Infosys Springboard 5.0 Internship

Project Report:

Predicting Sales Trends
for a
TATA Online Retail Dataset

Submitted By:

Sai Rithin Chakka

Arpan Dutta

Sathvika

Shiva Keshav Kumar

1

Table of Contents

2

1. Introduction

1.1 Problem Statement

Predicting sales trends is one of the most important business problems for any retail entity. If a business can predict the how much of each item it will sell in each month, it can manage its inventory better. Sales predictions also help in directing the marketing efforts in right direction to increase the chances of sale.

1.2 Objective

To Predict sales trends in the TATA online retail dataset.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

The dataset has been taken from Kaggle datasets :
http://www.kaggle.com/datasets/ishanshrivastava28/tata-online-retail-dataset

The dataset contains the following attributes:

Attribute name	Type	Description
InvoiceNo	Nominal	A 6-digit integral number uniquely assigned to each transaction.
StockCode	Nominal	A 5-6 digit integral number uniquely assigned to each distinct product
Description	Nominal	Product (item) name
Quantity	Numeric	The quantities of each product (item) per transaction
InvoiceDate	Numeric	The day and time when each transaction was generated
UnitPrice	Numeric	Product price per unit in sterling

CustomerID	Nominal	A 5-digit integral number uniquely assigned to each customer
Country	Nominal	Name of the country where each customer resides

2.2. Data Cleaning

Initial loading and inspection of datasets exposed some challenges in it for our study. We wrangled the data to make it fit for our analysis.

1. Handling Null Values in 'Description' and 'CustomerID' Columns
    - Each InvoiceNo should be linked to a single CustomerID. So we tried using InvoiceNo and CustomerId linkage to fill missing values. As we could not find the linkage, we dropped the null values
    - Deleting missing CustomerId removed all missing Description rows too.

2. Inconsistent Mapping Between 'StockCode' and 'Description'
    - Each StockCode should uniquely represent an item Description. But the original dataset has multiple Description for same StockCode. This is because there are data entry errors in the description as shown below

	InvoiceNo	StockCode	Description
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER
1	536365	71053	WHITE METAL LANTERN
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.

The data was wrangled to contain one to many mapping between StockCode and Description

3. Some CustomerID linked with 2 countries

    As per the data attribute description: 'Country' column is the name of the country where each customer resides. But we don't have any information on how is this data being captured. Is it through IP address of the country while creating account, or may be based on the shipping address, or may be something else.

Logically, each CustomerID should be linked to one country only. The reason for having more than one country could be:

a. Data entry error
b. Customer has moved to another country, and has got the address changed in his account
c. In case this attribute reflects the shipping address, the customer has shipped the order to an address different from his own.
d. In case this attribute is captured through the IP address while ordering, the customer might be ordering while travelling to another country.

Further analysis of data does not make it clear what is the reason behind 2 countries for a CustomerID, so for now, we are not making any changes in the CustomerID and country linkage.

2.3 Derived Features

We derived the following features from the existing InvoiceDate column to aid in our analysis:

1.TotalPrice : Derived Total price from unitPrice and Quantity (unitPrice * Quantity)

2.WeekDay: Extracted the weekday name from the InvoiceDate column using the %a format to represent abbreviated weekday names (e.g., Mon, Tue, etc.).

3.Day: Extracted the day of the month from the InvoiceDate column to help analyze patterns on specific days.

4.Month: Extracted the month as a numerical value from the InvoiceDate column to identify seasonal trends in the data.

5.Year: Extracted the year from the InvoiceDate column to distinguish data from different years and analyze long-term trends.

2.4. Data Cleaning Summary

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

Shape: (541909,8)

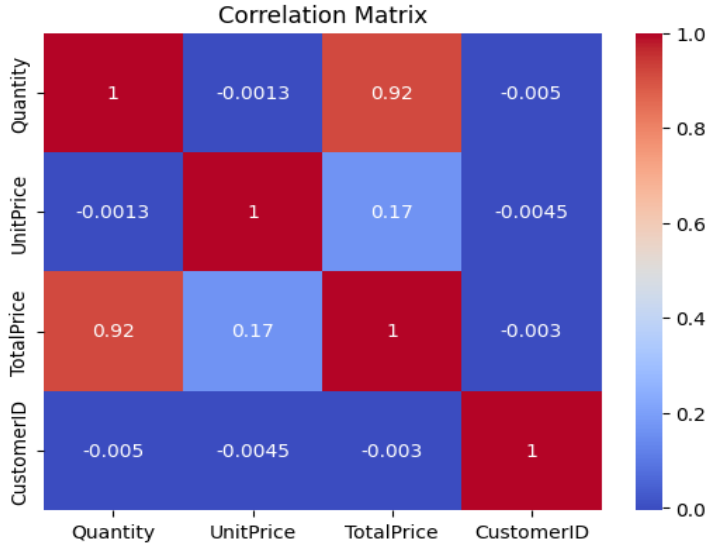
↓

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice	PriceCategory	WeekDay	Day	Month	Year
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	35	15.30	Low Price	5	1	12	2010
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	35	20.34	Low Price	5	1	12	2010
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	35	22.00	Low Price	5	1	12	2010
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	35	20.34	Low Price	5	1	12	2010

Shape: (404618,11)

The code for data acquisition and wrangling can be accessed in this python notebook .

2.5 Correlation matrix



1. **Quantity and TotalPrice Relationship**: There's a strong positive correlation (0.92) between Quantity and TotalPrice, indicating that as the quantity of items increases, the total price also

6

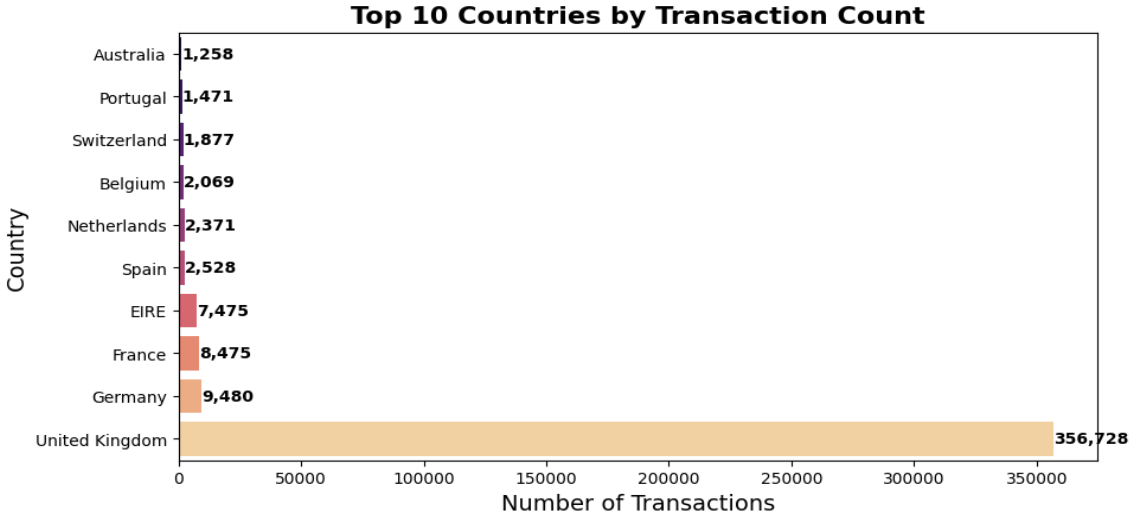
significantly rises.

2. **CustomerID's Weak Influence**: CustomerID shows very weak correlations with Quantity, UnitPrice, and TotalPrice, suggesting that it doesn't have a significant linear relationship with these variables.

3. Data Exploration

The code for EDA can be found in this Python notebook

3.1 Worldwide distributions



1. The **United Kingdom** dominates the transaction count with 356,728, indicating it as the primary market.

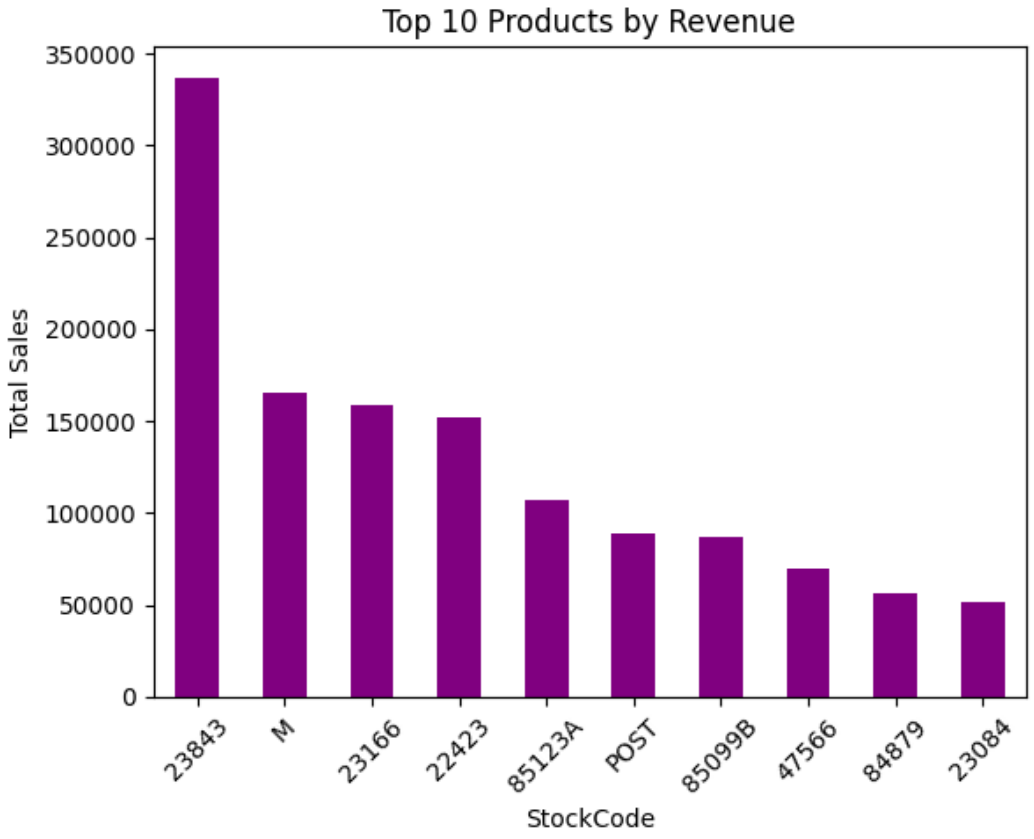
2. **Germany** (9,480) and **France** (8,475) follow as secondary contributors with significantly lower transaction volumes.

3. Other countries, such as **EIRE**, **Spain**, and **Netherlands**, show moderate transaction counts, while **Australia** and **Portugal** have the lowest among the top 10.

4. The chart highlights a clear geographical skew in transaction distribution, emphasizing the importance of localized strategies.

7

3.2 Best seller



1. Product **23843** generates the highest revenue, significantly outperforming other products in the top 10.

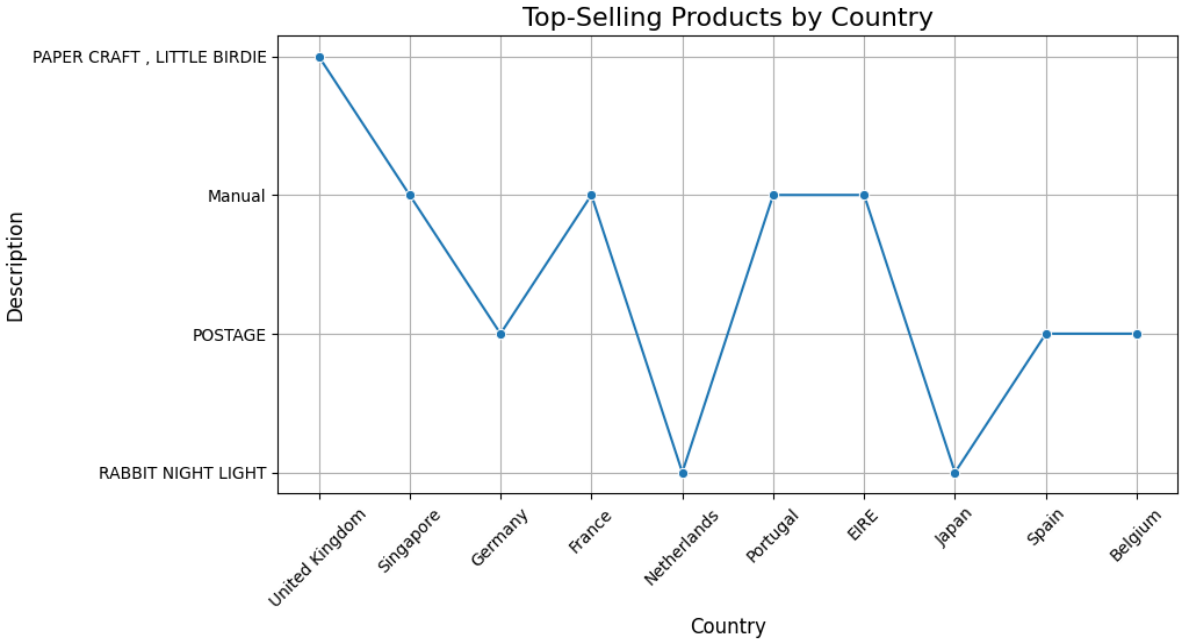
2. Products **M**, **23166**, and **22423** also contribute substantially to total sales, forming the next tier of high-revenue items.

3. Products **85123A**, **POST**, and others in the lower range show moderate sales, indicating varied demand levels within the top 10.

4. This chart highlights the disproportionate contribution of a few products, suggesting potential areas for inventory and sales focus.

8

3.3 Top Selling products



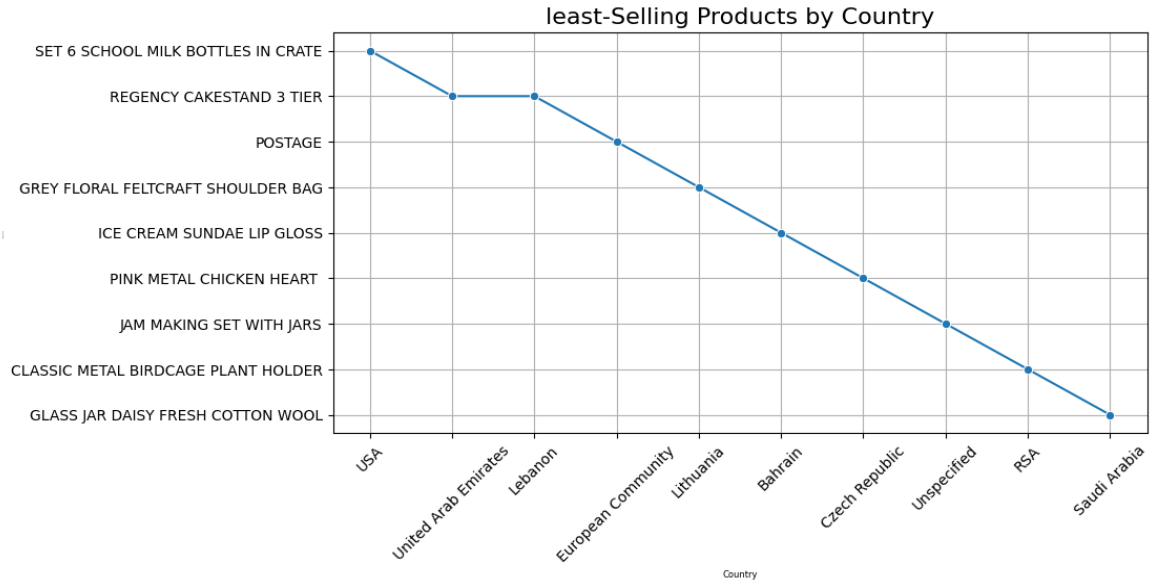
1. The chart showcases the top-selling products by country, revealing variations in product demand across regions.

2. The **United Kingdom** leads in sales for "PAPER CRAFT, LITTLE BIRDIE," indicating high local popularity for this item.

3. Products like **POSTAGE** and **Manual** have consistent presence in multiple countries, reflecting universal utility or appeal.

4. The demand for items like **RABBIT NIGHT LIGHT** varies significantly by region, with some countries showing little or no interest.

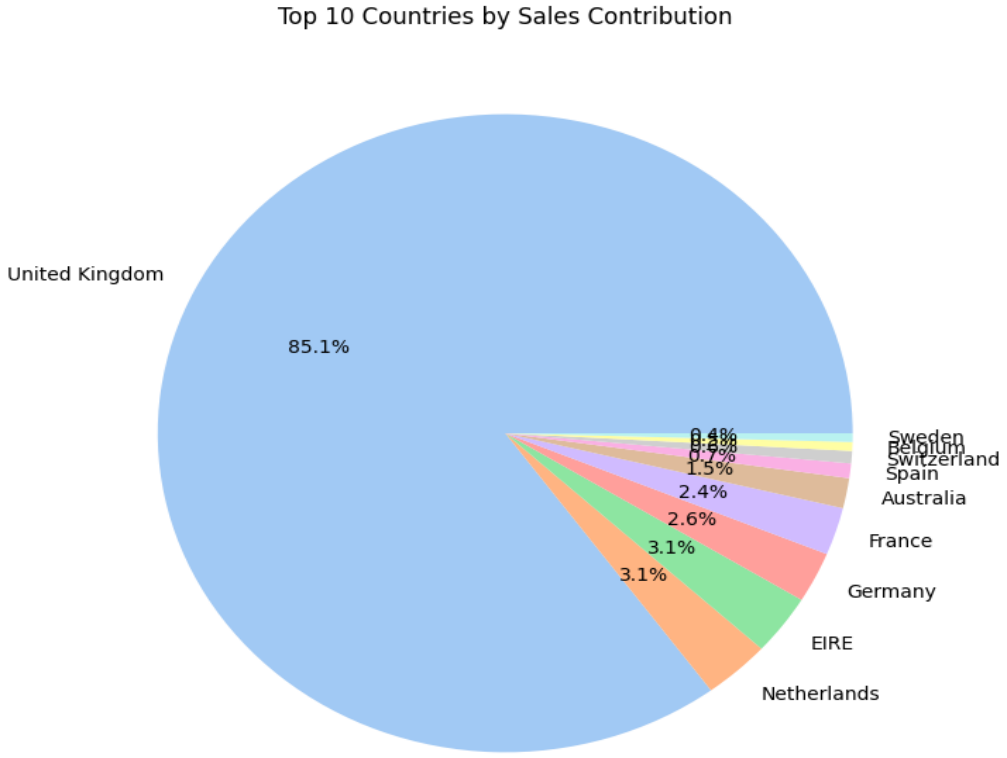
3.4 Least Selling Products



9

1. The chart highlights the least-selling products by country, shedding light on regional variations in product popularity.
2. The USA shows low sales for "REGENCY CAKESTAND 3 TIER," suggesting minimal demand for this item locally.
3. Products like "PINK METAL CHICKEN HEART" and "ICE CREAM SUNDAE LIP GLOSS" appear across multiple countries, indicating a broader, albeit niche, market.
4. Items such as "CLASSIC METAL BIRDCAGE PLANT HOLDER" have varied demand, with some countries showing little to no interest.

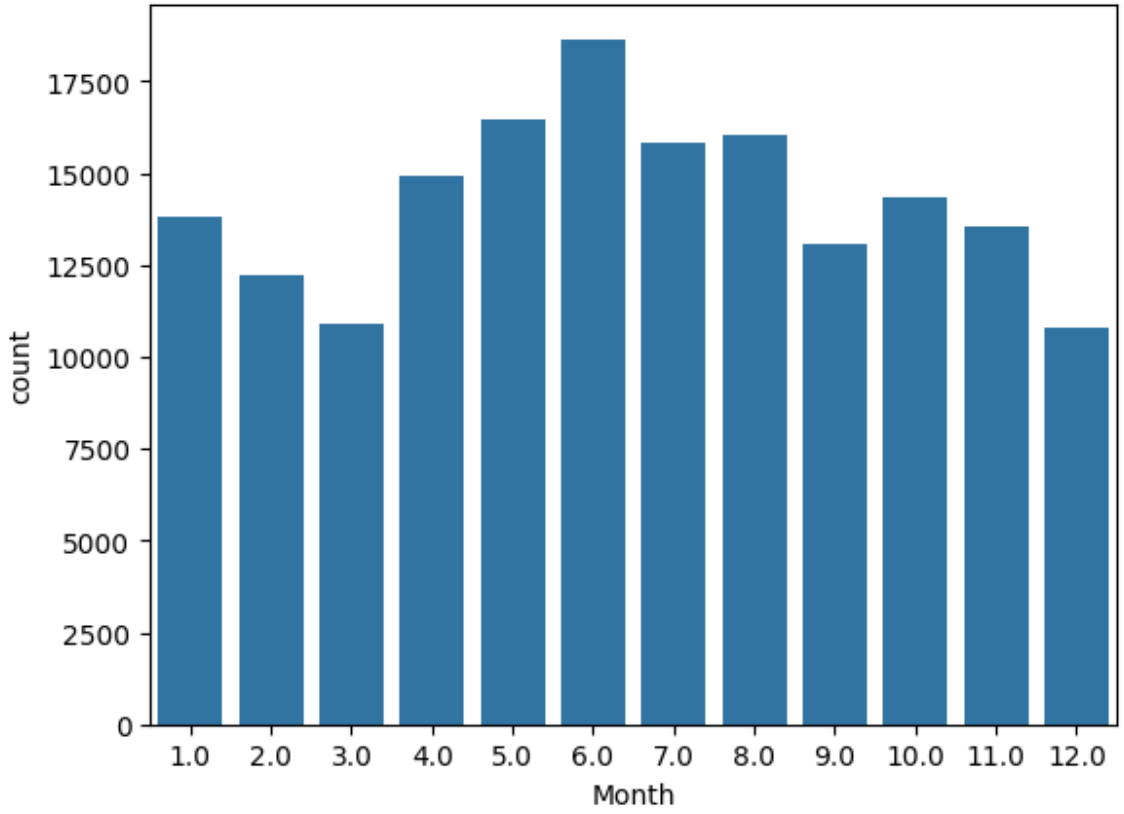
3.5 Top countries



10

1. The United Kingdom leads with a significant sales contribution of 85.1%, indicating a major market dominance.
2. The Netherlands and EIRE both contribute 3.1% each to the sales, highlighting them as notable secondary markets.
3. Countries like Germany (2.6%), France (2.4%), and Australia (1.5%) also make meaningful contributions, though substantially less compared to the UK.
4. Spain (1.0%), Switzerland (0.9%), Belgium (0.8%), and Sweden (0.4%) round out the top 10, reflecting smaller yet still significant market shares.

3.6 Monthly Activity Trends

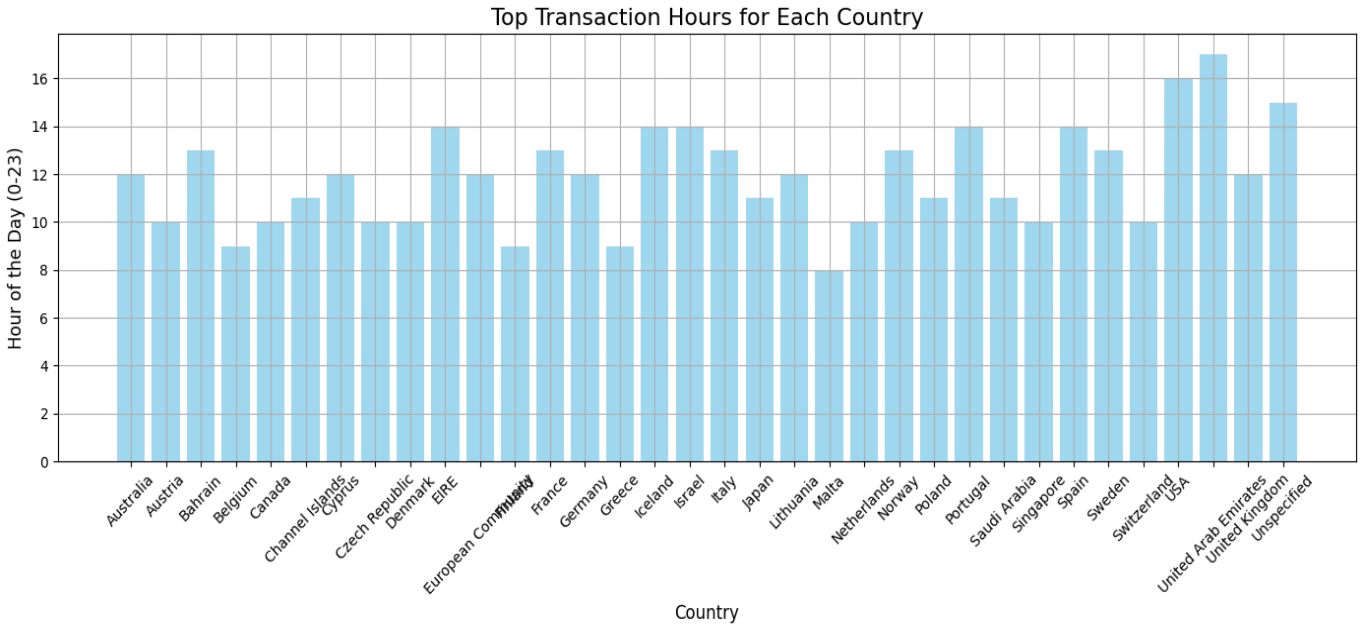


1. The highest activity is in June, with about 18,000 counts, likely due to seasonal factors.
2. May, June, July, and August have high activity, all above 15,000 counts. August is slightly lower at around 15,500.
3. Activity drops in winter, with counts of 13,500 in January, 12,000 in February, and 11,500 in December.

11

4. March, April, September, October, and November show moderate activity, between 11,000 and 14,500 counts.

3.7 Top Transaction Hours



1. Each country has a different peak transaction hour, indicating diverse consumer behavior patterns.
2. The USA sees the highest transactions at 16:00, reflecting a late afternoon peak.
3. Several countries, including Belgium and EIRE, peak at 08:00, suggesting early morning activity.
4. The peak transaction hours range from early morning to late afternoon across the countries, highlighting significant variations in transaction times globally.

4. Customer Segmentation using RFM Analysis

It is about customer segmentation using RFM analysis, which is a marketing technique used to identify and group customers based on their purchasing behavior. The three components of RFM are:

- **Monetary**: Total spending of the customer.
- **Frequency**: Number of transactions made by the customer.
- **Recency**: Days since the last purchase made by the customer.

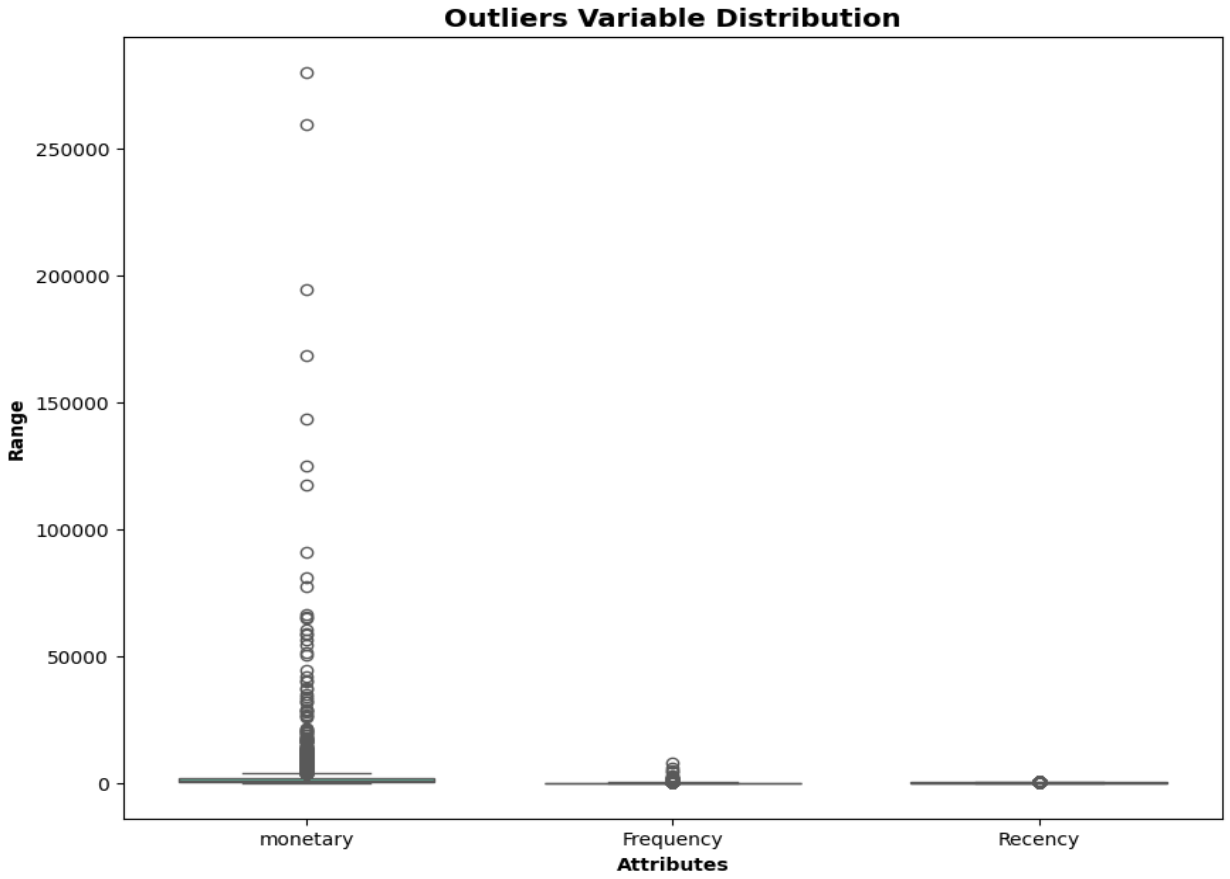
The goal of this analysis is to create a Data Frame that contains these three metrics for each customer, which can then be used for targeted marketing strategies.

12

4.1 RFM metrics calculation

In this analysis, we calculated the **Monetary** value by summing the **TotalPrice** for each customer, the **Frequency** by counting the number of **InvoiceNo** transactions, and the **Recency** by determining the days since the last purchase, all by grouping the data using **CustomerID**.

4.2 Outliers



13

1. The monetary variable has a substantial number of outliers, with values reaching up to around 275,000. This suggests a wide range of monetary values and potential extreme values in the dataset.

2. Both the Frequency and Recency variables have outliers, but they are much less extreme compared to the monetary variable. This indicates that while there are some unusual values, they are not as pronounced as those in the monetary variable.

To clean the dataset and remove extreme values, the "monetary," "Recency," and "Frequency" columns are filtered using the Interquartile Range (IQR) method. Refer to the code for details on the implementation. This process enhances the reliability of the data for further analysis.

4.3 Standardization

To scale the "monetary," "Frequency," and "Recency" columns, we used `StandardScaler`. This process standardizes the features by removing the mean and scaling to unit variance. Refer to the provided code for the implementation details. This ensures that all features contribute equally to the analysis, enhancing model performance.

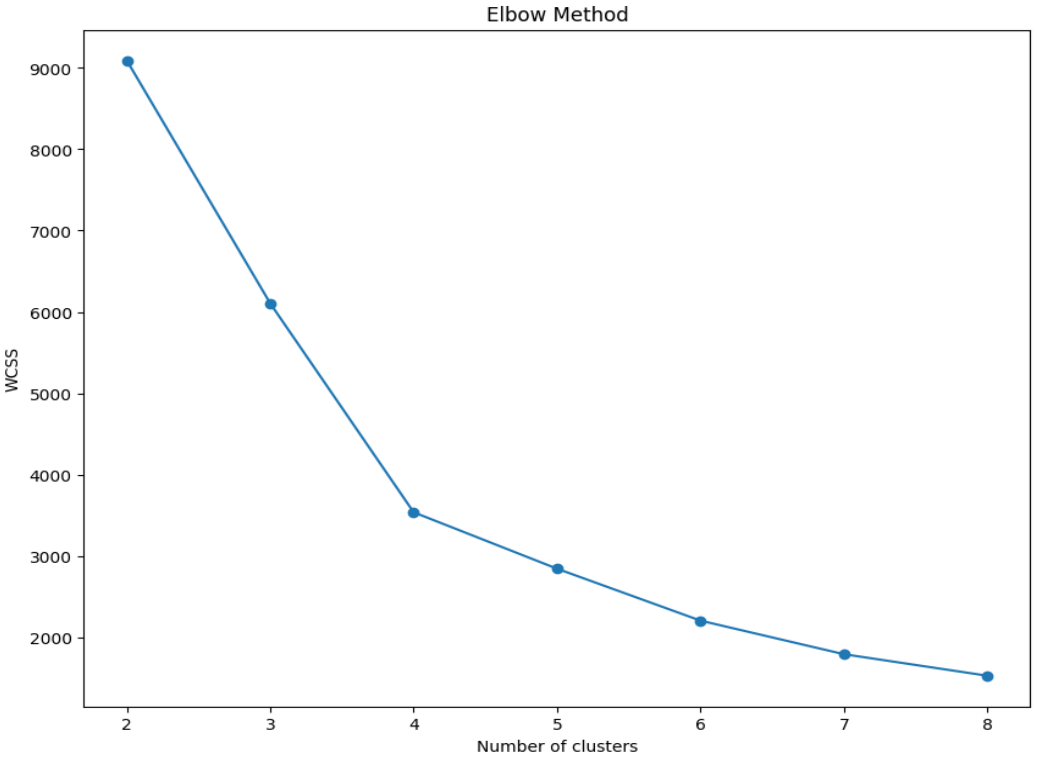
Data Frame after scaling:-

	monetary	Frequency	Recency
0	10.689355	-0.752318	2.326789
1	0.355086	1.008132	-0.910294
2	-0.001251	-0.460531	-0.180951
3	-0.006879	-0.052028	-0.740447
4	-0.208697	-0.596698	2.166934
...	...	...	...
4304	-0.230508	-0.664782	1.847222
4305	-0.244658	-0.693961	0.878095
4306	-0.230869	-0.645329	-0.850348
4307	0.033960	6.250579	-0.890312
4308	0.004428	-0.081207	-0.500663

4.4 K-means Clustering

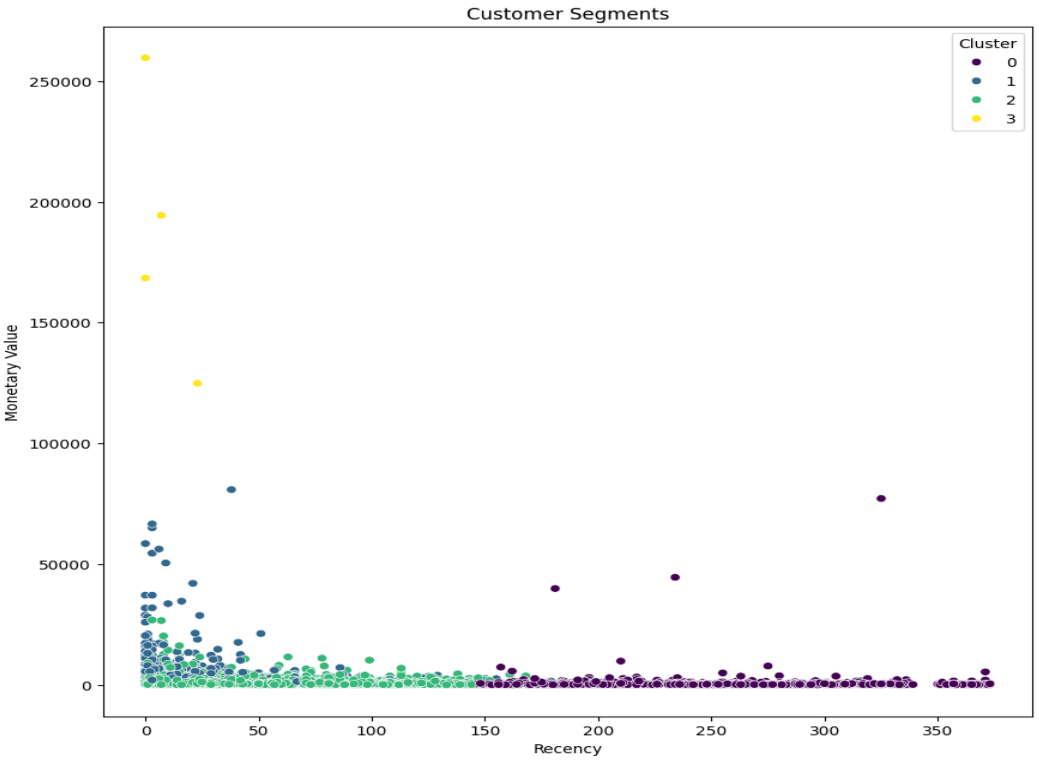
Applying K-means clustering for segmenting the customers

14



The graph shows a significant decline in WCSS from 2 to 4 clusters, so that 4 clusters will be the optimal choice for balancing simplicity.

Scatter Plot of Customer segmentation:



15

Here each color shows type of a customer. There are total of 4 clusters.

Recommendations for each cluster:

- **Cluster 0**: Focus on re-engaging inactive customers by offering tailored promotions and collecting their feedback to identify areas for improvement.
- **Cluster 1**:Appreciate and reward loyal customers with exclusive benefits and premium recommendations to maintain their loyalty.
- **Cluster 2**: Motivate consistent customers to spend more through attractive deals and reminders, encouraging them to increase their purchase frequency.
- **Cluster 3**: Provide top spenders with VIP treatment and personalized interactions to reinforce their loyalty and ensure continued high-value engagement.

Summary of Clusters:

Cluster	monetary	Frequency	Recency
0	624.671371	26.053283	249.014272
1	6910.016232	329.985782	18.189573
2	1222.569034	59.847105	45.016949
3	186858.780000	371.000000	7.500000

5.Predicting Sales using RFM data frame

5.1 Feature Selection

For feature selection and target, the features X include the "Recency," "Frequency," and "Cluster" columns from the rfm Data Frame, while the target variable y is the "monetary" column.

16

5.2 Train Test Split

Using the entire dataset to train our model might lead to data leakage and thus affect the performance of the trained model on unseen data. To address this issue, we split our dataset into train and test datasets wherein we train our model on the train dataset, and test its performance on the test dataset.

5.3 Random Forest Regressor

To train a Random Forest Regressor, the model is instantiated with 150 trees (estimators) and a random state. The model is then fitted to the training dataset, which includes the selected features and target variable. This process prepares the model to make accurate predictions based on the input data.

Here the input will be Recency, Frequency, Cluster and the output i.e., target variable is Monetary.

5.4 Performance Metrics

Regression performance metrics are essential for evaluating the accuracy and reliability of predictive models. These metrics provide insights into how well a model captures the relationship between input features and the target variable. Selecting appropriate metrics depends on the problem context, such as sensitivity to outliers or the need for interpretability.

Key metrics are:

1. Mean Absolute Error
2. Mean Squared Error
3. Root Mean Squared Error
4. R-Squared

- **Mean Absolute Error (MAE):**

  MAE calculates the average absolute difference between predicted and actual values. It provides an intuitive measure of average error in the same units as the target variable.

- **Mean Squared Error (MSE):**

  MSE computes the average squared differences between predictions and actual values, penalizing larger errors more heavily, which makes it sensitive to outliers.

- **Root Mean Squared Error (RMSE):**

  RMSE is the square root of MSE, offering an error measure in the same units as the target variable. It emphasizes large errors and is widely used for model comparison.

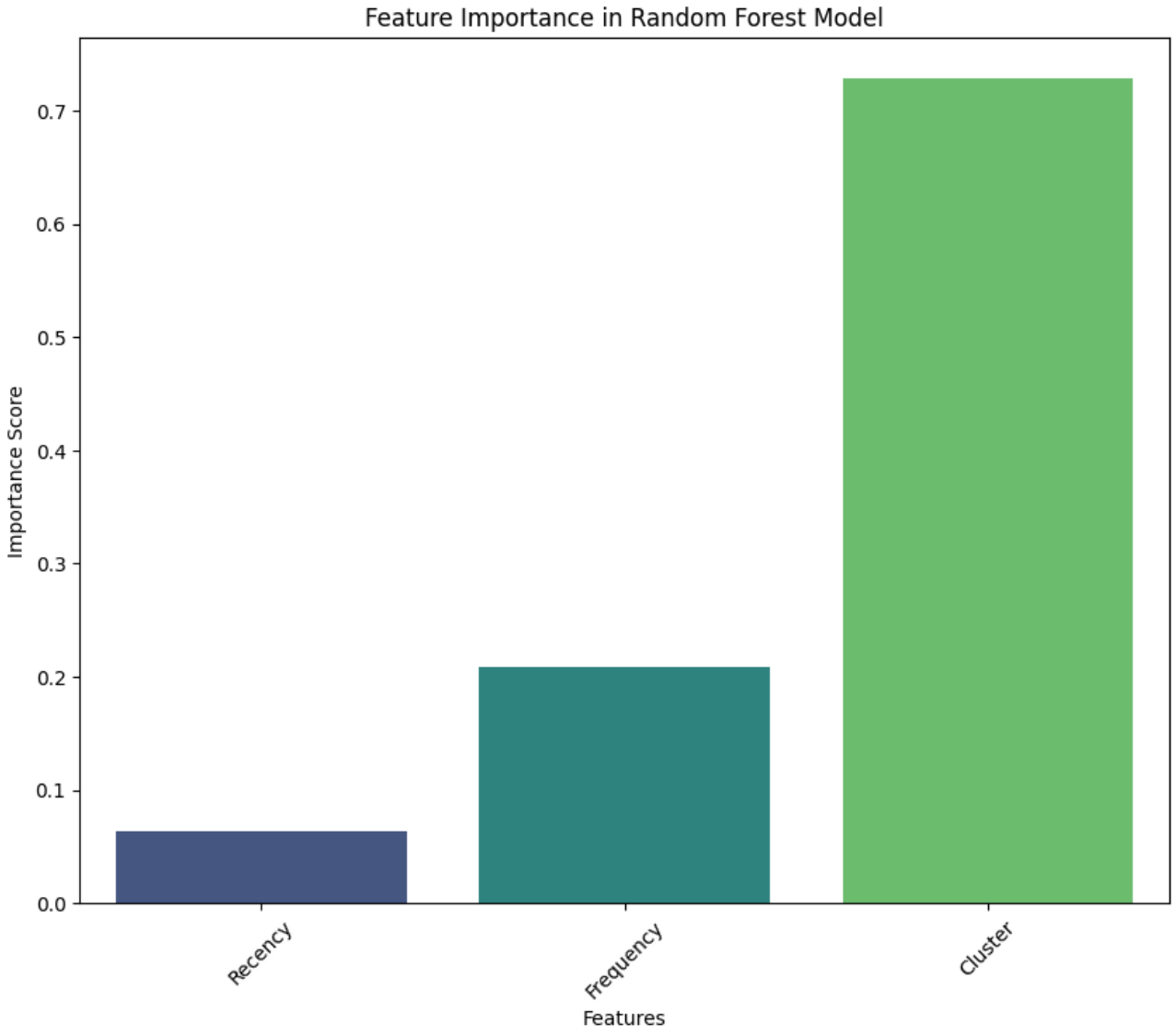
- **R-squared (R^2):**

  R^2 represents the proportion of variance in the target variable explained by the model. A value closer to 1 indicates a better fit, while 0 indicates no explanatory power.

17

For our model we got R2 score as 0.5.

5.5 Feature importance



Cluster: With an importance score just over 0.6, the "Cluster" feature plays a pivotal role in shaping the model's predictions.

Frequency: Carrying a score of around 0.35, the "Frequency" feature also adds substantial value to the model's performance.

Recency: Scoring just above 0.15, the "Recency" feature has a more modest influence on the predictions compared to the other features.

5.6 Support Vector Regression

18

The dataset is divided into training and testing sets, and feature scaling is applied using StandardScaler. A Support Vector Regressor (SVR) with specific hyperparameters is trained on the scaled training data. After training, the model's performance is evaluated on the test data using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. This process highlights the steps taken to build and assess the regression model's accuracy and fit.

6. Conclusion

In our analysis, we employed both Random Forest and Support Vector Regression (SVR) models to predict the target variable. After evaluating their performance, we found that the Random Forest model provided better results compared to the SVR model. This indicates that the Random Forest algorithm is more effective for this dataset, offering more accurate and reliable predictions.

19