Name

- Afrose Ahamed Fathima

Course

- Advanced Certification in Data Science and Artificial Intelligence

Capstone Project

Online Retail

Date: 01-10-2024

Table of Contents

Acknowledgment

I would like to express my sincere gratitude to the Intellipaat Team and IIT Madras for providing me with the opportunity to enhance my skills and knowledge through their comprehensive course. The curriculum, coupled with expert guidance, has significantly contributed to my understanding of data science, machine learning, and artificial intelligence. I appreciate the dedication and support of the instructors and the valuable resources that were made available throughout the program. I am eager to apply what I have learned in my future endeavours.

Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same. Using the above data, come up with useful insights that can be used by each of the stores to improve in various areas.

1. Using the above data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
2. Segment the customers based on their purchasing behavior

Problem Objective

The objective of this project is to analyze customer purchasing patterns for an online retail store and provide actionable insights by segmenting customers based on their buying behavior. The aim is to help the retailer:

1. Understand Customer Segments: Identify distinct customer groups based on their purchasing recency, frequency, and monetary value (RFM analysis) to inform marketing strategies.

2. Improve Customer Retention: Recognize which customers are at risk of churning and which ones are loyal, providing targeted strategies for retention and engagement.

3. Increase Customer Lifetime Value: Develop strategies to convert occasional buyers into loyal customers and increase the average monetary value of purchases by offering personalized promotions and recommendations.

4. Optimize Marketing Campaigns: Tailor marketing campaigns for each customer segment to improve the effectiveness of promotions, upselling, and cross-selling efforts.

5. Enhance Business Decision-Making: Leverage data-driven insights to prioritize business initiatives such as customer loyalty programs, win-back campaigns, and onboarding processes for new customers.

Data Description

Feature Name	Description	Data Type
Invoice	Invoice number	Integer
StockCode	Product ID	String
Description	Product Description	String
Quantity	Quantity of the product	Integer
InvoiceDate	Date of the invoice	Datetime
Price	Price of the product per unit	Float
CustomerID	Customer ID	Integer
Country	Region of Purchase	String

The data contains 541909 rows and 8 columns out of which 136534 are missing values.

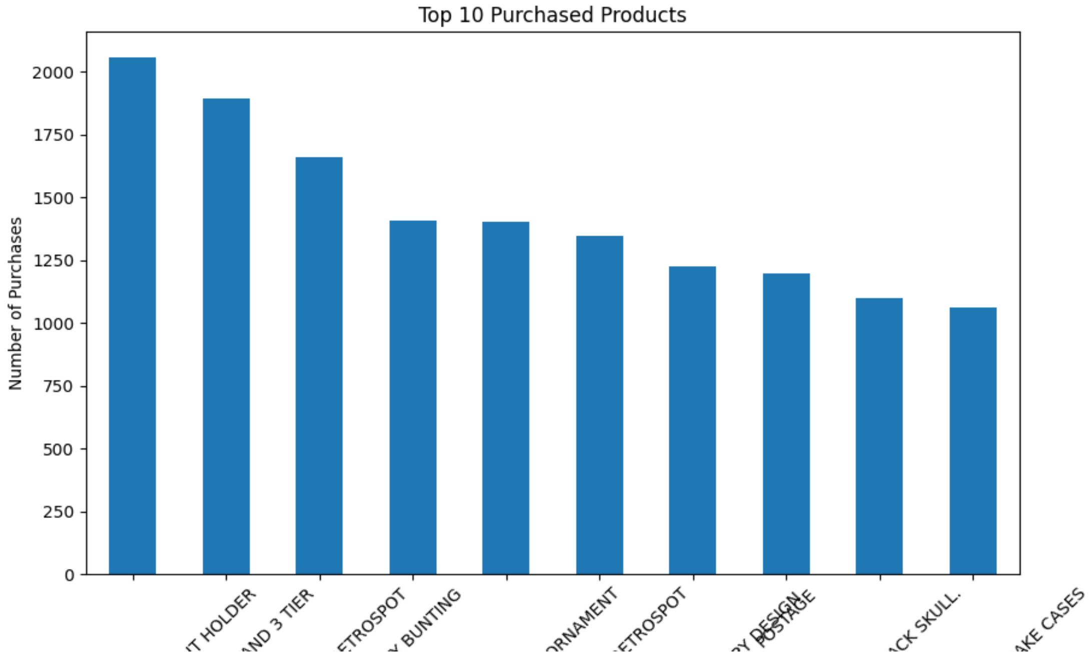
Data Pre-processing Steps and Inspiration

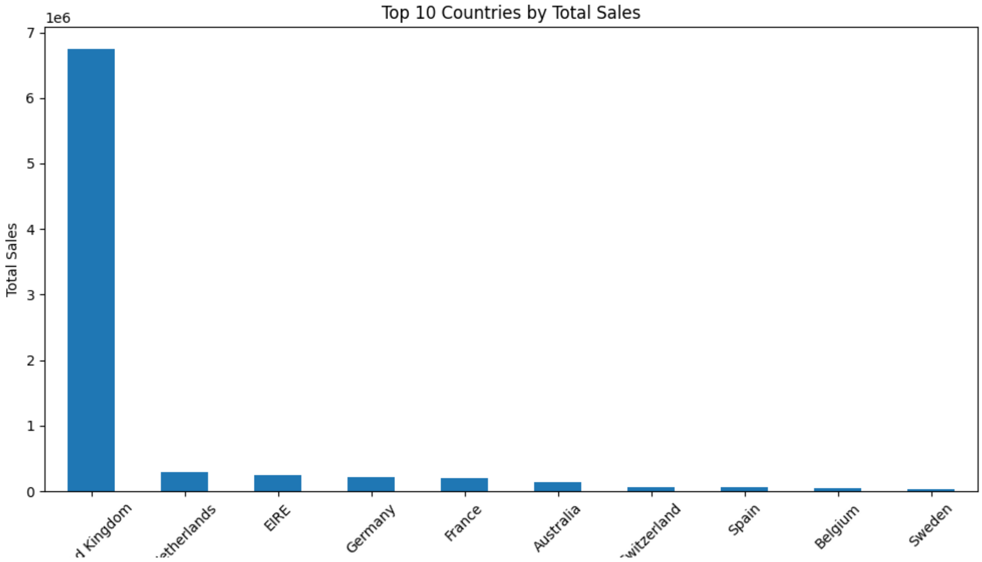
1. **Handling Missing Values:**
    - CustomerID: There are 135080 CustomerID missing values in the dataset. Since customer segmentation relies on unique customer identifiers, any rows without a CustomerID are removed to ensure proper analysis.
    - Product Description: 1454 missing descriptions are removed, as they do not affect the primary analysis focused on customer behavior.
2. **Handling Duplicates:**
    - Duplicate entries can distort the analysis. Removing duplicates ensures that each transaction is considered only once in customer behavior analysis.
3. **Data Type Conversion:**
    - InvoiceDate is converted to DateTime format for accurate time-based analysis (such as calculating recency).
4. **Handling Outliers:**
    - Techniques like boxplots are used to detect and manage outliers effectively. Outliers in Quantity and Unit price are treated by removing them.

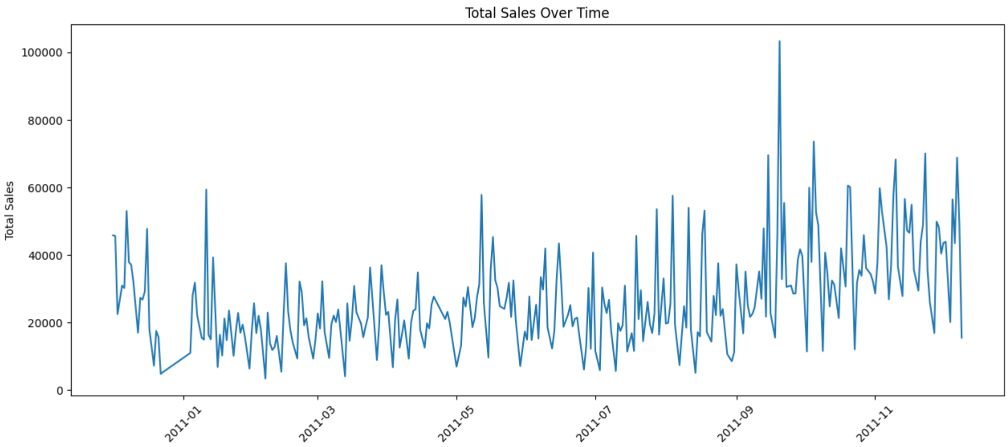
Inspiration

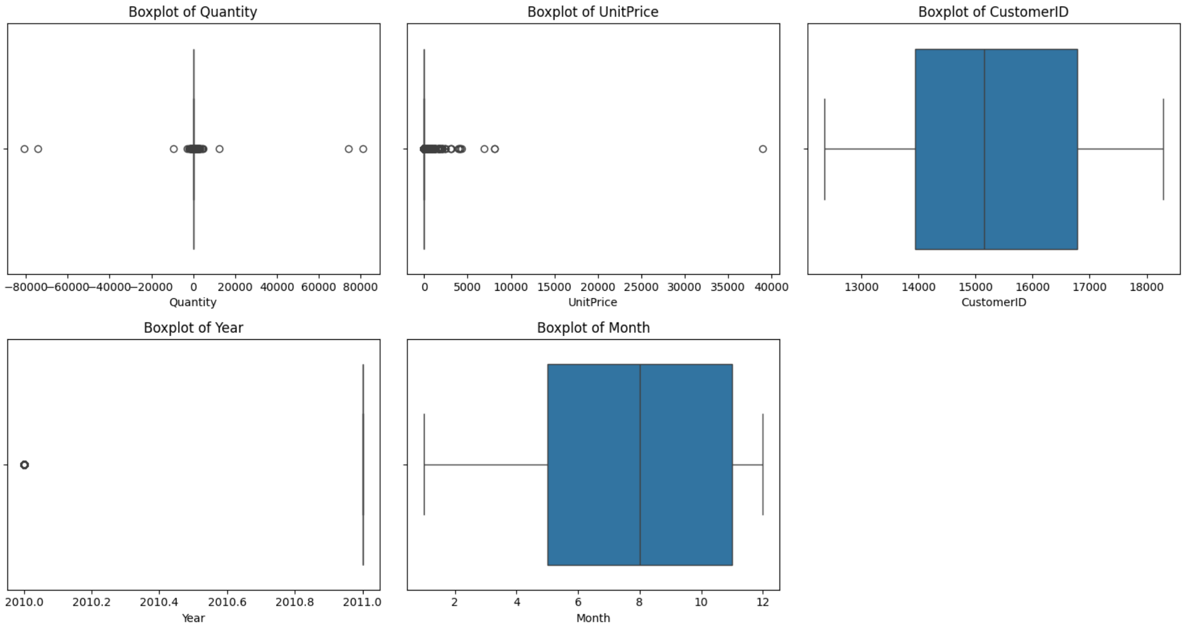
- **RFM Analysis:** The segmentation model is inspired by the classic Recency, Frequency, Monetary (RFM) analysis, which is widely used in marketing to categorize customers based on their purchasing behavior. This method helps businesses identify valuable customers and target them effectively.
- **Customer Segmentation:** The pre-processing aligns with best practices for customer segmentation, ensuring clean data is used to uncover insights about different customer segments. These insights help retailers understand customer loyalty, spending behavior, and the potential for personalized marketing strategies.

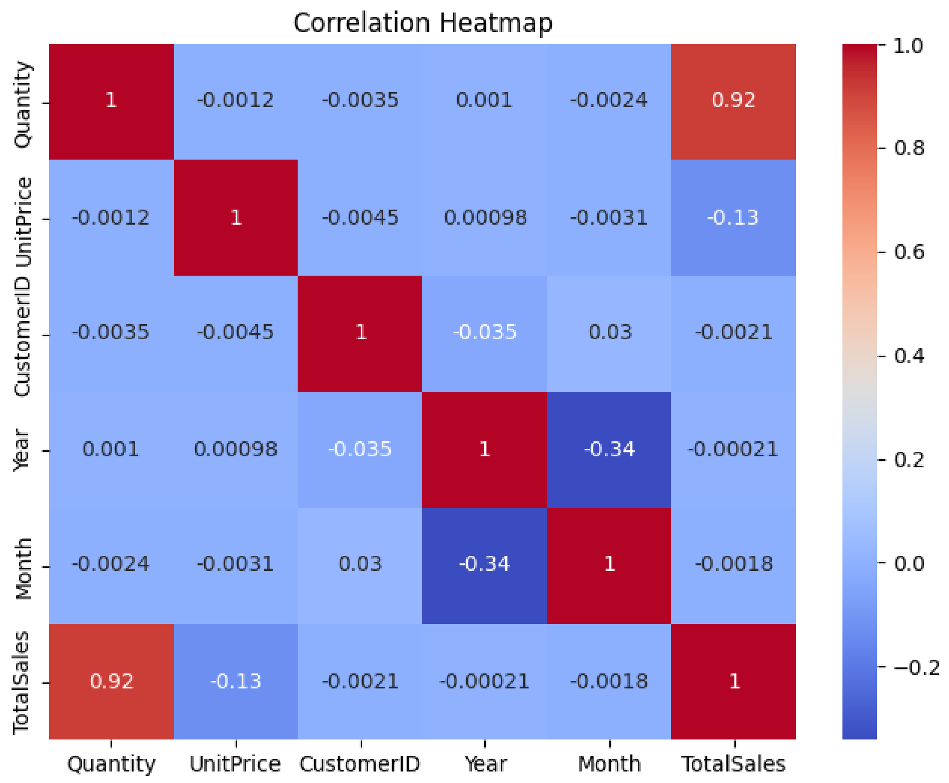
Data Visualization






Total Sales Over Time
Total Sales
0
20000
40000
60000
80000
100000
2011-01
2011-03
2011-05
2011-07
2011-09
2011-11


Boxplot of Quantity
−80000
−60000
−40000
−20000
0
20000
40000
60000
80000
Quantity
Boxplot of UnitPrice
0
5000
10000
15000
20000
25000
30000
35000
40000
UnitPrice
Boxplot of CustomerID
13000
14000
15000
16000
17000
18000
CustomerID
Boxplot of Year
2010.0
2010.2
2010.4
2010.6
2010.8
2011.0
Year
Boxplot of Month
2
4
6
8
10
12
Month



Choosing the Algorithm for the Project

K-Means: Chosen for its simplicity, scalability, and interpretability. Used to form distinct customer segments.

Hierarchical Clustering: Provides additional insights into the relationships between clusters and doesn't require pre-setting the number of clusters.

PCA: Used for dimensionality reduction and to visualize the clusters effectively.

Motivation and Reasons For Choosing the Algorithm

K-Means Clustering

Motivation:

- The primary goal is to group customers into distinct segments. K-Means is well-suited for this because it efficiently partitions customers into predefined clusters.
- K-Means is computationally efficient and works well on relatively large datasets. The dataset in this project has multiple rows (transactions) that need to be grouped, and K-Means can handle this task in a reasonable time frame without compromising performance.

Reasons for Choosing K-Means:

- K-Means produces well-separated clusters, making it easier to interpret the customer groups and their behavior.

- The **Elbow Method** allows us to choose the optimal number of clusters (k) by measuring the sum of squared errors (inertia). This ensures that the segments created are meaningful and not arbitrarily selected.

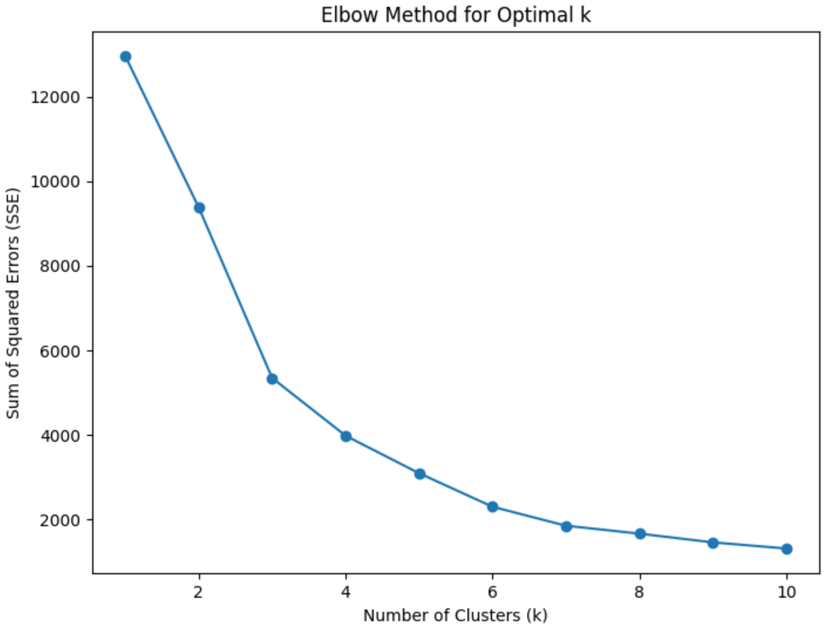
Assumptions

1. Assuming data is accurate, complete, and representative of customer behavior.
2. Assuming that Recency, Frequency, and Monetary value capture the most important aspects of customer behavior.
3. Assuming clusters are spherical and equidistant, with well-separated customer groups.
4. Assuming that reducing dimensionality via linear combinations captures the majority of important information.
5. Assuming customer behavior is stable and homogenous within segments.

Model Evaluation and Techniques

1. Elbow Method (for K-Means Clustering)

- **Purpose**: To determine the optimal number of clusters (k) for K-Means clustering.
- **How it Works**: The Elbow Method evaluates the sum of squared distances (inertia) between the data points and their corresponding cluster centroids. The goal is to identify the "elbow point" on the inertia curve, where adding more clusters no longer significantly reduces the inertia.
- **Interpretation**: The optimal k is typically chosen at the point where the inertia starts decreasing at a slower rate, indicating that additional clusters no longer provide substantial improvement.

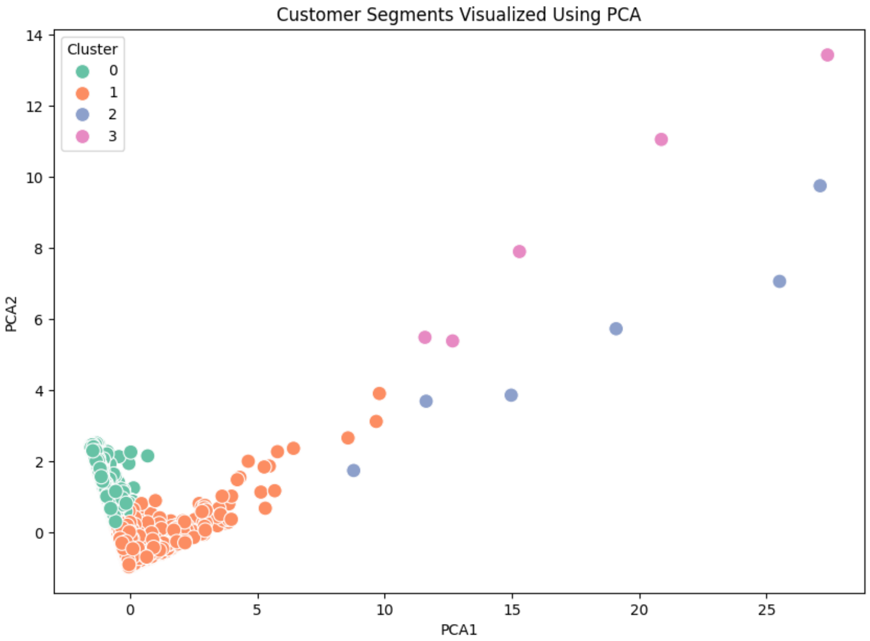


2. Silhouette Score

- **Purpose**: To assess the quality of the clusters created by K-Means or other clustering algorithms.
- **How it Works:** The silhouette score measures how similar each data point is to its own cluster compared to other clusters. The score ranges from -1 to 1:
    - A score close to 1 means that the data point is well-clustered and far from neighboring clusters.
    - A score close to 0 means that the data point is on or very close to the decision boundary between two clusters.
    - A negative score means that the data point has been misclassified and is closer to a neighboring cluster than its assigned cluster.
- **Interpretation**: A higher average silhouette score indicates better-defined clusters. Scores closer to 1 suggest well-separated clusters, while scores near 0 suggest overlap or poor separation between clusters.
    - From our model, we got a silhouette score of 0.602, which indicates that we have distinct and well-separated clusters.

1. Principal Component Analysis (PCA) for Visualization

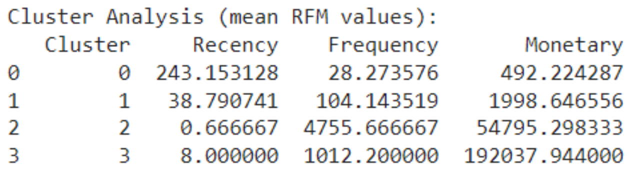
- Purpose: To visualize high-dimensional data and clustering results in a 2D or 3D space.
- **How it Works**: PCA reduces the dimensionality of the data by transforming it into a set of principal components. These components capture the majority of the variance in the data. The first two or three principal components are used to visualize the clustering results.
- Interpretation: If the clusters are well-separated in the reduced-dimensionality space, this provides a visual confirmation of the clustering effectiveness. PCA helps to validate whether the clusters are distinct and easily interpretable.



2. RFM Segmentation Evaluation

- **Purpose**: To ensure the quality and relevance of customer segments based on Recency, Frequency, and Monetary (RFM) values.
- **How it Works**: After clustering customers based on RFM metrics, the segments are evaluated in terms of their practical relevance. For example:
  - o Do the clusters correspond to meaningful business categories like "high spenders" or "churned customers"?
  - o Are there clear differences in average RFM values across the clusters?
- **Interpretation**: The success of the segmentation is judged by whether the clusters make sense from a business perspective and whether they can be used for targeted marketing or other strategic initiatives.

Cluster Analysis (mean RFM values):



```
Cluster Analysis (mean RFM values):
   Cluster     Recency    Frequency       Monetary
0        0  243.153128    28.273576     492.224287
1        1   38.790741   104.143519    1998.646556
2        2    0.666667  4755.666667   54795.298333
3        3    8.000000  1012.200000  192037.944000
```

Low Recency, High Frequency, High Monetary (Loyal Customers): These customers buy frequently, spend more, and purchased recently. **(Clusters 2 & 3)**

High Recency, Low Frequency, Low Monetary (Churning Customers): These customers haven't purchased recently and spend less. **(Cluster 0)**

Low Frequency, High Monetary (High-Value Shoppers): These customers don’t purchase often, but when they do, they spend a lot. **(Cluster 0)**

Inferences from the Same

1. **Customer segments** based on Recency, Frequency, and Monetary (RFM) analysis are distinct and actionable.
2. **High-value and churned customers** are key segments where targeted strategies like loyalty programs and reactivation campaigns can increase engagement and profitability.
3. **Visualization techniques (PCA)** confirm that customer behavior is well-clustered, providing confidence in the segmentation.
4. **An optimized number of clusters** ensures balanced and meaningful segmentation, helping the retailer focus on specific customer groups.

Future Possibilities of the Project

1. **Personalized Marketing Campaigns:**
   - Leverage customer segments to deliver personalized promotions, product recommendations, and tailored marketing messages. Target high-value customers with loyalty programs and re-engage churned customers through win-back campaigns.
2. Customer Lifetime Value (CLV) Prediction:
   - Use customer segmentation to predict CLV and identify potential high-value customers early on. This would help the retailer focus efforts on retaining and nurturing valuable customers.
3. Integration with Additional Data:

- Enrich the current model with external data (e.g., social media interactions, website activity) to build a more comprehensive view of customer behavior, enabling even more precise segmentation and personalized experiences.

4. Predictive Analytics and Behavior Forecasting:

- Extend the project to predict future customer behavior, such as predicting purchase frequency, likelihood of churn, or response to promotional offers. Predictive models can guide proactive marketing strategies.

5. Recommendation Systems:

- Implement a recommendation engine based on the segments, suggesting products that align with the purchasing behavior of each segment, improving customer engagement and increasing sales.

Conclusion

This project successfully segmented the online retailer's customer base using Recency, Frequency, and Monetary (RFM) metrics, identifying distinct customer groups with meaningful purchasing behaviors. Through data preprocessing, clustering, and visualization, we derived actionable insights that can guide personalized marketing, customer retention, and targeted engagement strategies.

The project provides a strong foundation for enhancing customer relationships, improving retention, and driving higher profitability through data-driven strategies. Additionally, it opens avenues for further development, such as real-time segmentation, predictive analytics, and personalized marketing.