Nithin G

Project 2

Online Retail

17-05-2023

Online Retail: Report

a. Problem Statement:

The problem is to analyze the customer purchase patterns for an online retail store and provide useful insights that can benefit the retailer. Additionally, the goal is to segment the customers based on their purchasing behavior.

b. Project Objectives

- Identify key insights and patterns in the customer purchase history.
- find useful insights about the customer purchasing history that can be an added advantage for the online retailer
- Provide actionable recommendations for the online retailer based on the identified insights.
- Segment the customers based on their purchasing behaviour.

c. Data Description:

The dataset provided for this project, online_retail.csv, contains the following information:

Invoice: Invoice: number associated with a purchase.

StockCode: Product ID or code for the purchased item.

Description: Description of the product.

Quantity: The quantity of the product purchased.

InvoiceDate: The date when the invoice was generated.

Price: The price of the product per unit.

CustomerID: Unique identifier for each customer.

Country: The region or country where the purchase was made.

The dataset consists of **387,961** rows and **8** columns, capturing relevant information about customer purchases made on the online retail store.

d. Data Pre-processing Steps and Inspiration

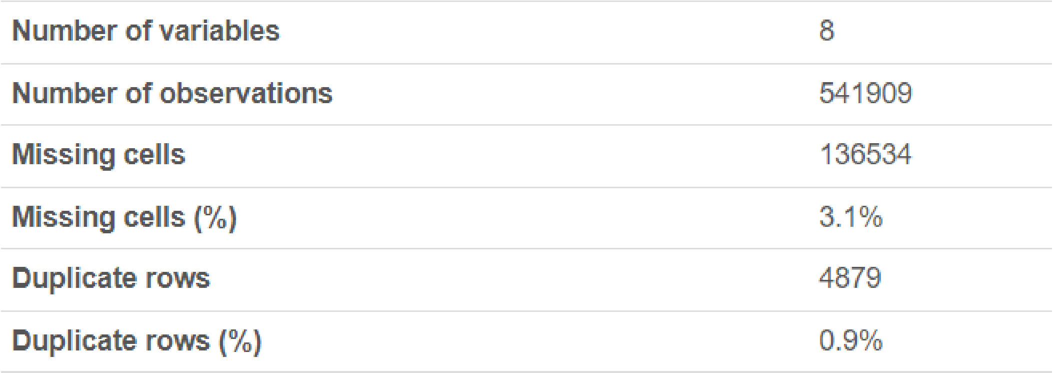
- Handling Missing Values:

  There are 24% of null values present in the customer id and .2 % in the description

  I removed the missing values using the pandas dropna() function

- Duplicate Rows:

  There are 4879 duplicated rows and dropped that columns



Number of variables	8
Number of observations	541909
Missing cells	136534
Missing cells (%)	3.1%
Duplicate rows	4879
Duplicate rows (%)	0.9%

- Data Transformation:

  Converted InvoiceDate datatype objective to DateTime and CustomerID as type str

- Feature Engineering:

  month and **year**: Extract from InvoiceDate

  Revenue: Quantity * UnitPrice

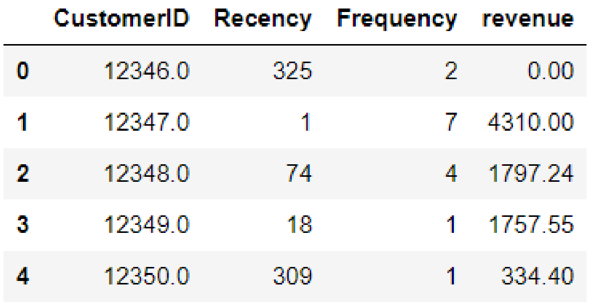
  DateDiff: Find the date difference last purchase date and the current date(max(date))

- Group by:

  Groupby by customerID, DateDiff(min), invoice no(unique), and revenue(sum)

  This will give insights into customer purchasing behaviour and identify high-value customers.

  And renamed it, customer_data.columns = ['CustomerID', 'Recency', 'Frequency', 'revenue']



	CustomerID	Recency	Frequency	revenue
0	12346.0	325	2	0.00
1	12347.0	1	7	4310.00
2	12348.0	74	4	1797.24
3	12349.0	18	1	1757.55
4	12350.0	309	1	334.40

- Outlayers Removal:

  Found and Removed outlayers from Recency, Frequency and revenue Columns

- Feature Selection:

  Selected These Features(Recency, Frequency and Revenue) For my model

e. Choosing the Algorithm for the Project

This project selects The K-means algorithm for customer segmentation due to its simplicity, scalability, and interpretability. It is suitable for analyzing customer purchase patterns without requiring pre-labeled data. By partitioning customers into clusters based on similarity, K-means helps identify distinct groups and understand their purchasing behavior.

f. Motivation and Reasons For Choosing the Algorithm

K-means algorithm lies in its simplicity, scalability, interpretability, ability to handle unlabeled data, and its established track record in customer segmentation tasks.

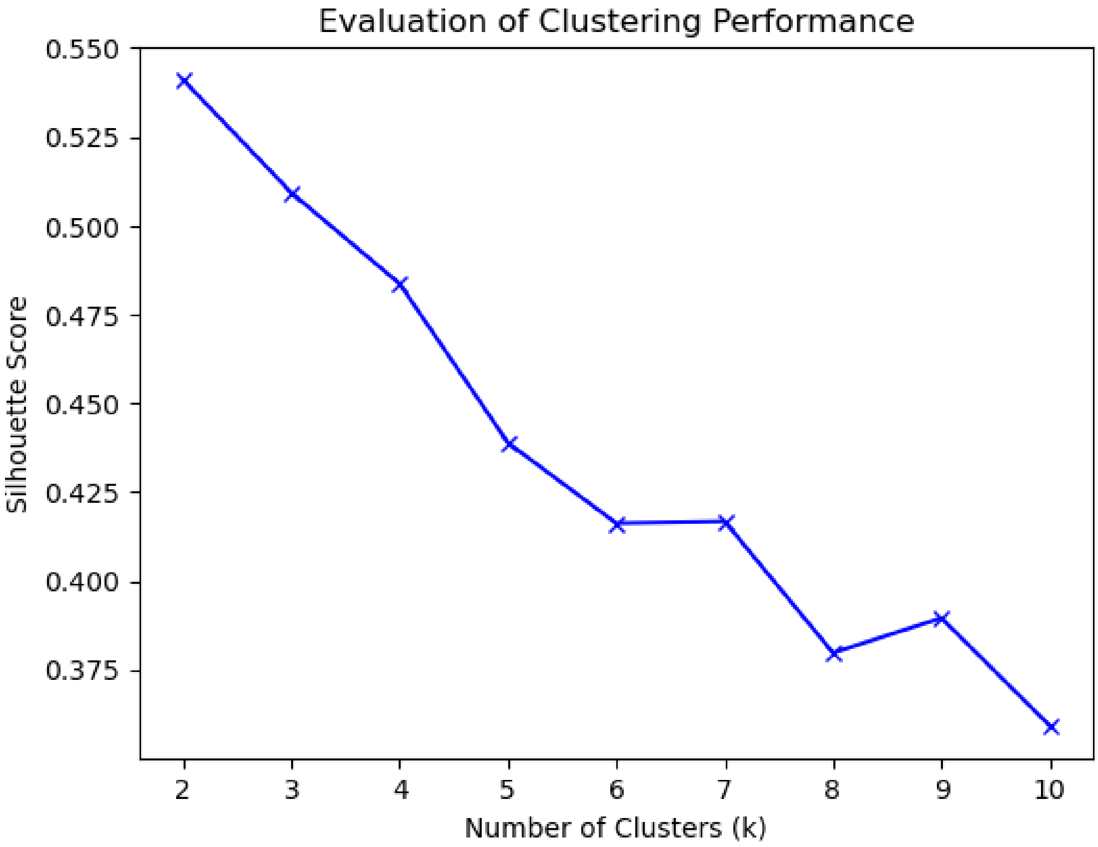
Additionally, I created an Arima model for finding the future buying quantity by customers

g. Assumptions

- The United Kingdom has the highest monthly sales, indicating that it is the most important market of the business
- Germany, France, Netherlands, and EIRE have high monthly sales, telling that these countries are also important markets for the business.
- Some countries have very low monthly sales, such as Saudi Arabia, Czech Republic, and Bahrain, which tells that these markets are not as important or may require further analysis to understand why sales are low.

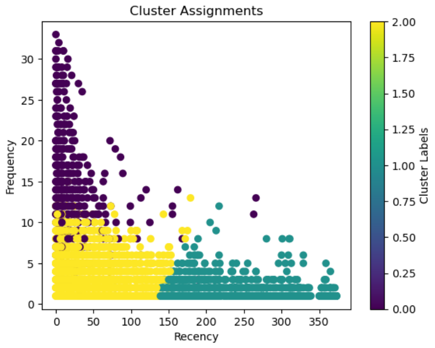
- I created 3 clusters here ie, separated Customers in 3 different Groups

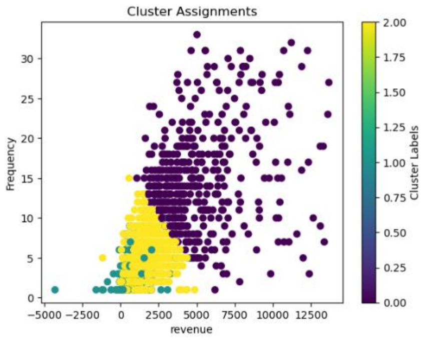
  Evaluate each cluster and its score 2 clusters give the best score, but 2 are small groups so I created 3 it easy when evaluating this groups



h. Model Evaluation and Techniques

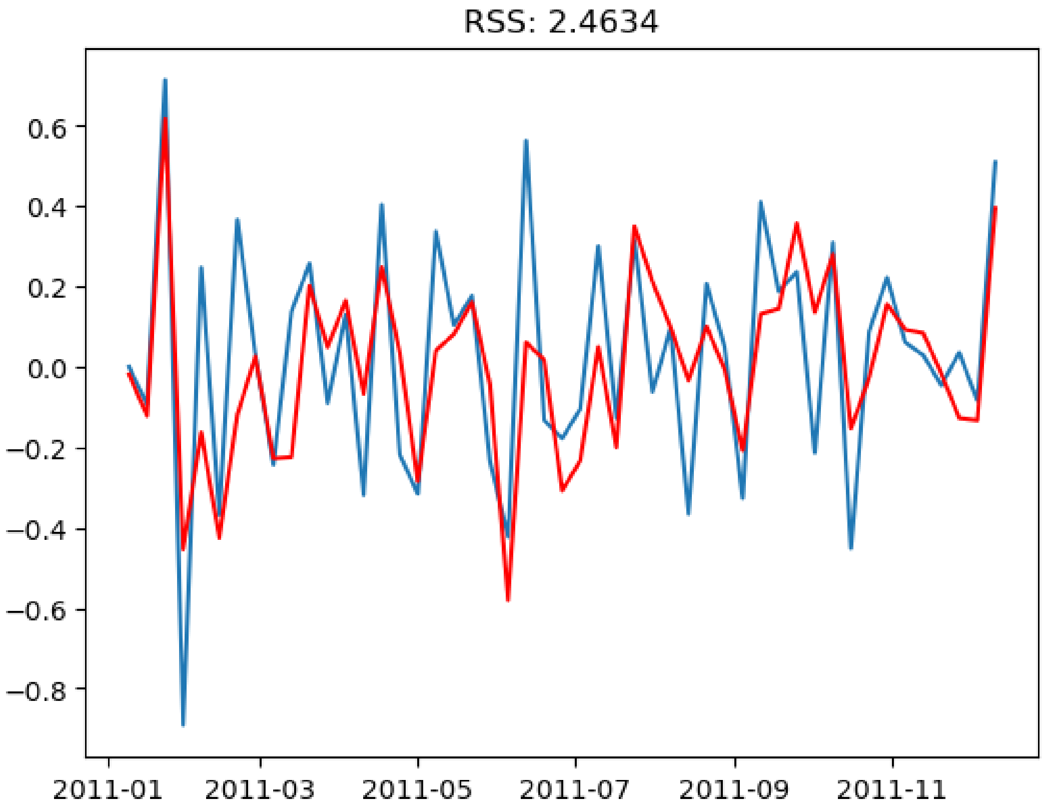
Using Silhouette Score Graph is above





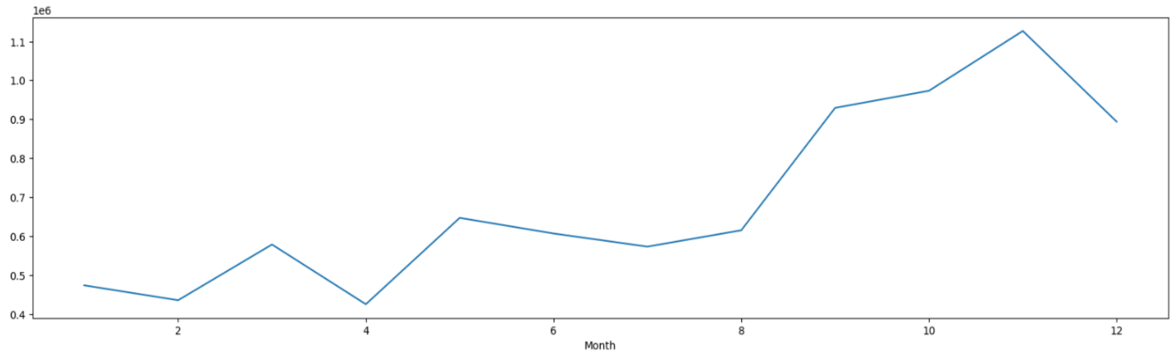
By Arima

RMSE = 0.211170

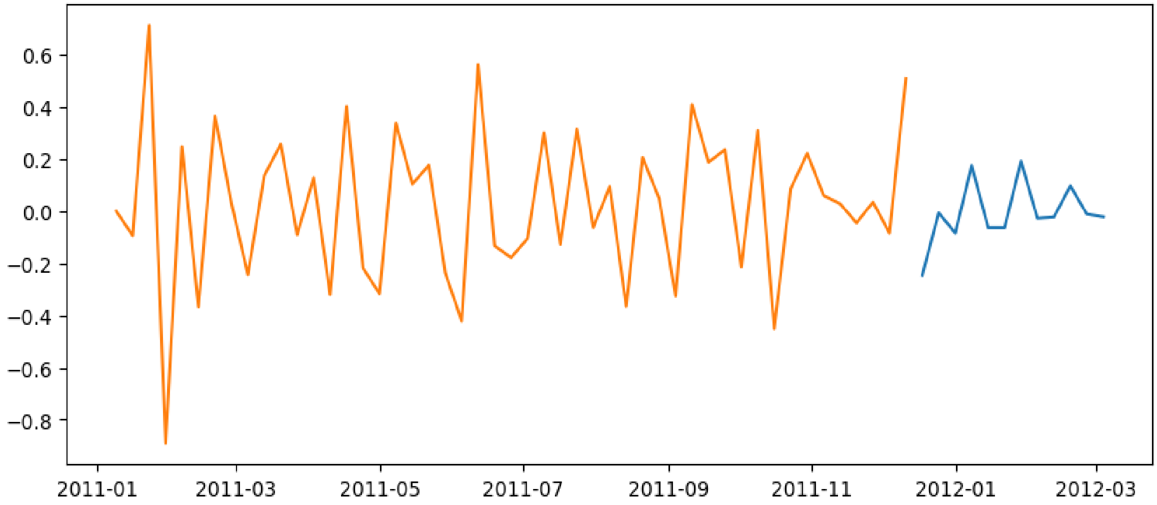


i. Inferences from the Same

- Customer Behavior Patterns: The identified customer segments provide a clear understanding of different behaviour patterns exhibited by customers, as seen above
- High-Value Customers: The segmentation analysis helps identify high-value customers who contribute significantly to the overall revenue
- Revenue by Month: find the revenue by each month



- Arima Prediction



j. Future Possibilities of the Project

- Personalized Product Recommendations: Utilize customer purchase history and segmentation insights to offer personalized product recommendations, enhancing cross-selling and upselling opportunities.

- Targeted Offers and Discounts: Tailor promotions and discounts based on customer segmentation, maximizing engagement and conversion rates.

- Country-Specific Strategies: Identify regions requiring attention based on purchasing patterns, enabling localized marketing efforts and customized product offerings.
- Demand Forecasting: Develop demand forecasting models using historical data to optimize inventory management and supply chain processes.