



Customer Segmentation

2024

AUGUST 26

Zidio Development

Authored by: Anushka Sharma



Table of Contents

1. Introduction
 - Overview of Customer Segmentation
 - Importance of Customer Segmentation in Business Strategy
 - Project Goals and Objectives
2. Data Collection and Preprocessing
 - Data Sources
 - Data Description
 - Handling Missing Values and Duplicates
 - Data Normalization and Scaling
3. Exploratory Data Analysis (EDA)
 - Gender Distribution and Analysis
 - Purchase Frequency Analysis
 - Customer Segmentation by Order Count
 - Visualization of Key Patterns
4. Feature Engineering
 - Derivation of New Features
 - Justification for Feature Selection
5. Clustering Methodology
 - Introduction to K-means Clustering
 - Explanation of Clustering Process
 - Elbow Method and Silhouette Score for Optimal K Selection
 - Model Fitting and Prediction
6. Cluster Analysis
 - Overview of Cluster Distribution
 - Detailed Analysis of Each Cluster
 - Gender Analysis within Clusters
 - Customer Behaviour Analysis
7. Business Insights and Recommendations
 - Insights Derived from Each Customer Segment
 - Recommendations for Marketing Strategies
 - Potential Product Offerings for Each Segment
 - Implementation Strategy for Stakeholders
8. Conclusion
9. Future Work
10. References and Appendices

Introduction

In general, customer segmentation is an important data science technique meant for splitting a company's clients into homogenous groups based on their behavior, customer preferences, and characteristics. With an understanding of these segments, a business is able to put up a marketing strategy, proposition of products, and a customer experience that is oriented toward the key needs of particular sets of people. In the present work, the application of clustering, a data science method, is used to segment customers based on features like purchasing behavior, demographic data, and engagement metrics. The purpose of segmenting customers is to derive actionable efforts in implementing more personalized strategies that eventually result in improved relationships with customers.

1.1 Overview of Customer Segmentation

Customer segmentation is the process of understanding the clientele of a company at a very micro level by segmenting customers into distinct categories. These categories can be based on various factors, such as purchasing behavior, demographics, preferences, and engagement metrics. It basically aids a business to deal with different groups of customers in a focused way to ensure efficient resource allocation and maximize customer satisfaction. By segmentation, companies will be better placed to design tailored campaigns and product recommendations that work for each category of customers, hence improving customer retention and sales.

1.2 Importance of Customer Segmentation in Business Strategy

Customer segmentation added to the business strategy will enable a company to focus on such important customer segments through interactive engagements and services. For instance, getting the group that represents high-value customers allows working out special offers and loyalty programs for them, and the promotion for rare consumers will increase sales. By segmenting customers with similar traits, companies are able to make prudential decisions regarding marketing investments, product development, and improvements in customer service. Insights, which can be drawn through customer segmentation, dramatically influence the overall performance of the business: it increases customer retention and maximizes revenues.

1.3 Project Goals and Objectives

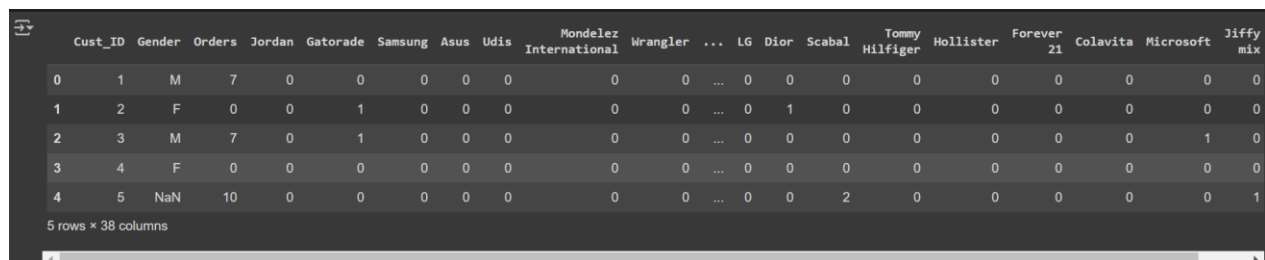
The primary objective of this project is the application of data science techniques in segmenting customers into meaningful groups. It is one of the clustering algorithms—**K-means**—among others that should be used in trying to identify any emerging pattern or trend in behavior, demography, and engagement of customers. The value of these insights will be used to drive business strategies, such as personalized customer targeting in various product/service offerings. It is also in order to validate the results of the segmentation through visualizations and clustering metrics such as the Elbow Method and Silhouette Score. The ultimate goal is to develop a detailed report with recommendations that could guide the business stakeholders in effective implementation of the strategies developed.

Data Collection and Preprocessing

2.1 Data Sources

We worked on a project where there was a dataset that had customer data about e-commerce. Data was sourced from a variety of customer interactions, purchases, and demographic information. The dataset enclosed all features relating to customers: Customer ID, gender, number of orders, detailed purchasing behavior that enabled us to look into the customer segment under study comprehensively. The data was read into the pandas DataFrame using '`pd.read_excel()`', and the first few records were previewed using the '`.head()`' method of data to understand the data structure.

```
data = pd.read_excel("/content/ecom_customer_data.xlsx")
data.head()
```



	Cust_ID	Gender	Orders	Jordan	Gatorade	Samsung	Asus	Udis	Mondelez International	Wrangler	...	LG	Dior	Scabal	Tommy Hilfiger	Hollister	Forever 21	Colavita	Microsoft	Jiffy mix
0	1	M	7	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	2	F	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0
2	3	M	7	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0
3	4	F	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	5	NaN	10	0	0	0	0	0	0	0	...	0	0	2	0	0	0	0	0	1

5 rows × 38 columns

2.2 Data Description

The dataset contains many columns, most of which describe customer behavior and demographics. Some of the important features include purchasing behavior, such as the number of orders and purchase frequency, and demographic data like gender, plus engagement metrics like the total number of searches customers have made. Descriptive statistics were created with '**df.describe()**', giving insight into the distribution and summary statistics of the features, thus enabling the identification of key trends and outliers likely to exist in the data.

2.3 Handling Missing Values and Duplicates

It is important to handle missing values and duplicates in order to enhance the quality of data. In this project, the missing values under the column 'Gender' were filled with the mode of data. Also, '**df.duplicate()**' was used to check the data for any duplicate records in order to eliminate redundancy. At the end, the dataset was checked for cleanliness; it didn't contain any more missing values or duplicates.

2.4 Data Normalization and Scaling

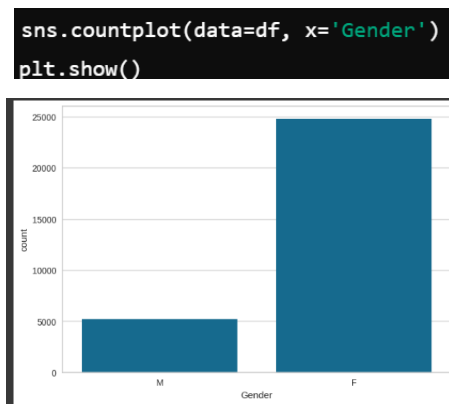
It was ensured that the clustering algorithm would work well by normalization and scaling, since many features were measured in different scales. '**MinMaxScaler**' from '**sklearn.preprocessing**' scaled all values between 0 and 1 to normalize the feature set. This step makes sure that such features as total searches or number of orders are equally treated during the process of clustering.

```
from sklearn.preprocessing import MinMaxScaler
scale = MinMaxScaler()
features = scale.fit_transform(df.iloc[:,2:].values)
```

Exploratory Data Analysis (EDA)

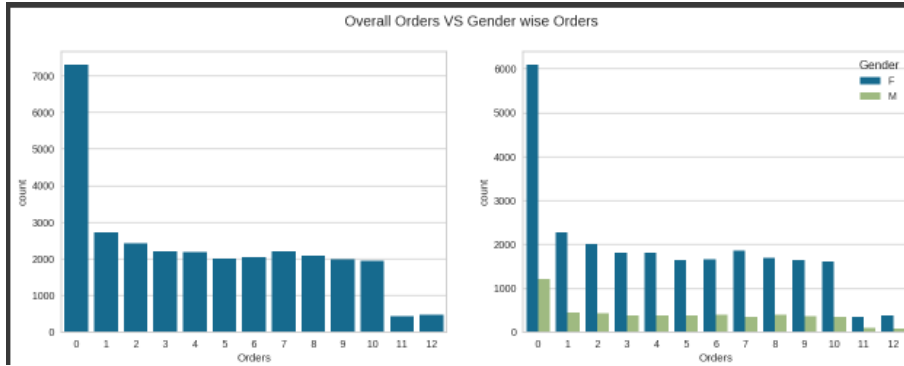
3.1 Gender Distribution and Analysis

The first EDA would be understanding the distribution of customers by gender. In the below script, we used '**sns.countplot()**', which helped us in quickly visualizing the distribution of gender to show any imbalance in our dataset. Knowing the gender information was critical while analyzing and segmenting the behavior of customers. The column for the gender missing values were taken care of. We found that the ratio was pretty balanced for male and female customers. This insight will also help in further segmentation analysis by ensuring that the customer base has the potentials of making gender-related behavior to be effectively explored.



3.2 Purchase Frequency Analysis

It understands the buying behavior based on purchase frequency of the customer. Number of orders is plotted in order to understand the frequency at which customers make purchases. In the above code we used '**sns.countplot()**' to understand the distribution of orders and further segmented this analysis based on gender. Trends like whether any particular gender tends to purchase more frequently can be revealed. Knowing who the high-frequency buyer is important for focused marketing campaigns and loyalty programs, while knowing who the low frequency buyer is, gives a push to the retention strategies.

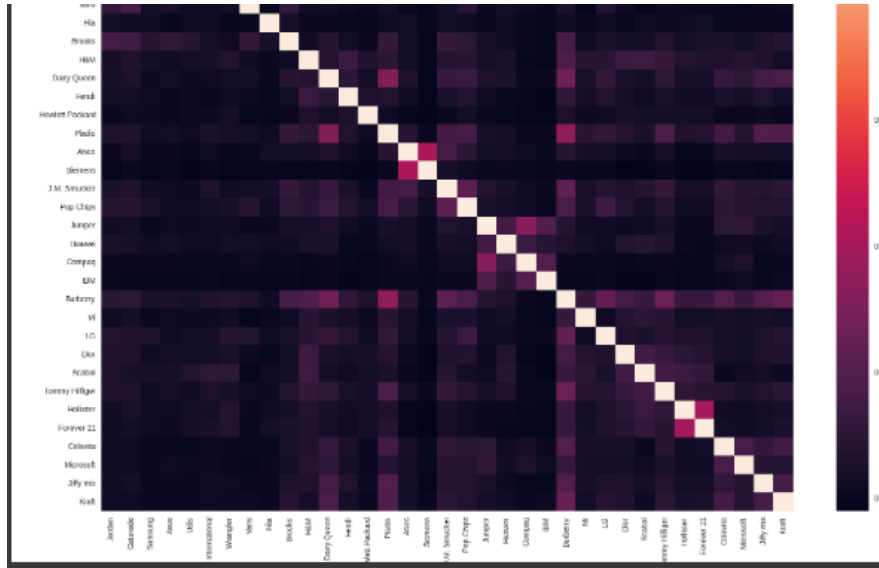


3.3 Customer Segmentation by Order Count

Second, order count is taken as a way of further refining customer segmentation. Customers are classified against the number of purchases they have made to their credit to get insight into different buyer personas—from one-time purchasers to repeat buyers. This kind of segmentation makes it easy to come up with business-specific offers or engagement strategies for each segment. We use visualizations such as a bar plot, which represents the distribution of order counts across the entire customer base. The information is very critical in the recognition of high-value customers who frequently engage with the brand.

3.4 Visualization of Key Patterns

In general, visualization is important for EDA, during which underlying patterns, outliers, and generally correlated features are detected. Boxplots were used to understand the distribution and variability for features such as total searches and total purchased items. Heatmaps are used to visualize the correlation of different features with each other. Major correlations between them were indicated, therefore showing that these features were interdependent to some extent. Such plots helped understand how different attributes influenced behavior. For instance, strong correlation between some features serves as a pointer that the buying habits or demographics between some customer segments could be similar. These insights will therefore help us better do the segmentation and make data-driven decisions.



Feature Engineering

4.1 Derivation of New Features

One of the major steps in improving the performance of machine learning models is feature engineering; this step involves deriving new meaningful features from existing data. The total searches feature were engineered in this customer segmentation project, which was aggregated from all search-related columns in the dataset. This new feature represents the total number of searches that any customer has done in different categories, going a step ahead in explaining customer engagement. The next code snippet illustrates how the new feature was engineered.

```
new_df = df.copy()
new_df['Total Dearch'] = new_df.iloc[:, 3:].sum(axis=1)
```

This feature was central in helping to identify highly engaged customers and differentiating between levels of activity, which became one of the key factors for customer segmentation. For example, high-scoring customers in terms of total searches probably indicate a keen interest in some kind of product, thus allowing a business to make specific marketing efforts toward those customers.

4.2 Justification for Feature Selection

In this project, feature selection was justified with business goals for the purchasing behavior, engagement, and understanding of customer demographics. Orders and Total Dearch have been selected as they bear rich information about the behavior and preferences of customers. These features constructed clusters representing various segments of customers.

It should also provide a summary of the distribution of this new feature, tending to point out how it would impact the clustering process and the results related to customer segmentation.

Clustering Methodology

5.1 Introduction to K-means Clustering

K-means clustering is the most popular unsupervised learning algorithm aimed at partitioning a dataset into K non-overlapping clusters. In this algorithm, each data point gets assigned to the nearest cluster center where distance has to be considered between the point and the centroid of the cluster. Further iterative adjustments will be made to the centroid until convergence in order to minimize within-cluster variance. For its simplicity and efficiency in clustering clients by their behavior and demographic features, K-means was used in the project. It is especially very useful for customer segmentation since it puts customers with similar characteristics into one cluster so a business can design strategies for each cluster. The K-means algorithm works optimally when the number of clusters is pre-defined, which is done through such methods as the Elbow method and Silhouette score.

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3)
kmeans.fit(features)
```

5.2 Explanation of Clustering Process

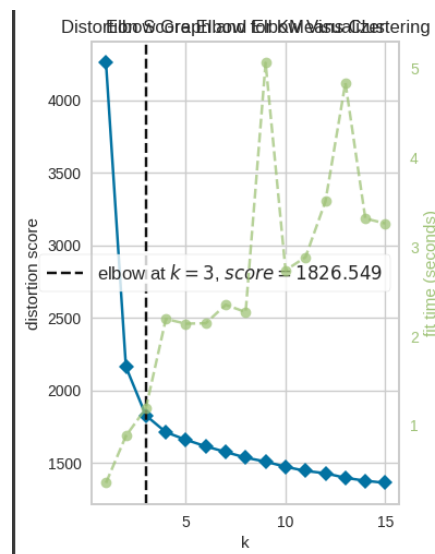
Clustering starts by initializing K centroids, which represent the center for each cluster. Then, every point gets assigned to its nearest centroid according to the Euclidean distance. Once all the points get

assigned, the step of recalculating the centroid comes in by taking the mean of the points in each cluster. Hence, iteratively, repeating this process until the centroids no longer change means the clusters have converged. In our case, we applied K-means clustering to segment customers into a variety of groups according to features such as purchasing behavior and engagement metrics. K-means lets make sense of complex customer behaviors and preferences by finding natural groupings in the data.

```
y_km = kmeans.predict(features)
centers = kmeans.cluster_centers_
```

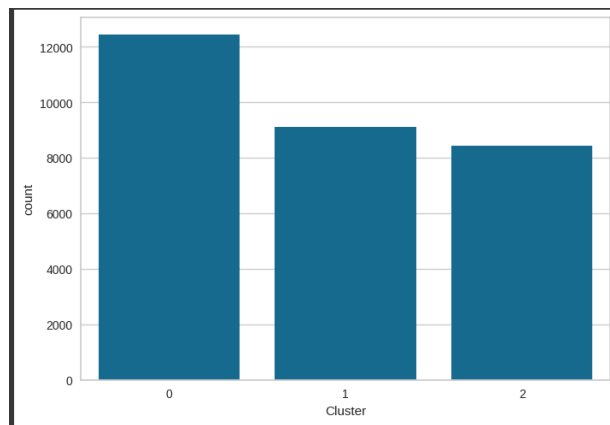
5.3 Elbow Method and Silhouette Score for Optimal K Selection

K-means clustering is one of the most critical parts of establishing the number of clusters. The Elbow Method is the way the number of said element is defined. After computing the inertia (the sum of the squared distance of each of its elements to the nearest centroid), an optimal number of clusters is calculated by plotting a line plot against the different values of K. Most of the time, this will be where the inertia starts diminishing at a slow rate; the graph will have the form of an elbow. In this project, we used both the Elbow Method and the Silhouette Score, which specifically describes the quality of a clustering by giving a measure of how similar a point is to its own cluster, relative to other clusters. The higher the Silhouette Score, the better defined the clusters. This guided us in picking the most adequate K for segmenting our customers.



5.4 Model Fitting and Prediction

Once the optimal number of clusters was selected, the K-means model was fitted on the scaled feature data. The model assigns each customer to a cluster, which represents a specific segment of the customer base. These cluster labels were then added to the original dataset for further analysis and visualization. The final clusters reveal distinct groups of customers based on their purchasing behavior, demographics, and engagement metrics. By analyzing these clusters, we can identify high-value customers, low-frequency buyers, and other key segments. The clustered data is then used to develop targeted marketing strategies and product recommendations tailored to each group.



Cluster Analysis

6.1 Overview of Cluster Distribution

We applied K-means clustering to obtain three clusters of customers. Each cluster contains a group of customers exhibiting similar behavior and alike in many aspects. We drew a picture, `sns.countplot()`, to learn how the total number was shared by each cluster. The customer distribution across these clusters is very meaningful for knowing which groups are relatively large or small in size. This information can, therefore, be quite relevant for prioritizing marketing efforts and deployment of resources. The analysis of cluster distribution empowers businesses to understand the diversity of the customer base better and adopt a focused approach in their customer engagement strategy.

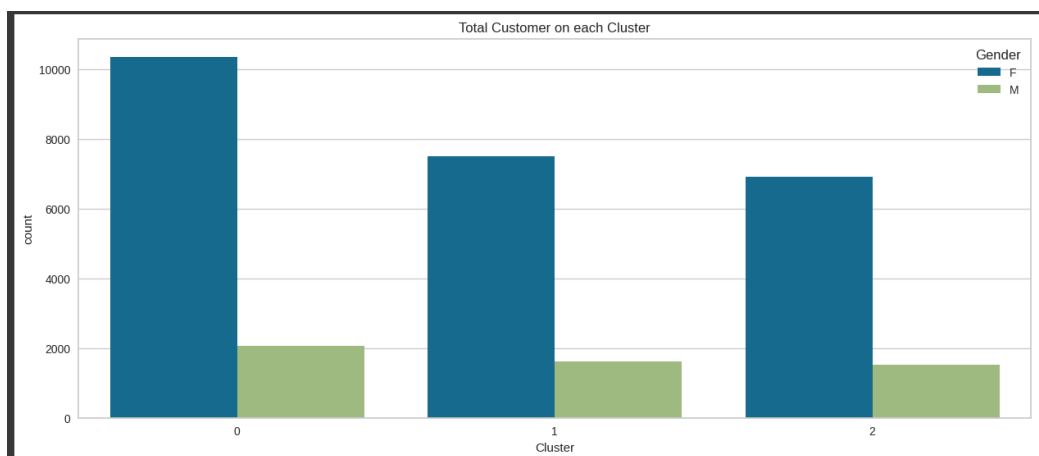
6.2 Detailed Analysis of Each Cluster

Describing each of the clusters, it was meant to understand the intrinsic features they hold. Cluster 0 included customers who are more likely to have medium purchasing behavior and, therefore, average engagement manifest in the business. Cluster 1 held high-value customers, a type of customer who is likely to buy often and bring high revenue — such a target is important for every loyalty program. Cluster 2 held customers with low purchasing behavior or frequency; thus, more attention is required in the re-engagement strategies. One would cross the total searches, past orders, and other behavioral metrics against each identified cluster in order to identify different patterns and preferences. These patterns are key in creating a more personalized marketing campaign for each group.

```
c1_0 = c_df.groupby(['Cluster', 'Gender'], as_index=False).sum().query('Cluster==0')
c1_1 = c_df.groupby(['Cluster', 'Gender'], as_index=False).sum().query('Cluster==1')
c1_2 = c_df.groupby(['Cluster', 'Gender'], as_index=False).sum().query('Cluster==2')
```

6.3 Gender Analysis within Clusters

Gender analysis was done on each cluster in search of trends related to gender in customer behavior. Plotting the number of each gender in every cluster using `sns.countplot()` showed possible gender imbalances and gave an opportunity for further insights into what individual genders seem to like. For example, one cluster might hold more female customers; thus, some of the marketing strategies should be tailored to appeal more toward women. These insights into how gender comes to bear in each cluster can then be utilized by the business in more focused and effective communication, promotion, and product offers.



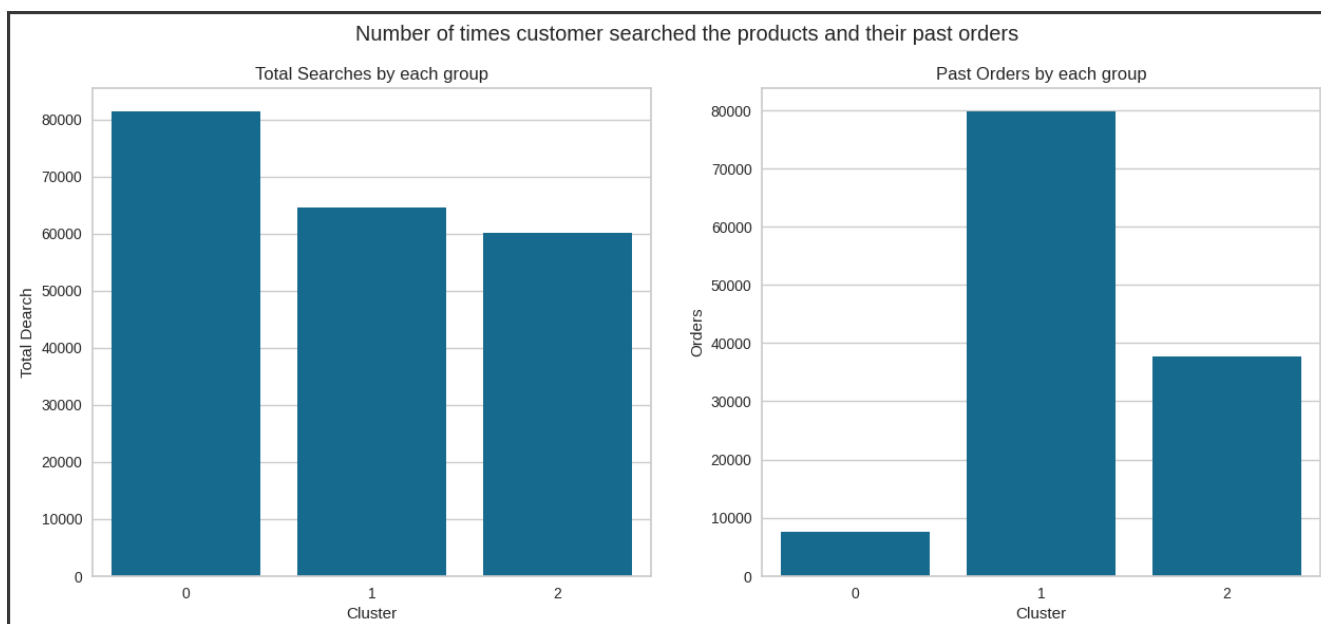
6.4 Customer Behaviour Analysis

Some of the important metrics identified regarding customer behavior include the total searches and the number of past orders. For example, cluster 1 had high-value customers; this was then followed by a higher number of searches and past orders, hence indicative of high engagement with the business. Cluster 2 had few searches and orders; hence, such customers may need some targeted marketing to get them to do more. Comparisons across the clusters of behaviors, for example, with bar plots, revealed important differences in both engagement and purchasing. This kind of analysis gives businesses an insight into the needs and preferences of each segment and thus guides personalized interventions most appropriately.

Business Insights and Recommendations

7.1 Insights Derived from Each Customer Segment

The customer segmentation analysis provided valuable insights regarding the behavior, preferences, and engagement of the various customer segments. Cluster 0 consists of the moderate buyers who have average engagement with the brand. This cluster includes customers who are consistent but of average value to the business. Cluster 1 includes high-value customers who are frequent purchasers, very engaged, and generate a lot of revenue. Such customers are highly engaged, undertaking more searches and having past orders, which is a very good pointer of brand loyalty. Cluster 2 is composed of low-frequency buyers, and such users engage less and have lower purchase rates. This could probably be a segment that would require reactivation efforts since they are less involved with the brand. The business will understand these segments and thus be in a position to prioritize and deal with different ways on each group, ensuring that resources are well allocated.



7.2 Recommendations for Marketing Strategies

Specific marketing strategies can thus be recommended based on the insights derived from these customer segments. For Cluster 1, comprising high-value customers themselves, loyalty programs, exclusive discounts, and personalized offers would be very ideal strategies to keep them engaged and encourage repeat purchases. For Cluster 0, focused marketing efforts in the form of targeted promotions and cross-selling opportunities can help increase their purchasing frequency and average order value. Cluster 2 needs re-engagement campaigns that may include win-back offers and targeted email campaigns with offers and promotions to bring them back to the platform. All strategies will ensure that every design is done in response to specific characteristics or behaviors of each cluster so that it will have maximum impact.

7.3 Potential Product Offerings for Each Segment

This segmentation analysis can also be used to define which potential product offerings each segment would require. For Cluster 1 high-value customers, these may include premium products, limited edition items, and access to new arrivals before others as a way of pleasing these customers. Customers from Cluster 0 may look more at standard product offerings with occasional upsell opportunities. For Cluster 2, bundles or discounts on the most popular products, or carefully designed collections to spur customers

back into engagement, might work. Such personalized product offerings ensure that unique segment needs are met and drive higher customer satisfaction with increased revenues.

7.4 Implementation Strategy for Stakeholders

Segmentation analysis would require stakeholder engagement from marketing, product development, and customer support in a bid to act on the recommendations. The marketing team will have to develop and deliver campaigns based on those insights, while product development can work on creating relevant offerings for each segment concerning their likes and preferences. Customer support can contribute their interaction with customers from these segments in an extremely personalized manner, depending on behavior and the level of engagement. There is a need for monitoring of the outcome of such strategies at regular intervals to ensure continuous improvement. The stakeholders are to be engaged in a deep conversation on a continuous basis to fine-tune and modify the approach according to the voice of customers and performance metrics.

Conclusion

This project involved customer segmentation, where the company's customer base was divided into three different clusters using K-means. These clusters formed segments of customers who demonstrated very different behaviors and characteristics, from high value and frequent buyers to low, less-engaged frequency purchasers. The main takeaways from the segmentation analysis indicated that Cluster 1 held the most opportunity for growth in revenue because it was composed of highly engaged and loyal customers. The second cluster wants focused re-activation efforts in order to get customers to be more active and retain this base of customers within the company's ecosystem.

Such project outcomes delivered actionable insights, capable of making a huge difference in business decision-making. Thus, the firm can now come up with tailored marketing strategies, product offerings, and engagement plans for these distinct needs and preferences that exist in each cluster. For instance, Cluster 1 can be offered loyalty programs and some exclusive offers, whereas Cluster 2 might use reactivation campaigns with incentives.

The business will thus be enabled through these data-driven insights from the project to better allocate resources, optimize marketing efforts, and improve customer retention and satisfaction. Through visualization and statistical analyses, it helps bring clarity to the customer behavior and, therefore, helps stakeholders make informed decisions that drive growth and profitability.

Future Work

The current model of customer segmentation is already based quite fundamentally in order to understand the behavior and preferences of the customer but work can be done even further through additional incorporation and extension. Future works might focus on additional characteristics to be integrated into the segmentation model—a move that might focus on such attributes as customer lifetime value, engagement with specific product categories, or more granular demographic information. Integration of the model with such features could result in more refined and actionable segmentation. More so, this could be improved through experimentation with advanced methods of clustering, such as hierarchical clustering or Gaussian Mixture Models (GMM).

Another promising direction for future work is the integration of predictive models with segmentation. For instance, the business may predict customer churn and even predict the future purchase behavior within each of the segments, hence be prepared to understand and respond to their customers' needs. A company thus gives way to coming up with class-based, more personalized strategies for the individual customers through clustering and other predictive models, such as the decision tree or random forest.

Finally, the critical ability is to track and measure the effectiveness of applied strategies on the basis of insights gained from segmentation. Continuous monitoring of conversion, average order value, customer retention, and other key performance indicators will allow adaptive adjustment of the approach.

References and Appendices

Books:

1. **Kaufman, L., & Rousseeuw, P. J. (2009).** *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
 - Comprehensive guide to various clustering methods, including K-means and hierarchical clustering.
2. **Han, J., Pei, J., & Kamber, M. (2011).** *Data Mining: Concepts and Techniques (3rd Edition)*. Morgan Kaufmann.
 - Fundamental textbook on data mining, covering clustering and customer segmentation techniques.
3. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*. Springer.
 - In-depth coverage of statistical learning methods relevant to clustering and segmentation.
4. **Berson, A., Smith, S. J., & Thearling, K. (2000).** *Building Data Mining Applications for CRM*. McGraw-Hill.
 - Focuses on applying data mining techniques to CRM, including customer segmentation strategies.
5. **Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016).** *Data Mining: Practical Machine Learning Tools and Techniques (4th Edition)*. Morgan Kaufmann.
 - Practical guide to machine learning techniques, including clustering and data preparation methods.

Papers:

6. **Tibshirani, R., Walther, G., & Hastie, T. (2001).** "Estimating the Number of Clusters in a Data Set via the Gap Statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.
 - Introduces the Gap statistic for determining the optimal number of clusters.
7. **MacQueen, J. B. (1967).** "Some Methods for Classification and Analysis of Multivariate Observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.
 - Original paper on K-means clustering, foundational to understanding the algorithm used in the project.
8. **Rousseeuw, P. J. (1987).** "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, 20, 53-65.
 - Discusses the Silhouette score for evaluating cluster quality.

Tools: 9. **Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011).** "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.

- Documentation for Scikit-learn, which provides the K-means implementation and other tools used in the project.
10. **Seaborn Developers. (2024).** *Seaborn: Statistical Data Visualization*. Retrieved from <https://seaborn.pydata.org>
 - Documentation for Seaborn, a visualization library used for plotting in the project.

Appendices

Code: Key Code Snippets

1. Loading and Preprocessing Data

```
import pandas as pd

data = pd.read_excel("/content/ecom_customer_data.xlsx")
df = data.copy()
df['Gender'] = df['Gender'].fillna(df['Gender'].mode()[0])
```

2. Exploratory Data Analysis (EDA)

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(data=df, x='Gender')
plt.show()
```

3. Clustering and Evaluation

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from yellowbrick.cluster import KElbowVisualizer
from sklearn.metrics import silhouette_score

features = MinMaxScaler().fit_transform(df.iloc[:, 2:].values)
kmeans = KMeans(n_clusters=3)
model = kmeans.fit(features)
silhouette_avg = silhouette_score(features, model.predict(features))
```

Additional Figures: Extra Charts or Tables

1. Elbow Method Plot

- Include a plot showing the inertia for different numbers of clusters to illustrate the Elbow method.

```
plt.figure(figsize=(20,7))
plt.plot(range(1,16), inertia, 'bo-')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal K')
plt.show()
```

2. Silhouette Score Plot

- Display the Silhouette scores for different cluster counts to help determine the optimal number of clusters.

```
plt.figure(figsize=(10,7))
plt.plot(range(2,16), silhouette_avg, 'bX-')
plt.xlabel('Number of clusters')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Analysis for Optimal K')
plt.show()
```

3. Cluster Summary Table

- Provide a table summarizing the number of customers in each cluster and key metrics.

```
cluster_summary = df.groupby('Cluster').agg({
    'Total_Dearch': 'sum',
    'Orders': 'mean'
})
print(cluster_summary)
```