Online Retail Exploratory Data Analysis

By:
Imran Basha S
10th June 2023

Outline

- Executive Summary
- Introduction
- Methodology
  - Dataset
  - Data Cleaning & Preprocessing
  - Analysis Approach
- Exploratory Data Analysis
- Results & Analysis
- Key Findings
- Suggestions & Recommendations
- Conclusion
- Limitations & Future Work
- Appendix

Executive Summary

- We have a large dataset of about 541909 transaction on Online Retail. The main focus of the project is data cleaning & Exploration, and Exploratory Data Analysis.
- After Initial data exploration, data has been cleaned and a lot of non sales entries have been removed, to make the dataset ready for further analysis.
- The sales are from 38 countries having 4362 unique customers, buying 3927 unique products.
- United Kingdom is largest market both in terms of sales value and the volumes of transactions.
- United Kingdom contributes an annual sales revenue of 8.28 Millions approx.
- After United Kingdom – Netherlands, EIRE, Germany & France have a little but leading contribution to sales revenue.
- Czech Republic, Bahrain, and Saudi Arabia contribute the least.

Introduction

In this project, we analyze transactional data from an online retail store to gain insights into sales trends, customer behavior, and popular products. The dataset details customer purchases, including product details, quantities, prices, and timestamps.

This Exploratory data analysis, aims to identify patterns, outliers, and correlations to enable data-driven decision making for stakeholders. We will uncover key trends like the top performing business regions, busiest sales months, best-selling products, and order frequency, using visualizations and statistical analysis.

Ultimately, this project aims to provide actionable insights that can drive strategic business decisions and enhance the store's overall performance in the competitive online retail market.

Methodology

Dataset

The dataset we will be working with is the "Online Retail" dataset. It contains transactional data of an online retail store from 2010 to 2011.

The dataset contains the following columns:

InvoiceNo: Invoice number of the transaction, if the code starts with letter 'c', it indicates a cancellation.

StockCode: Unique code of the product.

Description: Description of the product.

Quantity: Quantity of the product in the transaction.

InvoiceDate: Date and time of the transaction.

UnitPrice: Unit price of the product.

CustomerID: Unique identifier of the customer.

Country: Country where the transaction occurred.

The dataset is available as a .xlsx file named Online Retail.xlsx. This data file can also be downloaded here.

Data Cleaning & Preprocessing

- The initial raw data had 541909 rows & 8 columns of data.
- 5268 **duplicate entries** have been removed.
- 2906 **non sales entries** have been removed.
- 2496 entries having ‘UnitPrice’ equals ‘0’ have been removed.
- There are 1454 **missing Values** in the column ‘Description’ and 135080 values missing in the column ‘CustomerID’
  - Upon removal of non sales entries, the missing values in the column ‘Description’ were taken care of.
  - But, we still have 131580 missing values in ‘CustomerID’
- The data type of the column ‘CustomerID’ has been changed to ‘Int64’ and also the order of the columns have been rearranged for convenience.
- A new data frame sales_df was created, after all the changes, the dataframe now has **531239 rows & 8 columns** of data.

Note: For detailed explanation please refer to the **Jupyter Notebook** file.

Analysis Approach

- As this is a large dataset, the primary objective here in this project is to understand the data.
- Distribution of the data has been explored and evaluated before & after the data cleaning.
- Apart from regular data cleaning, many non sale entries had to be removed.
- Further, basic statistics and Exploratory Data Analysis (EDA) to have a better understanding of data was performed.
- Then, specific questions were explored to have more insights.
- Data Visualization has been utilized to derive inferences.
- The Analysis is based on:
  - Python version – 3.11.7
  - Numpy version – 1.26.2
  - Pandas version – 2.1.4
  - Matplotlib version – 3.8.0

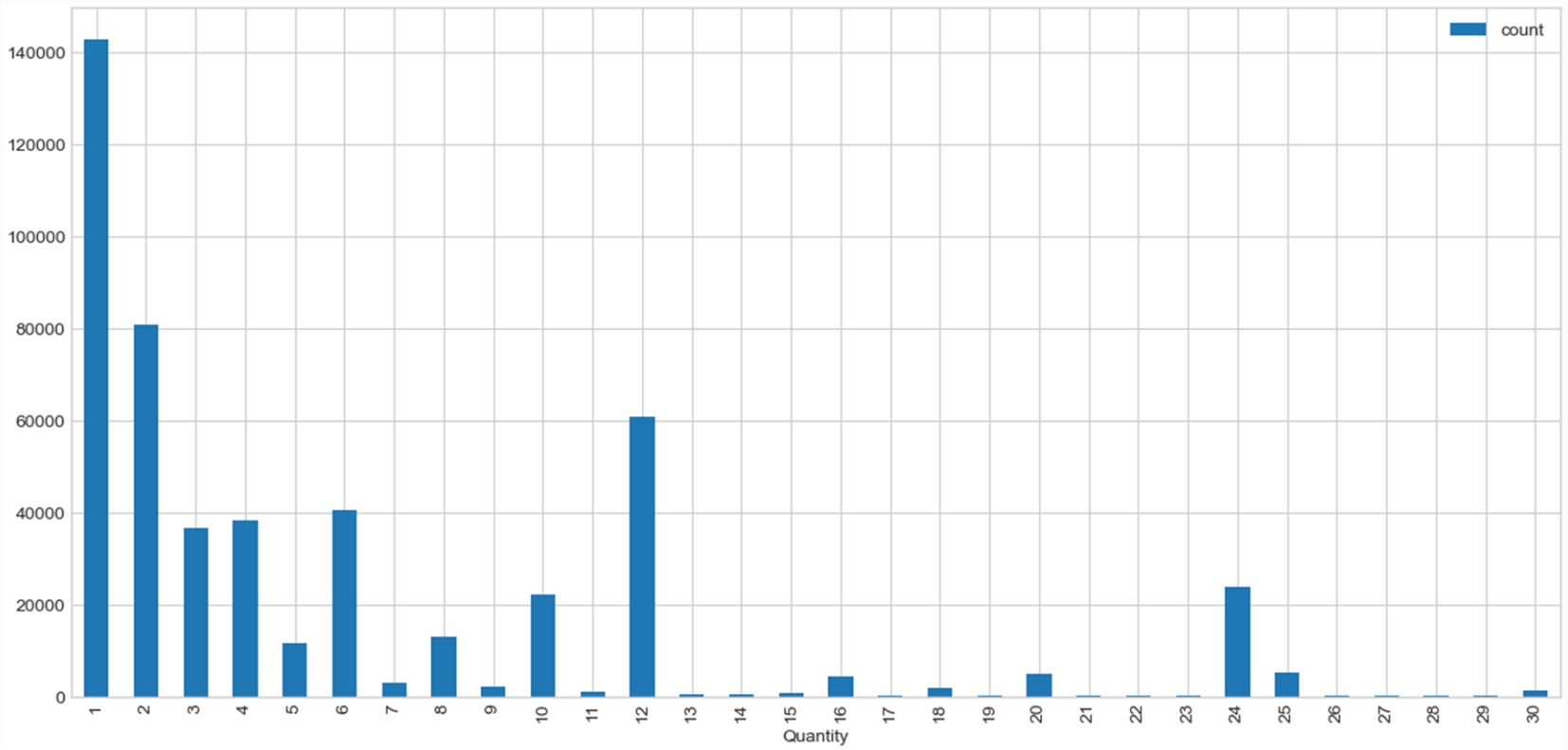
Exploratory Data Analysis

Data Exploration

- The cleaned dataset has **531239 rows & 8 columns** of data.
- Transactions are from 01-12-2010 to 09-12-2011
- There are 4362 unique customers.
- The customers are from 38 countries.
- There are 23195 unique transactions.
- 3927 unique stock codes or unique products.

Note: For detailed explanation please refer to the **Jupyter Notebook** file.

Distribution of Quantity Column

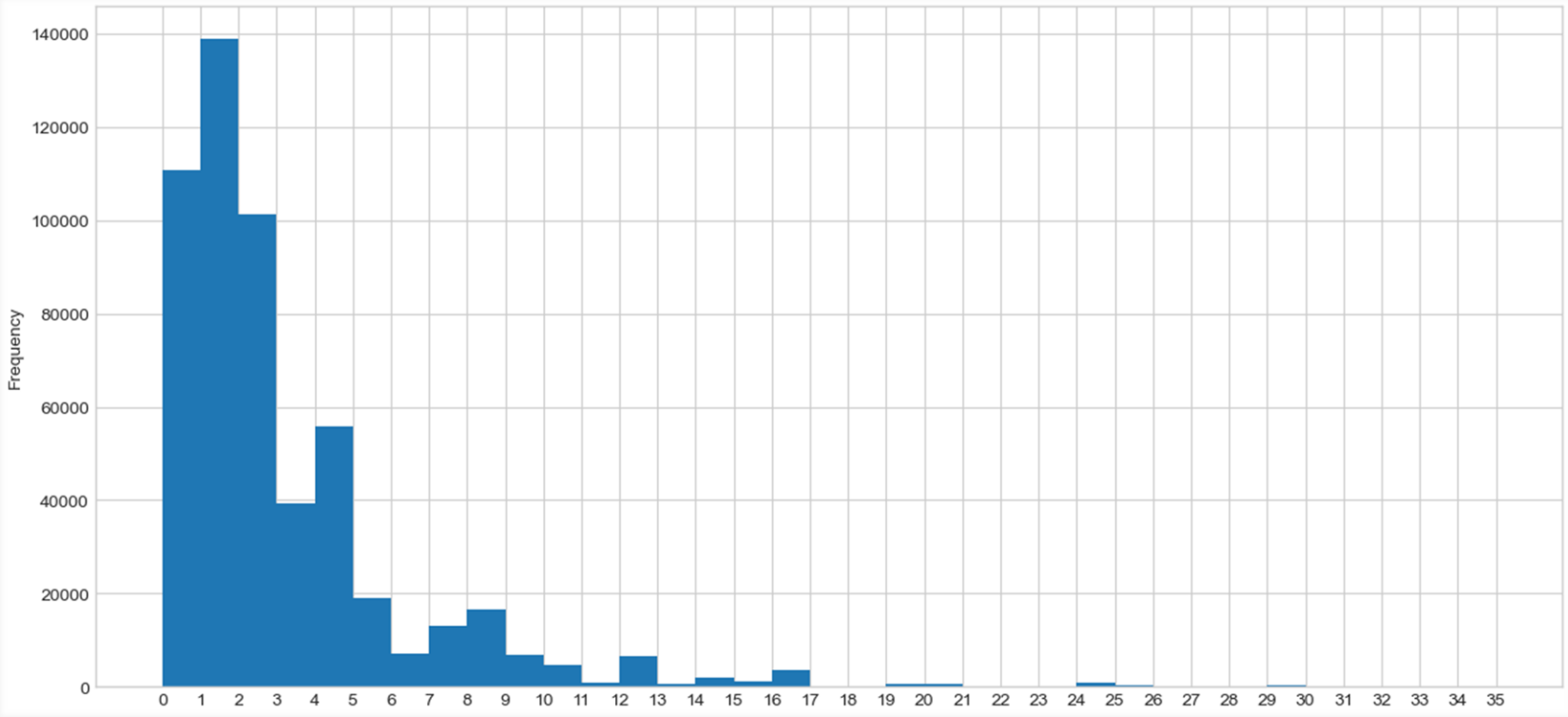


1. It can be observed that most of the sales have a purchase of **single** item.

2. Out of all the transactions, 140000 transactions have an ordered quantity equals “1” (i.e. 25% of the transactions).

3. 95% of data points do not have an ordered quantity more than 30.

Distribution of UnitPrice Column

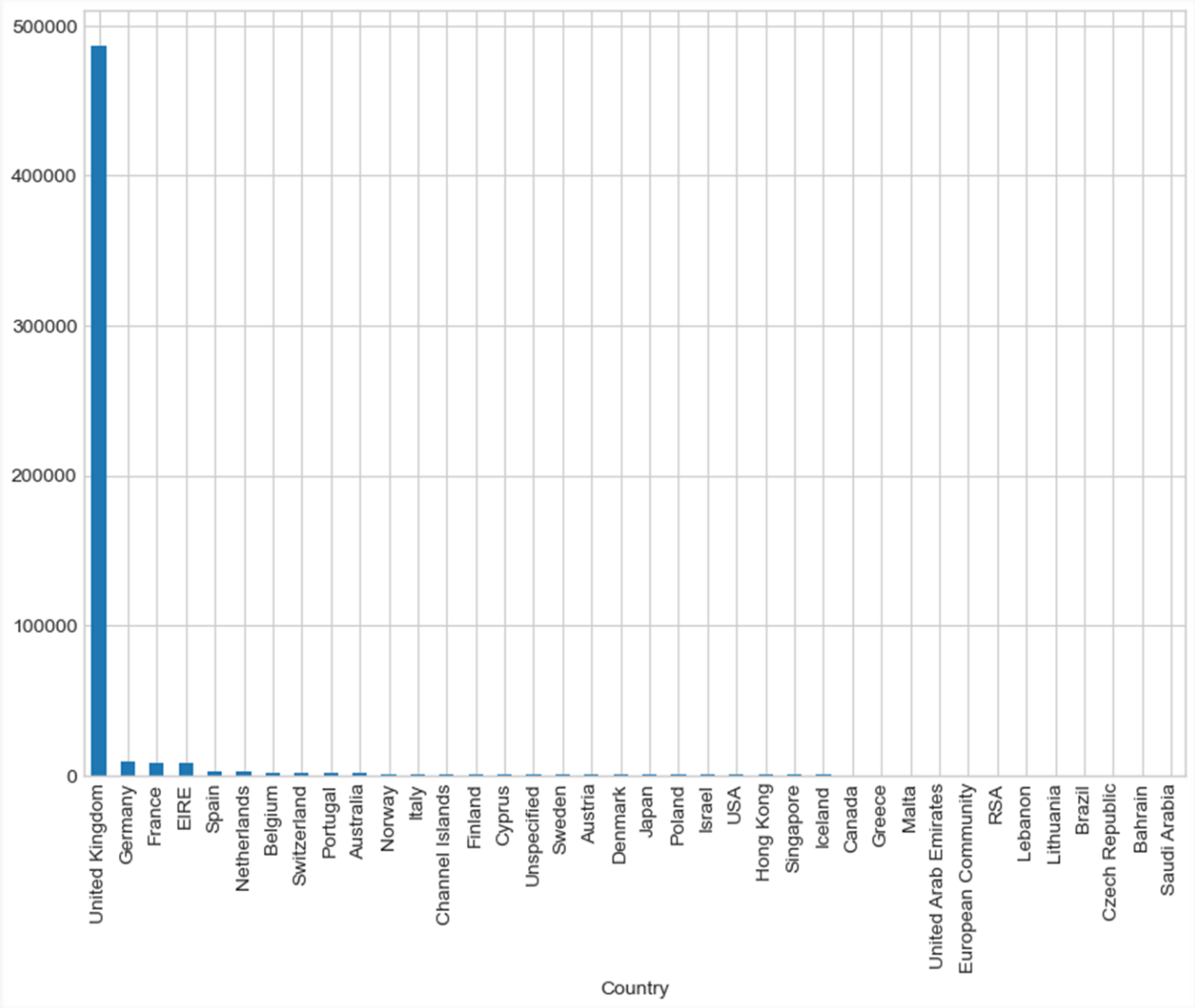


1. We can observe that 99.9% of the transactions have an unit price **less than 35**.

2. 50% of the transactions have an unit price within 2.08

3. And, 90% are within 7.65

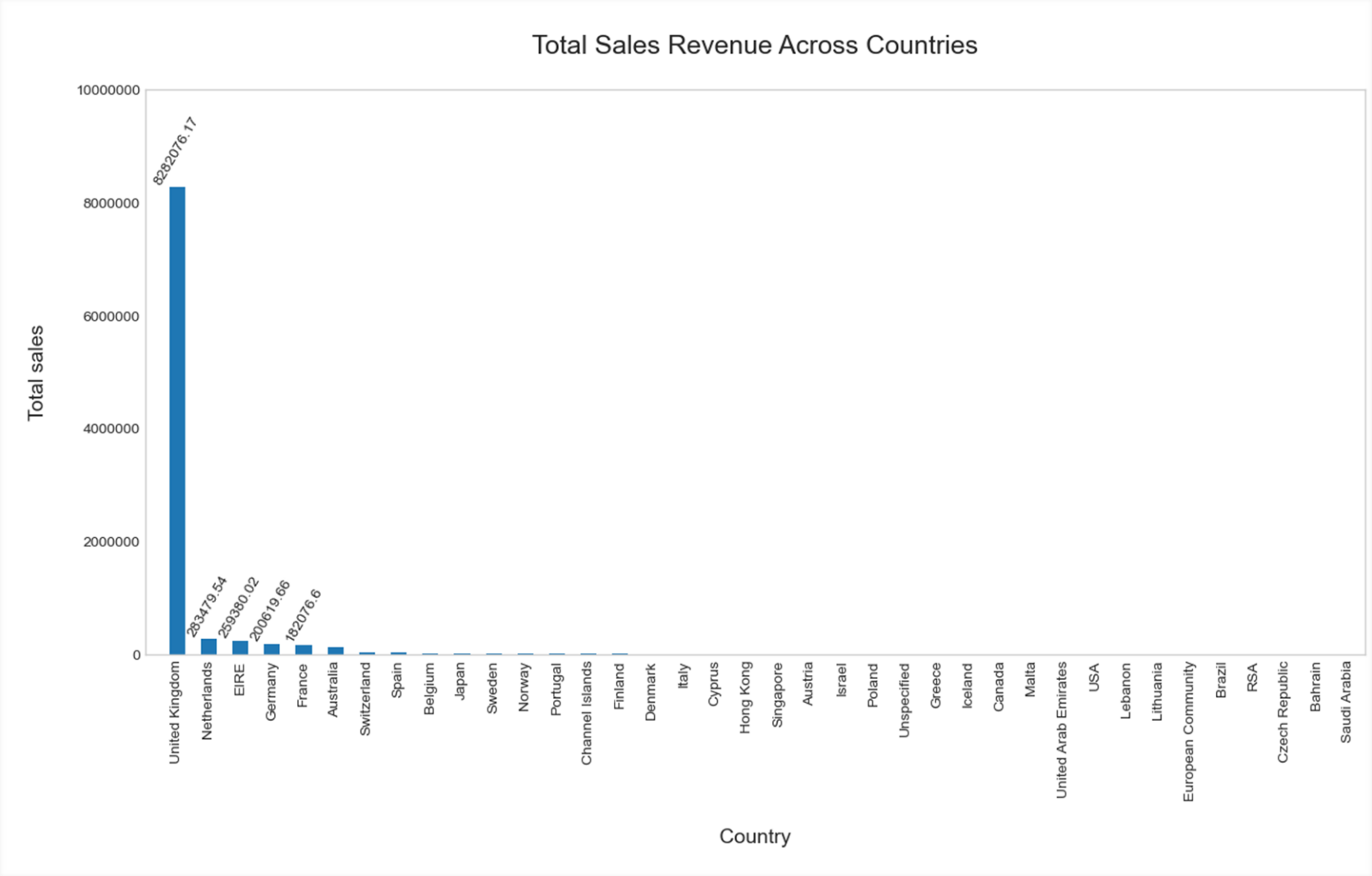
Distribution of Country Column



Majority of the transactions are from **United Kingdom**.

Results & Analysis

1. Total Sales Revenue Across Countries

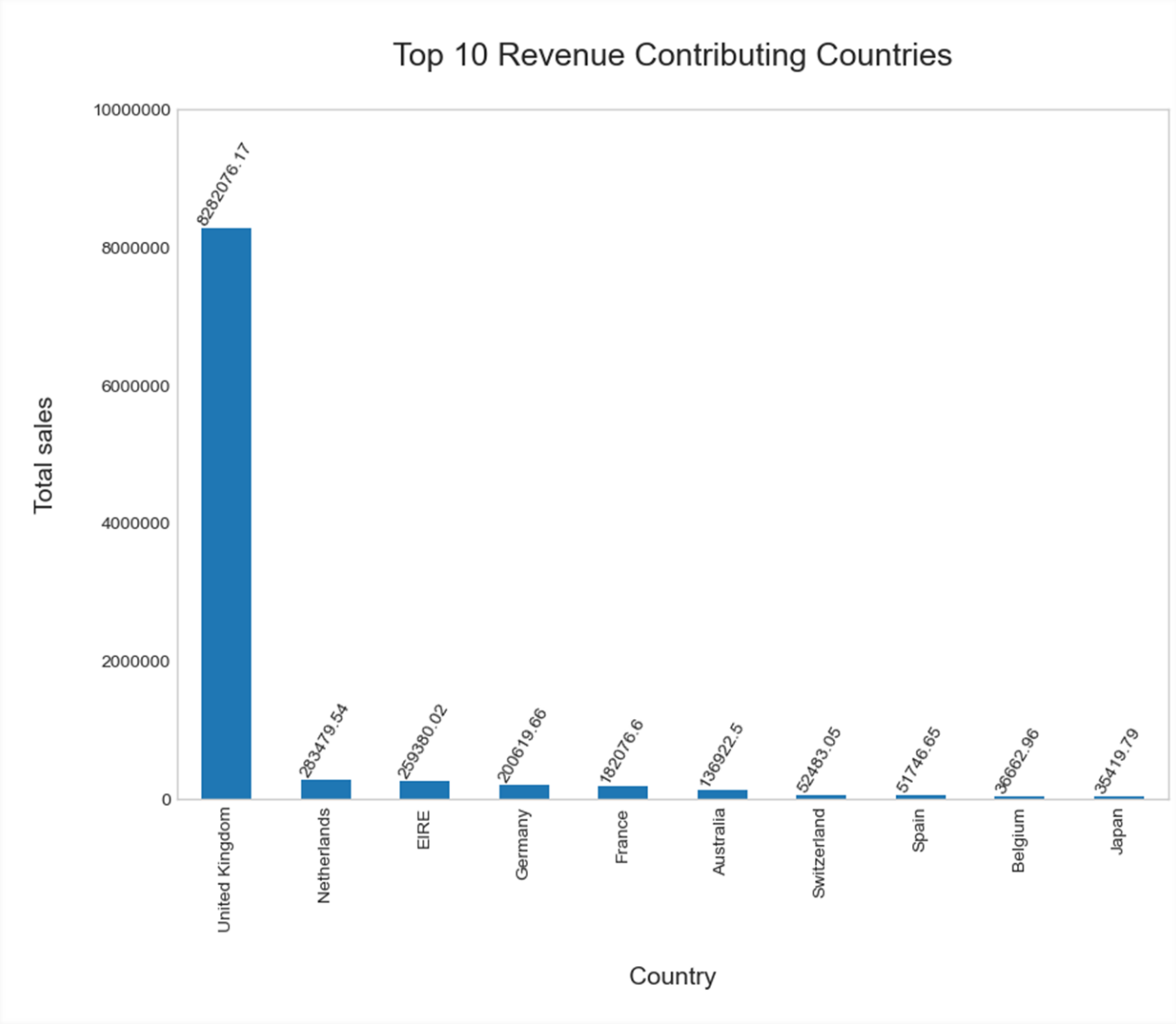


Inference:

1.We can see that **United Kingdom** brings in a major chunk of sales, 8282076 **(8.28 Millions approx)**.

2.It is followed by **Netherlands & EIRE** but the figures are meagre.

2a. Top 10 Revenue Contributing Countries



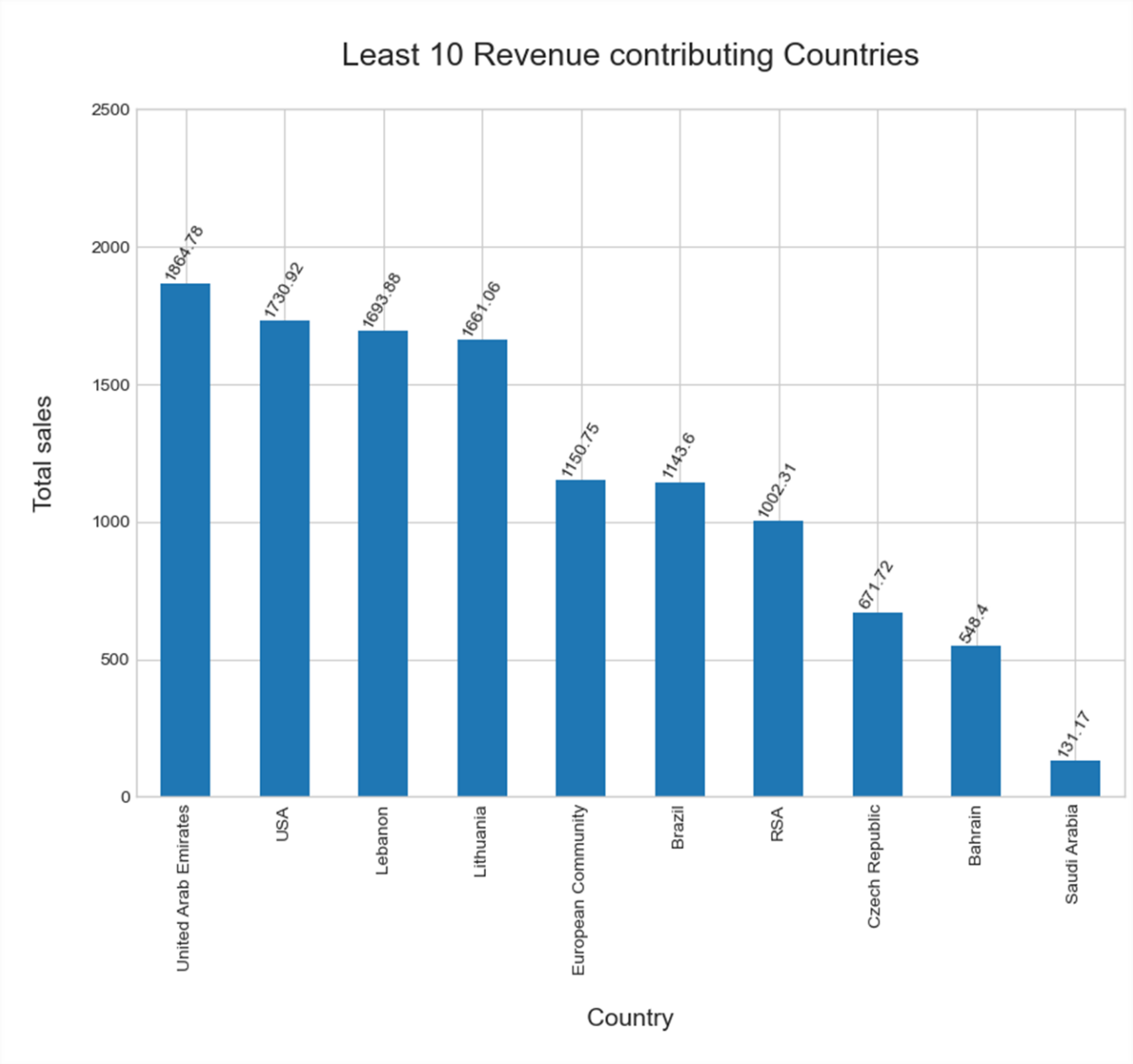
Inference:

1.We can easily say that **United Kingdom** is the biggest, perhaps only significant contributor to sales revenue.

2.The others in the list do not contribute much.

3.Apart from **United Kingdom**, the three other markets that can be considered are **Netherlands, EIRE, & Germany**. However, they are in a totally different frame.

2b. Least 10 Revenue Contributing Countries

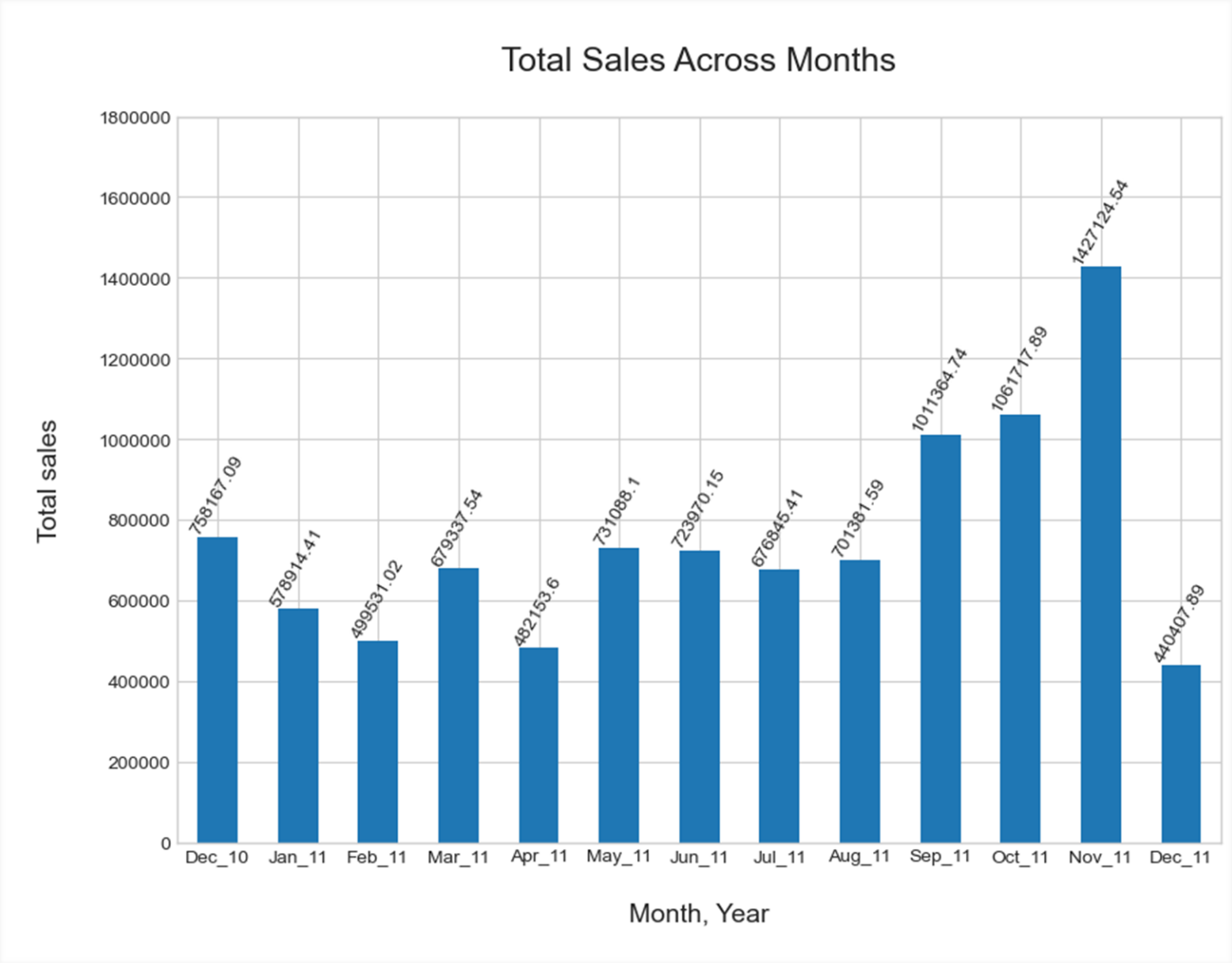


Inference:

1.We can see the figures of least sales contributing countries.

2.They are **Underwhelming**.

3. Total Sales Across Months

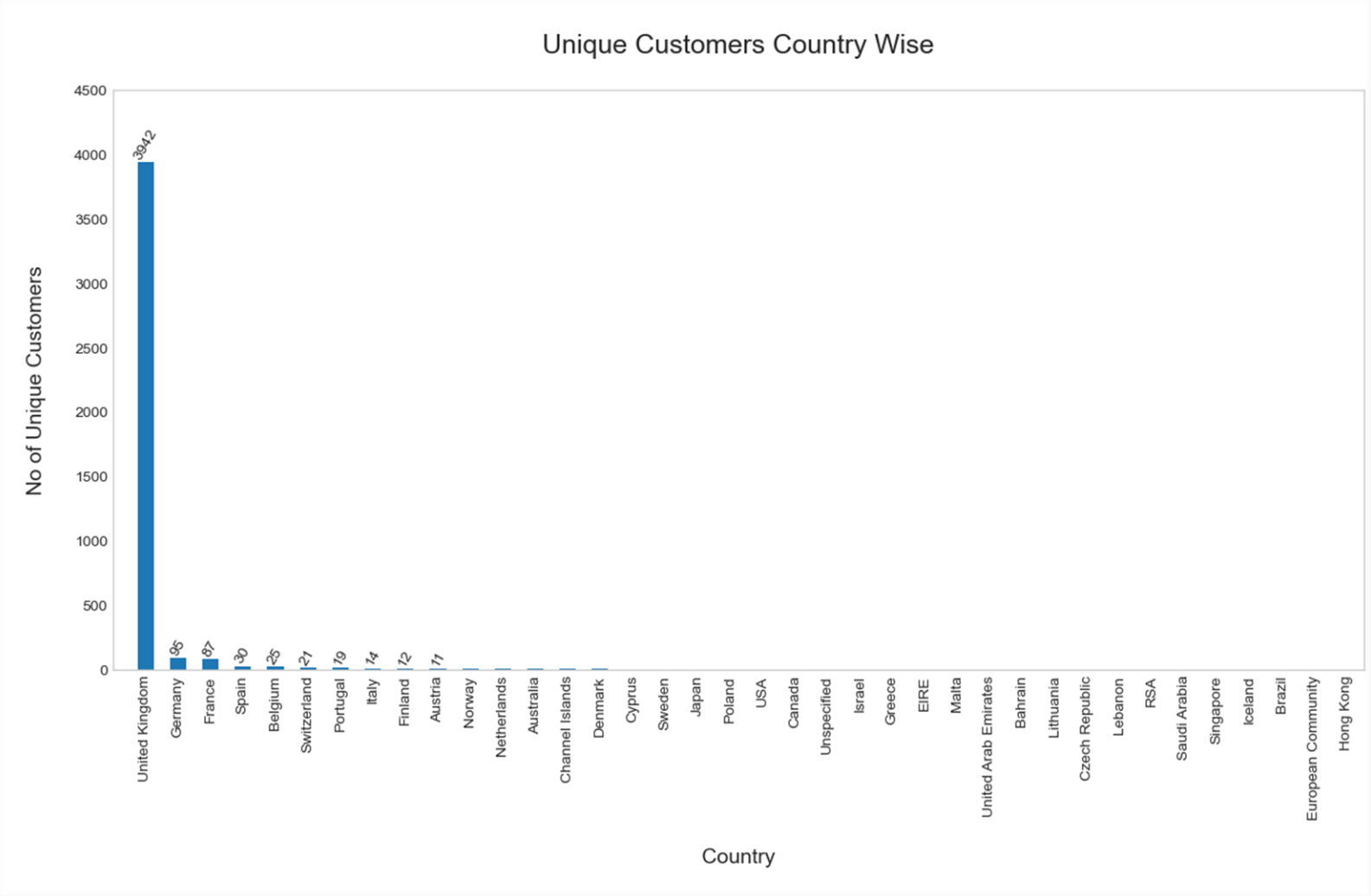


Inference:

1.Knowing the fact that we do not have complete data for the month of Dec_2011, we can say that **minimum sales were achieved in Apr_2011** and **maximum were achieved in Nov_2011**.

2.After the dip in Apr_11, there is a spike in sales observed from May_11. Although, there was a slow down in between, there is a gradual and considerable **increasing trend in sales**.

4. Unique Customers Country Wise



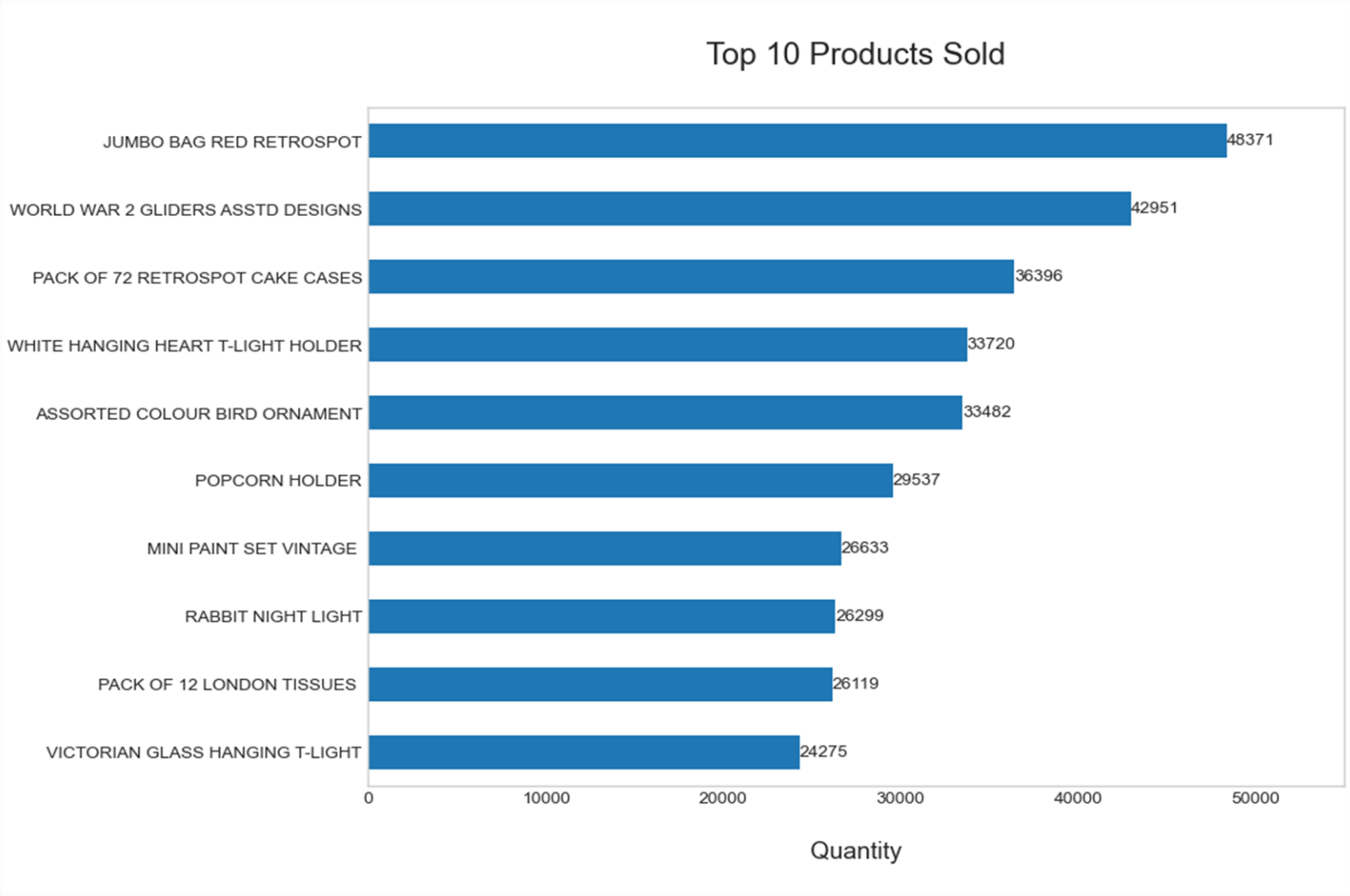
Inference:

1.As expected from earlier patterns, **United Kingdom** has the most number of **Unique Customers - 3942**.

2.In total sales we had observed that **Netherlands, Eire, & Germany** to be in 2nd, 3rd and 4th places. However, in terms of unique customers, **Germany with 95** occupies second place, followed by **France with 87**.

3.It can be inferred that the earlier observed Netherlands & EIRE situation could be due to customers buying in bulk. This can be verified in further analysis to see if these Customers can be offered different propositions, based on the scenario

5. Top 10 Products Sold



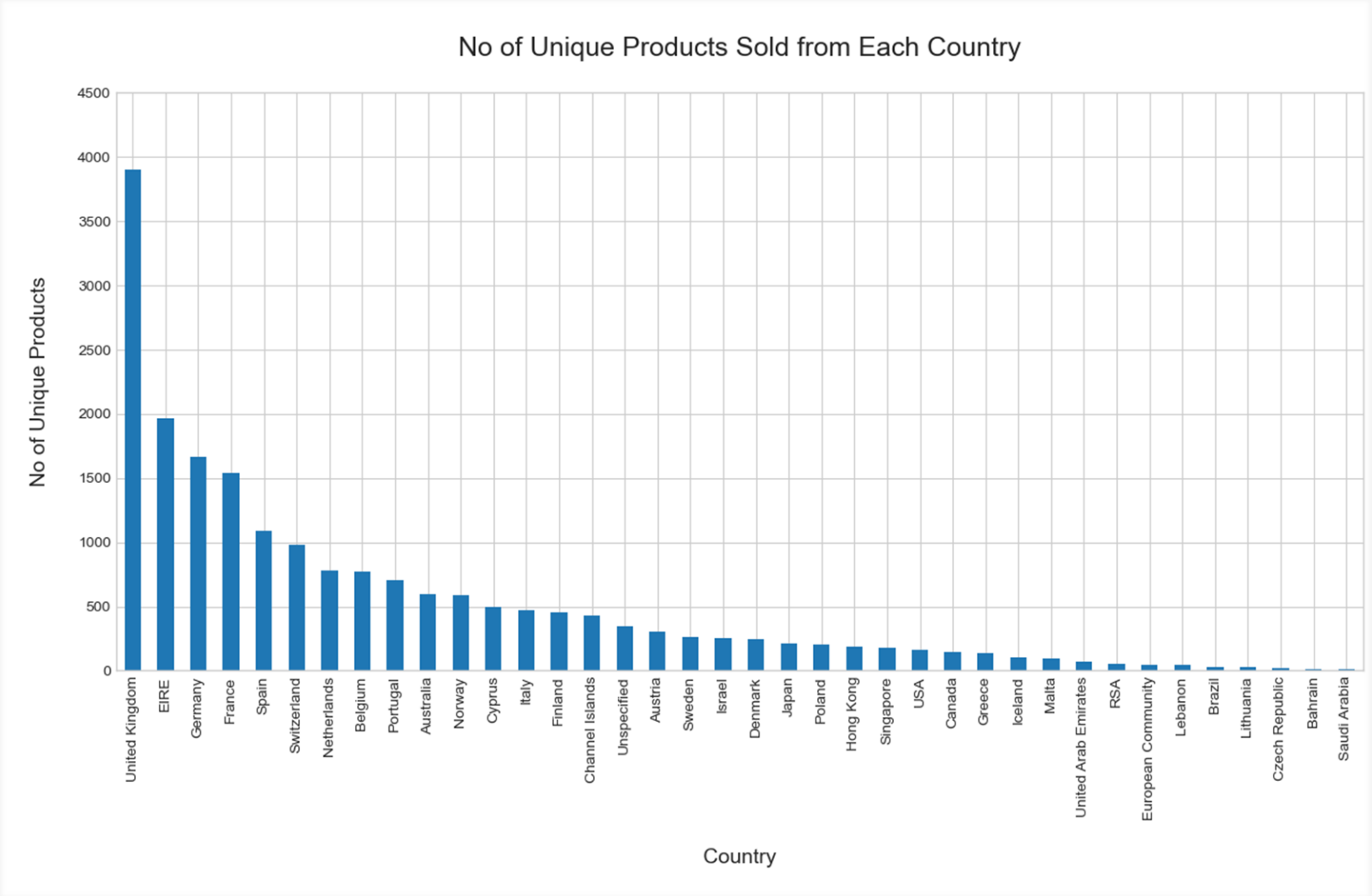
Inference:

1.It can be observed that the **Jumbo Bag Red Retrospot** is the most sold with numbers 48371.

2.The list also conveys the 9 other products.
Steps can be taken to ensure constant **supply and storage** of these products.

3.It can be evaluated if these products can also be promoted in other markets appropriately.

6. No of Unique Products Sold from Each Country



Inference:

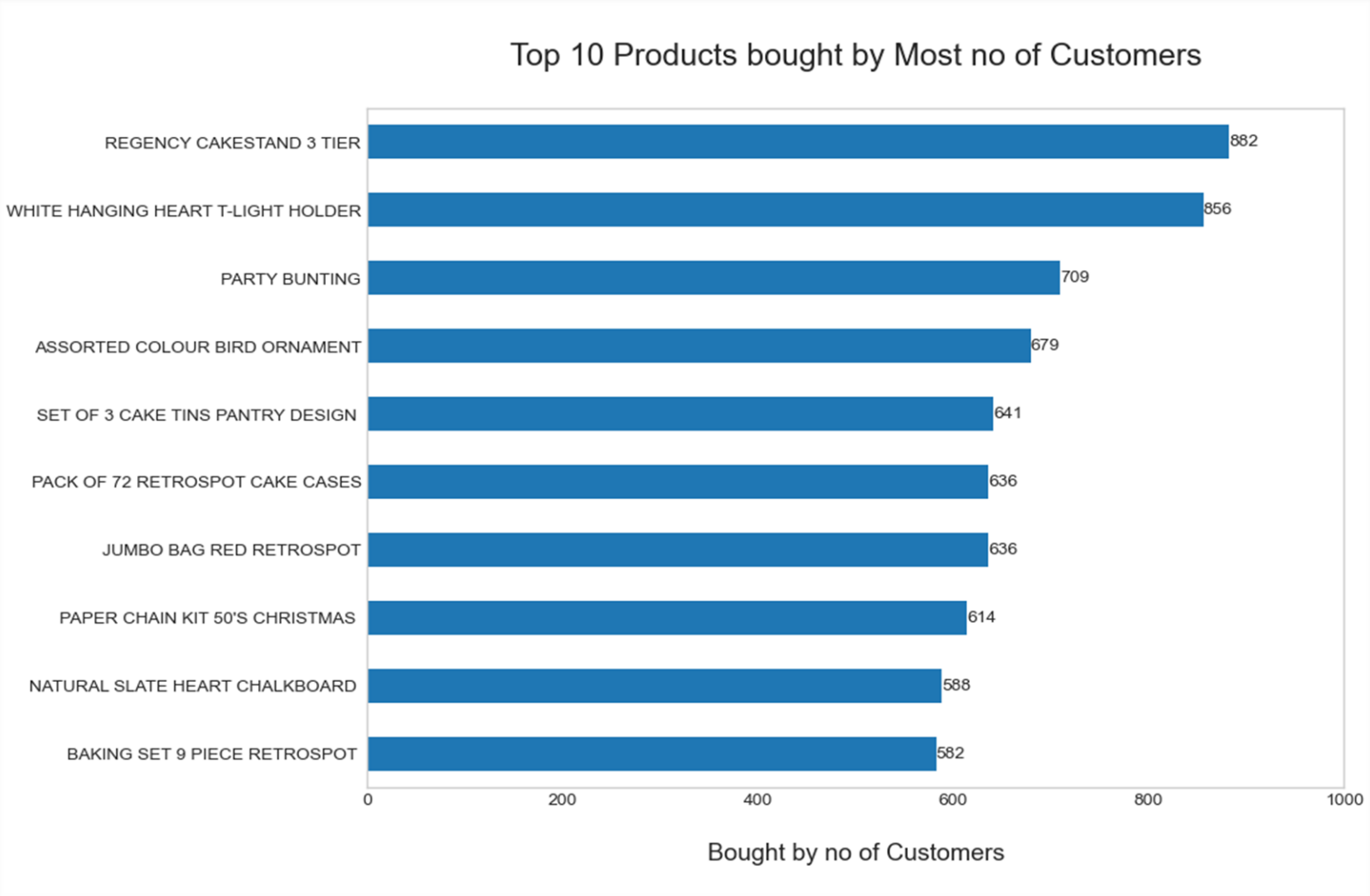
1.Here **United Kingdom** tops the list by selling the most no of Unique Products.

2.It is followed by **EIRE & Germany**.

3.Although other countries contribute least to the sales, we can see many countries have a demand in terms of no of **unique products**.

Further analysis can be considered, to check if these products have any special significance to these individual countries and accordingly marketing efforts may be altered

7. Top 10 Products Bought by Most no of Customers

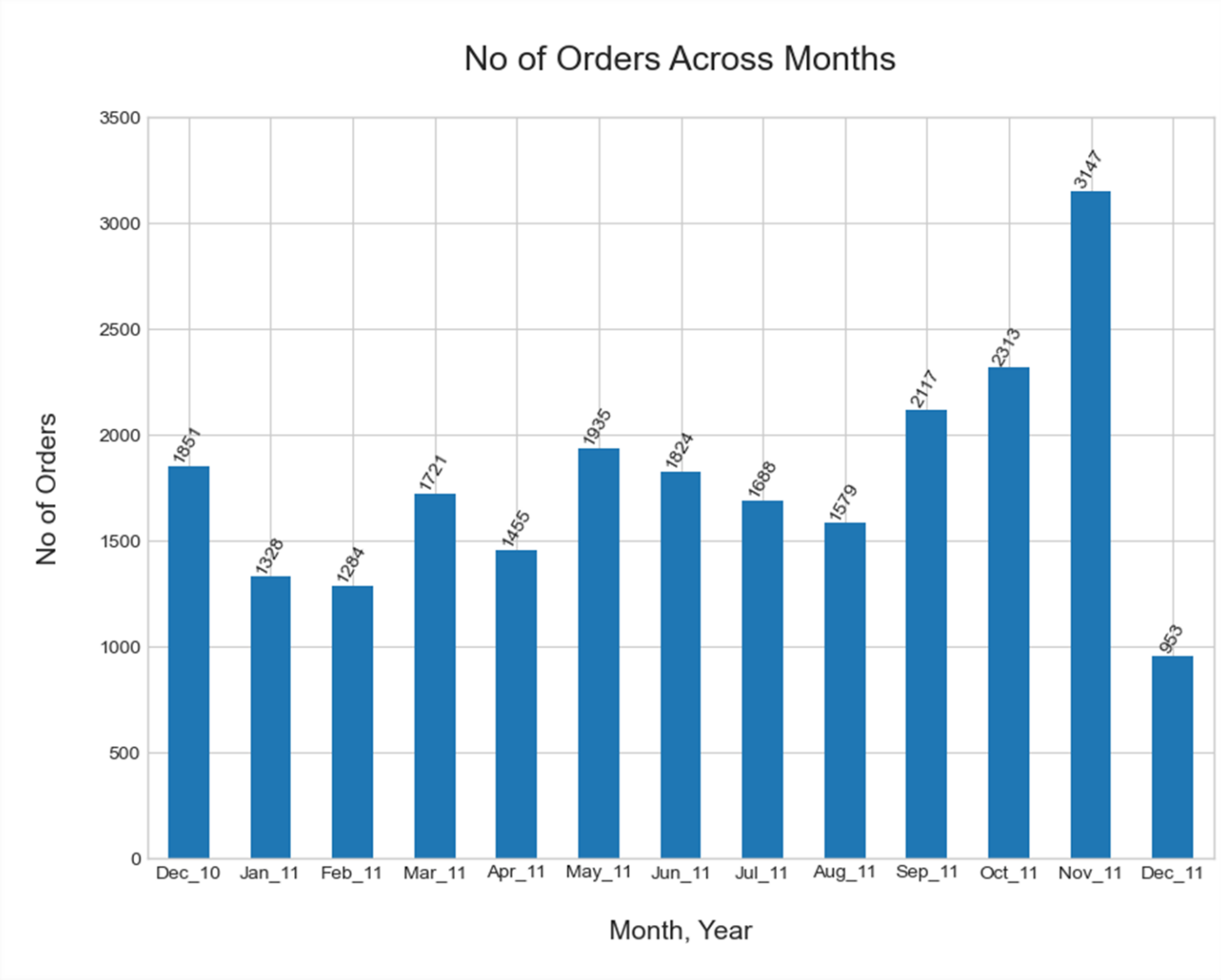


Inference:

1.We can see that **Regency Cakestand 3 tier** is bought by most of the customers(i.e. most popular among customers).

2.This chart gives us an idea about **most popular products**, which we need to keep in mind.

8. No of Orders Across Months

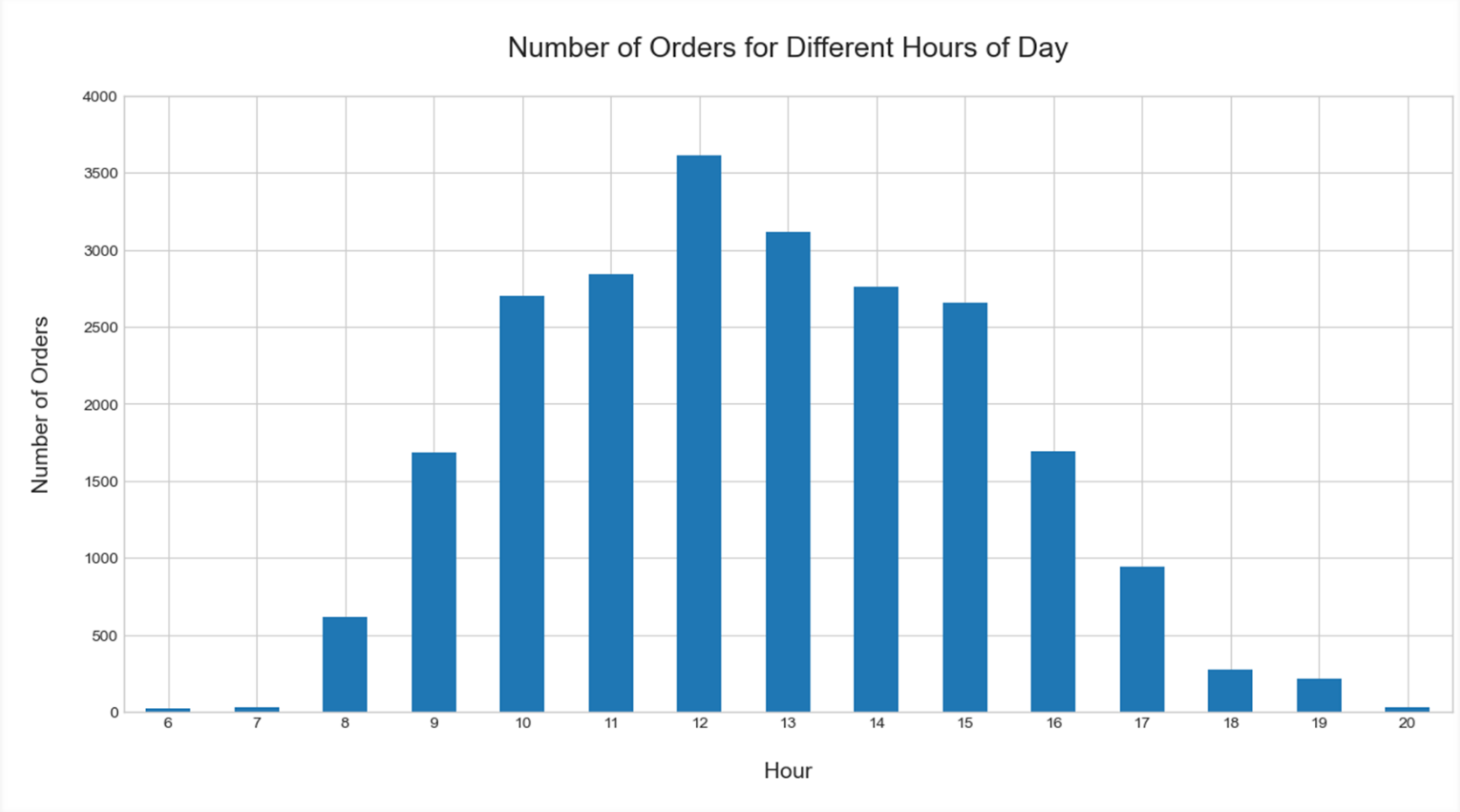


Inference:

1.The no of orders are following a **similar increasing trend** as the total sales. **Nov_11** has the most no of orders - **3147**.

2.We can safely discard Dec_11 as we do not have complete data for that month.

9. No of Orders for Different Hours of Day

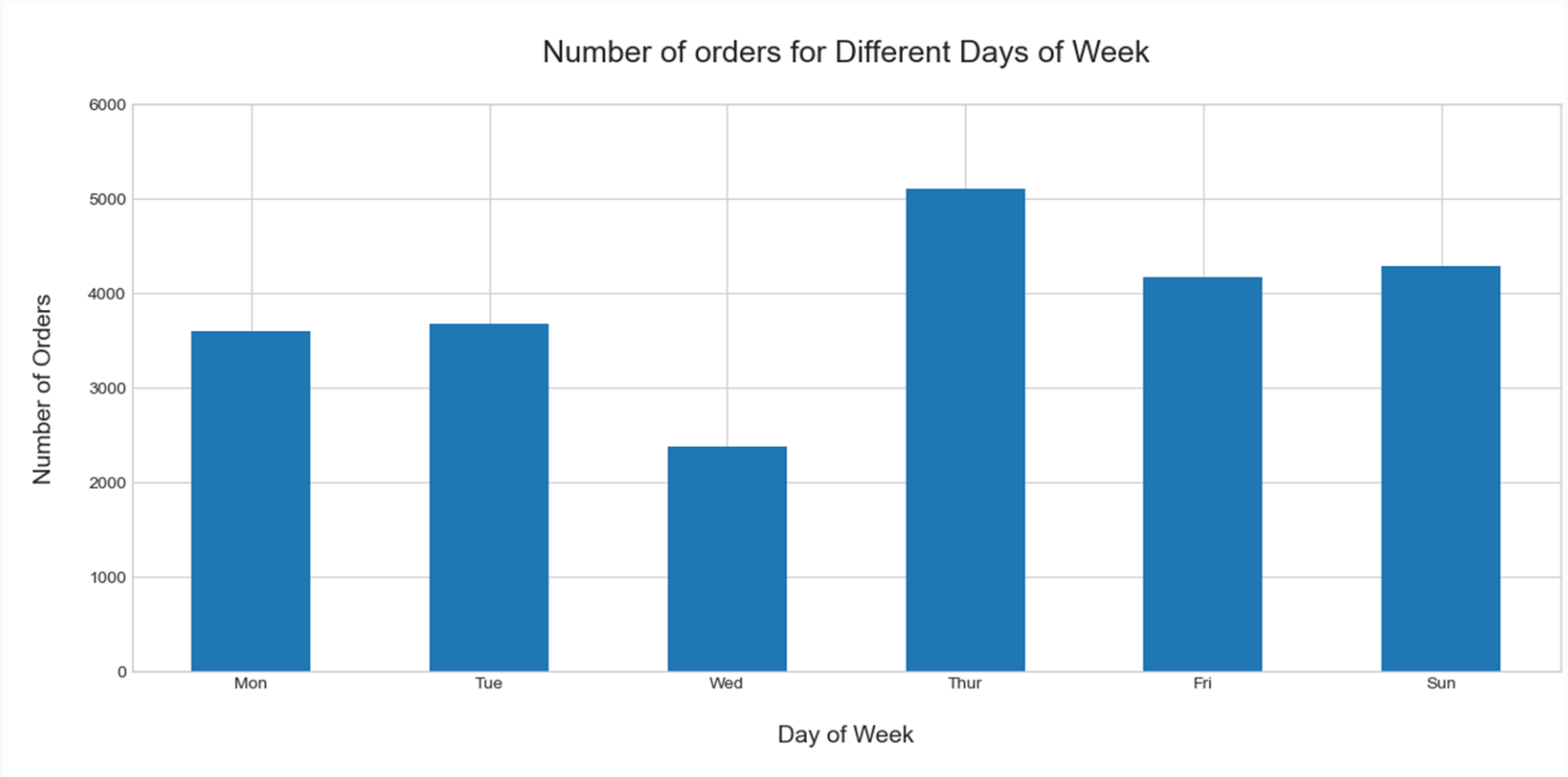


Inference:

1.We can clearly observe a pattern of orders increasing from **9am to 5pm** or **9 to 17 Hours**. With the orders peaking at 12 pm.

2.It can be assumed that the majority of customers place their orders **during their working hours**.
Further this can be verified to determine the **niche customers** and accordingly streamline the products offered for sales.

10. No of Orders for Different Days of Week



Inference:

1.**Thursday** is the day which brings in majority of sales.

2.There is no data for **Saturday** suggesting store is closed on Saturdays.

3.Also, we can see that sales are more across weekends.

4.Since, the least sales come from **Wednesday**, the store can have a weekly off on Wednesday instead of Saturday. This will insure sales are **continuous through the weekend**.

Key Findings

Key Findings

- United Kingdom brings major chunk of sales 8.28 Millions approx.
- Minimum sales(0.48 Millions) were achieved in Apr 2011 and maximum(1.43 Millions) in Nov 2011.
- An Increasing trend in sales is observed since Sep 2011.
- 90% of the Unique Customers are from United Kingdom.
- Although United Kingdom tops in each of the metrics, in terms of unique products sold, diversity is observed in the demand.
- Most of the orders were placed between 9am to 5pm, with max orders at 12pm.
- Sales are marginally higher on weekends.

Suggestions & Recommendations

Suggestions & Recommendations

- Netherlands, EIRE, Germany & France contribute 0.28, 0.25, 0.2 & 0.18 Millions annual sales respectively. These Markets can be considered for expansion efforts.
- Czech Republic, Bahrain & Saudi Arabia’s contribution is meagre, the continuation of sales in these regions needs review.
- Unique products ordered from various countries have a diverse demand, efforts can be made to focus on marketing activities centred around specific products accordingly.
- Efforts can be made to ensure storage and supply of in demand products.
- Data suggests a current week off on Saturdays. But, we observe an increased demand across weekends; to ensure the momentum of sales across weekend, a weekly off can planned on Wednesday(if required).

Conclusion

Conclusion

United Kingdom brings in a majority of sales both in terms of sales value and volume of transactions. And, there has been an increasing trend in sales observed since Sep 2011.

There is a varied demand for unique products from each of the countries, and marketing efforts can be altered accordingly.

Sales activity across weekends needs review.

Limitations & Future Work

Limitations & Future Work

Limitations:

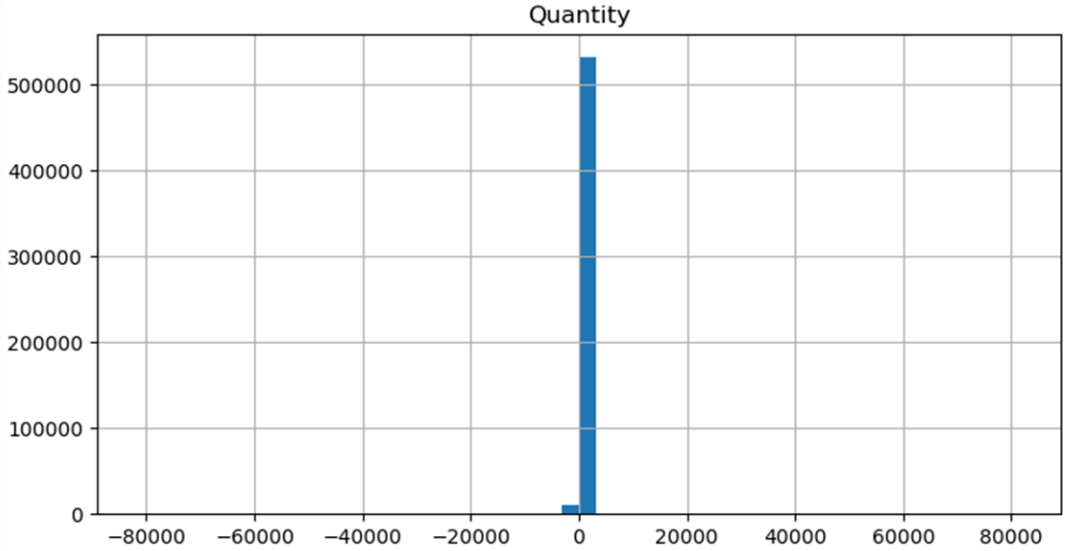
The dataset has values represented in negative to denote cancellations and also all the transactions other than sales have been clubbed together. This contributes to confusion and a potential for error. Further, data collection can be designed to avoid such occurrences, to ensure ease & accuracy of the analysis.

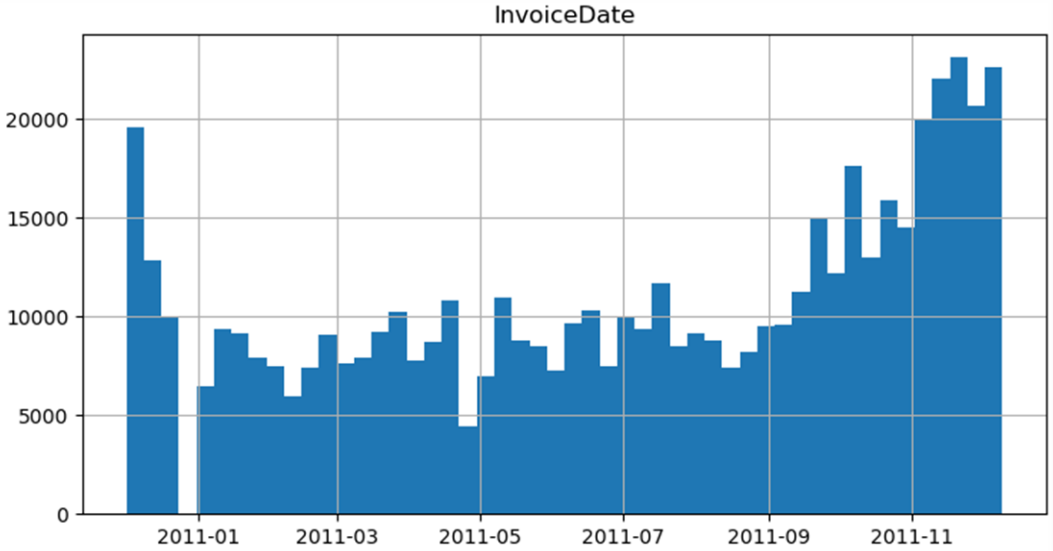
Future Work:

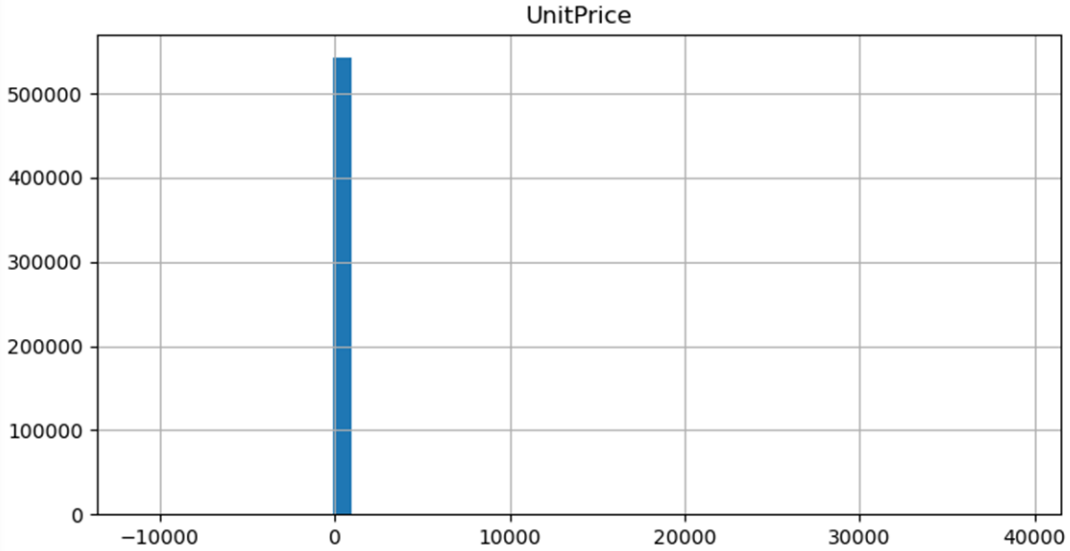
As the data is from 38 countries, country specific analysis can be undertaken. And, because of the shear volume of transactions an in depth Market basket analysis will uncover more actionable insights.

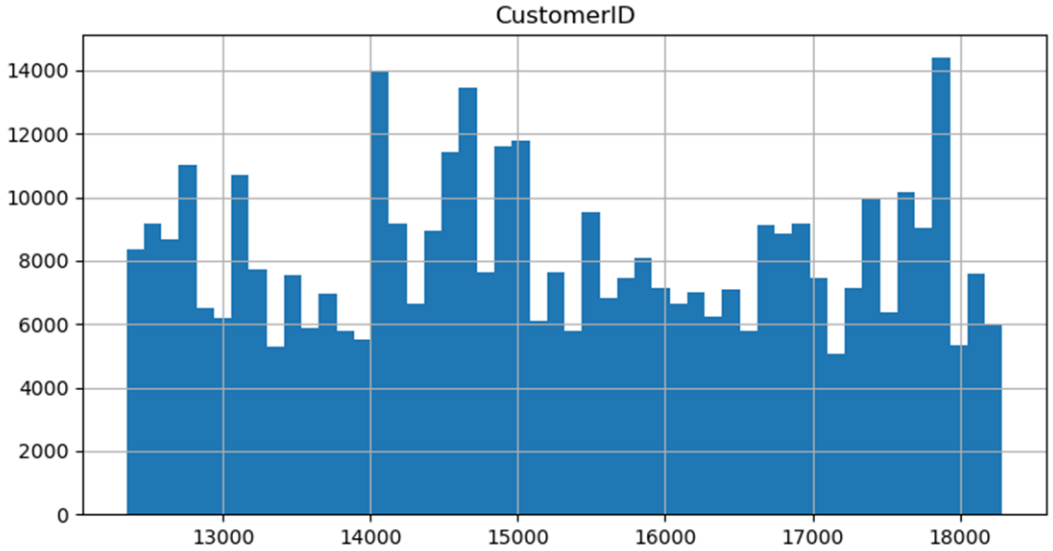
Appendix

Numerical Data Distribution Before Data Cleaning

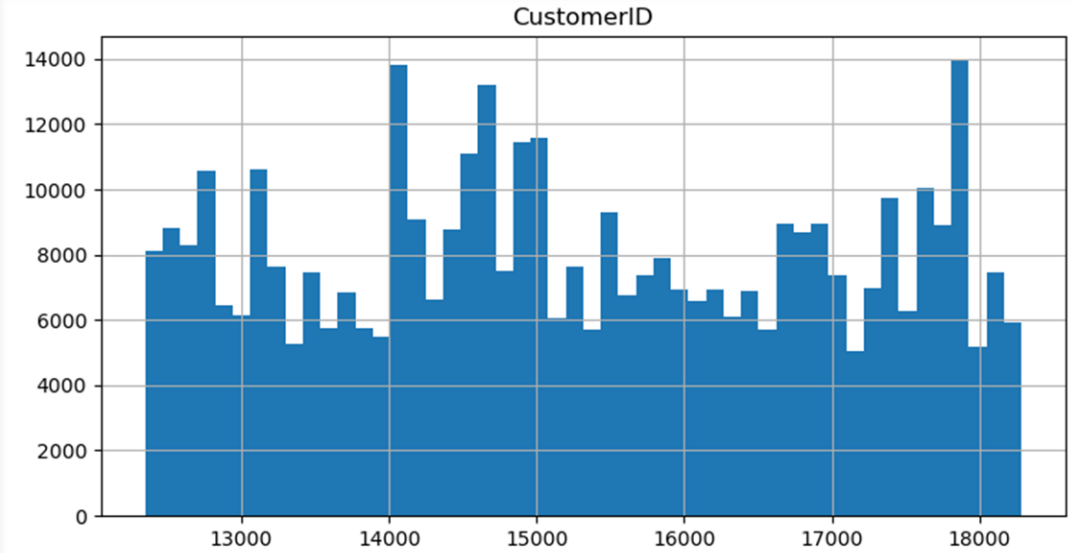


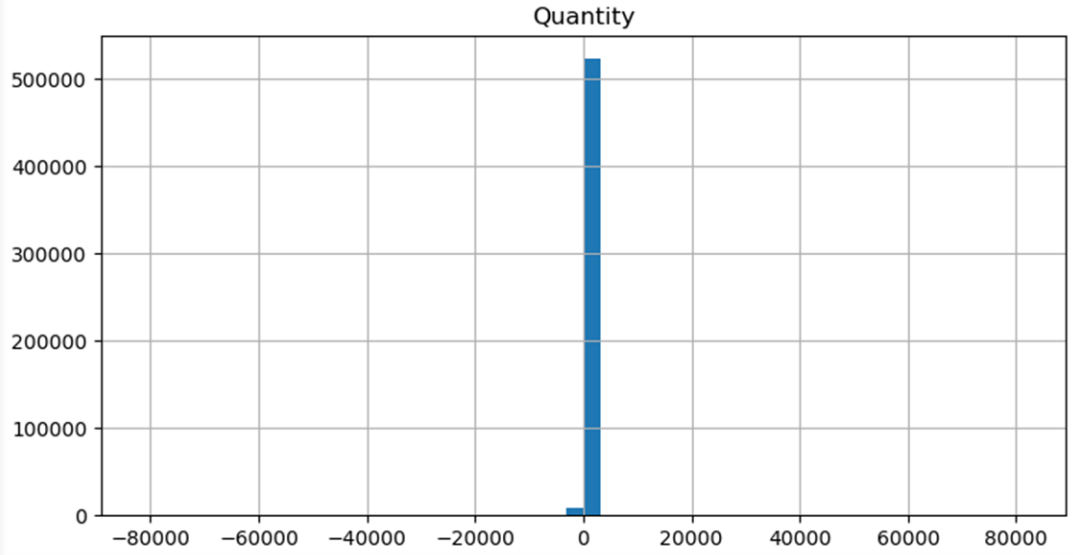


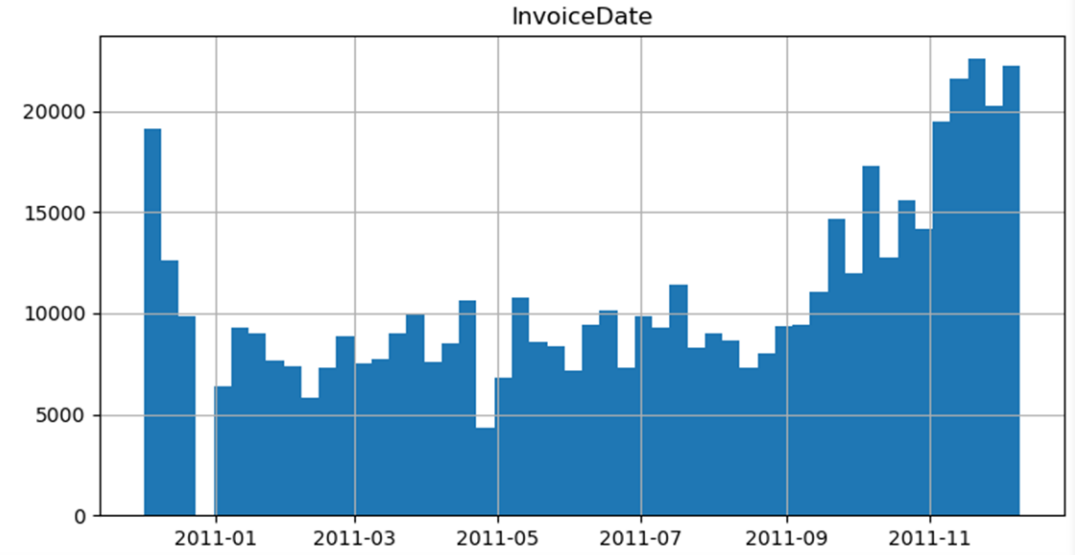




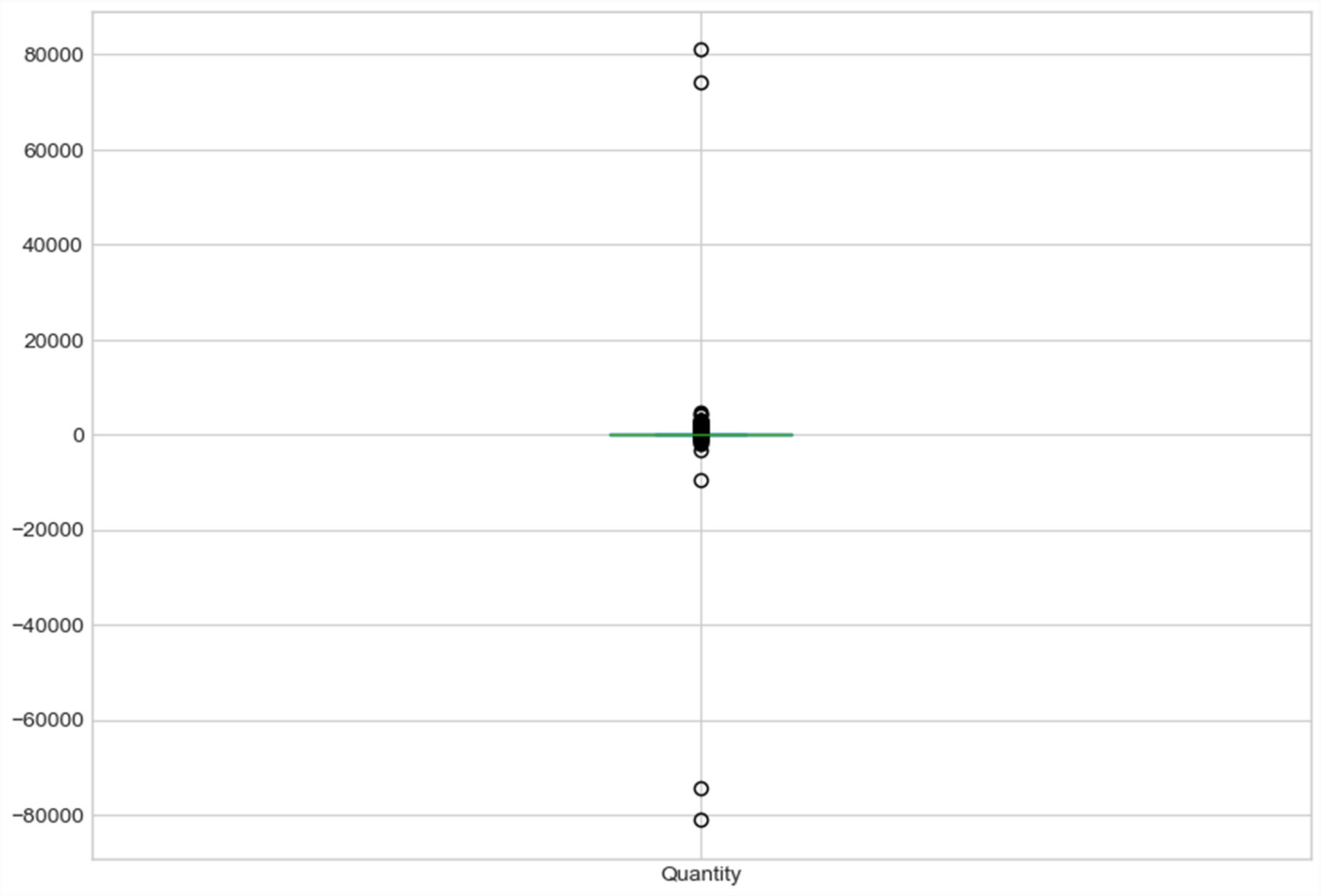
Numerical Data Distribution After Data Cleaning



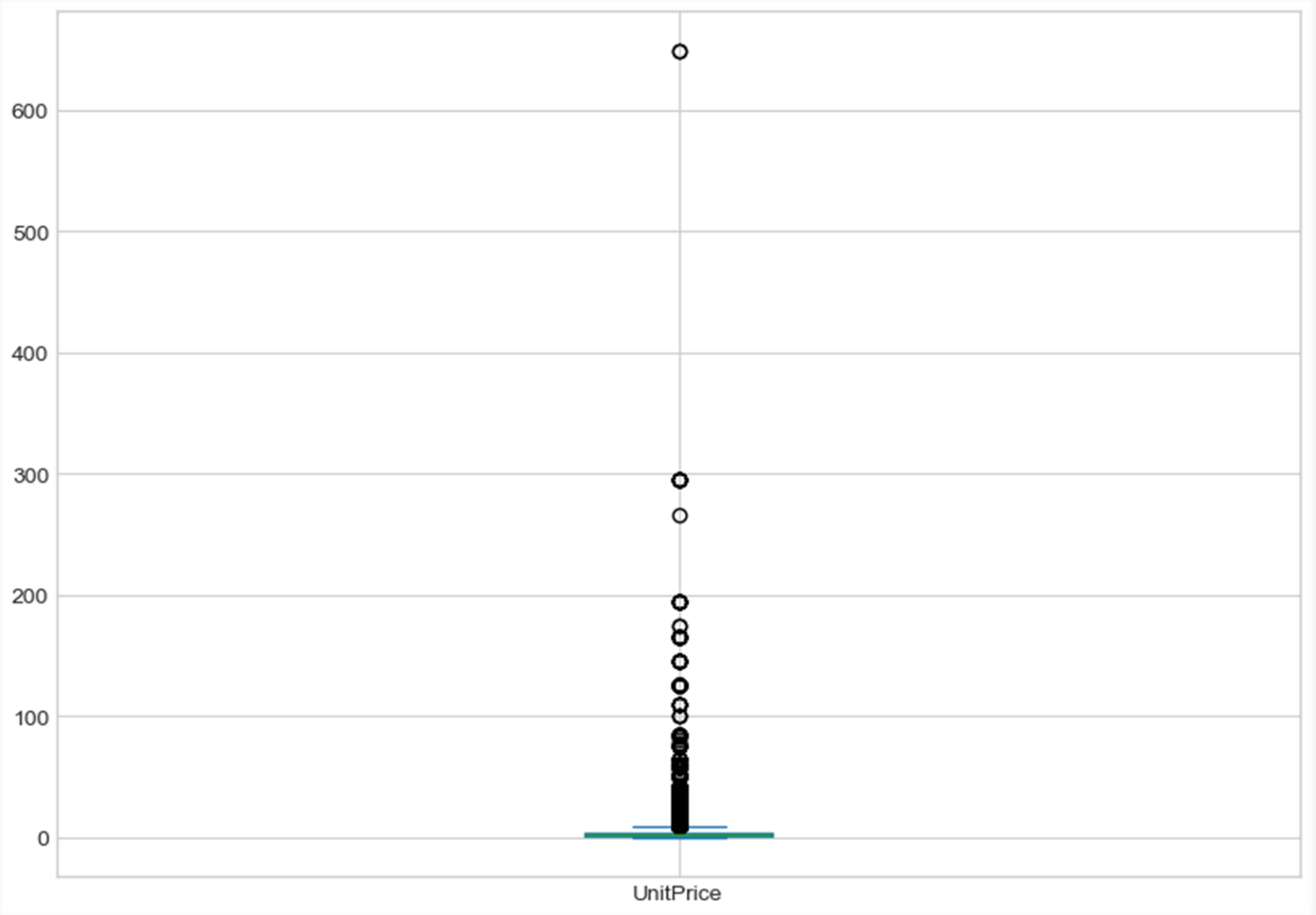




Box Plot of Quantity Column



Box Plot of UnitPrice Column



Thank You!