IEOR 4578: Forecasting Project

Process Documentation

PROJECT: Online Retail

Source: https://archive.ics.uci.edu/dataset/502/online+retail+ii

Final Deliverable

Authors: Ishita Pundir (ip2441), Saum Kothari (sbk2171), Tushar Bura (tb3077)

Table of Contents

1. Introduction

The advancement of machine learning (ML) in the field of retail analytics has revolutionized the way businesses forecast growth and turnover. This project is an ambitious endeavor aimed at leveraging the power of machine learning and statistical analysis to predict future sales trends for an online retail company. The primary objective is to develop a forecasting model that can accurately predict revenue based on historical sales data. This model is intended to serve as a strategic tool for the company, enabling it to make informed decisions on inventory management, marketing campaigns, and overall business strategy.

The project utilizes the "Online Retail II" dataset, a rich and detailed collection of transactional data from a UK-based non-store online retail platform. This dataset was sourced from the UCI Machine Learning repository and includes transactional records spanning two consecutive years (01/12/2009 to 09/12/2011). It is characterized by its multivariate nature, which includes quantitative data such as quantities sold and unit prices, as well as categorical data such as product descriptions and customer locations.

- InvoiceNo: A nominal 6-digit integral number that uniquely identifies each transaction, with a specific notation for cancellations.
- StockCode: A nominal 5-digit integral number, each corresponding to a unique product item.
- Description: The nominal name of the product item.
- Quantity: The numeric value representing the count of each product per transaction.
- InvoiceDate: The numeric timestamp of each transaction, precise to the day and time.
- UnitPrice: The numeric value of the product price per unit in sterling.
- CustomerID: A nominal 5-digit integral number uniquely assigned to each customer.
- Country: The nominal name of the customer's country of residence.

These elements were crucial in understanding the nuances of the retail company's sales dynamics. This dataset contains missing values, which will necessitate data preprocessing to ensure the quality and reliability of the forecasting models. Our analysis aims to utilize this rich dataset to forecast future sales, identify trends, and understand customer purchasing patterns. By leveraging the intricate details provided by the transactional data, this report will provide actionable insights that could be pivotal for strategic decision-making and optimizing the operational efficiency of the online retail business.

The project undertook several key steps:

- Data Preprocessing: This involved cleaning the data, handling missing values, removing duplicates, handling outliers, and filtering out irrelevant or erroneous entries.
- Exploratory Data Analysis (EDA): This step provided insights into the data distribution, patterns, and potential correlations between different variables.

- Feature Engineering: New predictive variables were created, and the data was transformed to extract meaningful attributes for the forecasting model.
- Model Development: Time-series analysis techniques were employed to build forecasting models, such as ARIMA and SARIMAX, which consider both trend and seasonality in the data.

1.1 Business Problem Definition

The primary business problem this project addresses is the need for accurate forecasting of sales for seasonal items in the retail sector. This encompasses predicting the volume of sales, understanding customer purchasing trends, and effectively managing inventory levels to meet demand without overstocking or stockouts. The ability to forecast with precision is critical for optimizing supply chain operations, aligning marketing strategies with consumer buying patterns, and ultimately ensuring profitability and customer satisfaction.

1.2 Value Creation

By leveraging ARIMA, SARIMAX, Random Forest Regressor, Exponential Smoothing, and Facebook Prophet, the project aims to provide a robust predictive framework capable of capturing complex behaviors inherent in time-series data, such as trend, seasonality, and cyclicality. The challenge lies in modeling these factors accurately to anticipate future sales, which are influenced by seasonal variations and potentially non-stationary trends. The value creation accomplished through this project is multifaceted, helping with:

a) Improved Inventory Management

By accurately forecasting demand, the project enables more efficient inventory management. It helps in maintaining optimal stock levels, reducing the costs associated with overstocking, such as warehousing expenses, and minimizing stockouts, ensuring that customer demands are met promptly. This balance improves cash flow and reduces lost sales opportunities, directly contributing to the bottom line.

b) Enhanced Strategic Planning

The insights gained from accurate sales forecasts empower decision-makers to plan more effectively for the future. This includes making informed decisions on marketing strategies, promotional activities, and resource allocation. For instance, knowing when to ramp up marketing efforts to capitalize on predicted sales peaks or when to scale back to conserve resources during slower periods.

c) Competitive Advantage

In a highly competitive retail environment, the ability to predict and react to market trends ahead of competitors can be a significant advantage. This project not only enhances

operational efficiencies but also enables the company to offer better customer service through reliable product availability and potentially more dynamic pricing strategies.

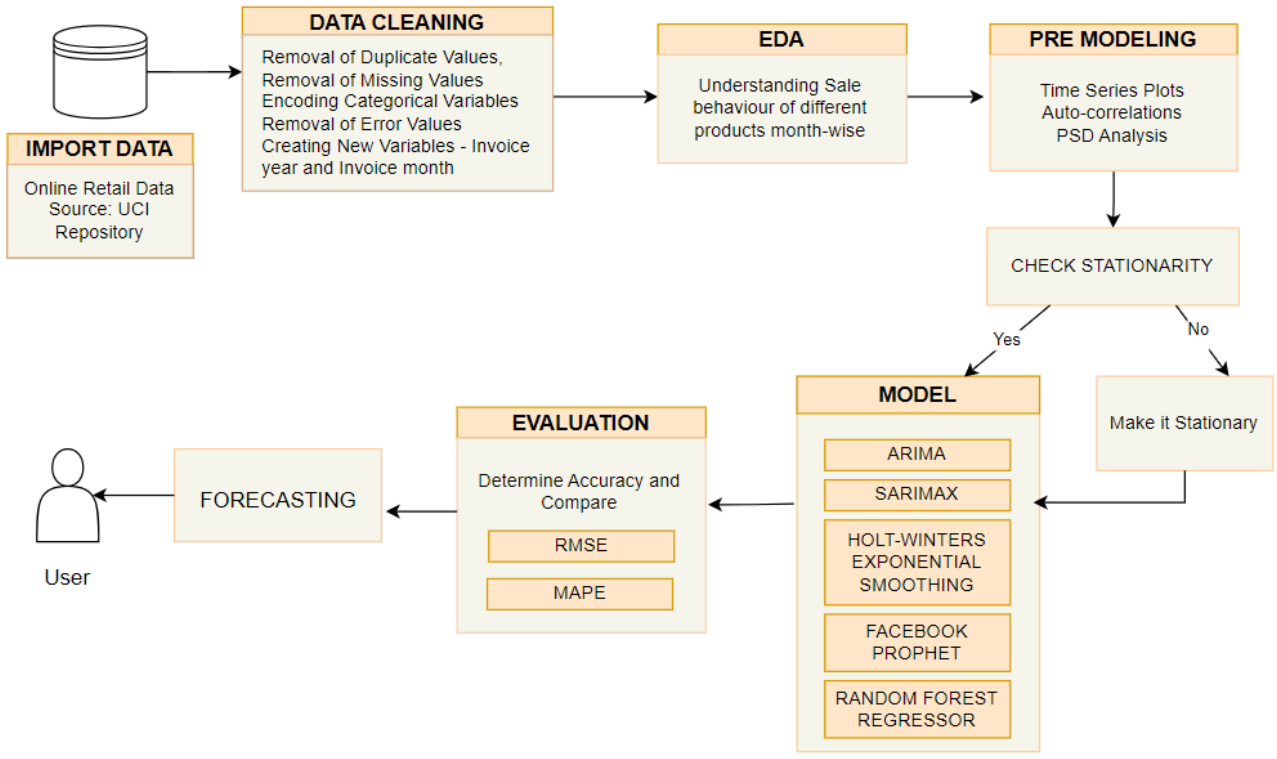
d) Customer Satisfaction and Loyalty

By ensuring products are available when and where they are needed, and by potentially using insights to tailor offerings more closely to customer preferences, the company can build stronger relationships with its customers. Satisfied customers are more likely to become repeat buyers and brand advocates, leading to increased loyalty and a stronger market position.

1.3 Need for Forecasting Analysis

In the fast-paced world of e-commerce, the ability to anticipate market trends and consumer behavior is invaluable. Forecasting analysis addresses this need by providing insights that can lead to enhanced operational efficiency, optimized stock levels, and improved customer satisfaction. The company must understand seasonal trends, customer preferences, and sales patterns to stay ahead of the competition and adapt to changing market conditions.

2. Process Flow



3. Data Extraction

The project began by importing necessary Python libraries for data manipulation and visualization. The project was initiated by importing a suite of Python libraries essential for handling the data:

- Pandas: A powerful data analysis and manipulation tool, which is a cornerstone for data science in Python. It provides data structures like DataFrames that make it straightforward to perform operations on tabular data.
- NumPy: This library adds support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays.
- Matplotlib: A plotting library that is very flexible and can generate a wide range of static, animated, and interactive visualizations.
- Seaborn: Built on top of Matplotlib, Seaborn is a statistical data visualization library that provides a high-level interface for drawing attractive and informative statistical graphics.

With the analytical environment prepared, the project moved to load the actual data:

- Data Loading: Using Pandas, the project loaded two Excel sheets into separate DataFrames. DataFrames are the primary data structure used in Pandas and are two-dimensional, size-mutable, and potentially heterogeneous tabular data structures with labeled axes (rows and columns). The excel sheets, containing data for the years 2010-2011 and 2009-2010 respectively, were read into these DataFrames, enabling the handling of data in a structured and efficient manner.
- Data Concatenation: Once the data was loaded, the two DataFrames were combined into a single DataFrame using the concat function from Pandas. This step was crucial to unify the data across the two years and prepare it for comprehensive analysis. Concatenation was performed along the rows, preserving all data points and aligning them into a singular dataset.
- Data Integrity: After concatenation, the unified DataFrame contained the complete transaction history across the specified period. This merging is integral to ensuring consistency and continuity in the data, which is essential for accurate time-series analysis.

The resulting unified dataset was thus a comprehensive representation of the company's transactions over the two-year period. It included every recorded sale, customer interaction, and product detail that could be leveraged to discern patterns and predict future trends. With this dataset in place, the project was well-positioned to proceed to the next phases of data cleaning, exploratory data analysis, and model building.

4. Data Processing

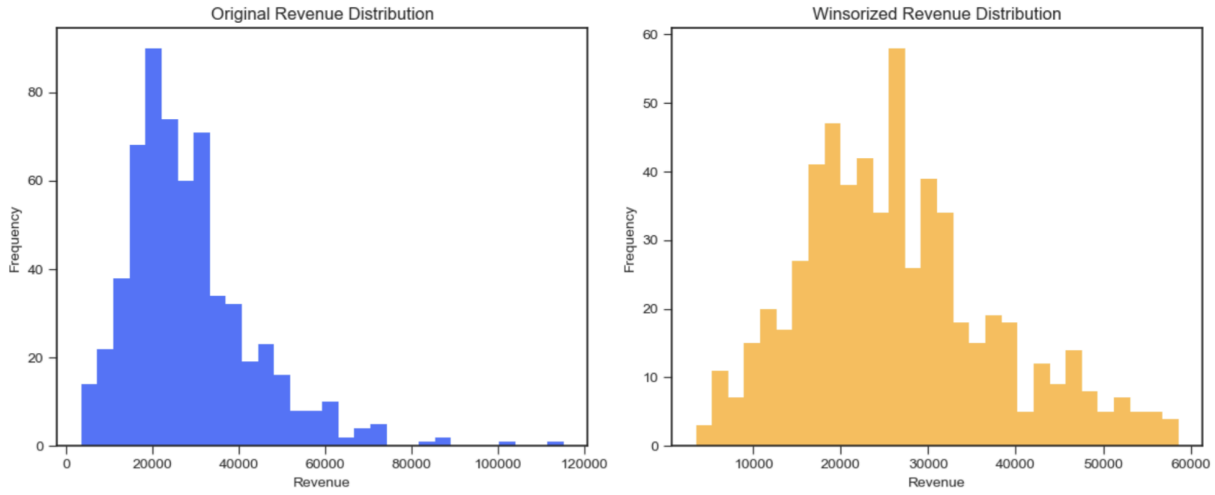
Preliminary Inspections

- **Dataset Summary:** The info() method was used to get a concise summary of the DataFrame, providing information about the index dtype and column dtypes, non-null values, and memory usage. This helps in understanding the structure of the DataFrame and the type of data in each column.
- **Uniqueness and Null Values:** The uniqueness check was performed by iterating through each column and printing out the number of unique values using len(sheets_combined[col].unique()). This helps identify the diversity within each column and can also reveal if any categorical data encoding is necessary. Missing descriptions were filled with a placeholder 'Unknown' to handle null values in the 'Description' field.
- **Focus on Complete Transactions:** By concentrating on rows with complete 'Customer ID', the analysis ensures that only verified transactions are included, which is essential for customer-level analysis and forecasting.
- **The data was limited** to the period between December 1st, 2009, and December 9th, 2011, to match the scope of the analysis.
- **Revenue** was calculated by multiplying 'Quantity' by 'UnitPrice', providing a key metric for the analysis.

Cleaning

- **Removal of Duplicates**: The Python code utilized the drop_duplicates() method from the pandas library. This method identifies and removes duplicate rows from the DataFrame, ensuring that each transaction is unique and not artificially inflating any of the statistical analyses or models.
- **Handling of Missing Values:** The dataset was inspected for missing values using the isnull().sum() method, which provides a count of null entries in each column. Specifically, rows with missing 'Customer ID' were considered incomplete and were removed using the dropna(subset=['Customer ID']) method. This ensures that the analysis is based on transactions where customer information is fully known.
- **Correction of Erroneous Entries:** Negative values in 'Quantity' and 'Price' columns, which might indicate cancellations or errors, were removed by filtering the DataFrame to only include non-negative values for these columns. This was done using boolean indexing: sheets_combined[(sheets_combined['Quantity'] >= 0) & (sheets_combined['Price'] >= 0)].
- **Handling outliers:** Detection and treatment of outliers were conducted to ensure the integrity of the dataset. Outliers can skew the analysis, leading to biased results and potentially misleading insights. To mitigate this, the **Winsorizing technique** was applied, which involves replacing extreme data points with less extreme values. The winsorized mean adjusts outliers by setting them to a specified percentile (e.g., the 5th and 95th percentiles) in the data distribution. This method retains the shape of the data distribution while reducing the impact of outliers.

We can see that after the handling of outliers, the data seems more to conform to the normal “bell-shaped” curve.



Aggregation

- **Temporal Granularity:** New columns 'InvoiceYear' and 'InvoiceMonth' from the 'InvoiceDate' using the dt accessor in pandas were created. This step is foundational for aggregating sales data at a monthly or yearly granularity, which aligns with typical forecasting models.

5. Target Variables

The target variable, also known as the dependent variable, is the primary metric that a predictive model aims to forecast or predict. It is the outcome interest, whose future values are to be determined based on other variables within the dataset. In predictive modeling, the target variable is what the model is trained to predict.

Revenue: The target variable, 'Revenue', was calculated by multiplying the 'Quantity' of items sold by their 'Unit Price'. This computation yields the total sales value for each transaction within the dataset. The choice of 'Revenue' as the target variable is strategic for several reasons:

- Financial Performance Indicator: Revenue is a direct measure of the financial health and performance of a retail company. By forecasting revenue, the company can gauge future financial states, aiding in budgeting, financial planning, and resource allocation.
- Comprehensive Metric: Unlike focusing solely on quantities sold or the number of transactions, revenue encapsulates both the volume of sales and the price point at which goods are sold. This comprehensive metric ensures that the forecasting model accounts for variations in both customer purchasing patterns and pricing strategies.
- Seasonal and Trend Analysis: Since revenue is affected by various factors including seasonal trends, economic conditions, and marketing efforts, forecasting this variable can provide insights into how these factors influence overall sales.

6. Predictive Variables

Predictive variables, also known as independent variables or features, are the inputs used by the model to make predictions about the target variable. These variables are presumed to influence or have a relationship with the target variable, and they are used to train the model on how changes in these variables are associated with changes in the target variable. InvoiceYear and InvoiceMonth: These time-related features were extracted from the 'InvoiceDate' field. They play a crucial role in the analysis for several reasons:

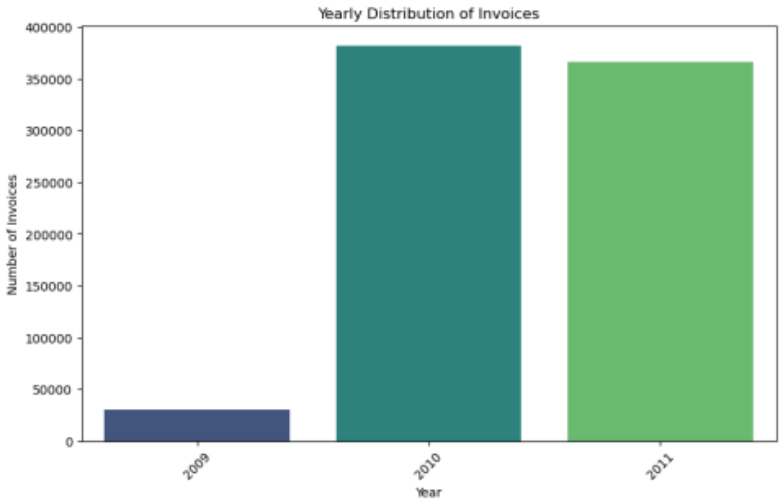
- Seasonality: Retail sales are profoundly influenced by the time of year, with certain periods (e.g., holidays, seasonal changes) driving higher sales. By including 'InvoiceYear' and 'InvoiceMonth', the model can learn these seasonal patterns and predict future revenue with greater accuracy.
- Trend Analysis: Long-term trends in sales can be identified through year-over-year and month-over-month comparisons. These trends might be driven by broader economic factors, changes in consumer behavior, or the company's growth trajectory.
- Granularity for Forecasting: The temporal granularity provided by these features allows for flexible forecasting models. Depending on the business need, the model can forecast revenue monthly or yearly, providing tailored insights for planning purposes.

Other variables within the dataset likely serve as predictive variables, including:

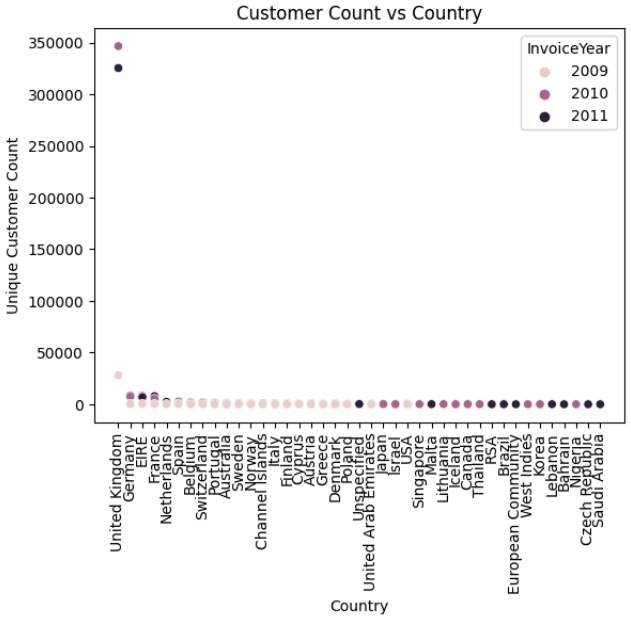
- Customer ID: Analysis of individual customer purchasing patterns can reveal insights into customer value, loyalty, and segmentation.
- Stock Code and Description: These product-related variables can help identify which items contribute most to revenue, seasonal product trends, and the impact of product mix on overall sales.
- Country: Geographical insights can be gained by analyzing sales by country, revealing regional trends and market penetration.

7. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) and visualization phase in this project plays a pivotal role in understanding the underlying structure and patterns within the Online Retail II dataset.

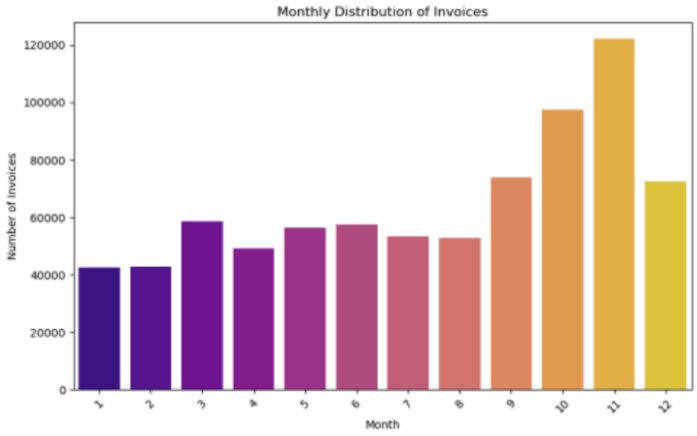


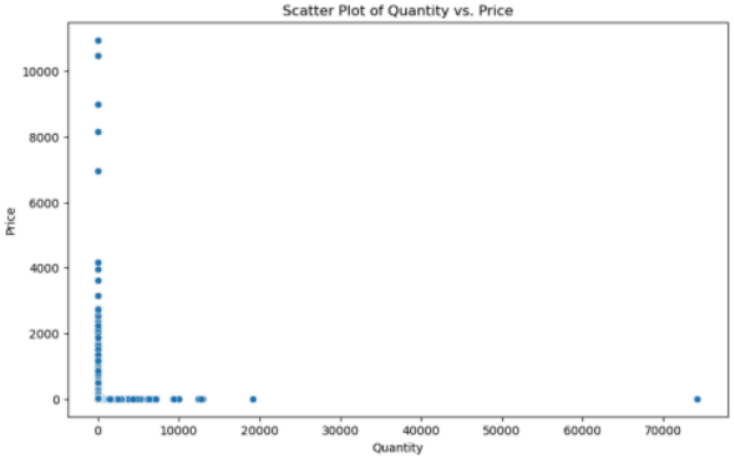
The number of invoices in 2009 is considerably lower than in the subsequent years, which may suggest that the dataset for 2009 is not complete or that the business started late in that year or was in its early stages of operation. There is a notable increase in invoices from 2009 to 2010, indicating a period of growth for the company, possibly due to business expansion, increased market penetration, or a successful marketing campaign. This is followed by a slight decrease in 2011, which may suggest market saturation, increased competition, changes in consumer behavior, or external economic factors.



The graph provides insights into the distribution of unique customer counts across different countries, segmented by the years 2009, 2010, and 2011, as indicated by the color-coded 'InvoiceYear'. The visualization shows a higher concentration of customers in certain countries, such as the United Kingdom, which is the retailer's primary market. By observing the distribution of colors, one can discern the growth or contraction of the customer base in each country over the three years. Countries with fewer data points, especially in the more recent years, might indicate emerging markets or areas with lower market penetration, which could be targets for future growth initiatives.

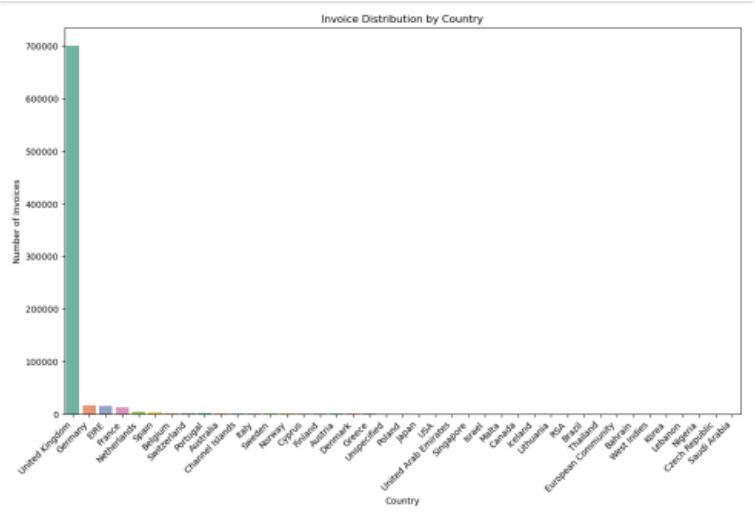
The "Monthly Invoice Distribution" bar chart indicates the number of invoices issued each month, with the highest numbers occurring in the later months of the year. This suggests a seasonal trend with invoice volume peaking towards year-end, possibly due to holiday shopping, or end-of-year sales promotions because of Black Friday, Cyber Monday, or Christmas.There is no significant spike in the middle months, which indicates a steady state of business operations without major fluctuations.

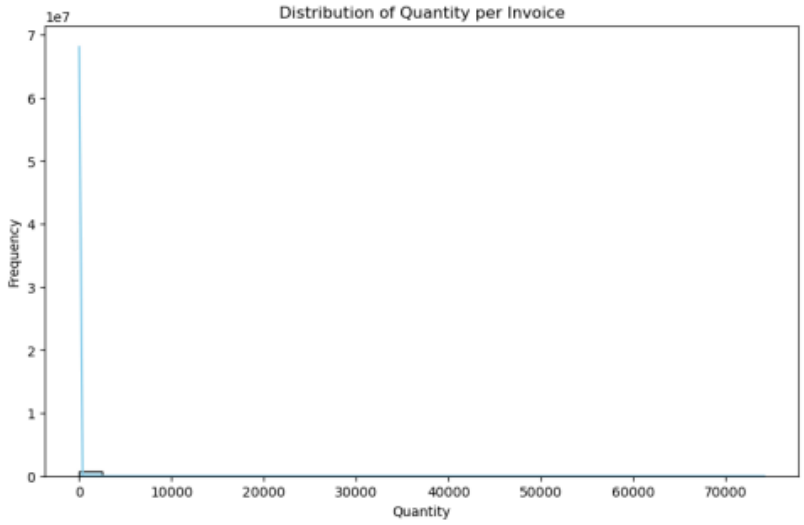




The plot displays a relationship between the quantity of items sold and their price. Most data points are clustered near the origin, indicating that a large number of transactions involve lower quantities and lower prices. This could suggest that the retailer deals primarily in low-cost items, or that bulk purchases are infrequent. As the quantity increases, the number of data points decreases sharply, which is typical in retail, where high-volume sales are less common than smaller transactions.

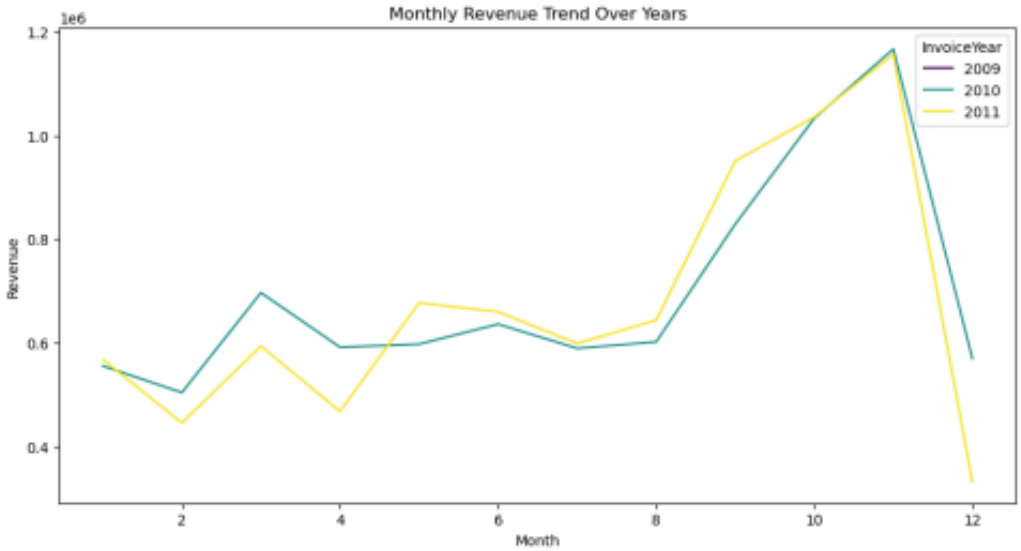
The bar chart presents the number of invoices issued for different countries. It is immediately apparent that the majority of invoices are attributed to the United Kingdom, dwarfing the invoice counts of other countries by a substantial margin. This suggests that the UK is the primary market for the company, with potentially a domestic base and a strong customer presence there.



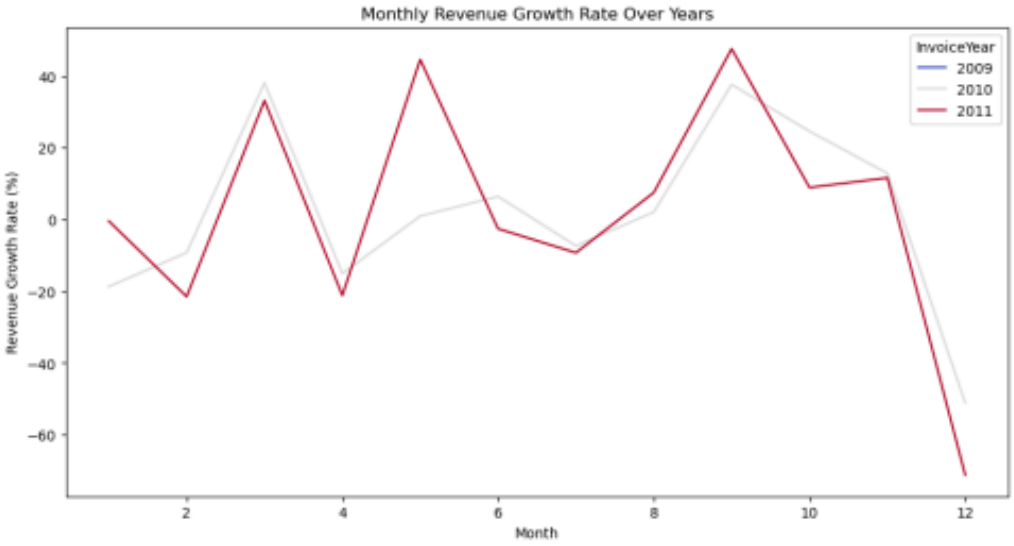


The histogram shows a highly skewed distribution with a concentration of frequency at the lower end of the quantity axis, indicating that most invoices contain a small number of items. This is common in consumer retail where small quantity purchases are more frequent. This suggests that the retailer should optimize for handling a high volume of small transactions efficiently.

The line chart illustrates the revenue generated by the company for each month across three different years: 2009, 2010, and 2011.

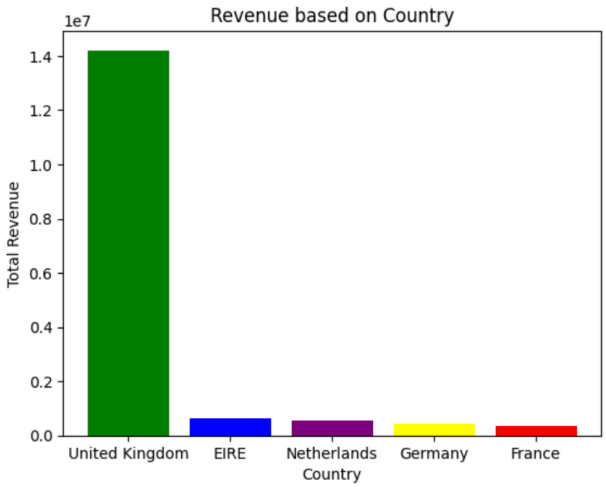


- Seasonal Pattern: There is a clear seasonal pattern with revenue peaking towards the end of each year, likely due to the holiday season, which typically sees an increase in consumer spending.
- Yearly Growth: There is an upward trend in revenue from 2009 to 2011, indicating overall growth in the business.
- Inconsistencies: The sharp decline in revenue at the end of 2011 may suggest incomplete data for that period, or it could be a result of external factors affecting sales negatively.
- Revenue Plateau: Mid-year, the revenue seems to plateau or slightly decrease, indicating possible seasonal slumps where the business could explore strategies to boost sales.

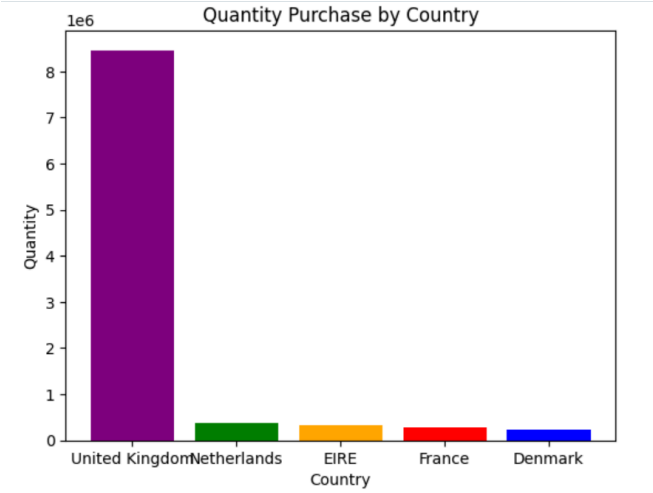


The line chart shows the % change in revenue month to month for the years 2009, 2010, and 2011.

- Volatility: There's significant volatility in revenue growth rates throughout each year, with notable peaks and troughs. This suggests that the business experiences significant fluctuations in revenue growth month to month.
- Growth Spikes: There are points where the growth rate spikes, particularly in the later months of each year, which could be attributed to seasonal sales events or holidays that boost purchasing.
- Negative Growth: There are also periods of negative growth rate where revenue has decreased from the previous month. This could indicate off-peak seasons or could be the result of stock issues, economic factors, or other business impacts.
- Comparative Performance: The lines for each year can be compared to understand the relative performance and stability of the company's revenue growth year over year.

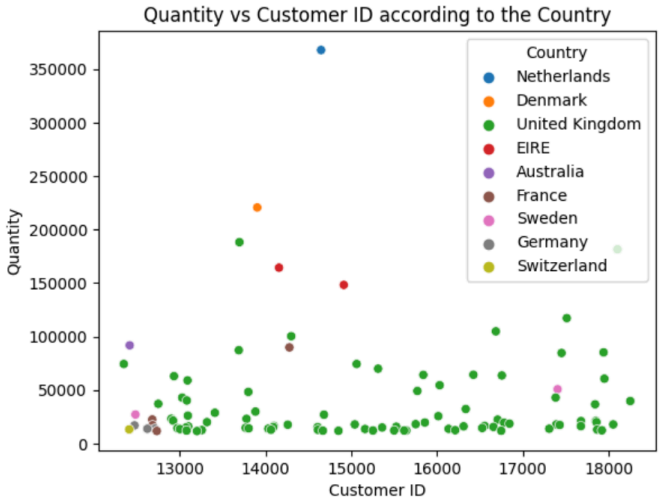


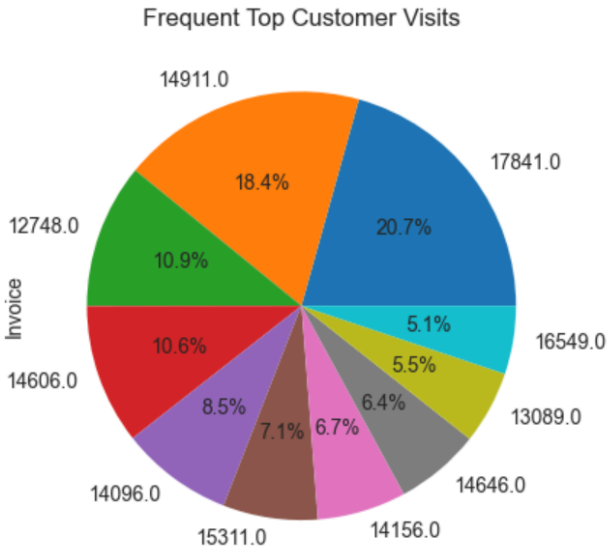
The UK bar is disproportionately higher than that of any other country, indicating that it is the dominant market for the company's sales. The visualization shows that the UK, EIRE (Ireland), the Netherlands, Germany, and France are the top markets. The company might use this data to focus its strategic efforts, such as marketing campaigns, expansion strategies, and customer engagement, on regions that offer the highest revenue potential.



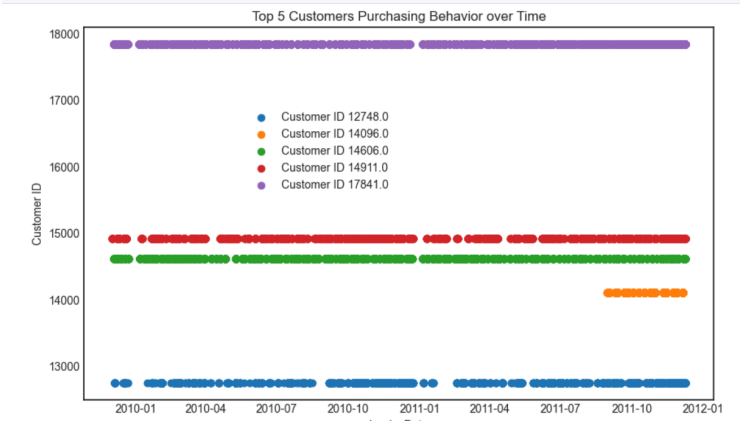
This represents the quantity of products purchased from an online retail company across different countries. The UK stands out with a significantly higher quantity of purchases than any other country. Understanding the volume of purchases by country can help the company manage its inventory more effectively.

There is a wide variation in the quantity of items purchased by individual customers. Most customers seem to purchase relatively small quantities, as indicated by the concentration of data points near the bottom of the chart. A few customers, particularly from the Netherlands, have made exceptionally large purchases. These could represent bulk buying or wholesale transactions. The United Kingdom has the most data points spread across the entire range of customer IDs, suggesting a diverse customer base with varying purchasing habits. This is consistent with the earlier inference that the UK is the company's primary market.



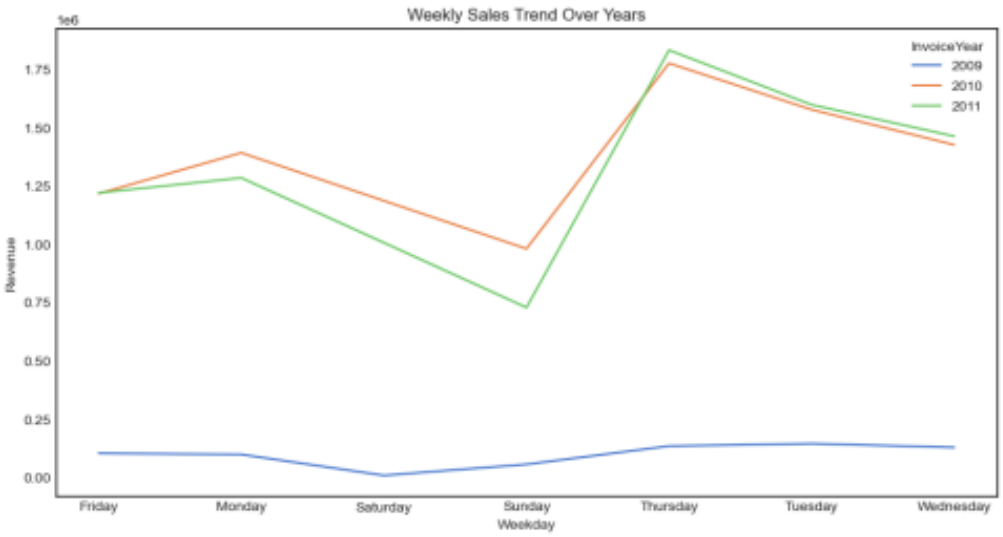


The image visualizes the distribution of visits or transactions made by the top customers of a company. The customer corresponding to the largest segment, with 20.7%, has the highest frequency of visits or transactions. Customers with a higher frequency of visits might be considered more loyal or engaged, and they could be targeted for retention strategies, loyalty programs, or upselling opportunities.



By examining the distribution of dots over time for each customer, we can assess the consistency of their purchasing behavior. A customer represented by dots evenly spread across the entire time frame might be considered a consistently active customer. If certain patterns or clusters emerge at regular intervals, this could suggest seasonality in the customers' purchasing behavior, such as increased activity around certain times of the year.

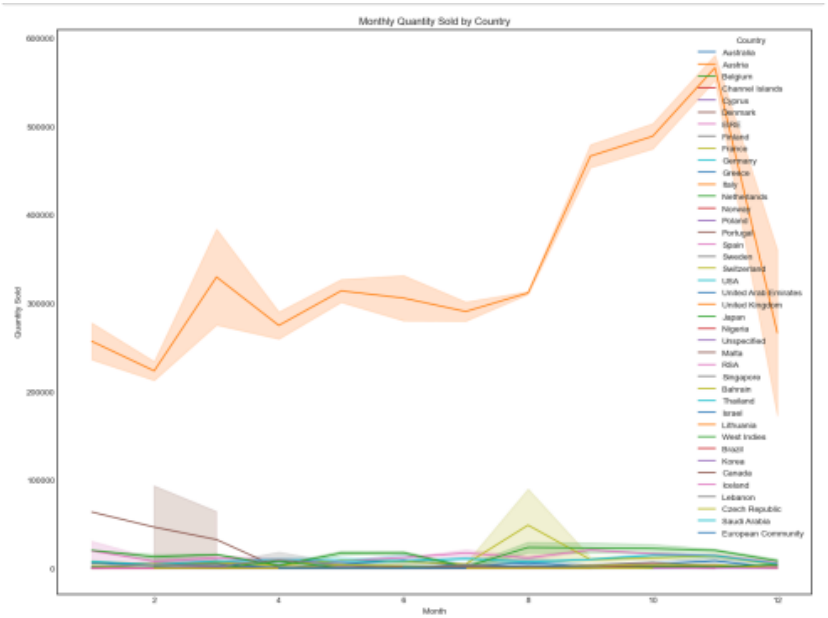
A significant number of customers purchase between 1 to 10 items, as shown by the tall yellow bar. This suggests that most customers are making small-scale purchases typical of retail consumers. The smallest bar, colored red, represents customers purchasing in large quantities (50 or more items), indicating these transactions are much less common.



The chart indicates:

- Midweek Peak: There's a prominent peak in sales revenue on Thursdays for all three years, suggesting that this day might have special significance such as weekly promotions or customer buying habits that lead to increased sales.
- Weekend Variation: Sales dip on the weekends, with Saturday showing a noticeable decrease in revenue. This might be due to the store's closure or reduced hours on weekends, or possibly because customers tend to shop less on Saturdays in this market.
- Consistency Across Years: The overall pattern appears to be consistent across the three years, with slight variations in the magnitude of revenue.
- End and Beginning of the Week: There's a rising trend in sales from the beginning of the week, reaching a peak on Thursday, followed by a fall towards the weekend.

The country represented by the top line (UK) has a sales volume that vastly exceeds the others, confirming its status as the company's primary market. There are noticeable peaks towards the end of the year, aligning with global retail trends that typically see a surge during the holiday season. The considerable variation in sales volume between countries highlights differing levels of market penetration, which could

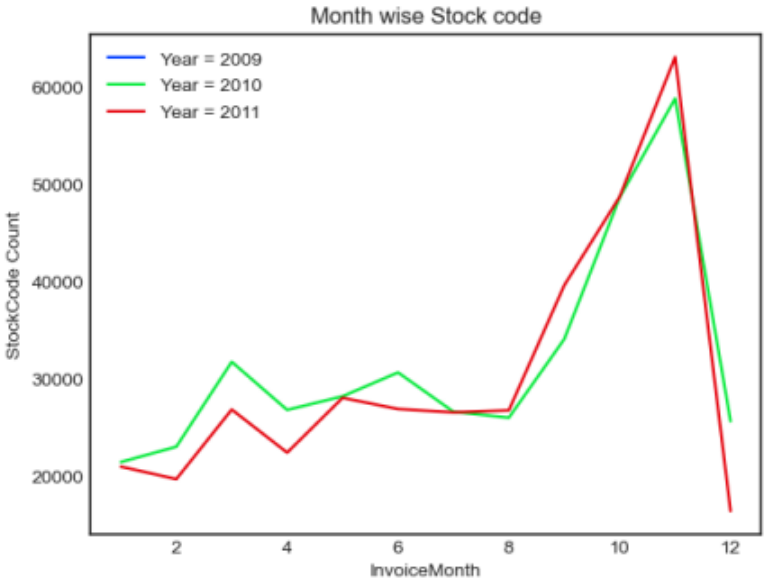


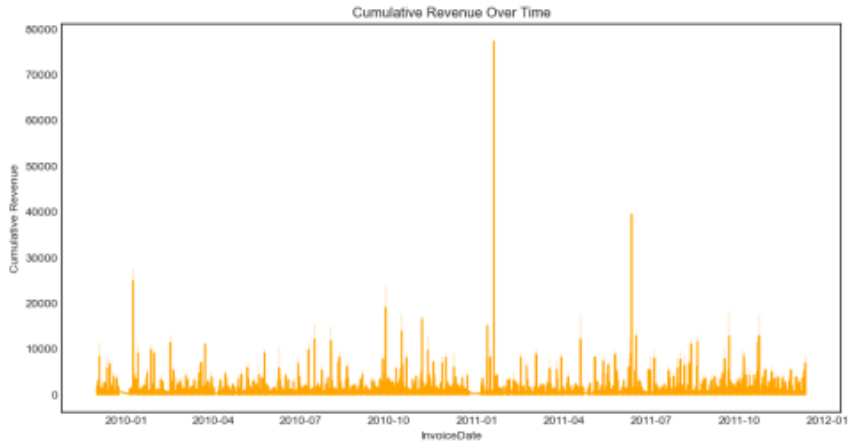
influence strategic decisions on resource allocation and targeted marketing. The closely packed lines at the bottom of the chart indicate that there are numerous countries with relatively similar and low sales volumes, which could suggest that these are emerging markets or areas with potential for growth.

There is a visible growth in the total quantity sold from 2009 to 2011, with each subsequent year starting at a higher point than the last. This suggests that the business is growing year over year. All three years show a trend of increased sales towards the end of the year, peaking around November, which could be related to holiday shopping. The sharp decline at the end of 2011 could be due to incomplete data for December, a significant drop in sales, or external factors affecting the market at that time.

There is a general upward trend in the variety of items sold each year, with the start of each year higher than the last, indicating an expanding inventory or a growing variety of products offered.

Similar to sales trends, there's a sharp increase in the stock variety towards the end of each year, particularly in November, possibly to cater to holiday season demand.



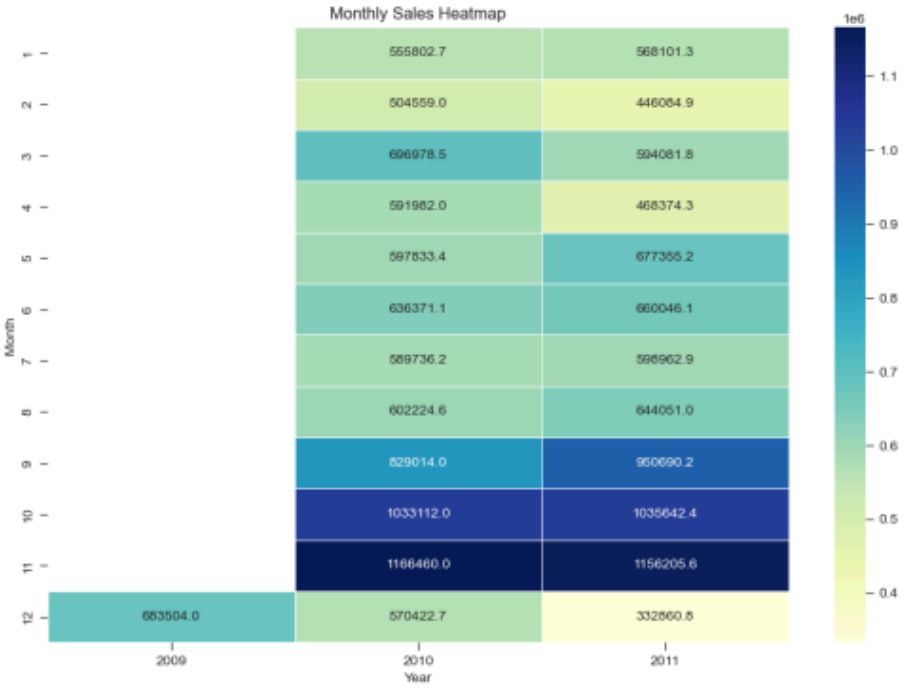


The overall trajectory is upwards, showing that the company's cumulative revenue is increasing over time, a healthy sign of business growth. There are sharp spikes at certain points, which could indicate days with exceptionally high sales.

The pattern of spikes could also reflect seasonality, with certain periods like year-end showing higher cumulative revenue increases, likely due to holiday shopping. The cumulative nature of the graph dampens the impact of short-term fluctuations and provides a clear view of the long-term trend in revenue.

Each cell in the heatmap corresponds to a specific month and year, with the color intensity and the number within the cell indicating the total sales for that month.

There's a clear pattern where sales increase towards the end of each year, which is particularly pronounced in 2010 and 2011, likely due to the holiday season.

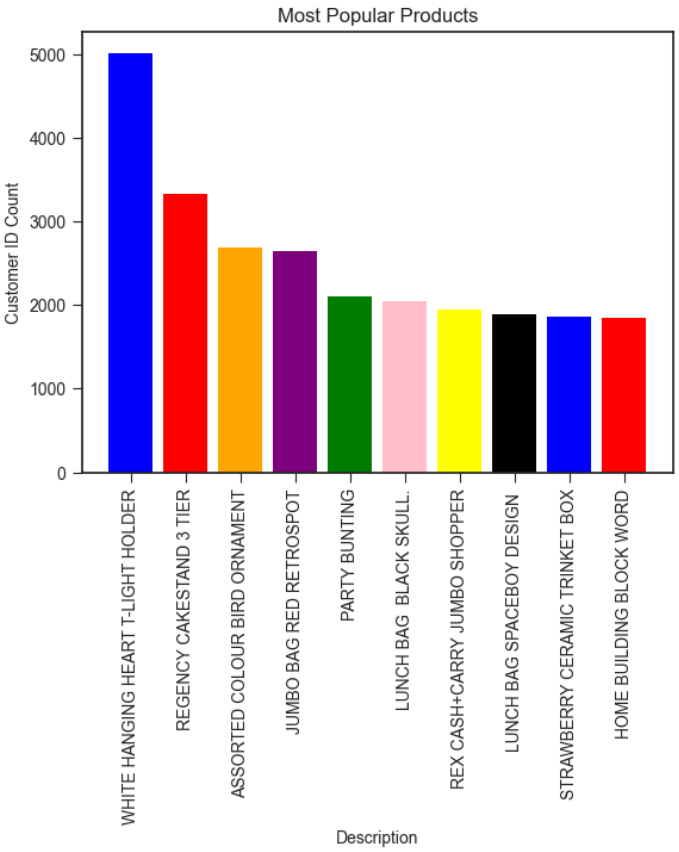


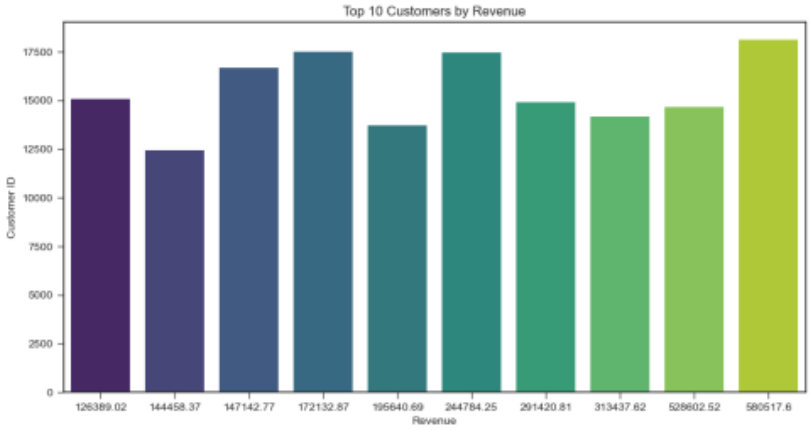
The varying color intensities indicate fluctuations in sales volume from month to month, with some months showing significantly higher sales than others.

Comparing the same months across different years, we can see that certain months show a year-over-year increase in sales, such as the end-of-year months.

There is a clear indication of price sensitivity among customers. The highest quantity of items sold is associated with the lower price of £1.25, suggesting that items within this price range are more popular or accessible to customers. There is a general trend of diminishing quantities with increased unit price. This suggests a negative correlation between price and quantity sold, which is a typical economic principle where demand decreases as the price increases.

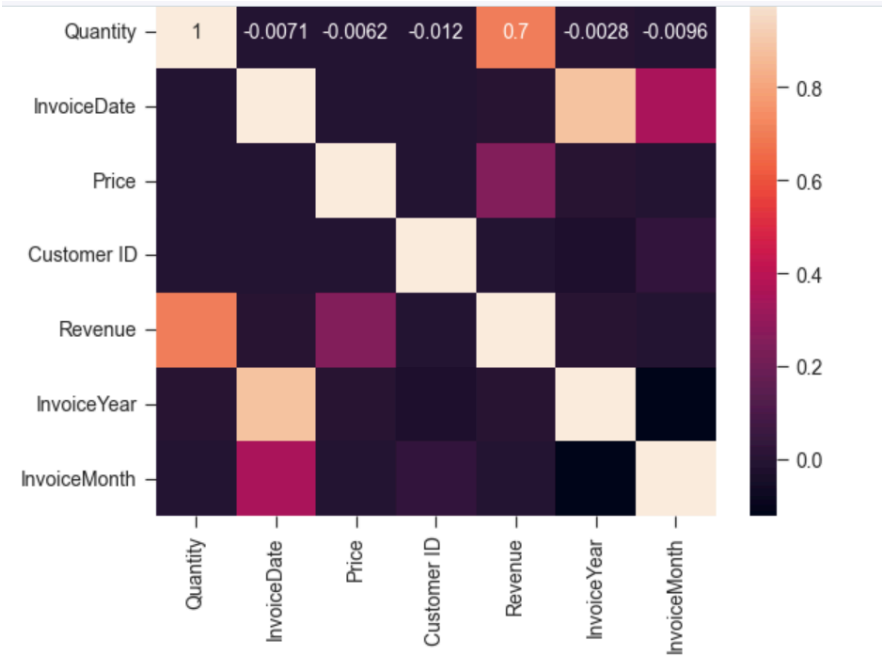
The chart ranks products by popularity based on the customer count. There is a significant drop in popularity from the first to the second product which could suggest that the "WHITE HANGING HEART T-LIGHT HOLDER" is an outlier in terms of popularity or may have been part of a promotion that drove its sales up. The popularity metric could be used for inventory decisions, stocking more of the popular items and considering phasing out or re-evaluating the marketing for less popular items.





There is a variation in revenue contribution among the top customers, with the rightmost customer generating the most revenue, indicating a valuable client for the company. The top few customers are possibly bulk buyers or frequent shoppers.

Identifying these top customers is critical for the company to maintain strong relationships, tailor services, and possibly create loyalty programs to retain these high-value clients. The disparity in revenue among the top customers could suggest different levels of engagement or purchasing power, which could be useful for segmentation and personalized marketing efforts.



The image is a correlation matrix heatmap that visualizes the relationship between different variables in a dataset.

The variables included in this correlation matrix are Quantity, InvoiceDate, Price, Customer ID, Revenue, InvoiceYear, and InvoiceMonth.

- Quantity and Revenue: The strong positive correlation between Quantity and Revenue suggests that as the quantity of items sold increases, the revenue also increases.
- Price and Revenue: There may be a less strong positive correlation between Price and Revenue, indicating that higher-priced items contribute to revenue, but possibly not as significantly as quantity sold.

- InvoiceMonth and Revenue: The correlation between InvoiceMonth and Revenue seems to be weak, indicating that there might not be a simple linear relationship between the month of the year and the revenue, or it could suggest that seasonality affects different products in different ways.
- Customer ID and Other Variables: Customer ID shows little to no correlation with other variables, which makes sense as it's a nominal variable that uniquely identifies a customer rather than a quantitative measure.

8. Time Series Analysis

A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.

Trend: In time-series analysis, the trend represents the long-term progression of the series. Trends can be upward, downward, or even sideways over time. A trend reflects the overall direction of the data points when looking at a graph over a long period. *Example*: If a company's sales have been increasing by an average of 5% per year over the last decade, this steady increase is the trend in the company's annual sales data.

Seasonality: Seasonality refers to periodic fluctuations that regularly occur in time-series data, which are predictable and repeat over a specific period, such as a day, week, month, or season. This is often due to factors like weather, holidays, and events. *Example*: Retail stores often experience seasonal patterns with spikes in sales during the Christmas holidays and other festive periods due to increased consumer purchasing.

Cyclicality: Cyclicality indicates the repeating patterns in time-series data that occur over variable periods, which are typically longer than a season. These cycles are not as predictable as seasonality because they don't have a fixed frequency. Economic cycles are a common example, where periods of expansion are followed by contractions over several years. *Example*: The business cycle, which consists of periods of economic growth (expansion), peak, contraction (recession), and recovery, shows cyclical behavior in economic time-series data such as GDP or employment rates. These are influenced by broader economic factors and do not follow a fixed calendar schedule.

Stationarity in a time series refers to the statistical property that the mean, variance, and autocorrelation structure of the series do not change over time. In simpler terms, a stationary time series retains its behavior and structure over time, which makes it predictable and hence easier to model for forecasting purposes. Most time series modeling techniques and forecasts are based on the assumption that the time series is stationary. If the time series exhibits trends, seasonality, or cyclicality, these can affect the stability of its statistical properties over time, and the series is said to be non-stationary. Non-stationarity can lead to unreliable and misleading models and forecasts.

To check stationarity, we have used the **Augmented Dickey Fuller** test:

According to the paper titled *"Time-series forecasting of seasonal items sales using machine learning – A comparative analysis" by Yasaman Ensafi, Saman Hassanzadeh Amin, Guoqing Zhang, and Bharat Shah, as published in the International Journal of Information Management Data Insights in 2022,* the objective of Augmented Dickey Fuller (ADF) test is to decide that the time-series is stationary or non-stationary by checking the presence of unit root in a time-series. This method observes the difference between the value level and the mean. If it was higher than

the mean, the next movement will be downward. Furthermore, if it was lower than the mean, the movement will be upward.

$$\Delta y(t) = \lambda y(t-1) + \mu + \beta t + \alpha_1 \Delta y(t-1) \pm \ldots + \alpha_k \Delta y(t-k) + \varepsilon_t$$

The equation above explains these value changes, where μ is a constant, β is the coefficient on a time trend, k is the lag order of the autoregressive process, and $\Delta y(t)$ can be defined as

$\Delta y(t) \equiv y(t) - y(t-1)$, $\Delta y(t-1) \equiv y(t-1) - y(t-2)$ (Corrius, 2018)

The null hypothesis states that the time-series in non-stationary ($\lambda = 0$). If this hypothesis is rejected, it shows that the next movement ($\Delta y(t)$) is not just a random value, and it depends on the current level $y(t-1)$. Thus, the time-series is stationary. In our case, the p-value is smaller than 0.05, and the time-series is stationary.

9. Pre-modeling

Before finally modeling the data, the data is split into train, validation, and test sets. It is a foundational practice in the development of forecasting models, including those used for time series analysis. This separation serves critical roles in model building, tuning, and evaluation, ensuring that the final model is both accurate and generalizable to new, unseen data.

- Training Set: used to train the model. It allows the model to learn the underlying patterns, trends, and relationships within the data. For time series, this usually comprises the earliest data.
- Validation Set: used to tune the model's hyperparameters and make decisions about model architecture without overfitting to the test data. It acts as a proxy for the test set during the development phase, providing feedback on how well the model generalizes to data it hasn't been trained on.
- Test Set: untouched part of the data reserved for evaluating the model's performance. It provides an unbiased assessment of how well the model will perform on future, unseen data. This step is crucial for understanding the model's predictive power and its potential effectiveness in real-world scenarios.

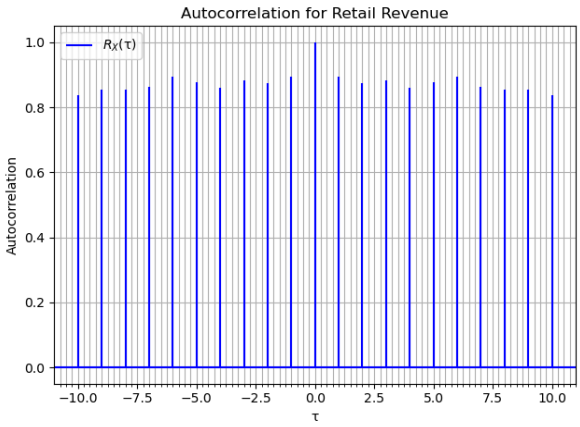
After establishing the train-validation-test split, further dividing the test set into four equal parts offers additional advantages for model evaluation and refinement, particularly in the context of time series forecasting:

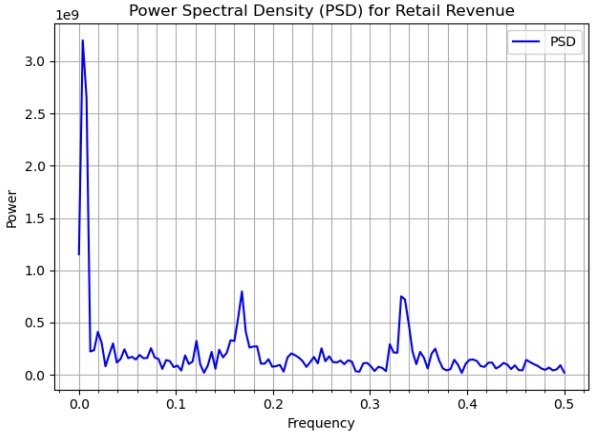
- Detailed Model Assessment: Segmenting the test set allows for a more granular analysis of the model's performance over different time intervals.
- Robustness to Time-Varying Patterns: Testing across multiple distinct periods helps ensure that the model's performance is stable across different conditions and not overly reliant on specific features or patterns present in a single time frame. This helps validate the model's robustness and reliability.
- Understanding Seasonality and Cyclic Behaviors: Time series data often contain seasonal and cyclic behaviors that may not be fully captured in a single test segment. By analyzing performance across several periods, the model's ability to generalize across different seasonal and cyclic patterns can be better assessed.
- Confidence in Deployment: Demonstrating consistent model performance across multiple segments of the test set builds confidence in the model's reliability and effectiveness in practical applications. This is vital for stakeholders who rely on model forecasts for making informed decisions.

In the pre-modelling phase, we analyze the online retail dataset through time series plots, autocorrelation, and Power Spectral Density (PSD) analysis to understand the underlying patterns and characteristics.

- Autocorrelation & PSD:

We aggregated the data by date which gave 603 unique rows of day-wise revenue. The resulting time series plot provides a visual representation of the retail revenue over time. Autocorrelation analysis was conducted to identify any correlation between the revenue values at different lags. The plot shows the autocorrelation values for different lag intervals. Additionally, the Power Spectral Density (PSD) analysis provides insights into the frequency components present in the data.

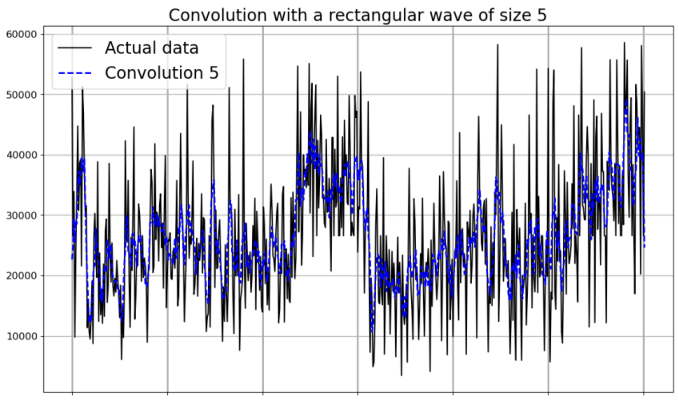


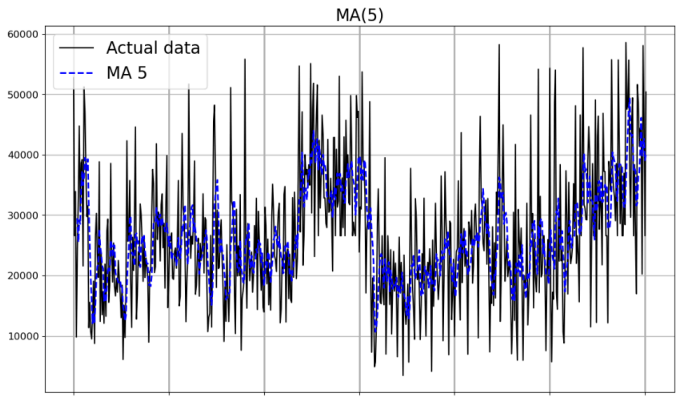


The Autocorrelation graph shows a cyclical pattern in retail revenue, suggesting a seasonal component. The Power Spectral Density (PSD) graph indicates a dominant frequency, which corresponds to the length of this seasonal cycle. These insights can inform your sales forecasting model.

- Convolution (5), MA(5), Butterworth Filter in Low Pass:

Similar convolution, MA, and Butterworth filter analyses were conducted on the dataset aggregate by date to explore the smoothing and filtering effects at the daily level. These techniques help in preparing the data for subsequent modeling steps, ensuring that the underlying patterns are more discernible and noise is effectively reduced.





The first graph, “Autocorrelation for Retail Revenue”, shows the correlation of the retail revenue with itself at different time lags. The second graph, “Power Spectral Density (PSD) for Retail Revenue”, shows the power of each frequency component in the retail revenue data. Both graphs suggest a strong periodic component in the retail revenue, which could be leveraged for more accurate forecasting.

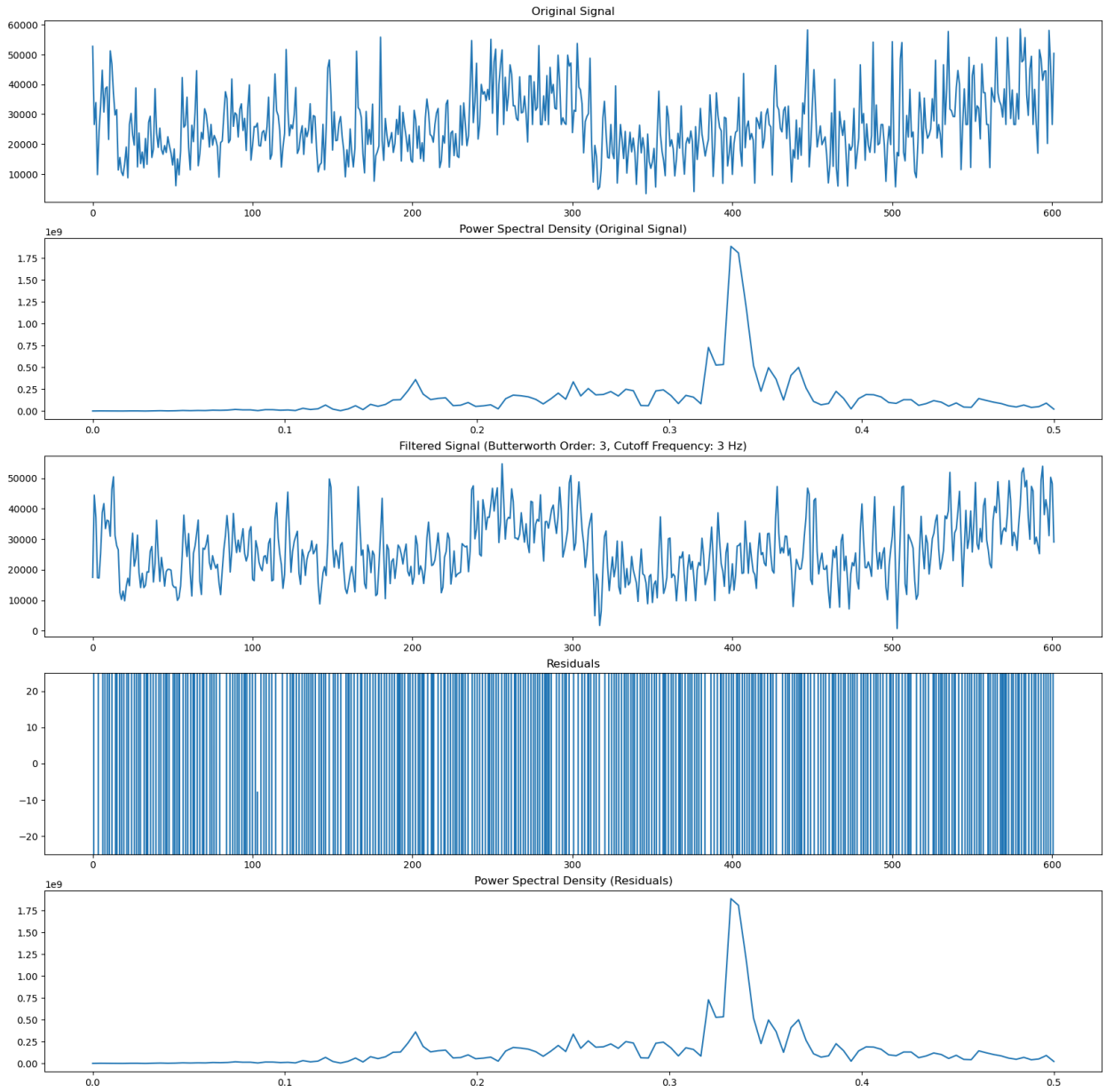


Fig. PSD and Butterworth Filter in Low Pass for data by-date

The Power Spectral Density (PSD) graph here shows the power distribution of the sales data over frequency, which helps in identifying patterns and noise. The Butterworth filter, a type of low-pass filter, is applied to this data to eliminate high-frequency noise. This results in a smoother signal, as seen in the “Filtered Signal” graph.

The “Original Signal” graph shows a noisy waveform. After applying a filter, as shown in the “Filtered Signal” graph, the noise is significantly reduced. The “Power Spectral Density” graphs before and after filtering show a clear peak, indicating a dominant frequency component in the signal.

This process is crucial for pre-modeling in sales forecasting as it reduces the variability in the data, making the underlying pattern more visible and thus, the forecasting model more accurate. The goal of this process is to extract meaningful insights from the data that can drive decision-making in the sales forecasting project.

10. Modeling

In the modeling phase, we explore the following models -

- Holt-Winters Exponential Smoothing
- ARIMA
- SARIMAX
- Random Forest Regressor
- Facebook Prophet,

on the online retail dataset to forecast future revenue trends. Below, we outline our approach, the choice of models, and their performance evaluation.

10.1 Holt-Winters Exponential Smoothing:

- **Model Description:**
  Holt-Winters Exponential Smoothing is a robust forecasting method that extends Exponential Smoothing to capture level, trend, and seasonality in time series data. It is characterized by the application of exponential weights, which decline exponentially over time. Unlike simple moving averages, exponential smoothing assigns more weight to recent observations while not discarding older observations entirely, which makes it adept at forecasting data with trends and seasonality.

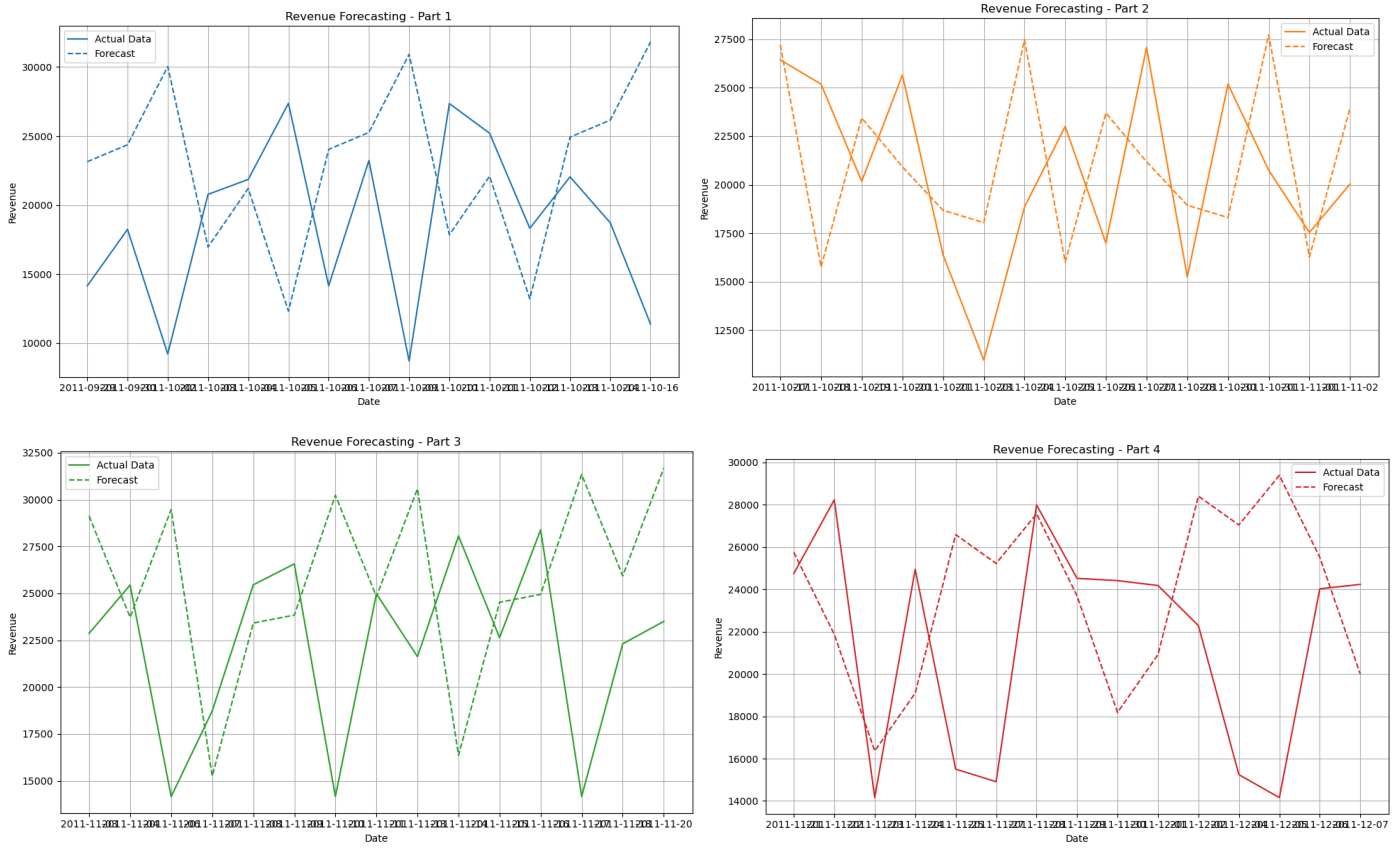
  - **Exponential Smoothing of the Level**: $\hat{L}_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(\hat{L}_{t-1} + \hat{T}_{t-1})$, where $\hat{L}_t$ is the estimated level at time t, y_t is the actual value at time t, S_{t-s} is the seasonal component at time t minus the length of the season, and $\hat{T}_{t-1}$ is the estimated trend at time t-1.
  - **Exponential Smoothing of the Trend**: $\hat{T}_t = \beta(\hat{L}_t - \hat{L}_{t-1}) + (1-\beta)\hat{T}_{t-1}$, where $\hat{T}_t$ is the estimated trend at time t, and β is the smoothing parameter for the trend.
  - **Exponential Smoothing of the Seasonal Component**: $\hat{S}_t = \gamma(y_t - \hat{L}_t) + (1-\gamma)S_{t-s}$, where $\hat{S}_t$ is the estimated seasonal component, and γ is the smoothing parameter for the seasonality.

  The forecast is then computed using these components depending on whether an additive or multiplicative model is used:

  - **Additive Forecast**: $\hat{y}_{t+m} = \hat{L}_t + m\hat{T}_t + \hat{S}_{t-s+1+(m-1)mods}$
  - **Multiplicative Forecast**: $\hat{y}_{t+m} = (\hat{L}_t + m\hat{T}_t) \times \hat{S}_{t-s+1+(m-1)mods}$

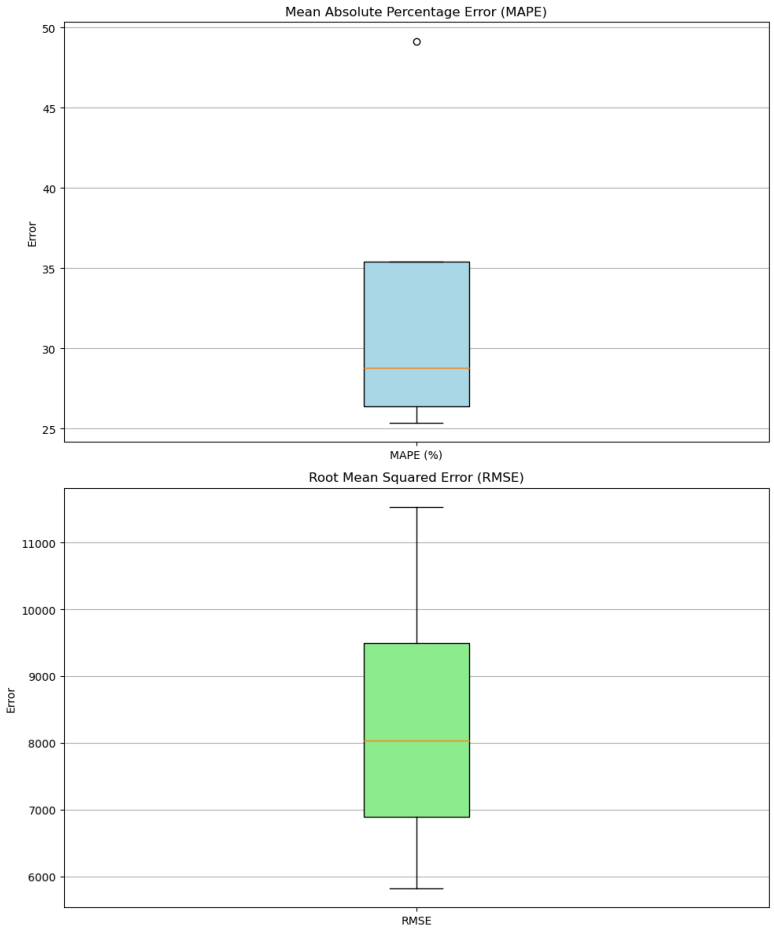
  Here, *m* is the forecasting horizon, *s* is the length of the seasonal cycle, and mod represents the modulo operation.
- **Approach:**
  - Utilized Holt-Winters Exponential Smoothing model for time series forecasting, incorporating additive trend and seasonal components with a seasonal period of 6 (representing months).
  - Fitted separate models to each part of the test data to capture seasonal variations and trends effectively.

- **Results:**
    - Mean Absolute Percentage Error (MAPE) ranged from 25.38% to 49.15% across different parts of the test data.
    - Root Mean Squared Error (RMSE) ranged from 5818.15 to 11530.80, indicating variations in forecast accuracy.
    - The model exhibited relatively high MAPE values, suggesting notable errors in predicting revenue trends, especially for certain parts of the test data.
    - Visual examination of forecasted values against actual data revealed discrepancies between forecasts and actuals, indicating potential limitations in capturing the underlying patterns in the data.



- **Performance Evaluation:**
    - Holt-Winters Exponential Smoothing model showed mixed performance in forecasting retail sales data, with varying levels of accuracy across different parts of the test data.
    - The relatively high MAPE values suggest potential challenges in capturing seasonality and trends effectively, especially for longer forecast horizons or periods with significant fluctuations.

- Further exploration into model parameters, such as adjusting the seasonal period or incorporating additional exogenous factors, could potentially enhance forecast accuracy and address the observed discrepancies.
- Despite the limitations, Holt-Winters Exponential Smoothing offers a flexible and interpretable approach for time series forecasting, providing insights into the underlying patterns and dynamics of retail sales data.



10.2 ARIMA (AutoRegressive Integrated Moving Average) Model:

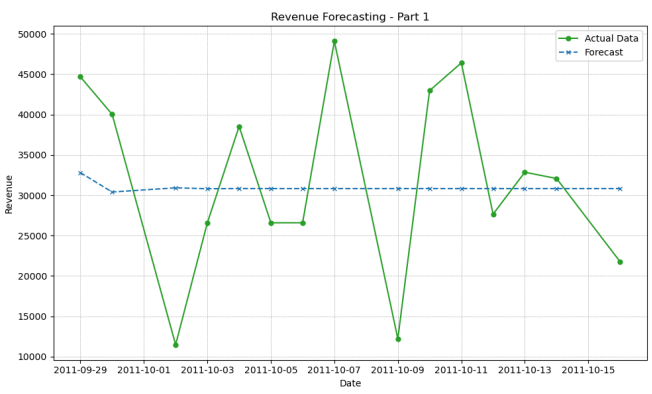
- **Model Description:**
  The ARIMA model synthesizes the complexity of real-world time-series data into a structured statistical framework. It leverages the auto-regressive (AR) method to utilize past data points—or 'lags'—as predictors for future values. The 'I' in ARIMA stands for 'integrated,' a vital feature that involves differencing the data series to achieve stationarity—a state where the mean, variance, and covariance are constant over time. This transformation is crucial as ARIMA models require stationary data to function correctly.

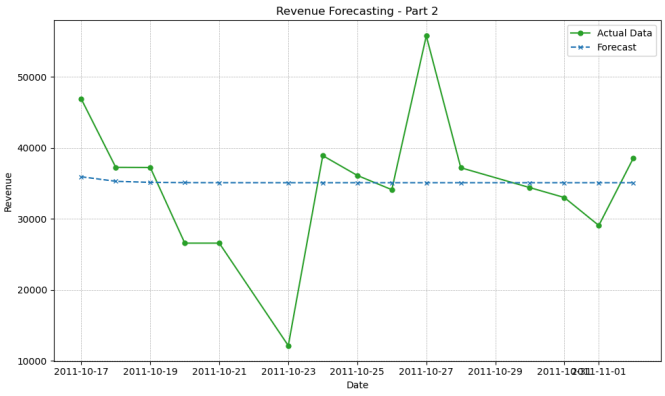
The 'MA' component, or moving average, uses past forecast errors to enhance predictions, contrasting the AR method's reliance on actual past values. Together, these elements form the ARIMA(p, d, q) model, where 'p' denotes the number of lags, 'd' indicates the degree of differencing needed for stationarity, and 'q' represents the number of past errors factored into the model.

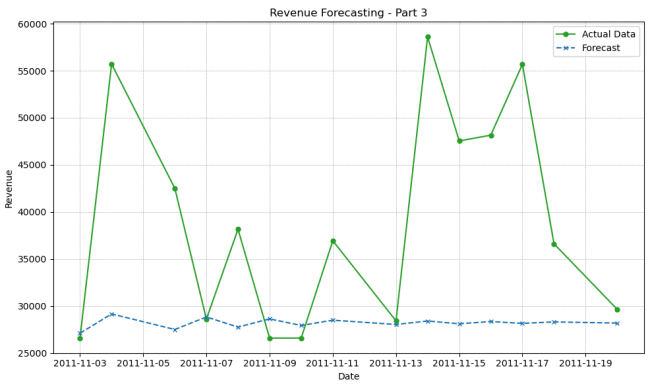
The strength of the ARIMA model lies in its integration process, which subtracts a data point from its predecessor, transforming non-stationary time series into a form that is suitable for robust analysis and forecasting. This approach allows the model to smooth out random fluctuations and pinpoint the underlying patterns in the time series data.

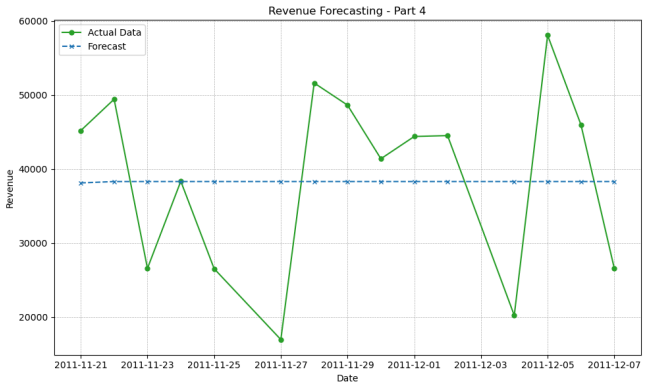
- **Approach:**
  - Utilized Autoregressive Integrated Moving Average (ARIMA) model for time series forecasting.
  - Conducted separate modeling and forecasting for each part of the test data.
  - Configured ARIMA model parameters with an order of (1,1,1) based on initial analysis.

- **Results:**
  - Mean Absolute Percentage Error (MAPE) ranged from 0.24% to 0.40% across different parts of the test data.
  - Visual inspection of forecasted values against actual data revealed generally accurate predictions, with forecasted values closely following the trend of actual sales.

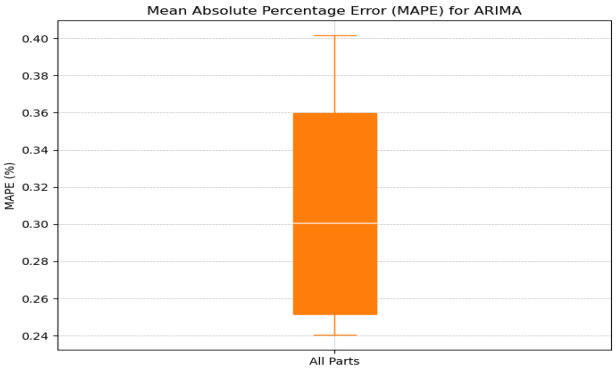








- **Performance Evaluation:**
  - ARIMA model demonstrated strong performance in forecasting retail sales data, with low MAPE values indicating high accuracy.
  - The visually appealing plots depicted the forecasted revenue alongside the actual data, providing a clear representation of model performance.
  - Despite the overall success, slight variations in MAPE across different parts of the test data suggest potential differences in underlying patterns or complexities that could be further explored for optimization.



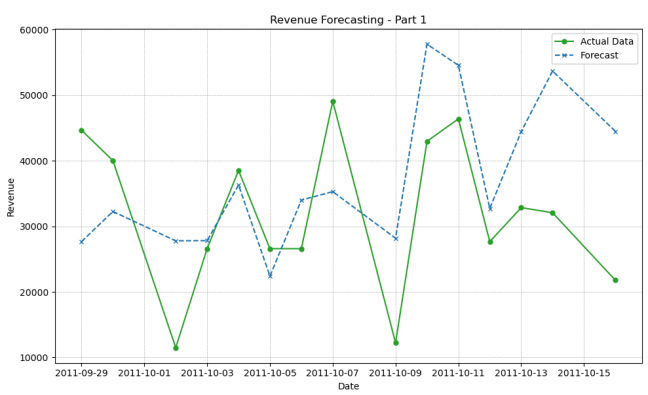
10.3 SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) Model:

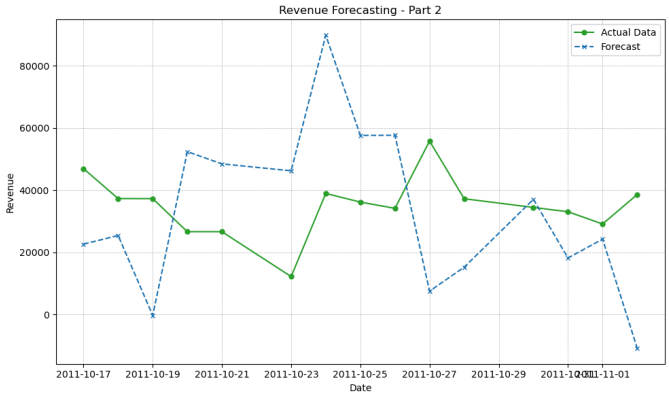
- **Model Description:**
  Expanding on the ARIMA model to account for seasonality leads to the SARIMA (Seasonal ARIMA) model. We use SARIMAX where 'X' represents exogenous variables that can be included in the model. This model incorporates both non-seasonal and seasonal elements of the series, and it's denoted by (p, d, q) (P, D, Q)m, where 'P', 'D', and 'Q' are the seasonal parts of the AR, differencing, and MA components, respectively, and 'm' denotes the number of periods in each season.

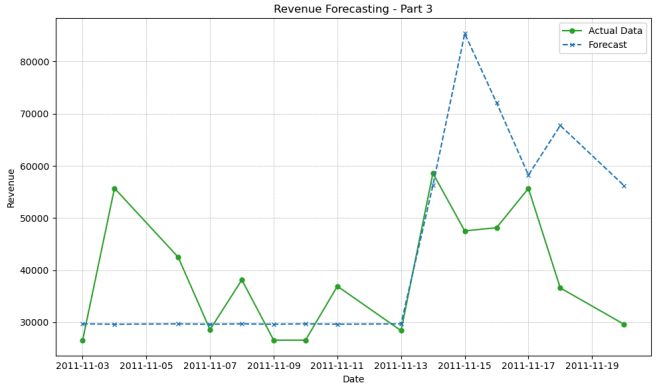
- **Approach:**

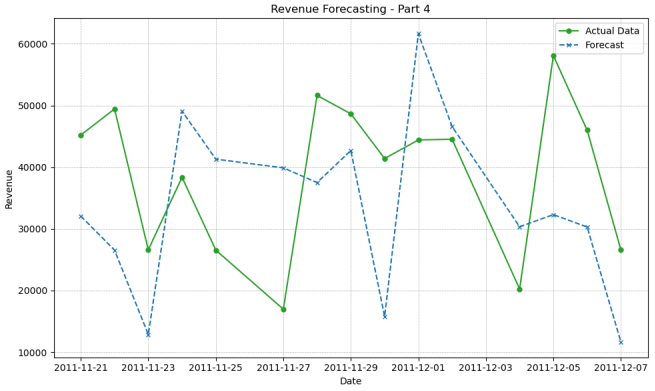
- Employed Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) model for time series forecasting, accommodating seasonality in the data.
- Configured SARIMAX model parameters with an order of (1,1,1) and seasonal order of (1,1,1,12) based on initial analysis.

- **Results:**
  - Mean Absolute Percentage Error (MAPE) ranged from 0.32% to 0.83% across different parts of the test data.
  - The MAPE values indicate relatively higher errors compared to the ARIMA model, suggesting potential challenges in capturing the seasonality or additional complexity introduced by exogenous factors.
  - Visual examination of forecasted values against actual data revealed varying degrees of accuracy, with some parts exhibiting closer alignment between forecasts and actuals than others.



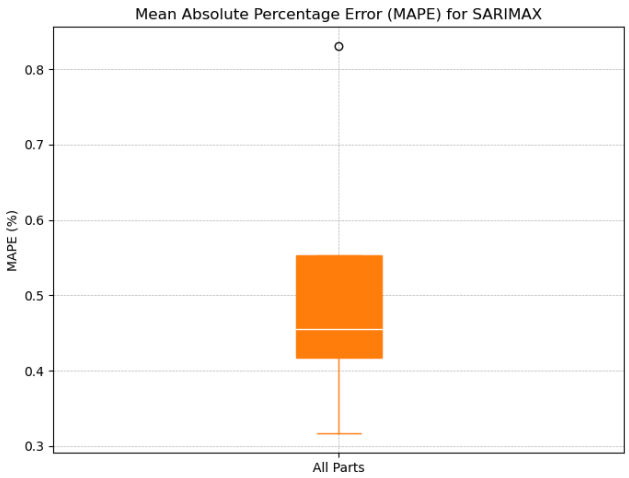






- **Performance Evaluation:**
  - While SARIMAX model demonstrated reasonable forecasting performance, the slightly higher MAPE values compared to ARIMA model suggest potential limitations in capturing the underlying patterns in the data effectively.

- Further investigation into the sources of variability and potential exogenous factors impacting sales data could help refine the model and improve forecasting accuracy.
- Despite the challenges, SARIMAX model offers a valuable framework for incorporating seasonality and external variables into the forecasting process, providing insights into the dynamic nature of retail sales data.

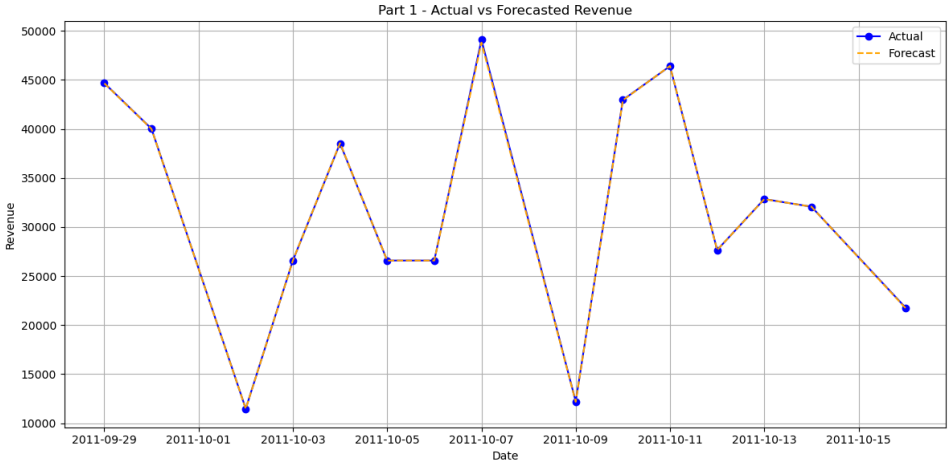


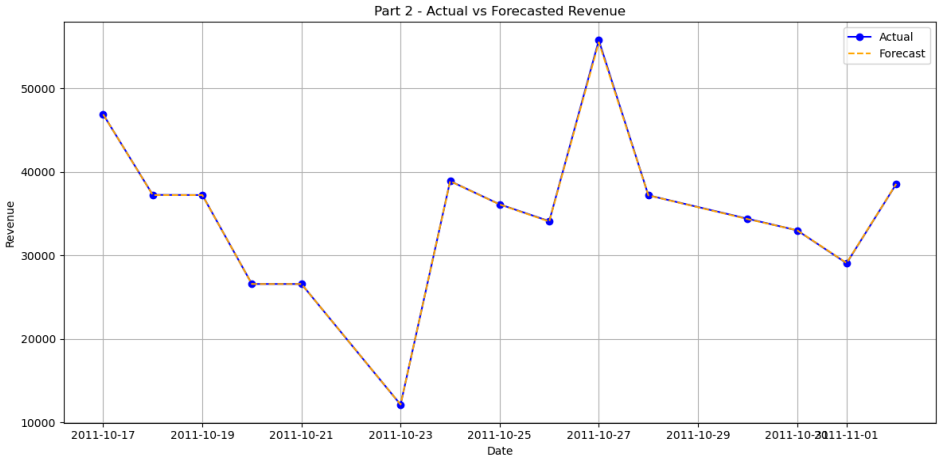
10.4 Random Forest Regressor:

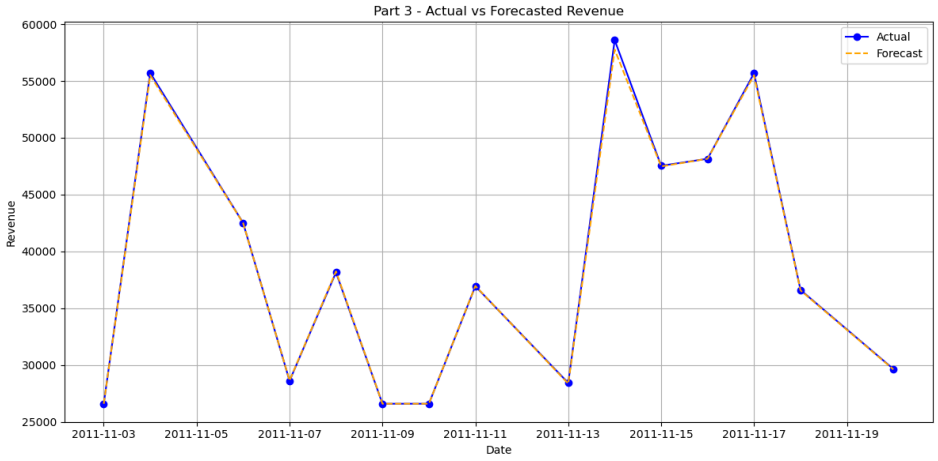
- **Model Description:**
  In time series analysis, the Random Forest Regressor is employed to forecast future values based on known historical data. This model is particularly effective due to its ability to handle the complex nonlinear relationships that are common in time-dependent data.
  When configuring a Random Forest for time series forecasting, it is essential to include lagged variables as predictors. These lagged variables are essentially previous time points in the series, which the model uses to learn how past values influence future ones. Past observations are used as input features, allowing the model to capture temporal dependencies similar to autoregressive models.
  Unlike ARIMA, which requires stationarity, Random Forest can inherently account for complex seasonal patterns and trends through feature engineering. The ensemble approach of Random Forest, where predictions are averaged over many decision trees, helps prevent overfitting, a common pitfall in time series forecasting.
  Random Forest provides insights into which features (e.g., specific lags, trend components, seasonal dummies) are most influential in predicting the target variable.
- **Approach:**
  - Utilized Random Forest Regressor model for time series forecasting, leveraging ensemble learning techniques to capture non-linear relationships and interactions in the data.

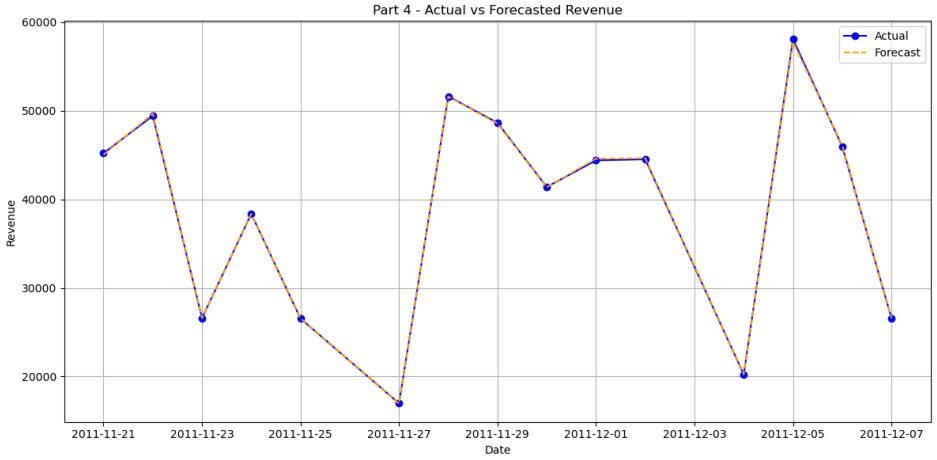
- Trained the model on combined training and validation sets, consisting of historical revenue data.

- **Results:**
  - Mean Absolute Percentage Error (MAPE) ranged from 0.08% to 0.55% across different parts of the test data.
  - Root Mean Squared Error (RMSE) ranged from 40.74 to 864.68, indicating variations in the forecast accuracy.
  - The model demonstrated relatively low MAPE values, suggesting high accuracy in predicting revenue trends for most parts of the test data.
  - However, there were notable differences in RMSE values across different parts, indicating potential challenges in capturing variations or outliers in the data.

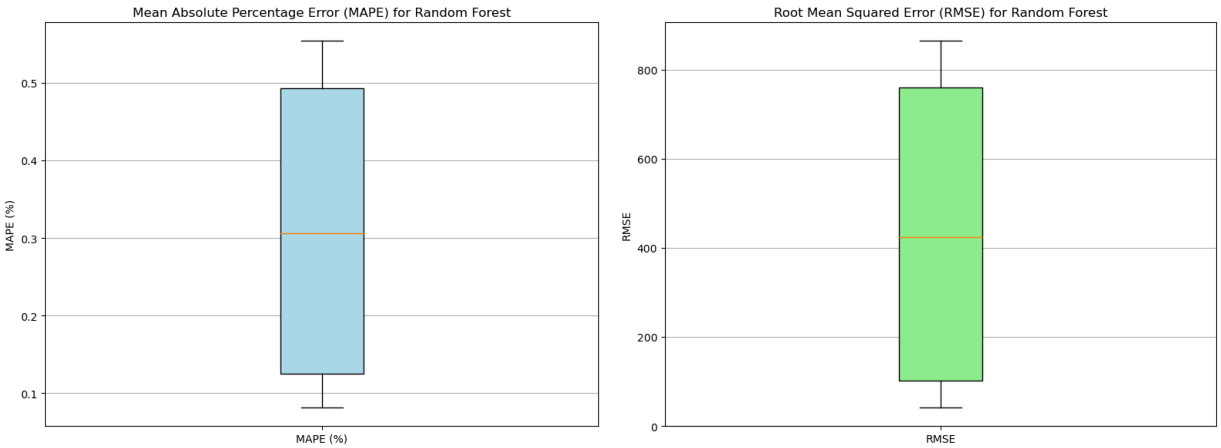








- **Performance Evaluation:**
  - Random Forest Regressor exhibited promising performance in forecasting retail sales data, with generally low MAPE values indicating high accuracy.
  - The model's ability to capture complex patterns and non-linear relationships in the data contributed to its effectiveness in generating accurate forecasts.
  - Despite the overall success, the higher RMSE values in certain parts suggest potential limitations in capturing extreme variations or outliers, warranting further investigation and refinement of the model parameters or feature engineering techniques.
  - Random Forest Regressor offers a flexible and robust framework for time series forecasting, providing insights into the dynamic nature of retail sales data and enabling informed decision-making for business planning and resource allocation.



10.5 Facebook Prophet:

- **Model Description:**
  Facebook Prophet is a forecasting tool designed for handling the intricacies of time series data with ease, particularly those with strong seasonal patterns.

  a) Trend Modeling: Prophet automatically fits non-linear trends with daily, weekly, and yearly seasonality, plus holiday effects, by using an additive model where non-linear trends are fit with yearly and weekly seasonality, plus holidays. It works well with time series that have strong seasonal effects and several seasons of historical data.
  b) Seasonality: It provides an intuitive and flexible approach to modeling seasonality by decomposing the time series into trend, seasonal, and holiday components. Prophet can accommodate both regular seasonality (such as weekly and yearly cycles) and irregular events (like holidays).
  c) Handling Missing Data: Prophet can handle missing data and trend changes well, and typically does not require any pre-processing to fill missing dates or values, making it convenient for real-world data which often comes with missing points.
  d) Robustness to Outliers: The tool is robust to outliers and shifts in the trend, and it typically manages them without manual intervention.
  e) Ease of Use: It provides sensible defaults for its hyperparameters, which the user can tune, but out-of-the-box settings often provide a solid starting point for many time series datasets.

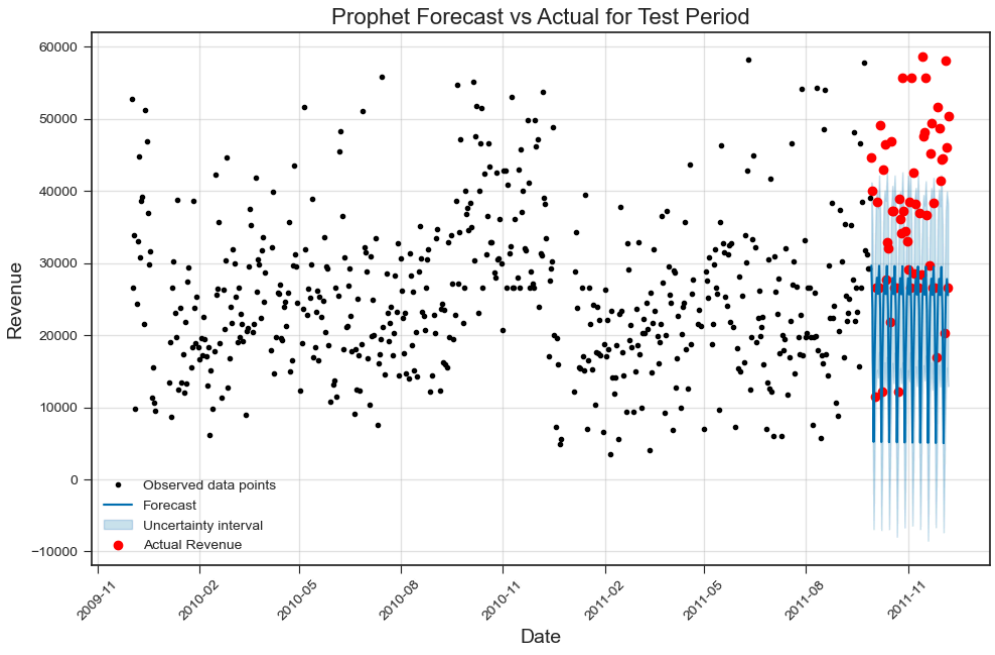
  The predictive model for Prophet is generally expressed as:

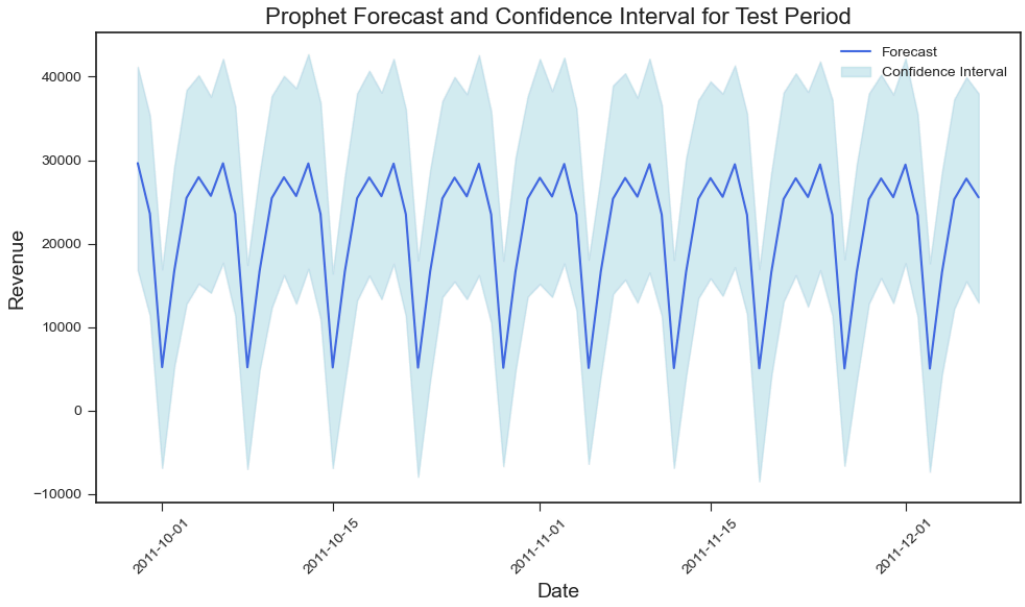
$$y(t) = g(t) + s(t) + h(t) + e(t)$$

Here,

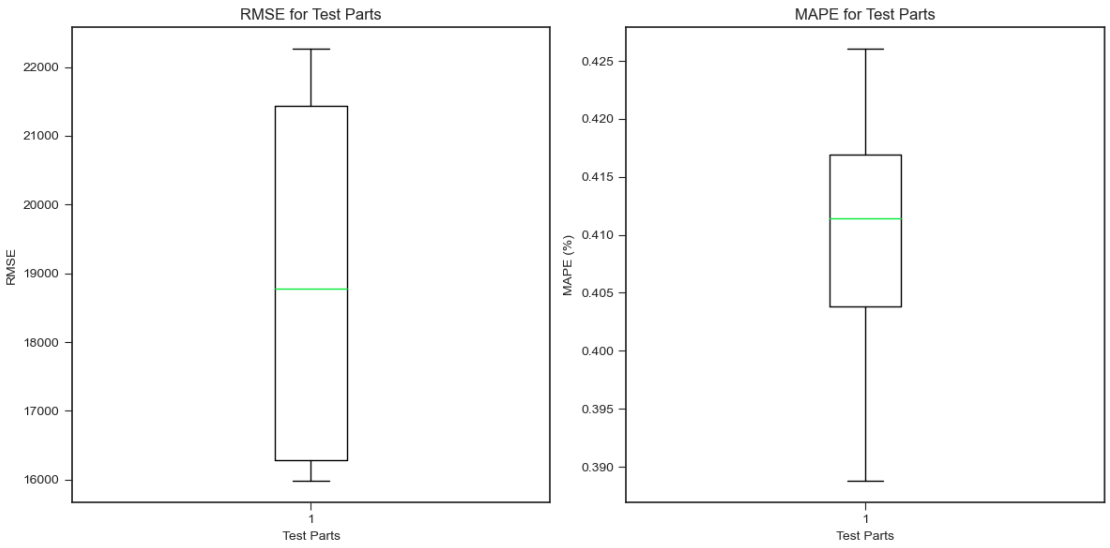
a) $g(t)$ represents the trend function which models non-periodic changes
b) $s(t)$ represents periodic changes (e.g., weekly, annual seasonality)
c) $h(t)$ represents the effects of holidays which occur on potentially irregular schedules over a year or several days.
d) $e(t)$ represents the error term.

- **Approach:**
  - Employed Facebook Prophet model for time series forecasting, utilizing its capabilities in handling seasonality, holidays, and trend changes automatically.
  - Fitted the model on combined training and validation data to capture underlying patterns and dynamics effectively.
- **Results:**
  - Mean Absolute Percentage Error (MAPE) ranged from 0.39% to 0.43% across different parts of the test data.
  - The model demonstrated consistently low MAPE values, indicating high accuracy in predicting revenue trends for all parts of the test data.
  - Visual examination of forecasted values against actual data revealed close alignment between forecasts and actuals, suggesting the model's effectiveness in capturing underlying patterns and dynamics accurately.





- **Performance Evaluation:**
  - Facebook Prophet exhibited strong performance in forecasting retail sales data, with consistently low MAPE values across different parts of the test data.
  - The model's ability to automatically handle seasonality, holidays, and trend changes contributed to its effectiveness in generating accurate forecasts.
  - The observed consistency in forecast accuracy underscores the robustness and reliability of Facebook Prophet in capturing the dynamic nature of retail sales data and providing valuable insights for decision-making and planning purposes.



11. Model Selection

Performance Measures:

Time-series forecasting performance measures indicate the capability of the models. We have assessed the model performance by two commonly used measures. **Root Mean Squared Error (RMSE)** is the standard deviation of the prediction errors, and it is scale-dependent. **Mean Absolute Percentage Error (MAPE)** is obtained as the mean absolute percentage error function for the prediction and the eventual outcomes. This error measure expresses error as a percentage. The formulas are as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2$$

$$RMSE = \sqrt{MSE}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(Y_t - F_t)}{Y_t}\right| * 100$$

In scenarios involving time-series forecasting of sales, or any metric where proportional errors are more intuitive and impactful for decision-making, MAPE can offer several advantages over RMSE, such as:

- As MAPE is a relative measure, it allows for the comparison of forecasts across different scales or units of measurement. This is particularly useful in diverse product portfolios where sales volumes can vary significantly across products
- MAPE expresses errors as a percentage, making it easier for business stakeholders to understand the magnitude of forecasting errors relative to actual values. This percentage error provides a more intuitive sense of accuracy, especially in business contexts where stakeholders are accustomed to thinking in terms of percentages.
- In many business scenarios, especially in sales and inventory management, the relative size of the error (i.e., how big an error is relative to the actual value) is more critical than the absolute size of the error. MAPE directly aligns with such objectives by quantifying the relative accuracy of forecasts.

Model Selection Narrative:

- Holt-Winters Exponential Smoothing:
    - Started with Holt-Winters Exponential Smoothing due to its simplicity and ability to capture seasonality and trends in time series data
    - Chose this model as an initial baseline to establish a benchmark for forecasting performance.

- ARIMA:
    - Transitioned to ARIMA (Autoregressive Integrated Moving Average) model to explore more sophisticated time series modeling techniques.
    - ARIMA is a widely-used approach known for its flexibility in capturing various patterns and dynamics in the data.

- SARIMAX:
    - Moved to SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) to incorporate additional exogenous factors that could influence retail sales data.
    - SARIMAX model offered the advantage of integrating external variables into the forecasting process, potentially enhancing predictive accuracy.

- Random Forest Regressor:
    - Experimented with Random Forest Regressor to explore the capabilities of ensemble learning techniques in time series forecasting.
    - Random Forest offered the advantage of capturing complex relationships and interactions in the data, potentially improving forecast accuracy, especially in the presence of non-linear patterns.

- Facebook Prophet:
    - Adopted Facebook Prophet as a final modeling approach to leverage its automated handling of seasonality, holidays, and trend changes.
    - Facebook Prophet's user-friendly interface and automatic feature selection made it an attractive choice for simplifying the modeling process while maintaining high forecasting accuracy.

- Overall Rationale:

  The progression from simpler models like Holt-Winters Exponential Smoothing to more complex ones like Facebook Prophet was driven by the need to explore and leverage different modeling techniques to capture the inherent complexities and dynamics of retail sales data.

  Each model was selected based on its unique strengths and capabilities, with the goal of continuously improving forecasting accuracy and providing valuable insights for business planning and decision-making.

Comparative Analysis of Forecasting Models: MAPE and RMSE Metrics

Model	RMSE (average)	MAPE (average)
Exponential Smoothing	18,216.15	41.69%
ARIMA	-	0.3125%
SARIMAX	-	0.515%
Random Forest Regressor	438.48	0.31%
Facebook Prophet	18,947.16	0.41%

Based on the provided MAPE results and the features of all the models, **Facebook Prophet emerges as the preferred choice** over Random Forest Regressor and others. Here's the reasoning for this choice:

- **Comparable MAPE:**
  - Facebook Prophet achieved a MAPE score of 0.41%, while the Random Forest Regressor scored slightly lower with a MAPE of 0.31%.
  - While Random Forest Regressor has a slightly lower MAPE score, the difference between the two models is minimal, indicating that both models provide accurate forecasts relative to the scale of the data.

- **Automatic Handling of Seasonality and Trends:**
  - Facebook Prophet offers automated handling of seasonality, holidays, and trend changes, reducing the need for manual feature engineering and parameter tuning.
  - This automatic feature selection and adjustment capability streamline the modeling process and reduce the risk of overfitting, making Facebook Prophet a more user-friendly and efficient choice for forecasting tasks.

- **Interpretability and Ease of Use:**
  - Facebook Prophet provides interpretable forecasts and easily interpretable trend and seasonality components, allowing stakeholders to understand the underlying patterns driving the forecasts.
  - The user-friendly interface and intuitive parameter settings of Facebook Prophet make it accessible to users with varying levels of expertise, enabling seamless integration into business planning and decision-making processes.

  In summary, while Random Forest Regressor may have achieved a slightly lower MAPE score, the minimal difference in performance, coupled with the automatic handling of seasonality and trends and the interpretability of forecasts, makes Facebook Prophet the preferred choice for forecasting retail sales data.

12. Future Scope

The project's journey in developing a predictive model for forecasting revenue in an online retail setting has carved out a robust framework for numerous future explorations and enhancements. The methodologies applied and the insights gathered pave the way for a diverse range of advanced analytical pursuits, with the potential to transcend the initial retail focus and find relevance across varied sectors.

Granular Forecasting:

- Product-Category Level Predictions: By delving into specific categories, we aim to uncover nuanced trends and performance metrics, facilitating strategies for inventory optimization, targeted marketing, and pricing adjustments tailored to product lines.
- Regional Sales Trends: We plan to broaden our analysis to dissect regional or geographical variations in sales patterns, identifying both underperforming segments and regions brimming with untapped potential. This will allow for the crafting of more localized and potent business strategies.

Data Enrichment and Model Retraining:

- Incorporating Additional Data Sources: Integrating external data such as economic indicators, social media sentiment, and weather data will significantly bolster the model's predictive accuracy by accounting for external factors affecting consumer behavior.
- Continuous Learning: A mechanism for ongoing model retraining with new sales data will ensure the model's relevance and adaptability to evolving market trends and consumer patterns.

Advanced Modeling Techniques:

- Deep Learning Models: The application of Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks will allow us to model complex, non-linear relationships and long-term dependencies in time-series data.
- Hybrid and Ensemble Models: By combining traditional statistical methods with advanced machine learning techniques, and aggregating predictions from multiple models, we aim to enhance forecast accuracy and reliability.

Cross-Sector Application:

- Adaptation to Different Retail Sectors: The versatile nature of our modeling framework will enable us to cater to the unique dynamics of various retail sectors.
- Transferability to Other Industries: The foundational principles of our forecasting model are poised for adaptation to sectors beyond retail, such as hospitality, manufacturing, or

services, where forecasting demand, revenue, or other key performance metrics is vital for strategic planning.

Technological and Data Infrastructure Enhancements:

- Real-time Analytics: Developing real-time data analysis capabilities will empower businesses to react swiftly to market changes, optimizing operational decisions and strategic initiatives.
- Big Data Technologies: Utilizing big data technologies and cloud computing will facilitate the handling of vast datasets more efficiently, enabling more complex analyses and the integration of diverse data sources.

13. Team Information

Group Name: TIS

Members:

1. Ishita Pundir (UNI: ip2441)
2. Saum Kothari (UNI: sbk2171)
3. Tushar Bura (UNI: tb3077)