MTH 765P Mini-project Report

Prapthi Harish
230028139

1 Introduction

In the realm of retail, understanding and dissecting sales data is crucial for strategic decision-making and future planning. This data analysis endeavors to unravel insights from a comprehensive sales data-set, shedding light on various aspects such as top-selling products, geographic sales distribution, customer behavior, and more.

The data-set at the heart of this analysis encapsulates a wealth of information, ranging from Invoice numbers,Stock Codes, Product Description etc.,This analysis starts with a thorough cleaning procedure to guarantee data integrity and then sets out to extract significant patterns, trends.

Each visual representation and statistical insight is a stepping stone toward a comprehensive understanding of the sales landscape, laying the groundwork for strategic decision-making and future business optimization

2 Obtaining and Pre-Processing the Data

2.1 Obtaining Data

The data-set used in this analysis is from the UCI Machine Learning Repository - Online Retail Data Set.In order to increase the volume of the data and, ultimately, aid in effective visualisation of data, I have combined two data-sets: online data-sets 1 and 2.Encompasses transactions recorded for registered, non-store online retail business.The data spans the period from December 1, 2009, to December 9, 2011. The company specializes in selling unique all-occasion gift-ware, and a substantial portion of its customer base consists of wholesalers.

Each transaction is assigned a 6-digit integral number in the **InvoiceNo** column; codes beginning with "c" denote cancellations. A five-digit integral number is used in the **StockCode** column to uniquely identify each product, and the corresponding product name is provided in the **Description** column. The quantity of every product used in a transaction is indicated in the **Quantity** column. The time and date of every transaction are shown in the **InvoiceDate** column. Product unit prices are entered in sterling in the **UnitPrice** column.

The **CustomerID** column, which has a 5-digit integral number assigned to it, makes customer identification easier. Finally, the **Country** column provides important demographic data by identifying the nation in which each customer resides. When combined, these columns provide a thorough picture of the data-set, allowing for in-depth examination of the retail transactions.

2.2 Pre-Processing Data

1.**Handling Missing Values:** Missing values in a data-set can significantly impact the results of any analysis.In our case, rather than imputing missing values, a decision was made to remove rows with any missing values.This approach was taken after careful consideration of the extent of missing data and its potential impact on the analysis.

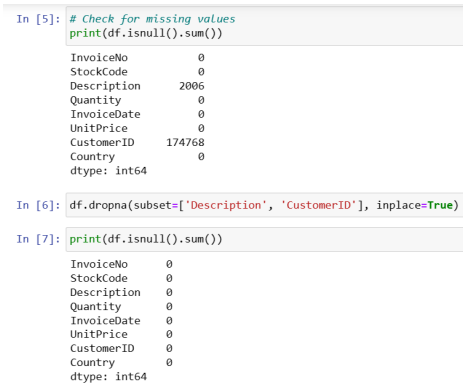

Figure 1: Python Code for Removing the Missing Values

2.**Identifying and Removing Duplicate Rows:** Duplicate entries can distort analytical outcomes, leading to biased results. To maintain the integrity of the data, duplicate rows were identified and subsequently removed. The process involved a meticulous examination of the to distinguish between intentional and erroneous duplicate entries.

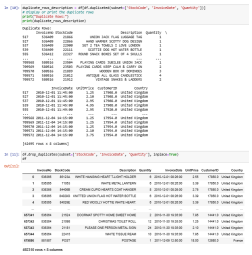

Figure 2: Python Code for Removing the Duplicate Rows

3.**Managing Irrelevant Stock Codes:** Stock code anomalies were examined, particularly those involving entries that were represented as strings. For additional examination, a thorough list of these stock codes was created. To ensure the accuracy and consistency of the analysis related to the product, decision was made to remove rows that were associated with the anomalies based on their nature.

```
In [13]: df = df[~(df['StockCode'].notna() & df['StockCode'].str.isalpha())]

         # Reset the index after dropping rows
         df.reset_index(drop=True, inplace=True)

         # Display the DataFrame after dropping rows
         print("DataFrame after dropping rows with 'StockCode' as a string:")
         df.shape[0]

         DataFrame after dropping rows with 'StockCode' as a string:
Out[13]: 490291
```

Figure 3: Python Code for removing Incorrect Stock Codes

4.**Addressing Incorrect Unit Prices**: The mode of unit prices for each stock code was calculated, ensuring a representative value. Through an iterative process, the data-set was then updated, replacing incorrect unit prices with their corresponding modes.The emphasis on utilizing mode as a measure of central tendency underscored its effectiveness in capturing prevailing unit prices, fostering data integrity.

```
In [15]: #Incorrect Prices
         StockList = df.StockCode.unique()
         CalculatedMode = map(lambda x: df.UnitPrice[df.StockCode == x].mode()[0],StockList)
         StockModes = list(CalculatedMode)
         for i,v in enumerate(StockList):
             df.loc[df['StockCode']== v, 'UnitPrice'] = StockModes[i]
```

Figure 4: Python Code for handling Incorrect Unit Prices

3 Analysis

This analysis dives deep into an online retail data-set that has been meticulously cleaned in an effort to uncover significant insights using a carefully chosen set of questions. In addition to the standard prepossessing stages that require careful data cleaning, this project seeks to go beyond traditional analyses by addressing important questions with a variety of visualisations.

It's important to note that the figures presented here are based on the selected dataset, and the Python program is designed to seamlessly accommodate and analyze alternative datasets.

Top 10 Selling Products

Marketing tactics can be informed by data on best-selling products. To get the most out of these products, businesses might concentrate their marketing, promotion, or advertising campaigns on them. The visualisation assists businesses in determining which specific products have a significant impact on overall sales and revenue by showcasing the top-selling products. Gaining insight into consumer preferences can be achieved by identifying the top-selling products. By matching product offers with customer interests, this information can be used to improve customer satisfaction.

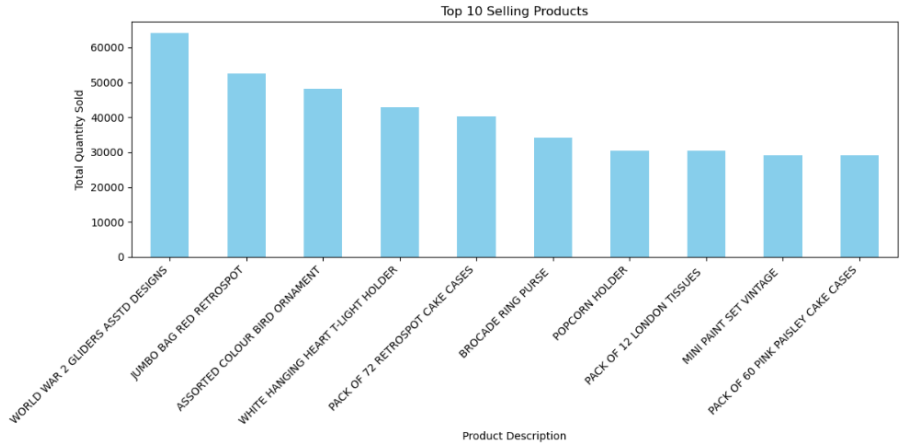

Figure 5: Top 10 Selling Products

- The bars represent individual products, and their lengths are proportional to the total quantity sold of each product.
- The product with the highest quantity sold is "WORLD WAR 2 GLIDERS ASSORTED DESIGNS," with just over 50,000 units sold.
- The quantities sold for the last three items (PACK OF 12 LONDON TISSUES, MINI PAINT SET VINTAGE, PACK OF 60 PINK PAISLEY CAKE CASES) are similar, each approaching the 30,000 units mark.
- The bar chart is a straightforward way to compare the relative sales volumes of these products at a glance, showing a clear difference between the best-selling item and the tenth best-selling item.

Top 10 Countries By Sale

Geographic trends in consumer behaviour can be found by looking at the top nations. Future business strategies can be informed by knowing which regions are more receptive to the offered products or services.It assists in determining

which nations make up the largest portion of total sales. Acknowledging these key players is essential for making strategic choices, concentrating marketing efforts, and allocating resources as efficiently as possible.

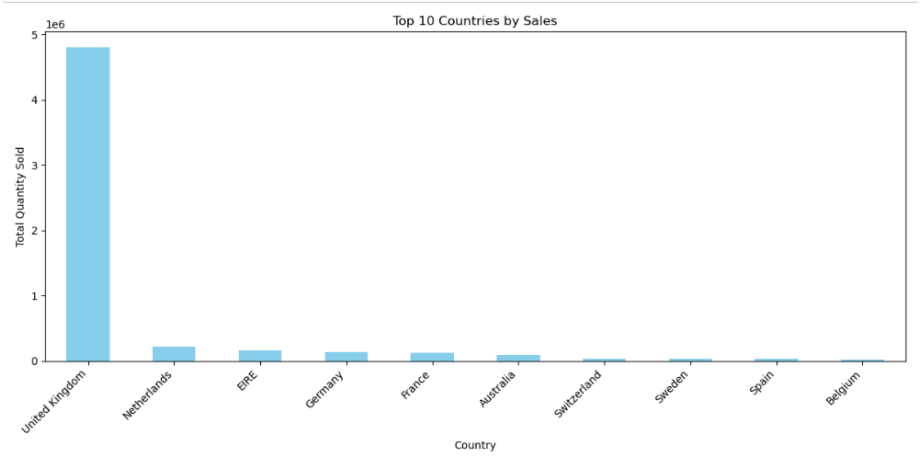

Figure 6: Top 10 Countries By Sale

- The bars represent the total quantity sold in each country, and their lengths are proportional to the reported sales figures.
- The United Kingdom (UK) has the longest bar by a significant margin, indicating that it has the highest sales among the countries listed, with sales reaching 5 million units.
- The sales in the Netherlands, EIRE, Germany, France, Australia, Switzerland, Sweden, Spain, and Belgium are much smaller in comparison to the UK.
- The chart demonstrates a significant disparity between the market size of the UK and the other countries, suggesting the UK is the primary market for the products.

Top 20 Products by Quantity Sold and Monetary Benefit

The analysis assists in determining which products are selling in the largest quantities. Understanding consumer preferences and maintaining a sufficient supply of popular items are made possible with the help of this information.Businesses can concentrate on items that make a substantial contribution to revenue by looking at the top products in terms of financial benefit.

This knowledge is helpful for maximising profitability and refining pricing strategies.Having an understanding of the best-selling products facilitates efficient inventory management. By making sure that popular items are properly

stocked to meet demand, businesses can avoid stock outs and possible revenue loss.Accurate revenue forecasting is aided by knowing how each product contributes to overall revenue. Companies are able to set reasonable objectives and make well-informed financial projections.

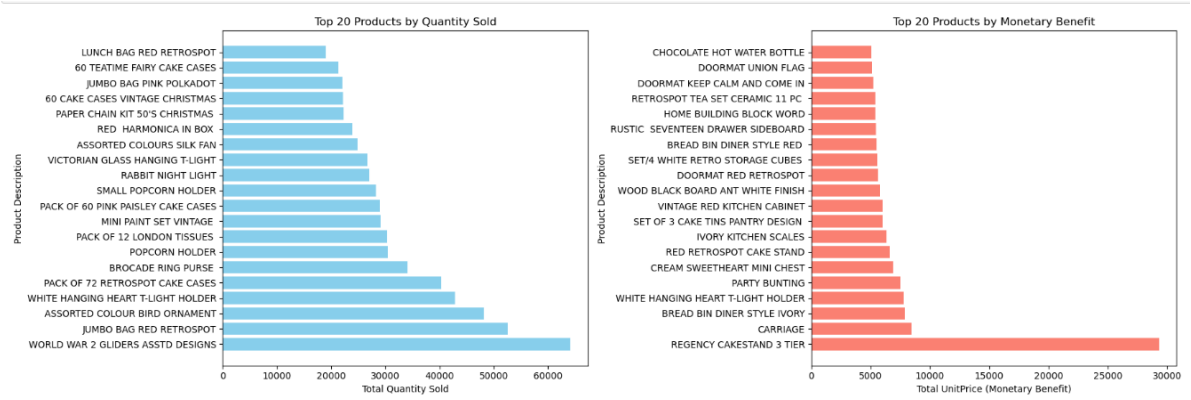

Figure 7: Top 20 Products by Quantity Sold and Monetary Benefit

- The product 'WORLD WAR 2 GLIDERS ASSORTED DESIGNS' has the highest quantity sold, approaching 60,000 units. Other notable products include 'JUMBO BAG RED RETROSPOT' and 'WHITE HANGING HEART T-LIGHT HOLDER', both of which have sold over 40,000 units. The quantities gradually decrease as we move down the chart.
- The 'REGENCY CAKESTAND 3 TIER' stands out as the product with the highest monetary benefit, totaling close to 30,000 in unit price. Other items with significant monetary benefit include 'PARTY BUNTING' and 'WHITE HANGING HEART T-LIGHT HOLDER'. Like the first chart, the monetary value decreases as we move down the list.
- Notably, some products appear on both charts, such as the 'WHITE HANGING HEART T-LIGHT HOLDER', indicating that they are not only popular in terms of quantity sold but also generate significant revenue. However, most products differ between the two charts, suggesting that some items sell in high volumes but may not necessarily provide the highest monetary benefit and vice versa.

Most Returned Items and Customers with Most Returns

The first subplot's bar chart shows the top ten items that customers have returned the most. Quality assurance and inventory management can benefit from this information. It assists the company in determining which products

might be more likely to be returned, giving them the opportunity to handle possible problems like product flaws, inconsistent product descriptions, or unhappy customers.

The second subplot's bar chart displays the top 10 customers with the greatest number of returned items, along with the corresponding countries for each customer.Consumers who send things back frequently might not be happy with what they bought. There could be a number of reasons for this dissatisfaction, including differences between the product description and what was received, improper product shipments, or poor product quality.Returns may also be linked to shipping-related issues, like items that break in transit or arrive late. Consumers who have problems with the shipping procedure might be more inclined to send back their purchases.

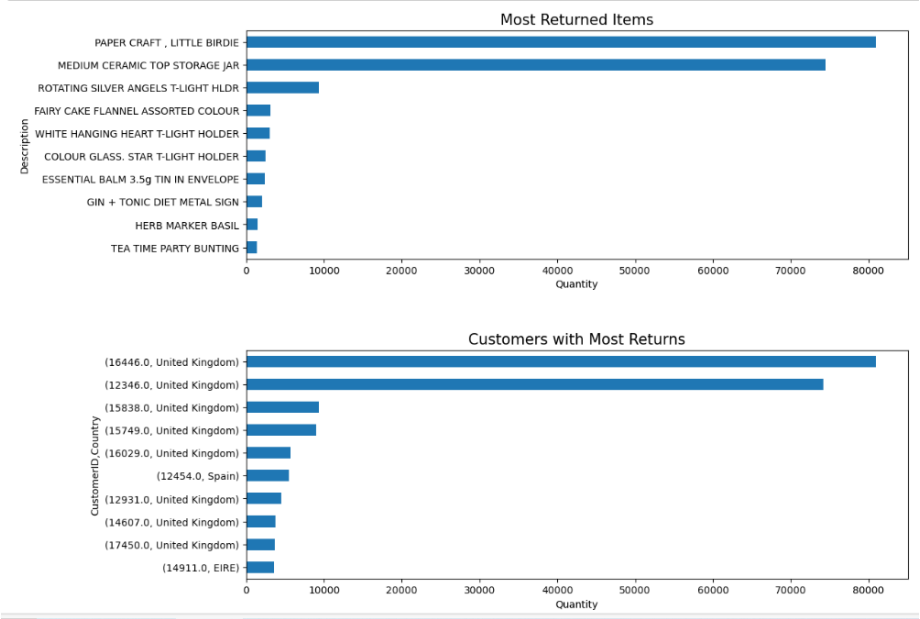

Figure 8: Most Returned Products and Customers with Most Returns

- The product with the most returns is "PAPER CRAFT , LITTLE BIRDIE", which has been returned over 80,000 times. The quantities of returns decrease as we move down the chart, with other items such as "MEDIUM CERAMIC TOP STORAGE JAR" and "ROTATING SILVER ANGELS T-LIGHT HLDR" having significantly fewer returns but still quite substantial in number.
- The customer with the most returns, denoted by the ID (16446.0, United Kingdom), has returned items approximately 80,000 times. The scale of

returns decreases as we move down the list, with the lowest displayed customer (14911.0, EIRE) still having a significant number of returns. Notably, most of the customers are from the United Kingdom, with one from Spain and one from EIRE (Ireland), suggesting that the data may be predominantly for a UK-based company or market.

Weekly Sales and Returns Quantity

Trends and patterns in sales and returns over time can be found with the aid of the visualisations. This may indicate whether certain days, weeks, or months have steadily higher or lower sales than others, offering information about seasonality or outside variables influencing consumer behaviour.Recognising any reoccurring patterns is made easier by analysing weekly returns.It enables companies to determine whether there are particular days or weeks when returns occur more frequently. To address possible problems with product quality, customer satisfaction, or fulfilment procedures, it is essential to understand return patterns.

Visualisations of weekly sales and returns help to improve comprehension of customer behaviour. Companies can determine when customers shop the most and then customise incentives or promotions to increase sales on those days. Furthermore, knowing return trends aids in enhancing customer service, quality assurance, and product descriptions.

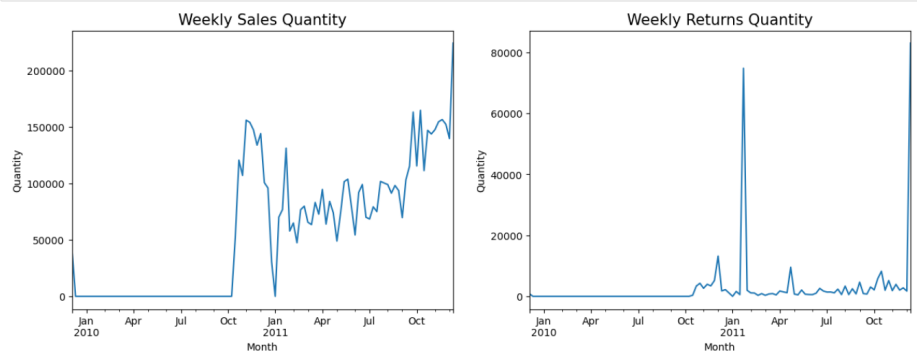

Figure 9: Weekly Sales and Returns Quantity

- The sales quantity appears to start at a low point in January 2010 and experiences a significant spike early in the graph's timeline.Sales then show some volatility but generally trend upwards over time, with several noticeable peaks and troughs.The highest peaks in sales quantity occur around April 2010, late in 2010, and the highest peak at the end of the displayed data, around October 2011, indicating potentially the highest weekly sales volume in the entire period.

- The "Weekly Returns Quantity" graph is much less volatile than the sales quantity graph, with many weeks showing very low return quantities.There is one extremely prominent spike in returns that towers over the rest of the data, which occurs around January 2011.Aside from this spike, there are a few smaller peaks in returns, but they are relatively minor compared to the major spike.

- The spike in returns in January 2011 could be related to post-holiday returns, as it is common for returns to increase after the holiday season due to unwanted gifts or exchanges.

- The correlation between sales and returns is not directly clear from these graphs, but the significant spike in returns does not seem to have a corresponding drop in sales, which might have been expected if the returns were due to a singular event affecting product quality or customer satisfaction.

Repeat Customers

A quantitative indicator of customer loyalty is the percentage of returning customers. Businesses can quickly assess the success of their customer retention initiatives by visualising this percentage. A higher percentage denotes a base of devoted customers, while a lower percentage might lead companies to look into ways to increase customer loyalty.Knowing how top repeat customers are distributed across national borders enables companies to pinpoint the areas in which their marketing campaigns or loyalty programs have had the most success.

Allocating resources, creating customised promotions, and modifying tactics for particular regional markets all benefit from this information.Visualisations' insights can be used as a feedback loop for ongoing development. Businesses can improve their goods, services, and customer engagement tactics to encourage enduring loyalty by studying the traits of repeat customers and how they are distributed.

- The chart indicates that 71.3 percentage of the customers are repeat customers, which is a significant majority.Non-repeat customers make up 28.7 percentage of the total customer base.The significant proportion of repeat customers could indicate customer satisfaction and loyalty.

- The United Kingdom has the highest number of repeat customers, as indicated by the tallest bar, significantly surpassing the other two countries.EIRE (Ireland) is shown next with a much smaller number of repeat customers.The Netherlands has the fewest repeat customers among the three, as represented by the shortest bar.

- A large percentage of repeat customers could be indicative of a successful business strategy in terms of product offerings, customer service, or other factors that contribute to customer retention.The concentration of repeat customers in the United Kingdom suggests either a larger market presence

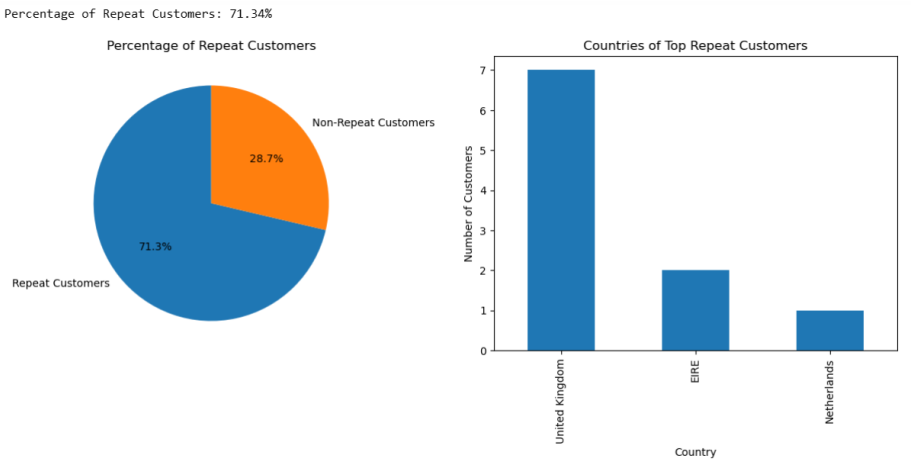

Figure 10: Percentage of Repeat Customers and Countries with Most Repeat Customers

or better market strategies in place there compared to Ireland and the Netherlands.

- The data presented could be used to evaluate market penetration and customer loyalty programs in each country.The substantial difference between the United Kingdom and the other two countries could prompt further investigation into market conditions or business practices in the respective regions.

4 Future Improvements

These could be some suggestions for additional research and dataset extensions in the future to improve comprehension:

- Investigating more sophisticated methods of customer segmentation according to location, purchasing patterns, or other pertinent variables. This can assist in customising marketing plans for particular clientele.
- Building predictive models to predict future customer behaviour, sales, and returns. Proactive decision-making can be facilitated by the application of machine learning algorithms or time series forecasting to predict future trends and obstacles.
- Integrating sentiment analysis into any available customer reviews or feedback. Recognising consumer sentiment can reveal important information about product satisfaction and possible areas for development.
- Incorporating external datasets, such as economic indicators, weather data, or seasonal trends, to assess their impact on sales and returns. This can offer a more thorough comprehension of the outside variables affecting consumer behaviour.
- Implementing more advanced fraud detection algorithms to spot strange trends in returns or transactions. Unusual patterns could point to fraud, and a sophisticated model could improve the accuracy of the detection.