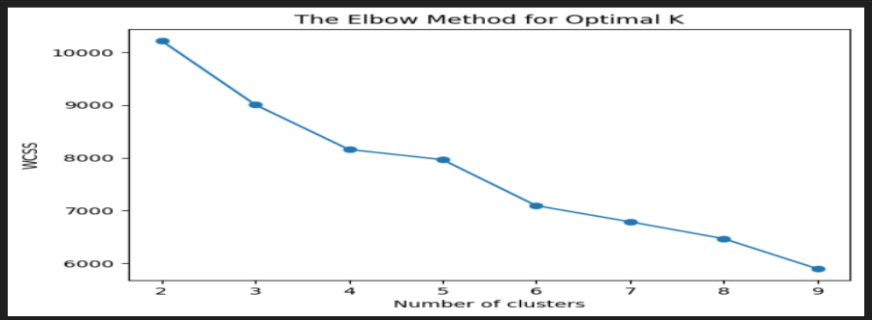
MODELLING K-MEANS CLUSTERING :

Applied StandardScaler for normalization and Selected relevant features for clustering. Finally our features are,

features = ["Age", "Income", "Total_Spending", "NumWebPurchases","NumStorePurchases", "NumWebVisitsMonth", "Recency"]

- Used KMeans from sklearn.cluster.
- Chose optimal number of clusters using **WCSS** and **Elbow Method.**
- Applied clustering and visualized the segments.

```
plt.plot(range(2,10), WCSS, marker="o")
plt.title("The Elbow Method for Optimal K")
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

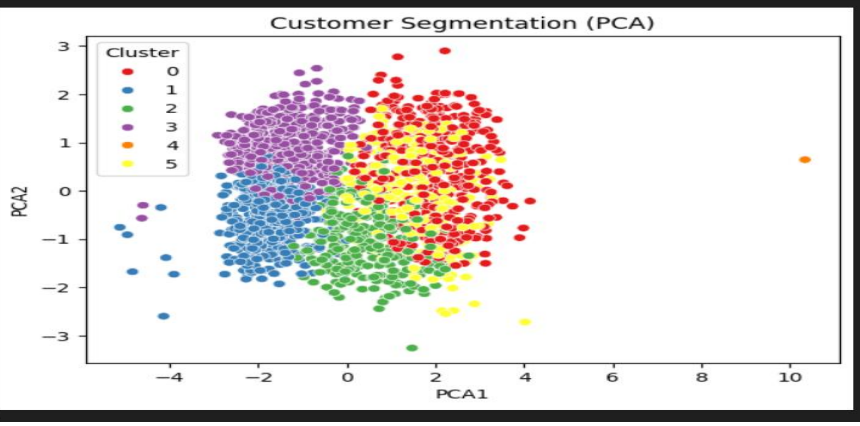



```
cluster_summary
```

Cluster	Age	Income	Total_Spending	NumWebPurchases	NumStorePurchases	NumWebVisitsMonth	Recency
0	55.988395	37082.862669	135.058027	2.446809	3.491296	6.369439	75.835590
1	46.122257	79012.852665	1320.576803	4.564263	8.551724	2.598746	48.952978
2	69.945392	74080.593857	1182.880546	4.436860	8.337884	2.443686	54.180887
3	58.910653	61402.103093	1007.597938	8.333333	8.491409	6.292096	62.384880
4	50.675676	31671.886100	81.113900	1.870656	2.980695	6.646718	27.264479
5	60.964029	55479.517986	620.064748	5.892086	6.705036	6.025180	20.276978

Results and Insights :

```
sns.scatterplot(x="PCA1", y="PCA2", hue="Cluster", data =dataset, palette="Set1")
plt.title("Customer Segmentation (PCA)")
plt.show()
```



Cluster 0: Budget Customers - This cluster has the lowest income & spending. They are also older and have not made a purchase recently.

Cluster 1: Premium Customers - This is the highest-earning cluster, and they are also the highest spenders. They are younger and have made a purchase recently.

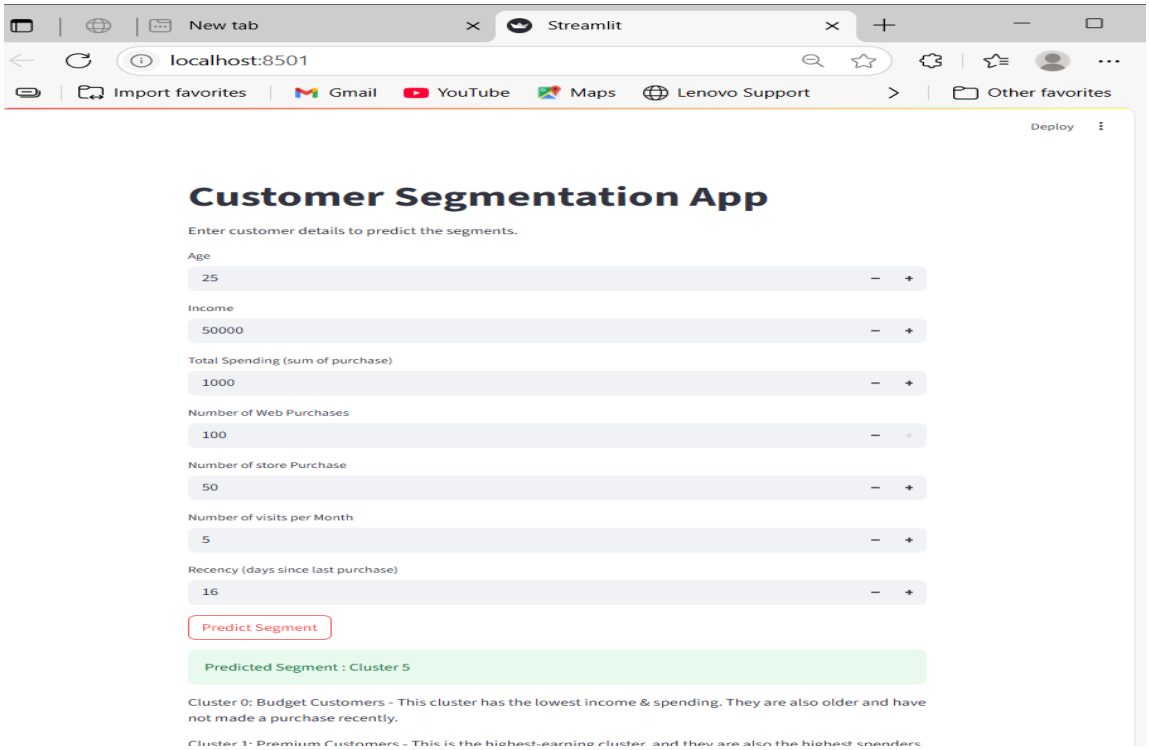
Cluster 2: Senior Spenders - This cluster is the oldest and has the second-highest income and spending.

Cluster 3: Digital Buyers - This group has a high number of web purchases and a low number of in-store purchases. They have a high income and are of average age.

Cluster 4: Inactive Customers - This cluster has the lowest recency, meaning they have not made a purchase in a long time. They also have low income and spending.

Cluster 5: Frequent Buyers - This cluster has the highest number of purchases, both online and in-store. They have an average income and are of average age.

FINAL DEPLOYMENT IN STREAMLIT:



CONCLUSION :

K-Means clustering helped uncover patterns in customer data. The segments can improve decision-making for marketing, product recommendations, and customer engagement.

Links

GitHub: https://github.com/YogaPriya2000/Customer-Segmentation-KMeans-Clustering-

Linkedin: https://www.linkedin.com/feed/update/urn:li:activity:7355552242793373696/

YouTube: https://lnkd.in/gSih6NUF