Implementing KMeans Clustering

Bosco M Morales

Module 2

CAP4767-2243-4701

Spring 2024

Wednesday, March 20 2024

Table of Contents

1. Introduction……..3
2. Dataset……..3
3. Methodology……..3
4. Visualizations and Analysis
    a. Frequency Vs Total Spend……..4
    b. Recency Vs Total Spend……. …….5
    c. Cluster Zero……..6
    d. Cluster One……...7
    e. Cluster Two……8
5. Summary……..9
6. Recommendations…… ……11
7. Conclusion…… ……13
8. References…… ……13

1. Introduction

The ‘Module 2 - Implementing KMeans Clustering’ notebook aims to implement KMeans clustering for customer segmentation using a dataset from an online retail company.

2. Dataset

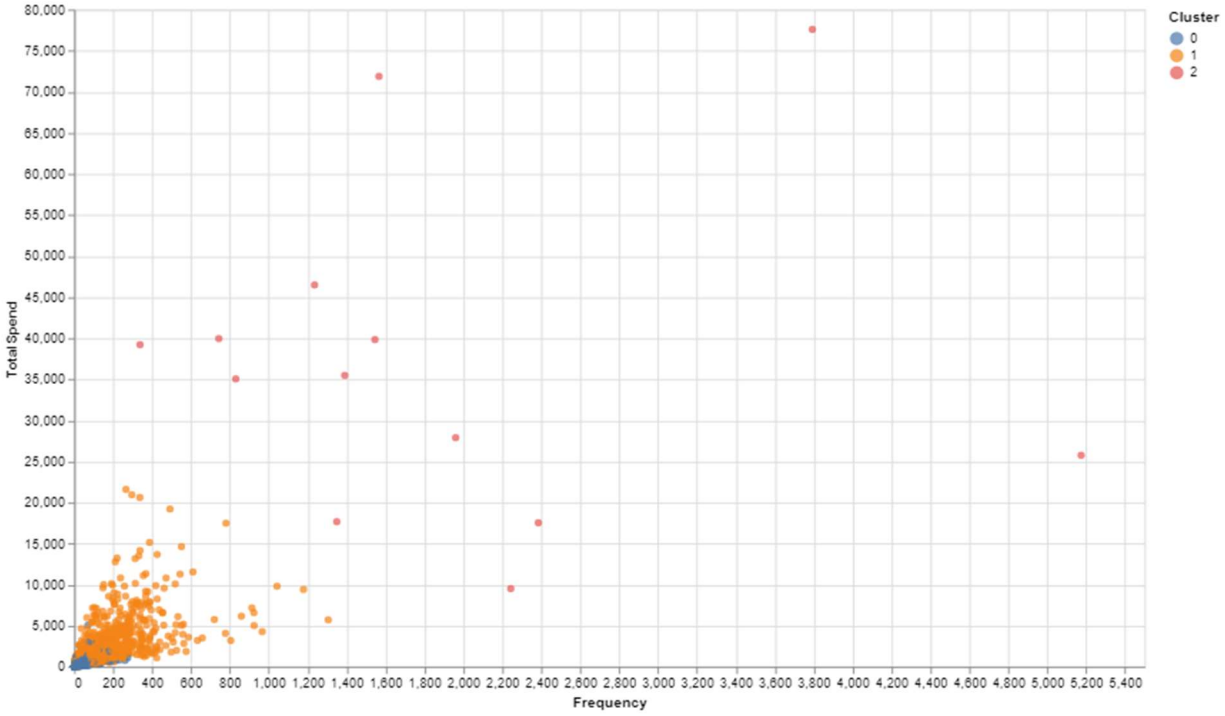
The dataset is a substantial collection with 541,909 entries across eight columns. All the transactions occurred between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.. Each entry represents a transaction.

3. Methodology

- ***Data Sourcing***: The dataset was sourced from Kaggle, covering transactions from a UK-based online retail company from December 2010 to December 2011.
- ***Exploratory Data Analysis (EDA)***: Initial data analysis involved loading the data, examining its structure, checking for missing values, and performing statistical summaries to understand distributions.
- ***Data Preprocessing***: This included outlier treatment, handling duplicates, addressing missing values, and removing irrelevant columns. The features considered essential for the analysis were 'InvoiceNo', 'InvoiceDate', 'CustomerID', 'Quantity', and 'UnitPrice'.
- ***Feature Selection***: Features for segmentation were chosen, including purchase frequency, recency of purchases, and total spend. New features were created as necessary, and data was aggregated at the customer level.
- ***Normalization***: The selected features were normalized using the StandardScaler.
- ***Elbow Method***: The Elbow Method was applied to determine the optimal number of clusters, which was identified as three based on the WCSS (within-cluster sum of squares) plot.
- ***KMeans Clustering***: The KMeans algorithm was implemented using the identified number of clusters. The model was fitted to the scaled features, and clusters were predicted and appended to the original dataframe.

4. Visualizations and Analysis

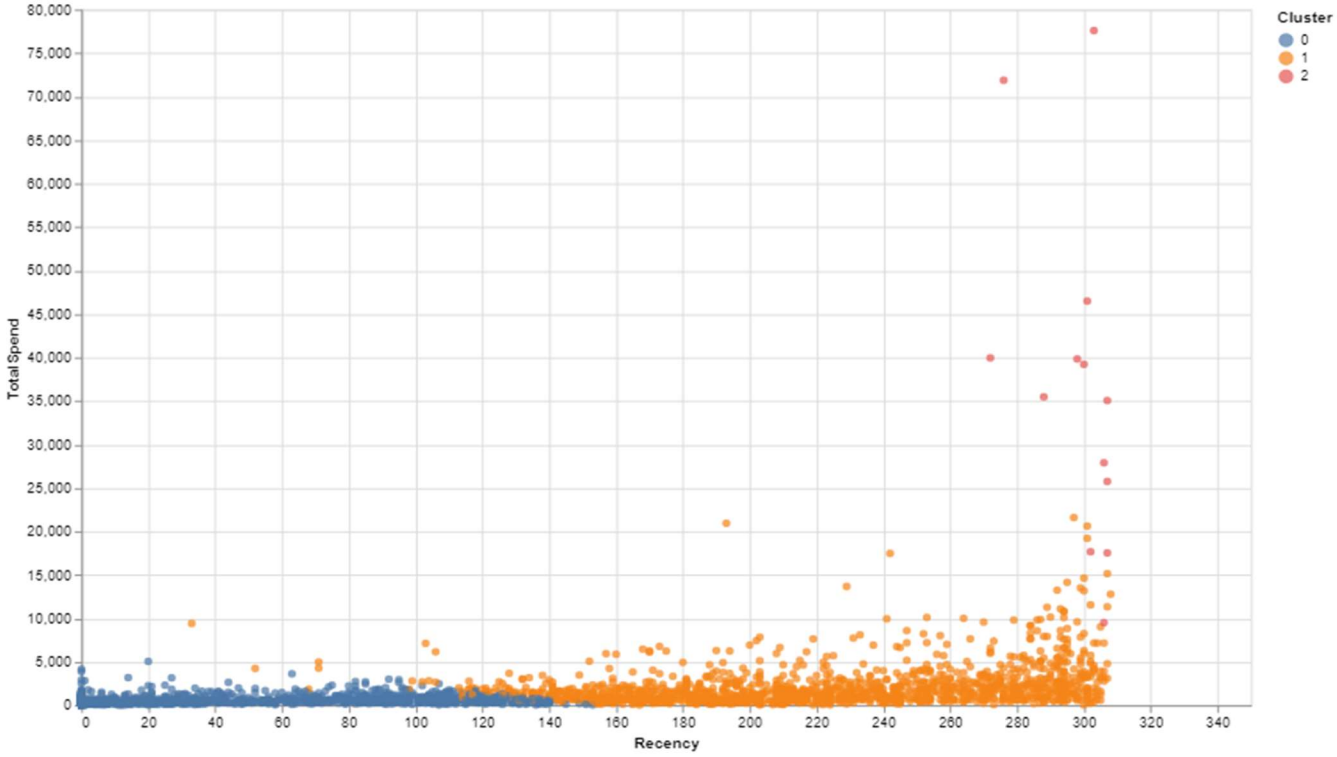
A. Frequency Vs Total Spend



This chart visualizes the relationship between 'Frequency' (on the x-axis) and 'Total Spend' (on the y-axis). The data points are color-coded to represent different clusters—labeled as 0, 1, and 2. Here are some observations:

- **Cluster 0 (Blue)**: These customers have the lowest frequency and spend; they are grouped at the bottom left of the plot. This cluster represents occasional or one-time customers who spend the least amount.
- **Cluster 1 (Orange)**: This group has a lower frequency range (up to around 500) and the total spend varies widely, from low to the highest values on the plot. It includes customers who don't purchase often but have a wide range of spending, indicating some may be making large one-time purchases.
- **Cluster 2 (Red)**: This cluster has customers across the entire frequency spectrum but generally at a higher spend level than Cluster 1, though less than some of the high-spend individuals in Cluster 1. These could be regular customers with consistent, but not necessarily high, transaction values.

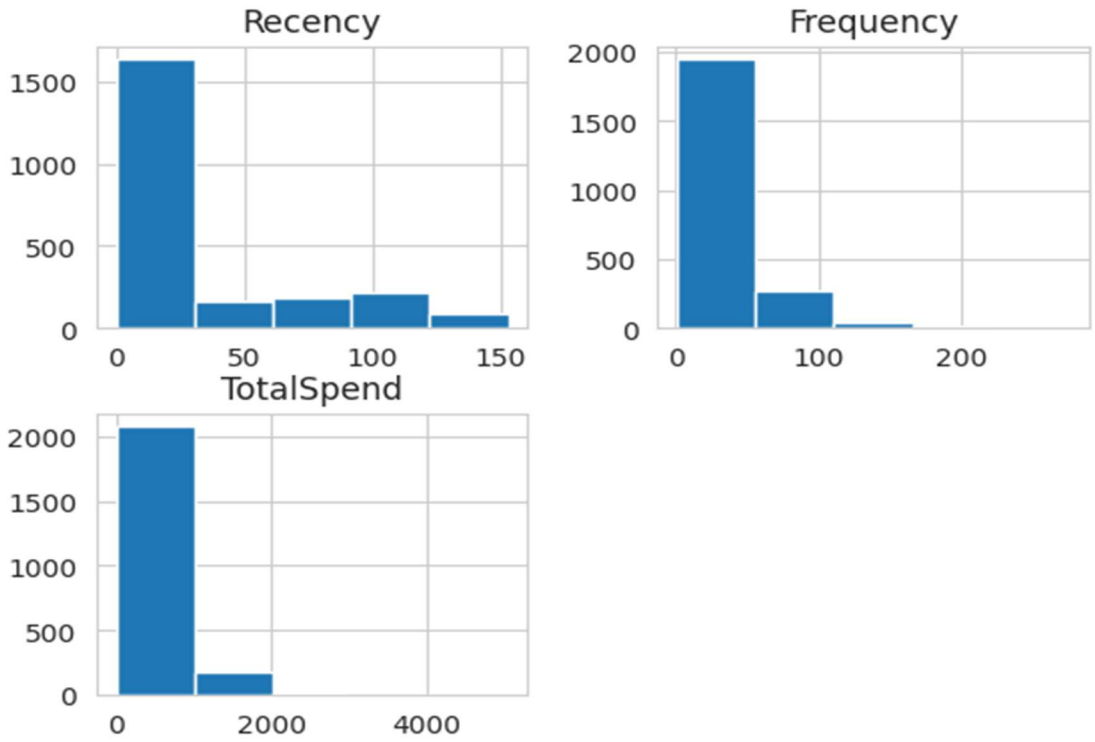
B. Recency Vs Total Spend



This scatter plot represents customer clustering based on 'Recency' (on the x-axis) and 'Total Spend' (on the y-axis). The data points are color-coded to represent different customer clusters—labeled as 0, 1, and 2.

- **Cluster 0 (Blue)**: Shows a wide distribution across the "Recency" axis but is concentrated at the lower end of the "Total Spend" axis, suggesting this group consists of either infrequent shoppers or those who spend less in total.
- **Cluster 1 (Orange)**: It is also spread out in terms of "Recency" but with a higher "Total Spend," hinting that this group might be regular customers with moderate to high spending.
- **Cluster 2 (Red)**: It is quite sparse and mostly found at the higher end of both axes, suggesting these are customers who have not made recent purchases but have a high total spend, which could imply they made significant one-time purchases or used to be high spenders.

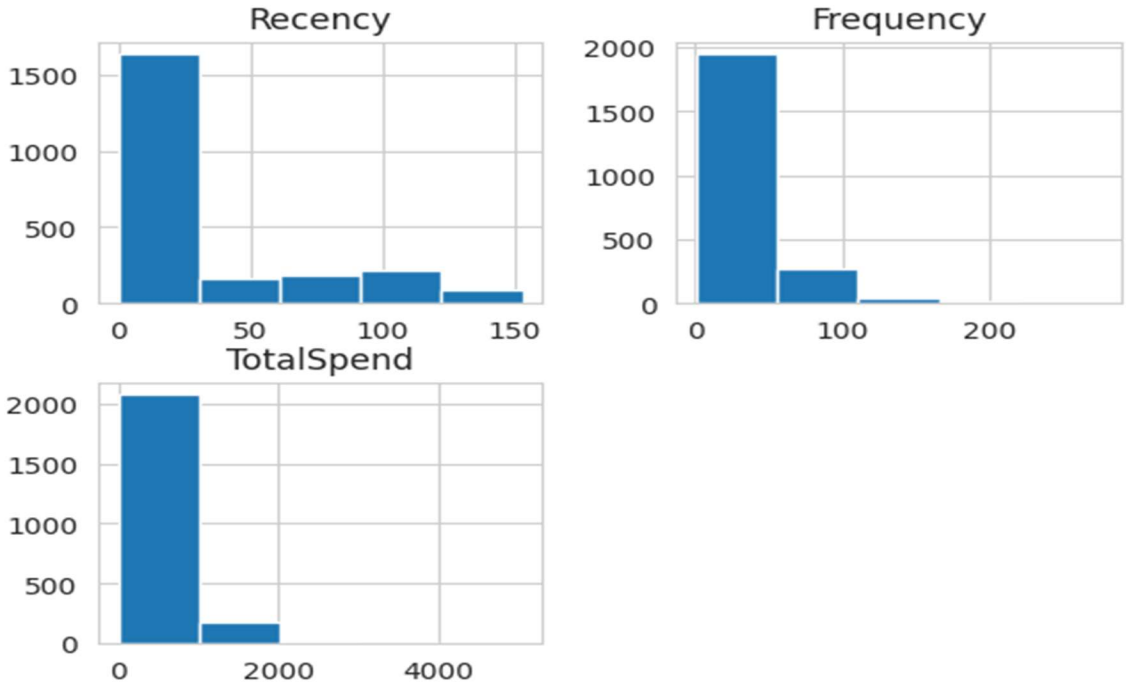
C. Cluster Zero



- ***Recency Histogram***: Shows that a large number of customers have a low recency value, indicating they have engaged with the business recently. The frequency count dramatically decreases as the recency value increases, which suggests fewer customers have not interacted with the business for a longer period. The data is heavily skewed to the right.
- ***Frequency Histogram***: Illustrates that most customers have a low frequency count, with the number of customers decreasing as the frequency increases. This indicates that while there are a lot of one-time or infrequent customers, regular repeat customers are less common. Again, the distribution is right-skewed.
- ***Total Spend Histogram***: Depicts that the majority of customers have a relatively low total spend, with very few customers having a high total spend. The data is right-skewed, suggesting that while most customers spend smaller amounts, there are a few customers who spend much more.

In summary, these histograms suggest that Cluster Zero consists of customers who are generally recent and either one-time or infrequent shoppers with lower total spending. The skewness of the data implies that there are outliers in each of the variables — customers with very high recency values, customers who make purchases frequently, and customers who spend a lot in total. These outliers are few in number compared to the majority of the customer base within this cluster.

D. Cluster One



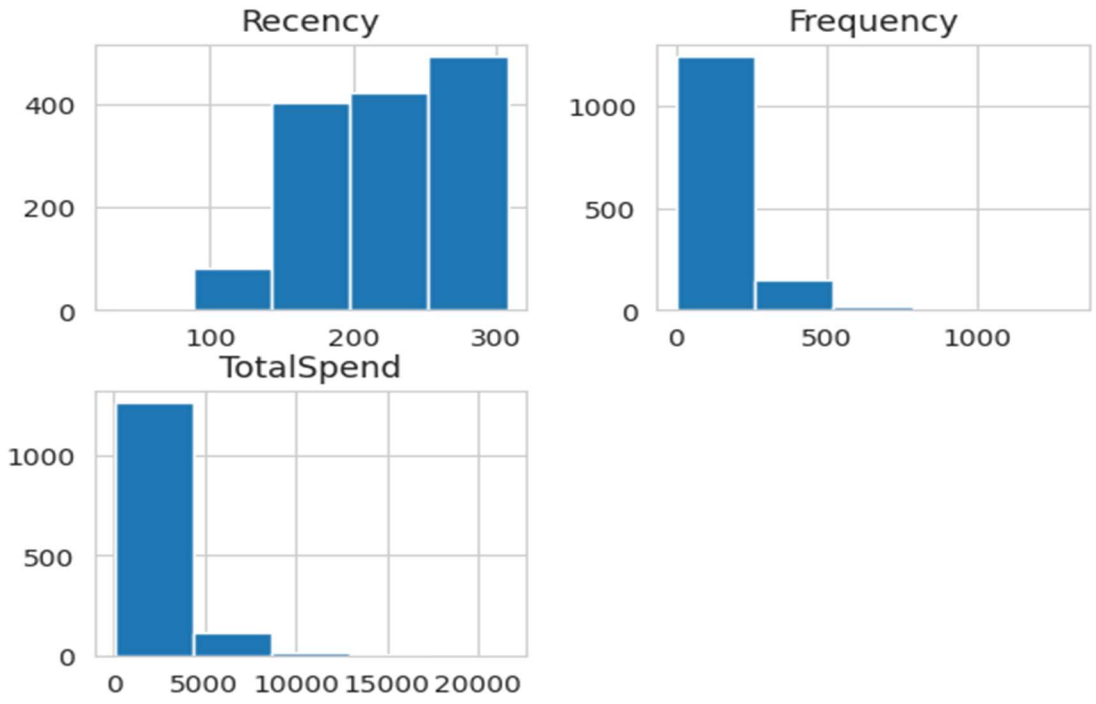
- ***Recency Histogram*:** Shows that the majority of customers in this cluster have very low recency values, suggesting they have recently interacted with the business. There is a significant drop as recency increases, which suggests that fewer customers have gone longer without interaction. The distribution is heavily skewed to the right, indicating that recent activity is more common in this customer segment.
- ***Frequency Histogram***: Illustrates that a vast majority of customers have a low frequency of interaction, with the number sharply decreasing for higher frequency

values. This cluster is predominantly composed of customers who interact infrequently with the business. The distribution is also right-skewed, suggesting infrequent transactions are more common among this group.

- ***Total Spend Histogram***: There is a large number of customers with low total spending. There is a steep decline in the number of customers as the total spend increases, indicating that high spenders are rare within this cluster. This distribution is right-skewed, with most customers in the cluster spending smaller amounts.

In summary, Cluster One is characterized by customers who have recently interacted with the business and generally engage in a lower frequency of transactions with modest spending. The cluster reflects typical customer behavior with a few outliers that either interact more frequently or spend more in total than the average customer in this segment.

E. Cluster Two



- ***Recency Histogram***: It shows that customers are somewhat evenly spread out across different recency values. This indicates that there are nearly as many

customers who have recently interacted with the business as there are those whose last interaction was not as recent.

- ***Frequency Histogram***: Most customers have low frequency counts, with numbers dropping precipitously for higher frequency counts. This indicates that in this cluster, most customers transact infrequently. The distribution is heavily right-skewed, showing that high-frequency customers are very rare within this group.
- ***Total Spend Histogram***: It exhibits a large number of customers with low total spend, and a rapid decrease in customers as total spend increases. There is a very long tail extending to the right, suggesting a few customers with exceptionally high total spend. This indicates that while most customers in this cluster spend small amounts, there are outliers who spend much more.

In summary, these histograms suggest that Cluster Two is characterized by a customer base that has a fairly even distribution of recent to less recent interactions with the business, a general tendency towards infrequent purchases, and typically low total spending, albeit with a few significant outliers who spend substantially more than others. The overall shape of these distributions indicates that there are a few exceptional customers in terms of both frequency and total spending, while the majority show more moderate commercial behavior.

5. Summary

Here's a summary of each cluster:

A. Cluster 0:

- ***Recency***: Average of 26.23, which suggests that customers in this cluster have had recent interactions with the business.
- ***Frequency***: Average of 30.36, indicating that customers in this cluster have a relatively low frequency of transactions.

- ***Total Spend***: Average of 439.28, suggesting that the total amount spent by customers in this cluster is moderate.

B. Cluster 1:

- ***Recency***: Average of 223.33, indicating that customers in this cluster have not interacted with the business recently.
- ***Frequency***: Average of 130.37, which is higher than Cluster 0, suggesting these customers interact more frequently.
- ***Total Spend***: Average of 2144.57, showing that customers in this cluster spend significantly more than those in Cluster 0.

C. Cluster 2:

- ***Recency***: Average of 297.92, which is the highest among the clusters, suggesting that these customers have the longest time since their last interaction.
- ***Frequency***: A very high average of 1888.54, indicating that despite the longer time since last interaction, when these customers do interact, they do so very frequently.
- ***Total Spend***: A very high average of 37212.87, showing that these customers spend far more than those in the other clusters.

From this summary, we can interpret that Cluster 0 might represent new or occasional shoppers, Cluster 1 might consist of more regular customers who spend more per visit, and Cluster 2 could represent high-value customers who, despite infrequent interactions, spend large amounts and have high transaction frequency when they do engage.

6. Recommendations

Here are some recommendations for each cluster:

A. **Cluster 0 - Emerging Customers:**

- ***Recency***: Since these customers have interacted recently, prompt follow-up communications can help reinforce their purchasing behavior.
- ***Frequency***: With a lower frequency, these customers might benefit from incentives to shop more often, such as loyalty programs or frequency rewards.
- ***Total Spend***: Moderate spenders might be sensitive to promotions or special offers that encourage them to increase their basket size.

Strategies:

- Implement a loyalty rewards program to encourage repeat purchases.
- Send personalized follow-up emails or notifications suggesting related products or upcoming deals shortly after a purchase.
- Offer bundled promotions or discounts on higher-margin items to increase average order value.

B. **Cluster 1 - Lapsed Regulars:**

- ***Recency***: These customers haven't made recent purchases, indicating they may be lapsing or have turned to competitors.
- ***Frequency***: A previously higher frequency suggests they were once regular shoppers.
- ***Total Spend***: Their high total spend indicates significant past engagement and potential value if re-engaged.

Strategies:

- Initiate win-back campaigns with personalized offers based on past purchase history to rekindle their interest.
- Conduct customer satisfaction surveys with incentives to complete them, which can provide insights into why they may have lapsed.
- Use time-limited offers to create urgency and bring them back to the business.

C. Cluster 2 - High-Value VIPs:

- ***Recency***: The longest time since last purchase, suggesting potential dissatisfaction or completion of a sales cycle.
- ***Frequency***: High frequency when they do engage indicates they are likely to make multiple purchases.
- ***Total Spend***: Very high spenders, possibly bulk buyers or business clients, contributing significant revenue.

Strategies:

- Create an exclusive VIP program offering premium services, early access to new products, or bulk buying discounts.
- Establish a relationship management program with dedicated account managers to provide personalized service and maintain engagement.
- Regularly update them with tailored communications about premium or exclusive offerings, ensuring the brand remains top-of-mind.

General Recommendations:

- Data-driven Personalization: Leverage data analytics to understand individual customer preferences and tailor offers accordingly.
- Multi-channel Engagement: Engage customers through their preferred channels, whether it be email, social media, direct mail, or in-person interactions.
- Feedback Loop: Implement a system to gather customer feedback after promotional campaigns to continually refine and optimize the marketing strategy.

These strategies aim to not only retain customers but also to increase their lifetime value by encouraging more frequent interactions and higher spending.

7. Conclusion

The analysis conducted through KMeans clustering provides a strong foundation for understanding customer value and behavior. This method enables the company to make data-driven decisions in its marketing strategies and customer relationship management, which can lead to enhanced customer engagement and retention, ultimately improving overall profitability. By segmenting customers into distinct clusters, the organization can tailor its approach to target different customer groups effectively. Implementing the recommended strategies, such as personalized communication, loyalty rewards, and exclusive VIP services, could significantly enhance the customers' lifetime value and ensure sustained revenue growth.

9. References

Essam. (n.d.). E-commerce Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/shedai/retail-data-set?resource=download