**CUSTOMER SEGMENTATION**

**DATA SCIENCE PROJECT**

NAME: Alhan Charan Beshra

**PROJECT REPORT**

**Project Overview**

The Customer Segmentation project aims to analyze and segment customers based on their purchasing behaviors and characteristics. This segmentation allows for targeted marketing, improved customer service, and more informed business decisions.

**Data Preprocessing**

It involves cleaning and transforming raw data into a format suitable for analysis.

**Data Cleaning:**

* **Handling Missing Values:** Missing values can be imputed (filled in) or removed based on the nature and amount of missing data. Common imputation methods include filling with the mean, median, or mode, or using more sophisticated methods such as interpolation or machine learning techniques.
* **Removing Duplicates:** Duplicate records can distort the results and should be identified and removed.

**Initial Data Exploration:**

* **Understanding the Data:**
  + **Shape of Data:** Check the number of rows and columns to understand the dataset's size.
  + **Data Types:** Examine the data types of each column to ensure they are appropriate for analysis (e.g., numerical, categorical).
* **Descriptive Statistics:**
  + **Summary Statistics:** Compute summary statistics like mean, median, and standard deviation to get an idea of the data distribution and variability.
  + **Value Counts:** For categorical columns (e.g., 'Gender'), view the distribution of categories using value counts.

**Data Visualization**

It is a crucial step in data analysis that involves creating graphical representations of data to uncover patterns, trends, and insights. Effective visualization helps in understanding complex datasets, communicating findings, and guiding decision-making.

**Gender Distribution:**

A graph with a bar and a number of columns

Description automatically generated with medium confidence

**Orders Distribution:**

A graph of different sizes and colors

Description automatically generated

**Brands Search and Orders Analysis:**

A screenshot of a graph

Description automatically generated

**Correlation Heatmap:**

A graph of a number of data

Description automatically generated with medium confidence

**Total Searches Visualization:**

A chart of a graph

Description automatically generated with medium confidence

**Data Scaling**

It is a crucial preprocessing step in machine learning and data analysis, especially when dealing with algorithms that are sensitive to the scale of input features. Scaling transforms features to a common scale without distorting differences in the range of values.

* **Min-Max Scaling (Normalization):** Min-Max Scaling transforms features to a fixed range, usually [0, 1]. This is useful when the features need to be bounded within a specific range.

**Elbow Method**

It is used to determine the optimal number of clusters in a dataset for clustering algorithms like K-Means. It helps in identifying the point at which adding more clusters yields diminishing returns in terms of within-cluster variance.

**Elbow Graph:**

A graph with a line

Description automatically generated

**K-Elbow Visualiser:**

A graph of a distorsion

Description automatically generated

Optimal Clusters: 3

**Silhouette Method**

It is used to evaluate the quality of clusters in clustering algorithms such as K-Means. It provides a measure of how similar each data point is to its own cluster compared to other clusters, helping to determine the optimal number of clusters.

**Silhouette Graph:**

A graph with blue lines

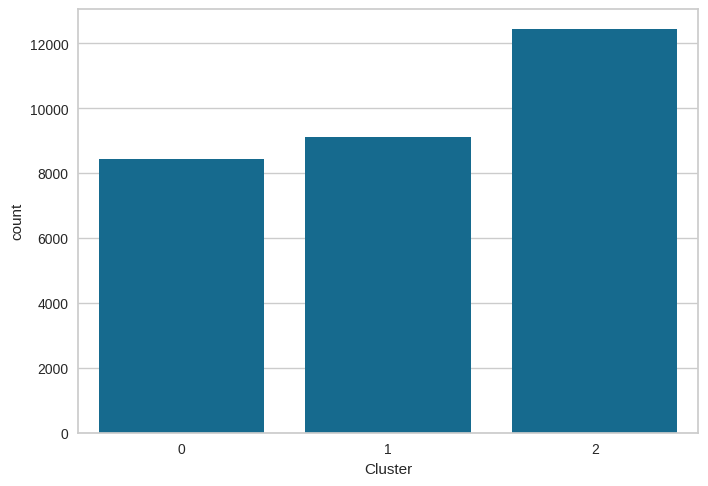
Description automatically generated

Optimal Clusters: 3

**K-Means Model**

It is used to partition a dataset into K distinct, non-overlapping groups or clusters. Each cluster is defined by its centroid, and the algorithm iterates to minimize the variance within each cluster.

Here K=3, so we get:



**Cluster Analysis**

After performing K-Means clustering algorithm, we analyze and interpret the clusters to understand their characteristics and significance.

**Cluster 0:**

A close-up of a graph

Description automatically generated

**Cluster 1:**

A close-up of a graph

Description automatically generated

**Cluster 2:**

A screenshot of a graph

Description automatically generated

**Overall Analysis**

We can understand the impact of the clustering process, evaluating the performance of the clustering algorithm, and deriving actionable insights from the segmented data through the plots.

A graph with a bar and a number of columns

Description automatically generated with medium confidence

A graph of blue rectangular bars

Description automatically generated with medium confidence

**Conclusion**

The customer segmentation analysis has successfully categorized customers into 3 distinct clusters based on their purchasing behavior and search activity. This segmentation will aid in developing targeted marketing strategies and enhancing customer engagement by understanding the specific needs and behaviors of each group.