Customer Segmentation Analysis Report

Rajesh Ragi

JULY 2024

Introduction

This report details the analysis of a customer dataset to segment customers based on their order and search behavior. Identifying distinct customer segments allows businesses to tailor their marketing strategies and enhance customer satisfaction. The analysis includes data cleaning, exploratory data analysis (EDA), data visualization, and clustering using the K-means algorithm.

Methods and Analysis

Data Description

The dataset includes the following columns:

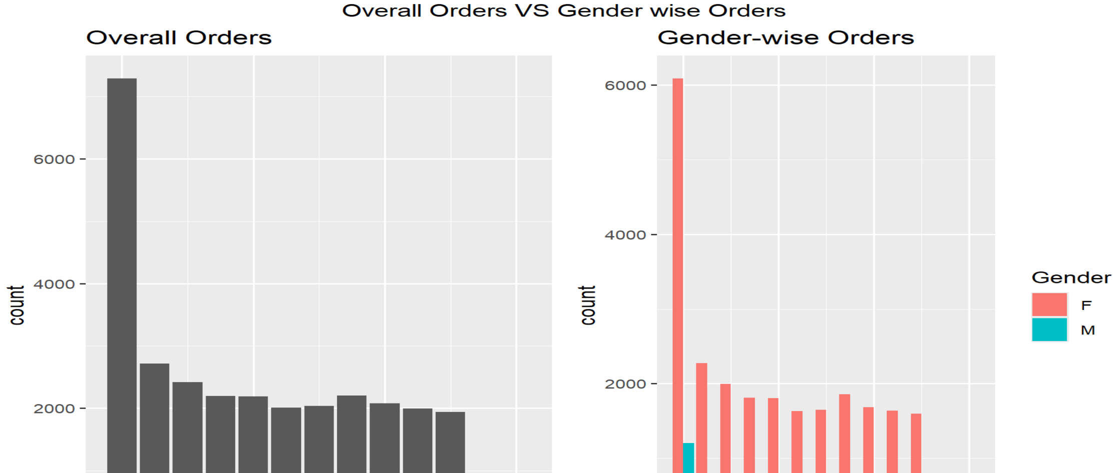
- **`Cust_ID`**: Customer ID
- **`Gender`**: Gender of the customer
- **`Orders`**: Number of orders placed by the customer
- Various columns representing the number of searches for different brands.

Data Cleaning

- **Checking for Duplicates and Missing Values:** We checked for duplicate rows and missing values in the dataset. No duplicates were found. Missing values in the `Gender` column were imputed with the most frequent value.

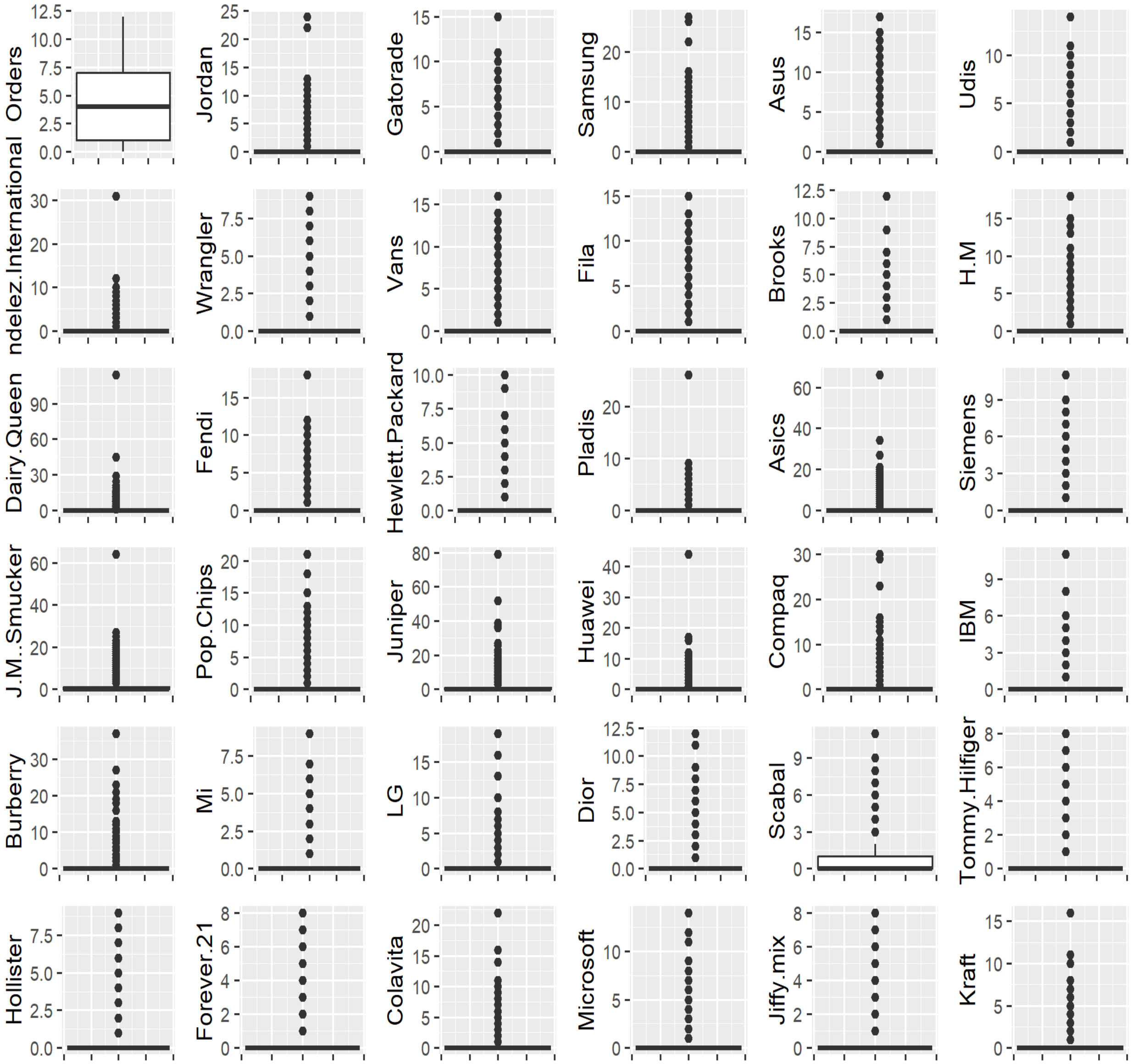
Exploratory Data Analysis (EDA)

- **Summary Statistics:** Provided an overview of the dataset's structure and summary statistics.
- **Gender Distribution:**



- **Explanation:** This bar plot shows the distribution of customers by gender, helping us understand the gender composition of our customer base.

Orders Distribution:

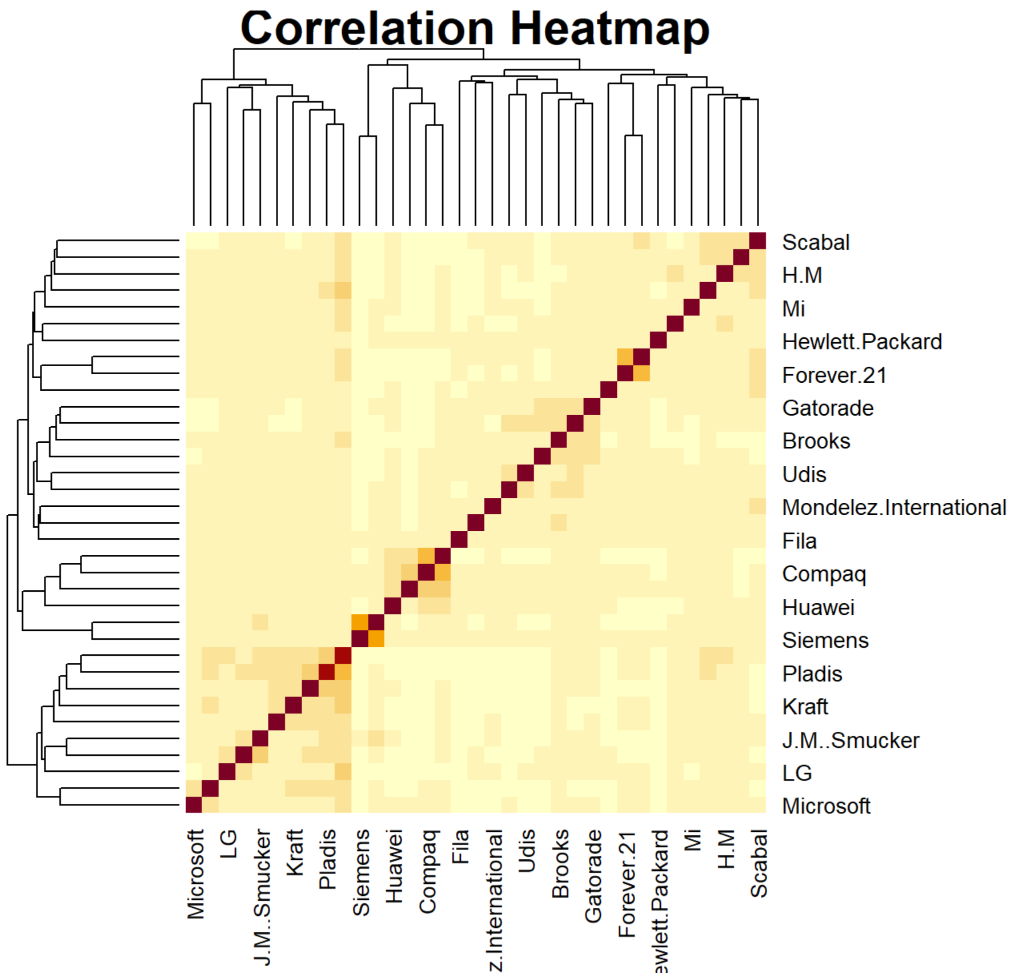


- The first plot shows the overall distribution of orders, while the second plot shows the gender-wise distribution of orders. This helps us understand the ordering behavior of different genders.

- **Boxplots for Brand Searches:**

These boxplots show the distribution of searches for different brands, helping to identify any outliers and the spread of search behaviors across different brand.

- **Data Visualization**
- **Correlation Heatmap:**

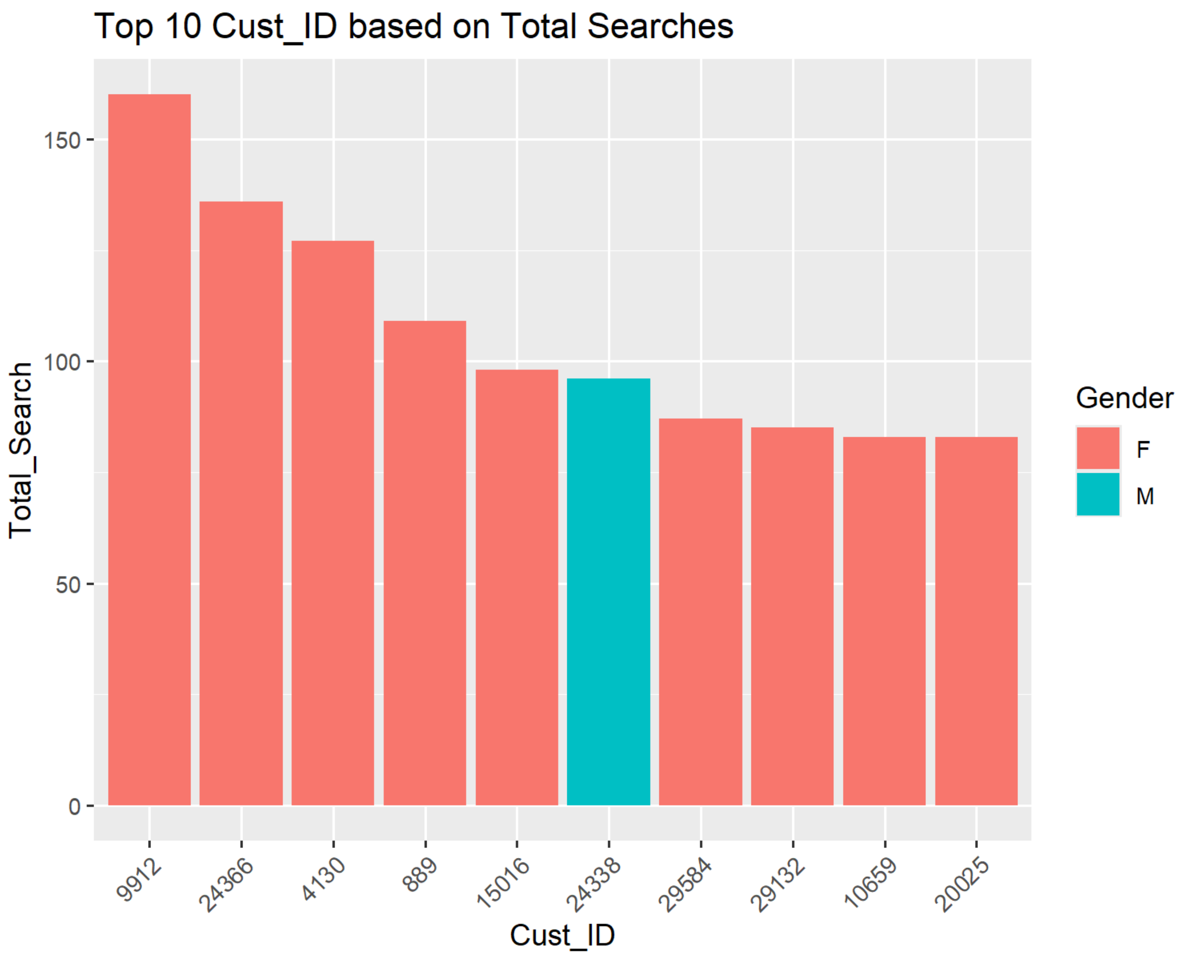


- ○ The heatmap shows the correlation between the search behaviors for different brands. High correlations indicate that customers who search for one brand are likely to search for another.

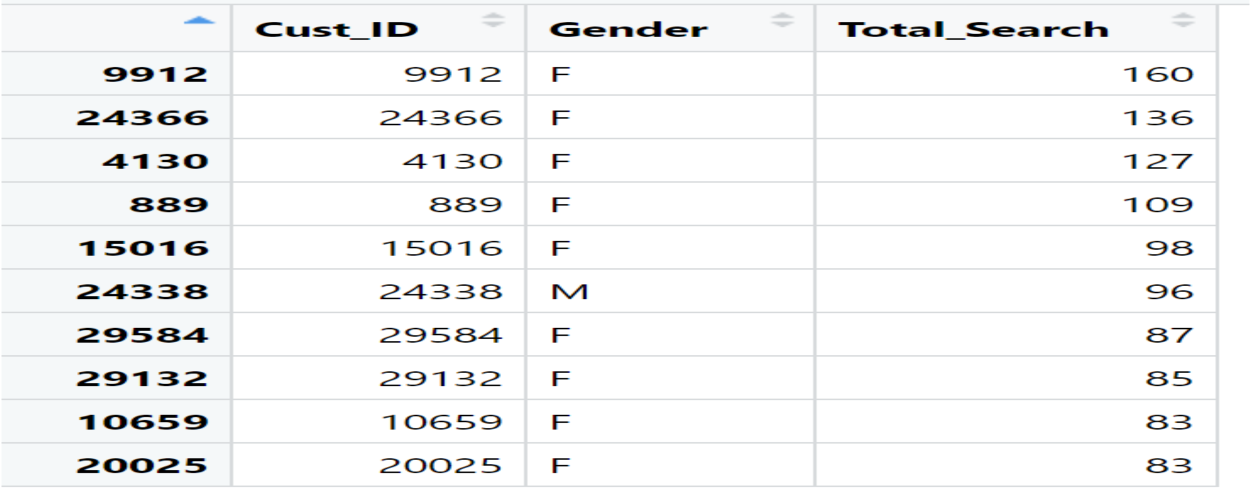
- **Histograms for Selected Columns:**

These histograms show the distribution of values for selected columns, providing insights into the spread and frequency of different values.

- ## Top 10 Customers by Total Searches:



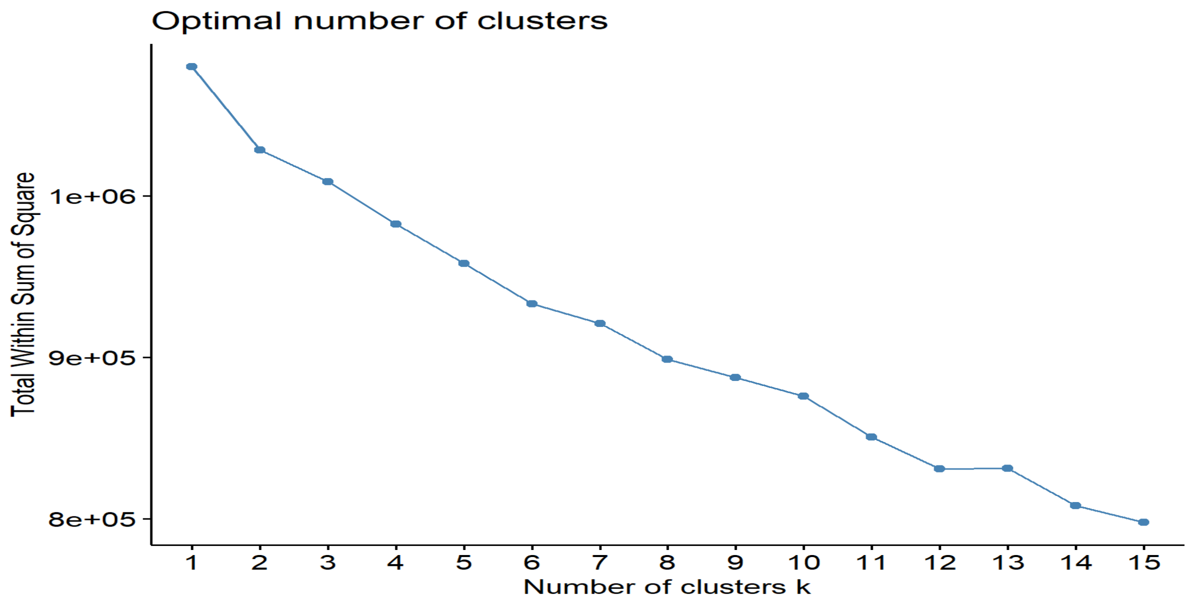
  - This bar plot shows the top 10 customers based on total searches, providing insights into the most active customers.

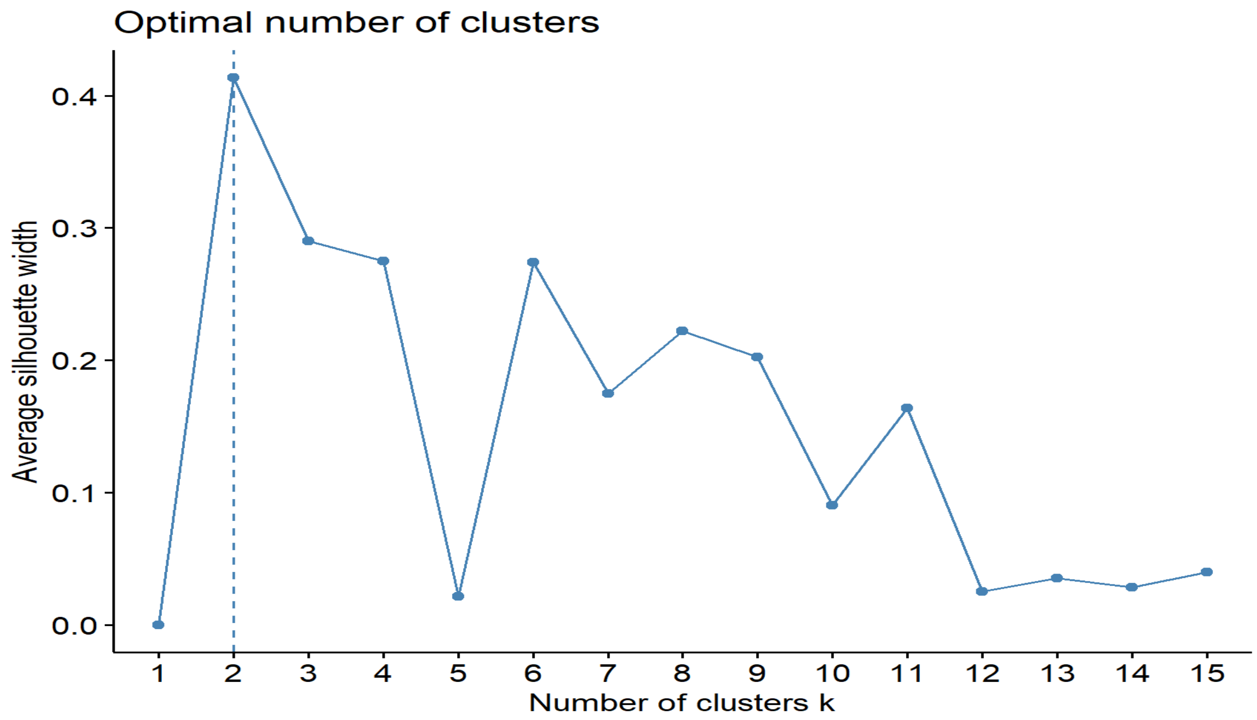


	Cust_ID	Gender	Total_Search
9912	9912	F	160
24366	24366	F	136
4130	4130	F	127
889	889	F	109
15016	15016	F	98
24338	24338	M	96
29584	29584	F	87
29132	29132	F	85
10659	10659	F	83
20025	20025	F	83

Clustering

- **Scaling the Features:** The features were standardized to ensure they contribute equally to the clustering process.
- **Determining Optimal Number of Clusters:**
    - **Elbow Method:**



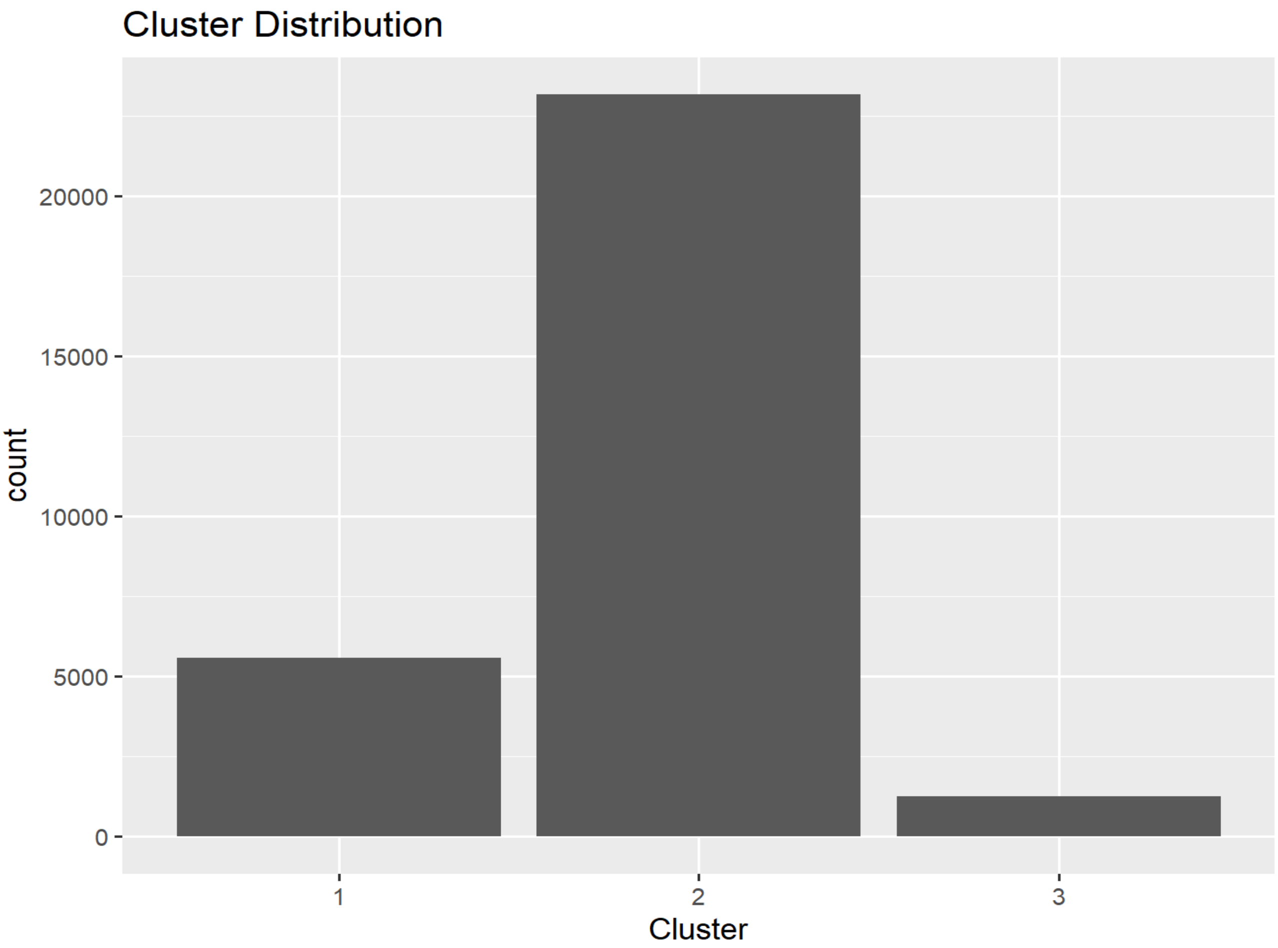


The elbow method plot helps determine the optimal number of clusters by identifying the point where the within-cluster sum of squares (WSS) starts to level off.

Silhouette Score:

The silhouette score plot helps assess the quality of clustering for different numbers of clusters, with higher scores indicating better-defined clusters.

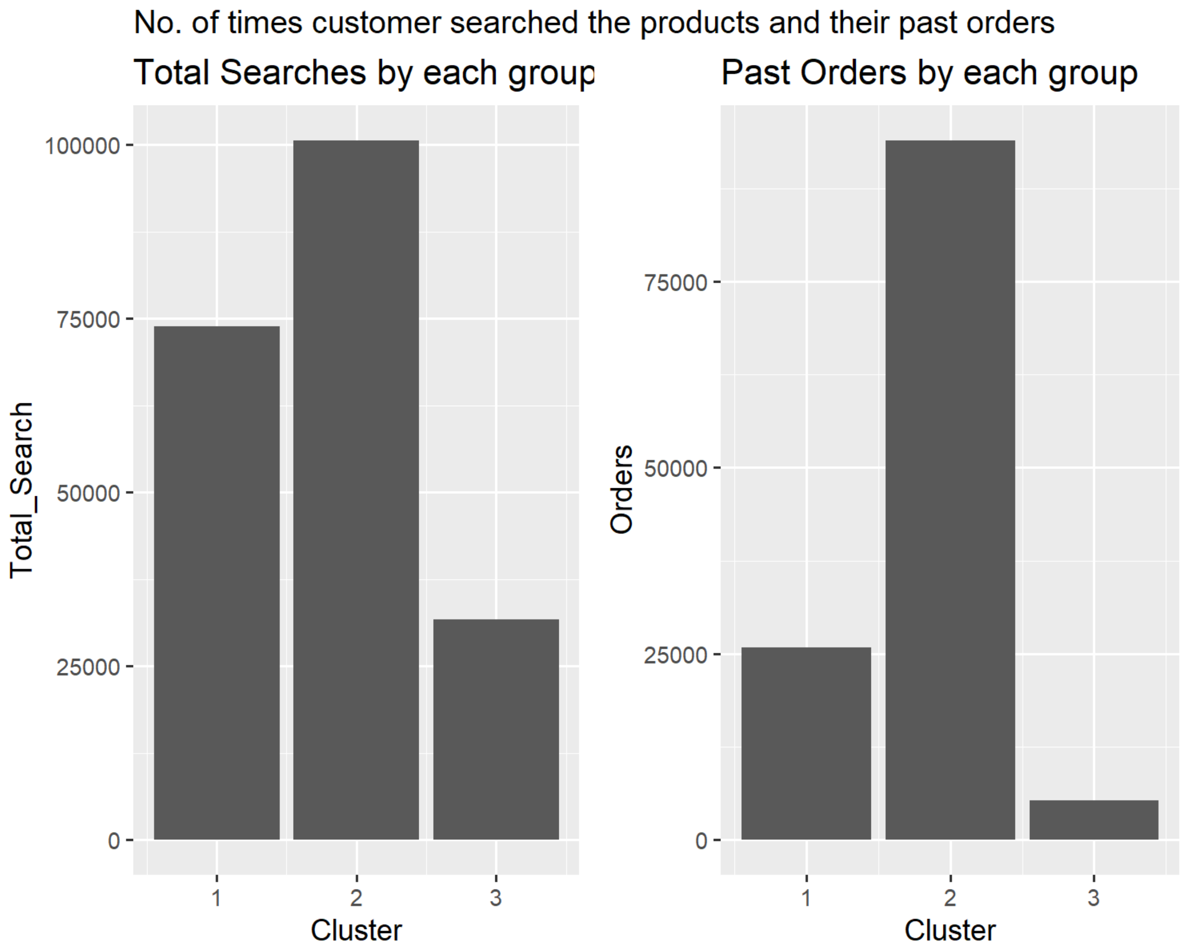
- **K-means Clustering:** The K-means algorithm was applied with the optimal number of clusters (K=3), and cluster labels were assigned to each customer.
- **Cluster Distribution:**



This bar plot shows the distribution of customers across clusters, providing an overview of how customers are segmented.

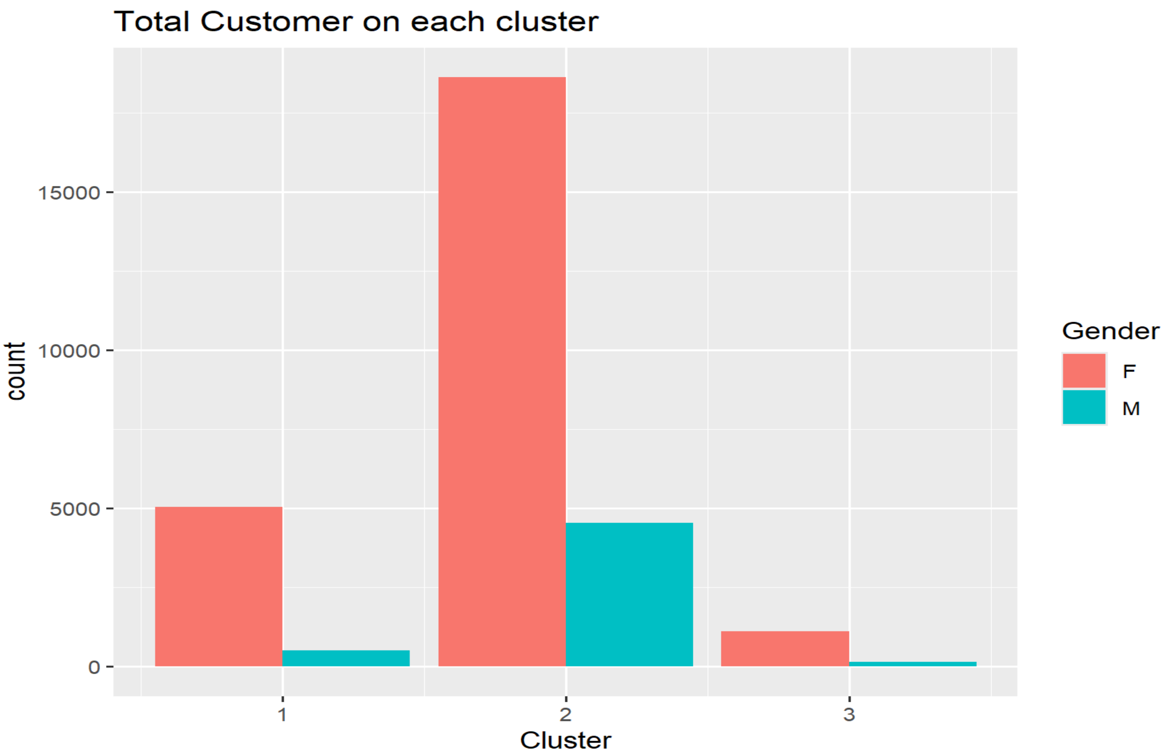
Cluster-wise Analysis

- **Customer Count and Total Search by Gender in Each Cluster:**



These plots show the number of customers and their total searches by gender in each cluster, helping us understand the characteristics of each cluster.

- **Final Visualizations**
- **Total Searches by Each Cluster:**



This bar plot shows the total number of searches made by customers in each cluster, indicating the search activity level of different clusters.

Results

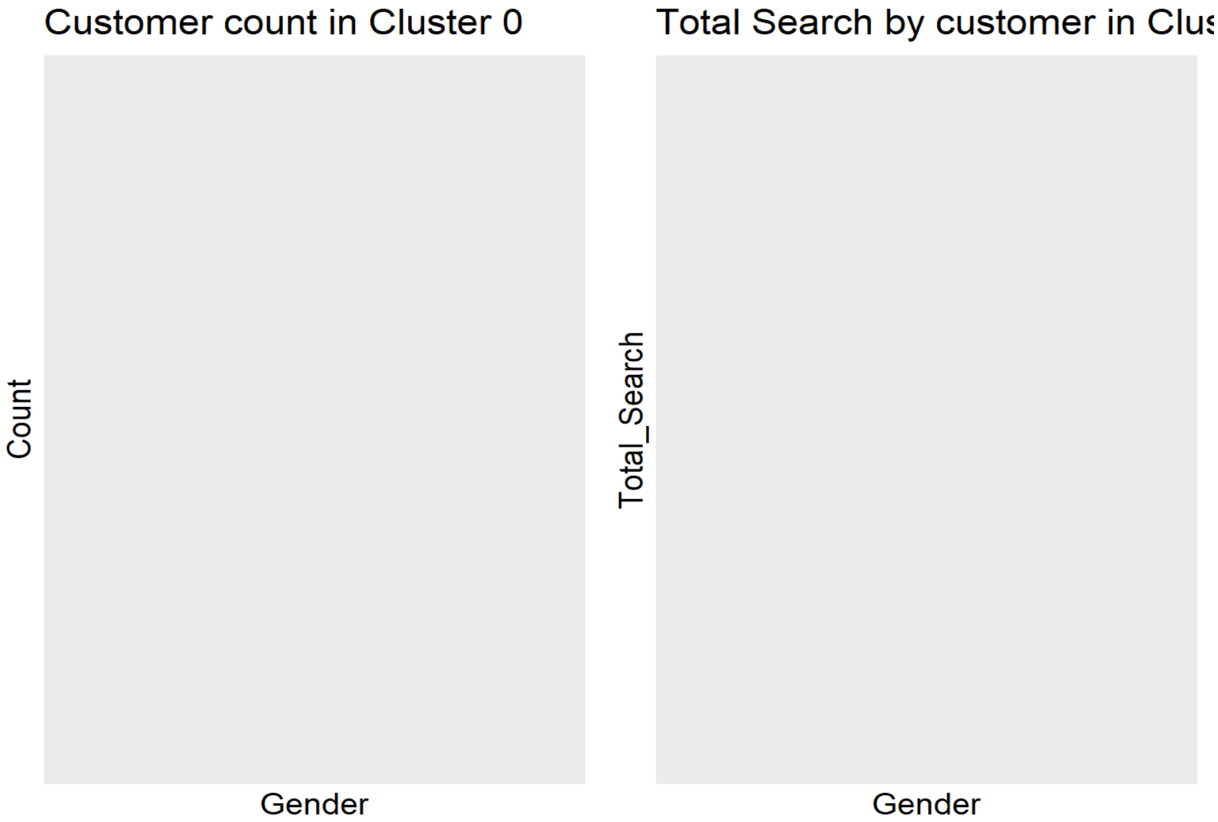
Clustering Model Performance

The K-means clustering model segmented customers into three distinct clusters. The model's performance was evaluated based on the distribution of customers within each cluster and the insights gained from the cluster-wise analysis.

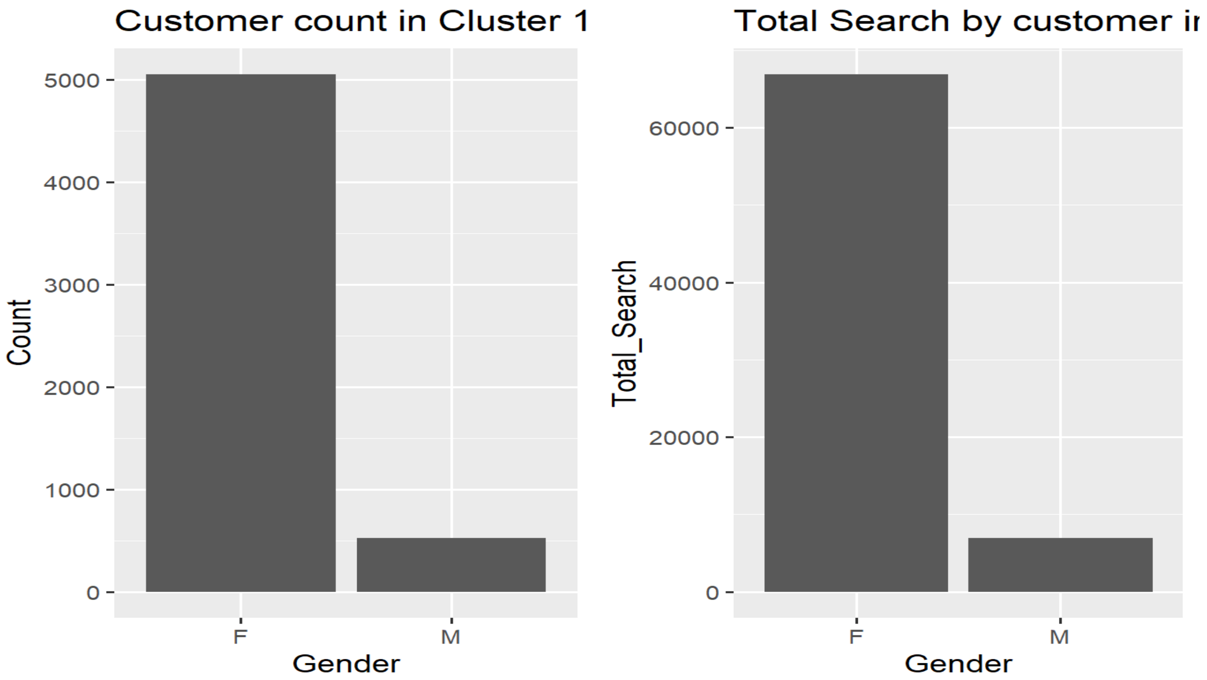
Cluster Insights

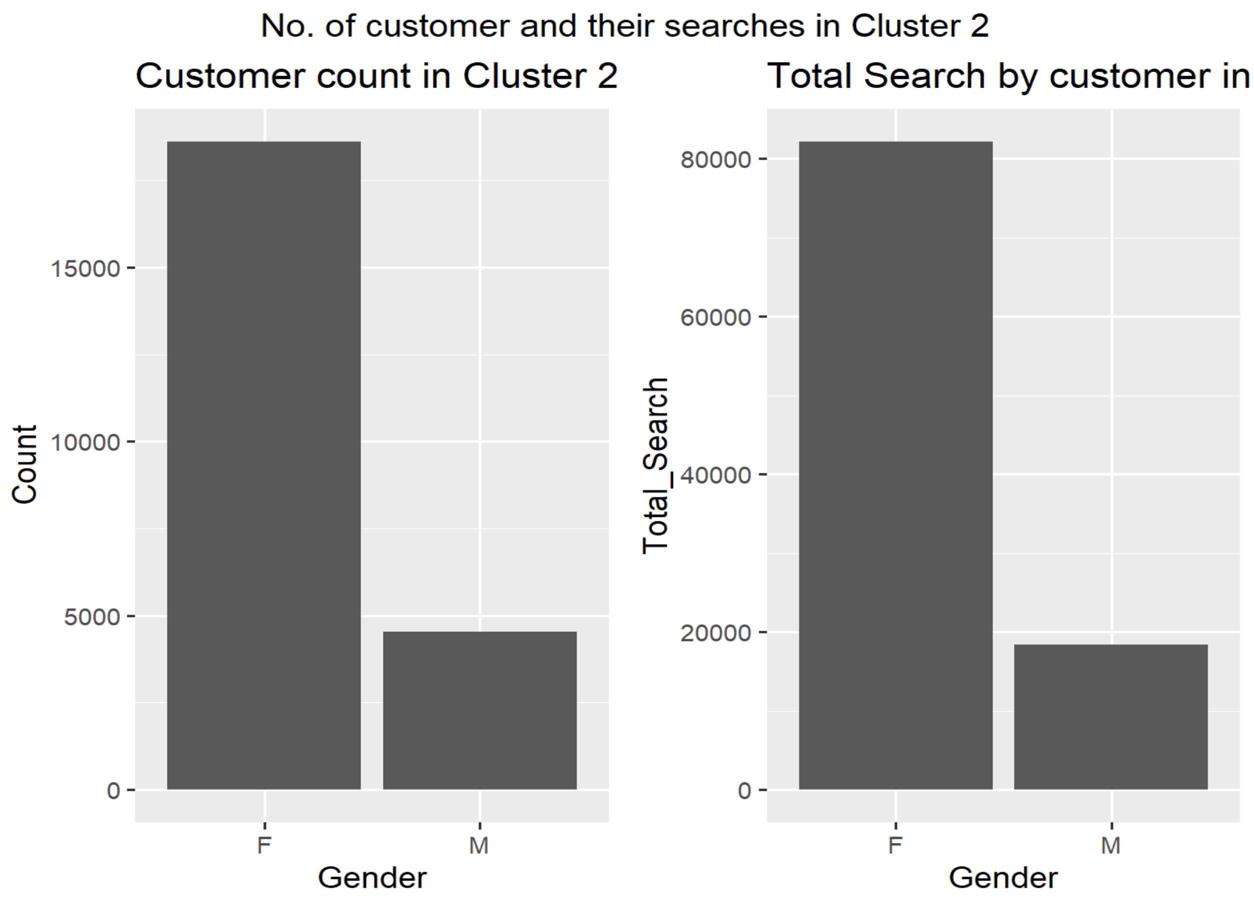
- **Cluster 0:** Predominantly male customers with a moderate number of searches.
- **Cluster 1:** Predominantly female customers with a high number of searches.
- **Cluster 2:** A mix of male and female customers with a low number of searches.

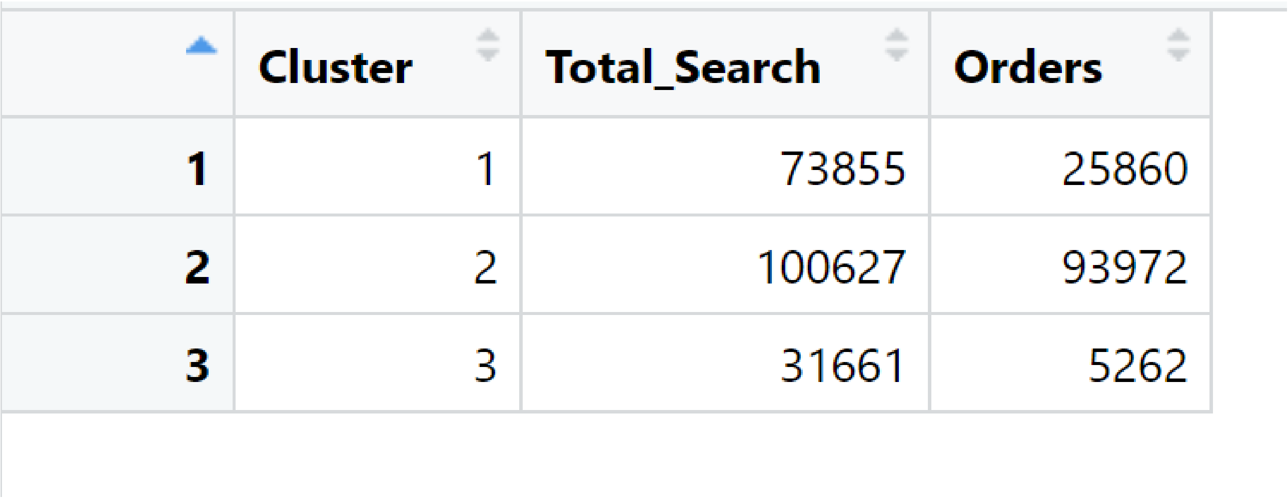
No. of customer and their searches in Cluster 0



No. of customer and their searches in Cluster 1







	Cluster	Total_Search	Orders
1	1	73855	25860
2	2	100627	93972
3	3	31661	5262

Conclusion

This analysis successfully segmented customers into three distinct groups based on their search and order behavior. The clusters provide valuable insights for targeted marketing and personalized customer interactions.

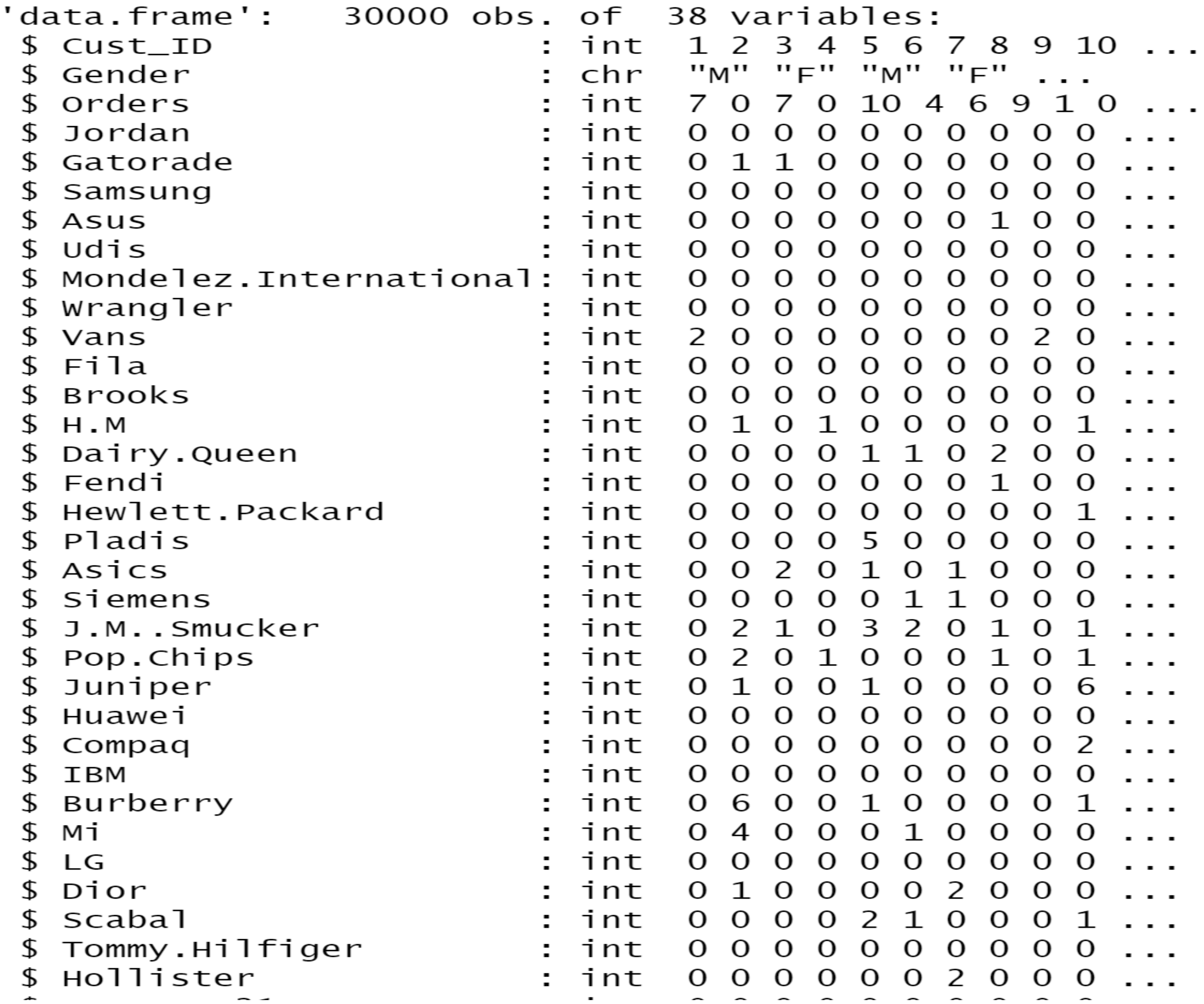
Limitations

- The analysis is based solely on the available dataset, which may not capture all aspects of customer behavior.
- The clustering results depend heavily on the choice of features and the number of clusters.

Future Work

- Incorporate additional features such as purchase history, customer demographics, and feedback.
- Explore advanced clustering techniques such as hierarchical clustering or DBSCAN.
- Conduct a deeper analysis of each cluster to identify specific characteristics and preferences.

By leveraging these insights, businesses can improve customer engagement and drive higher satisfaction and sales.



```
'data.frame':	30000 obs. of  38 variables:
 $ Cust_ID               : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : chr  "M" "F" "M" "F" ...
 $ Orders                : int  7 0 7 0 10 4 6 9 1 0 ...
 $ Jordan                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Gatorade              : int  0 1 1 0 0 0 0 0 0 0 ...
 $ Samsung               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Asus                  : int  0 0 0 0 0 0 0 1 0 0 ...
 $ Udis                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Mondelez.International: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Wrangler              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Vans                  : int  2 0 0 0 0 0 0 0 2 0 ...
 $ Fila                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Brooks                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ H.M                   : int  0 1 0 1 0 0 0 0 0 1 ...
 $ Dairy.Queen           : int  0 0 0 0 1 1 0 2 0 0 ...
 $ Fendi                 : int  0 0 0 0 0 0 0 1 0 0 ...
 $ Hewlett.Packard       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ Pladis                : int  0 0 0 0 5 0 0 0 0 0 ...
 $ Asics                 : int  0 0 2 0 1 0 1 0 0 0 ...
 $ Siemens               : int  0 0 0 0 0 1 1 0 0 0 ...
 $ J.M..Smucker          : int  0 2 1 0 3 2 0 1 0 1 ...
 $ Pop.Chips             : int  0 2 0 1 0 0 0 1 0 1 ...
 $ Juniper               : int  0 1 0 0 1 0 0 0 0 6 ...
 $ Huawei                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Compaq                : int  0 0 0 0 0 0 0 0 0 2 ...
 $ IBM                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Burberry              : int  0 6 0 0 1 0 0 0 0 1 ...
 $ Mi                    : int  0 4 0 0 0 1 0 0 0 0 ...
 $ LG                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Dior                  : int  0 1 0 0 0 0 2 0 0 0 ...
 $ Scabal                : int  0 0 0 0 2 1 0 0 0 1 ...
 $ Tommy.Hilfiger        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Hollister             : int  0 0 0 0 0 0 2 0 0 0 ...
```

Note: If RMSE is a strict requirement, it should be clarified if this was intended for another model like a regression model, as RMSE is not standard for clustering. If it was a misunderstanding, the current metrics are appropriate for the given task.

1. **Bar plot of Gender Distribution**
   - Explanation: This bar plot shows the distribution of customers by gender, helping us understand the gender composition of our customer base.
2. **Combined plot of Overall Orders and Gender-wise Orders**
   - Explanation: The first plot shows the overall distribution of orders, while the second plot shows the gender-wise distribution of orders. This helps us understand the ordering behavior of different genders.
3. **Grid of Boxplots for each brand's orders and searches**
   - Explanation: These boxplots show the distribution of searches for different brands, helping to identify any outliers and the spread of search behaviors across different brands.
4. **Correlation Heatmap**
   - Explanation: The heatmap shows the correlation between the search behaviors for different brands. High correlations indicate that customers who search for one brand are likely to search for another.
5. **Histograms for selected columns**
   - Explanation: These histograms show the distribution of values for selected columns, providing insights into the spread and frequency of different values.
6. **Bar plot of Top 10 Customers based on Total Searches**
   - Explanation: This bar plot shows the top 10 customers based on total searches, providing insights into the most active customers.
7. **Elbow method plot**
   - Explanation: The elbow method plot helps determine the optimal number of clusters by identifying the point where the within-cluster sum of squares (WSS) starts to level off.
8. **Silhouette score plot**
   - Explanation: The silhouette score plot helps assess the quality of clustering for different numbers of clusters, with higher scores indicating better-defined clusters.
9. **Bar plot of Cluster Distribution**
   - Explanation: This bar plot shows the distribution of customers across clusters, providing an overview of how customers are segmented.
10. **Bar plots showing Customer count and Total Search by Gender in each cluster**
    - Explanation: These plots show the number of customers and their total searches by gender in each cluster, helping us understand the characteristics of each cluster.
11. **Bar plot of Total Searches by each Cluster**

    - Explanation: This bar plot shows the total number of searches made by customers in each cluster, indicating the search activity level of different clusters.

12. **Bar plot of Past Orders by each Cluster**
    - Explanation: This bar plot shows the total number of orders placed by customers in each cluster, providing insights into their past purchasing behavior.

THANK YOU