Forecasting Project
Process Documentation
V1.0

Team Name: Data Wizards

Shenova Davis (ssd2184)
Shachi Hardi (sh4615)
Ritayan Patra(rp3247)

Introduction

This project aims to analyze and forecast UK sales data using time series modeling techniques. The dataset is sourced from the "Online Retail II" dataset, which contains transactional records of purchases made by customers between 2009 and 2011. The objective of the project is to analyze sales trends, identify seasonal patterns, and develop an accurate forecasting model for future sales. This analysis provides insights for businesses in terms of inventory planning, promotional activities, and revenue forecasting.

Data Description

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.

The data has some the following the feature variables (or columns) =>

1. InvoiceNo: Invoice number, a 6-digit integral number uniquely assigned to each transaction. If it starts with 'C', it indicates a cancellation. (Nominal)
2. StockCode: Product (item) code, a 5-digit integral number uniquely assigned to each distinct product. (Nominal)
3. Description: Product (item) name. (Nominal)
4. Quantity: The quantities of each product (item) per transaction. (Numeric)
5. InvoiceDate: The date and time when a transaction was generated. (Numeric)
6. Price: Unit price, representing the product price per unit in sterling (£). (Numeric)
7. CustomerID: Customer number, a 5-digit integral number uniquely assigned to each customer. (Nominal)
8. Country: The name of the country where a customer resides. (Nominal)

Data Extraction

The data is downloaded from the website manually in the form of excel files. The excel file contains two separate sheets (Year 2009-2010) and (Year 2010-2011).

Each excel sheet has around 550K rows where each row represents a transaction. Each transaction either shows the buyer putting an order for a product or cancelling an order.

Data Processing

For processing the data, we are using pandas. Since we have two separate sheets in excel, we have read each sheet in separate variables and then contacted the two dataframes together.
Below is the code snippet for reading the data

Reading the data

```
[14] import openpyxl
     path = 'online_retail_II.xlsx'
     sheet1 = pd.read_excel(path, sheet_name=0, engine=None)
     sheet2 = pd.read_excel(path, sheet_name=1, engine=None)
     df = pd.concat([sheet1, sheet2], axis=0, ignore_index=True)
```

There are **238625** missing values in the Customer ID column.

There are multiple records in the data where the Customer ID is missing and the values of other prominent columns like Quantity and Price do not make any sense.

- There are **768** records where Price is **0.0** but the Quantity is less than **0**.
- There are **981** records while Price is equal to **0.0** but Quantity is greater than **0**.
- There are **5** records where the Price is negative.
- Finally, there are **749** records where Price is greater than **0.0** but Quantity is < 0

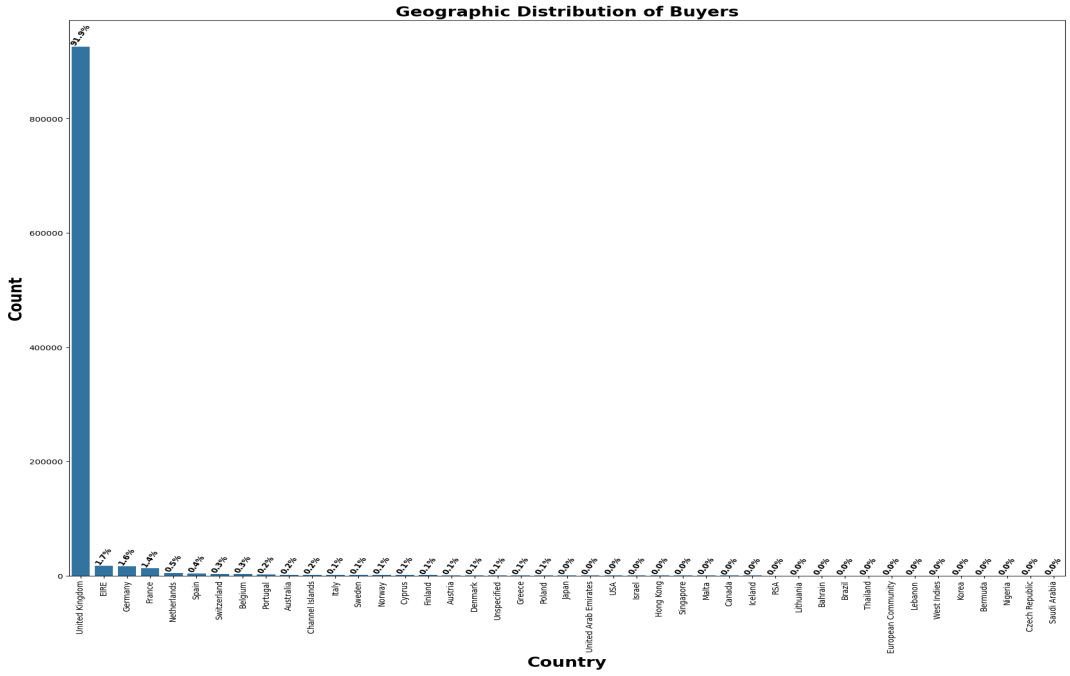
These ambiguous records which hold no logic are dropped.

There are 236122 records which are logical but the Customer ID is missing can be handled by imputing with some common value like Unknown Customer ID.

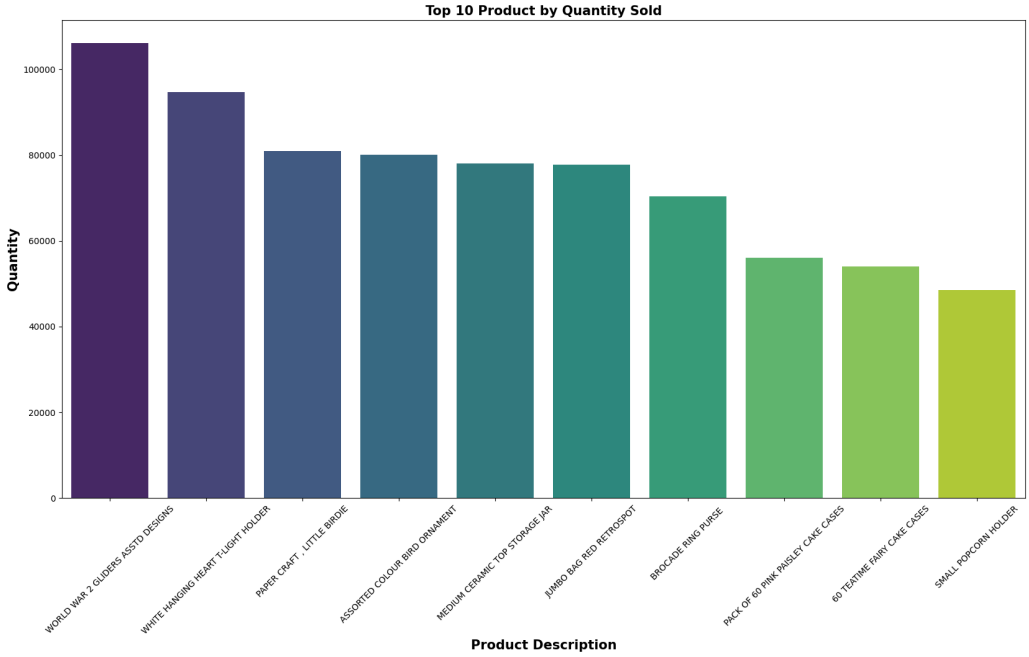
There are some Invoice IDs beginning with 'C' which represent cancelled orders. These records have negative quantity values. These records are also separated from the actual data.

A 'Sales' column is generated by multiplying 'Quantity' and 'Price', which will be used for model building and prediction.

Exploratory Data Analysis



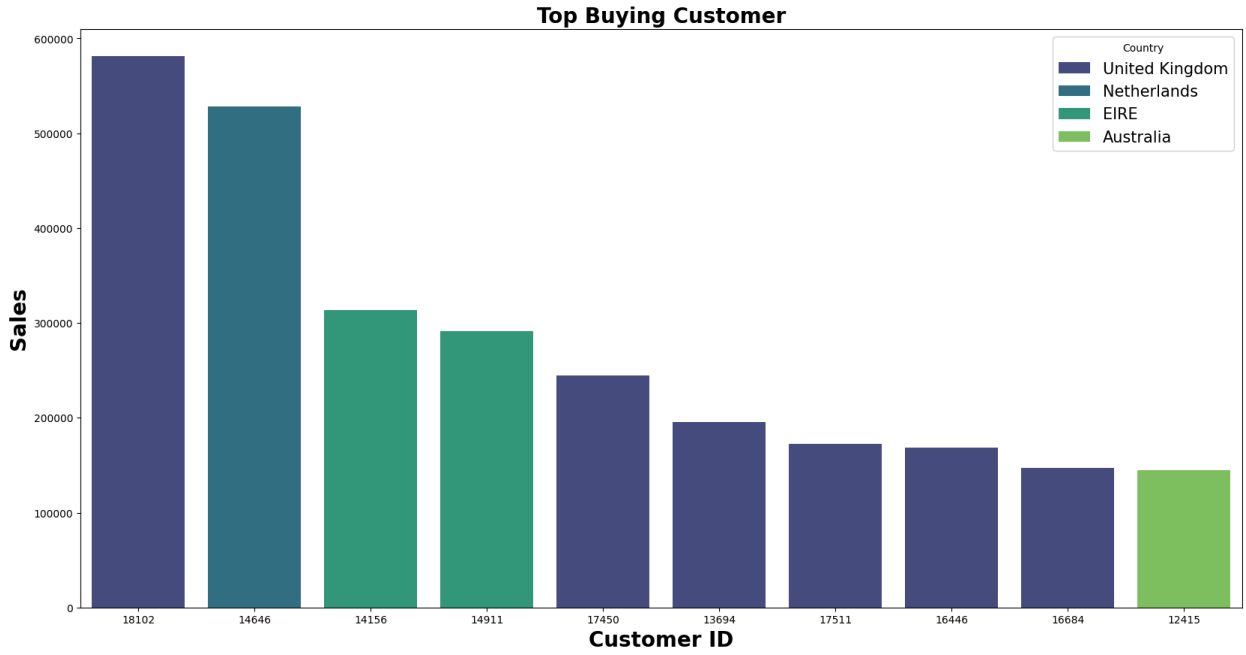
After cleaning the dataset, the total number of records reduces to around 1 million records. The United Kingdom accounts for a staggering 91.9% of all buyers, as indicated by the overwhelming height of the bar compared to other regions. This suggests that the majority of the buyers are concentrated in this region. The remaining regions, such as Germany, France, Netherlands, and others, contribute significantly less to the buyer count. Each of these regions accounts for less than 1% of the total buyers. A large number of regions (Switzerland, Belgium, Portugal, and others) contribute very small percentages (0.5% or lower). Many regions have a negligible or near-zero representation. This distribution suggests that the business or platform being analyzed is heavily focused on or successful in the UK market. There may be opportunities for growth in underrepresented regions if expansion is a goal.



The top selling products by the online retail store are:

- WORLD WAR 2 GLIDERS ASSTD DESIGNS
- WHITE HANGING HEART T-LIGHT HOLDER
- PAPER CRAFT , LITTLE BIRDIE
- ASSORTED COLOUR BIRD ORNAMENT
- MEDIUM CERAMIC TOP STORAGE JAR
- JUMBO BAG RED RETROSPOT
- BROCADE RING PURSE
- PACK OF 60 PINK PAISLEY CAKE CASES
- 60 TEATIME FAIRY CAKE CASES
- SMALL POPCORN HOLDER

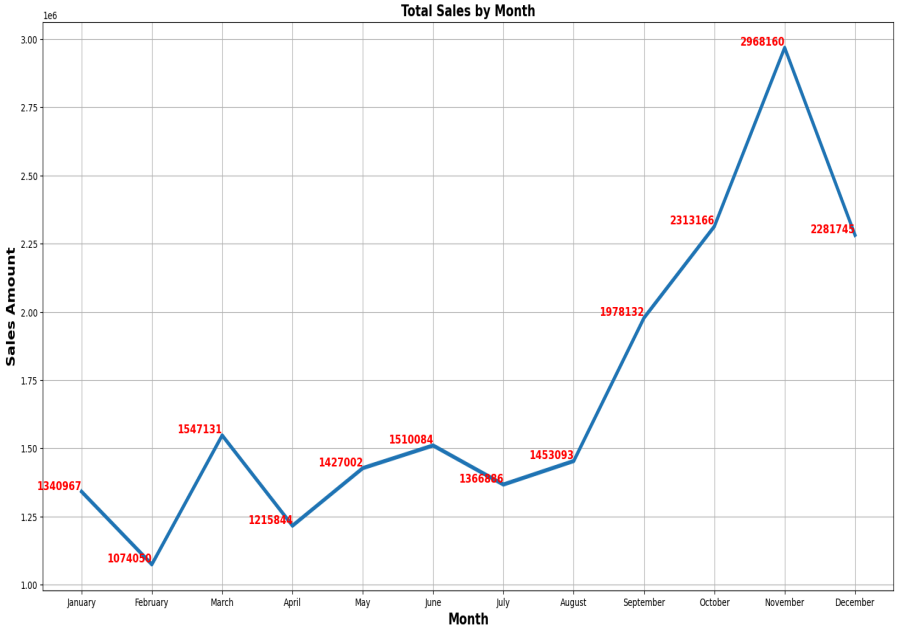
This chart gives an impression of what most of the customers prefer to buy from this online retail store. It tells the store to keep these items in stock for customer satisfaction.



The top 10 buying customers in terms of total sales are from countries United Kingdom, Netherlands, EIRE (Ireland) and Australia. The customer IDs of such customers are:

- 18102
- 14646
- 14156
- 14911
- 17459
- 13694
- 17511
- 16446
- 16684
- 12415

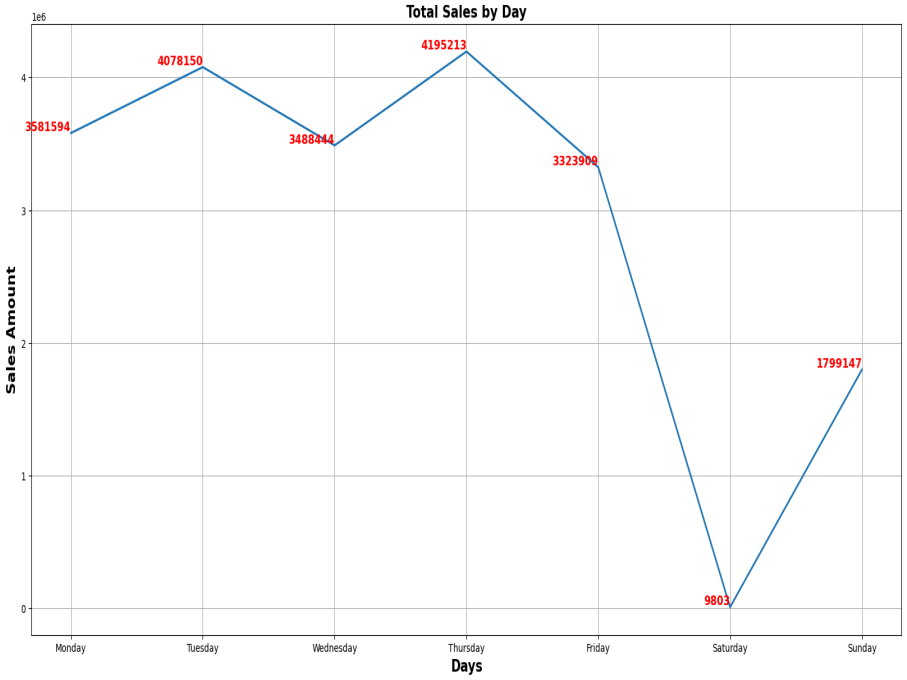
These customers have bought a significant amount of products from this online store. While most top buyers are from the United Kingdom, there is notable representation from other countries like the Netherlands, EIRE, and Australia. This chart can help businesses identify their most valuable customers and understand their geographic distribution. Such insights can guide marketing strategies, customer relationship management, and regional focus for future growth.



Sales tend to be relatively low at the beginning of the year, showing a slow and steady increase as the months progress toward mid-year. By July, sales have gained momentum, and from August onward, there is a noticeable upward trend. This growth continues consistently, with sales reaching their highest point in November.

The peak in November can be largely attributed to the festive season, which plays a significant role in driving consumer purchases. Various celebrations and events, such as Black Friday, Christmas, and New Year's, create a heightened sense of enthusiasm among shoppers. Retailers often introduce attractive discounts, promotional campaigns, and exclusive deals during this time, further fueling the increase in sales. Consumers are more inclined to spend on gifts, decorations, and other holiday-related items, contributing to the surge in market activity.

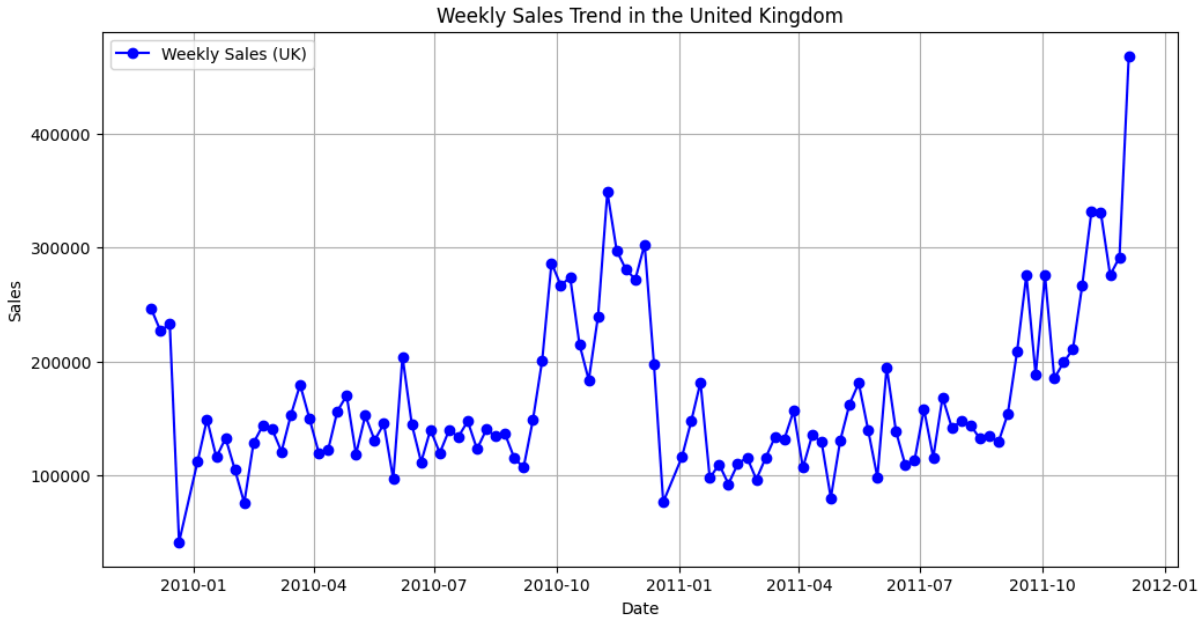
Overall, there is a clear seasonal trend in sales, with a gradual build-up throughout the year and a strong spike in the final quarter. This pattern underscores the impact of festive occasions on consumer behavior and highlights the importance of strategic planning for businesses to maximize revenue during peak shopping periods.



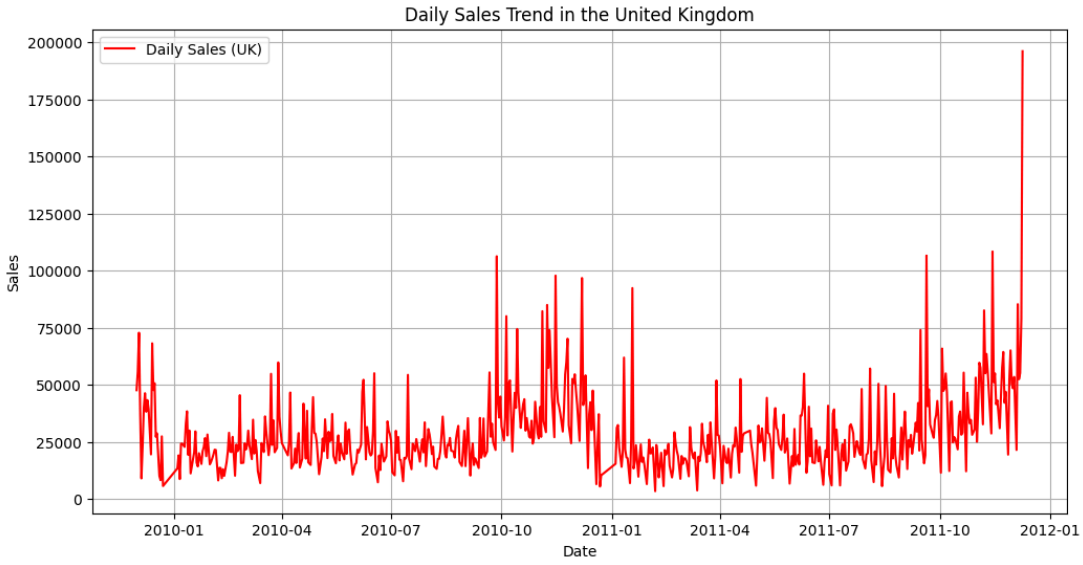
From the graph, it is evident that sales started at approximately $3.58 million on Monday. On Tuesday, there was an increase to $4.08 million, indicating growth in sales activity. However, Wednesday saw a slight dip to $3.49 million. Thursday marked the peak of the week, with sales reaching their highest point at $4.20 million. This suggests that Thursday was the most successful day for generating revenue.

Following this peak, sales began to decline on Friday, dropping to $3.32 million. Saturday experienced a dramatic drop in sales, plummeting to just $9,803, being the lowest figure of the week by a significant margin. This sharp decline could indicate an operational issue, reduced customer demand, or other anomalies affecting performance. On Sunday, sales rebounded significantly to approximately $1.80 million but remained below weekday levels.

In summary, Thursday was the best-performing day of the week, while Saturday saw an unusually low performance that requires further investigation. The weekend as a whole underperformed compared to weekdays, highlighting a potential area for improvement through targeted strategies like promotions or marketing campaigns to boost weekend sales.



The weekly sales trend in the UK shows noticeable ups and downs over time. Early on, there's a sharp peak followed by a quick drop, likely reflecting a seasonal surge in demand. After that, sales remained fairly steady with moderate fluctuations throughout most of 2010. However, in the last few months of the year, there's a clear spike, probably driven by holiday shopping. A similar pattern appeared in late 2011, with a gradual increase leading to another sharp rise at the end of the year. This suggests a strong seasonal influence, with sales peaking around major events like Black Friday and Christmas. Overall, the trend points to not just seasonal variations but also a steady increase in sales over time, hinting at potential business growth or rising customer demand.

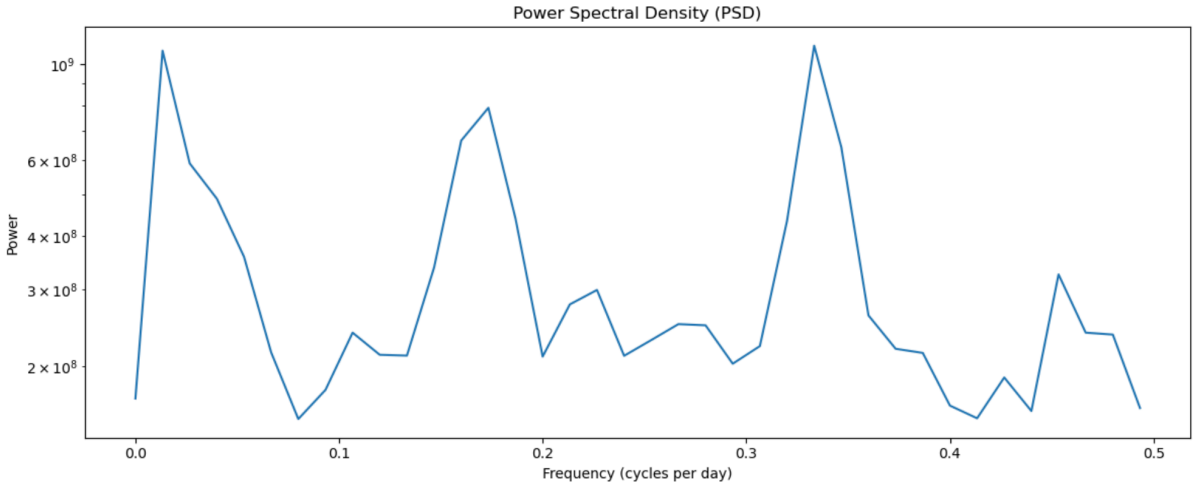


Daily sales in the UK show a lot of ups and downs, with frequent spikes suggesting that demand can change quickly, possibly due to promotions, special events, or seasonal shopping trends. Over time, there's a clear upward trend, with sales spikes becoming more frequent and

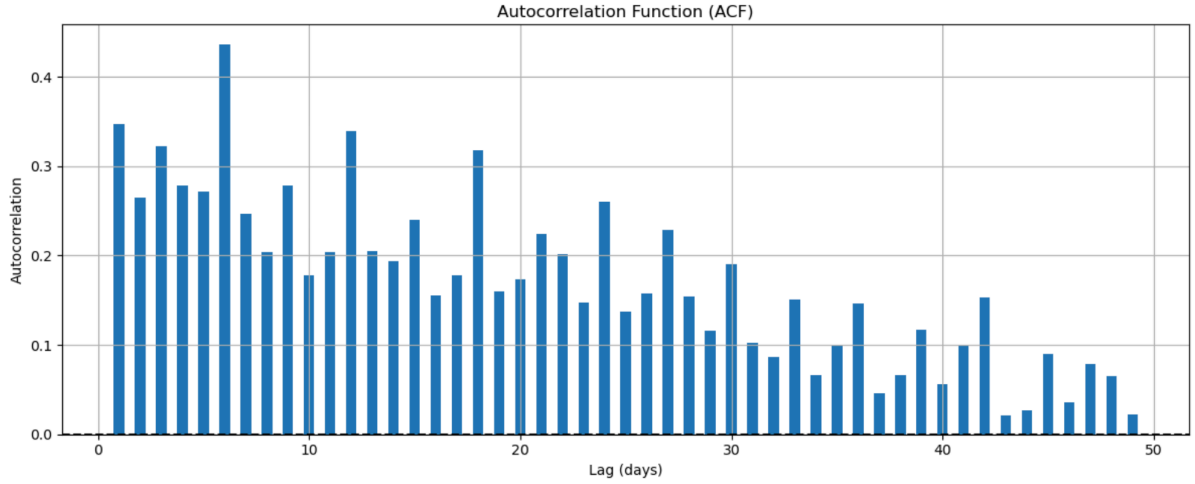
intense, especially in late 2011. This could mean more customer engagement, a growing market, or an increase in transactions. Toward the end of the timeline, there's a major surge in sales, likely driven by the holiday shopping season. The high variability in daily sales suggests they are highly influenced by external factors, so looking at weekly trends or using smoothing techniques might help reveal more consistent patterns.

Pre Modelling

We calculated the PSD and ACF of the signal to detect trends and seasonality in the data.



The spectral analysis provides valuable insights into sales patterns. The strongest peaks at lower frequencies highlight important long-term cycles, with a key frequency of 0.14 corresponding to a 7 day pattern, confirming a strong weekly sales cycle. There are also notable peaks below 0.1, indicating possible monthly trends. Additionally, a peak near 0.3 suggests a shorter 3-4 day cycle, hinting at more frequent fluctuations in demand. In contrast, higher-frequency variations appear to be random daily fluctuations rather than meaningful patterns, helping to separate true cyclical trends from background noise in sales activity.

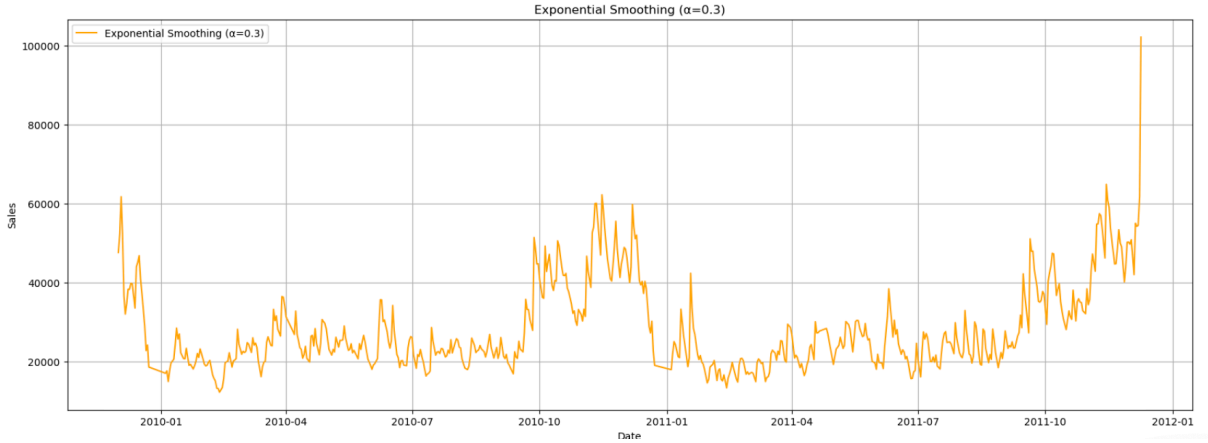


The sales data shows a strong connection between recent and current sales, especially within the first 10 days. This short-term dependence creates consistent weekly patterns, with noticeable spikes in correlation at around 7, 14, and 21 days, matching the weekly cycles seen

in the spectral analysis. As the time gap increases, this relationship fades, meaning past sales become less reliable for predicting future trends. There are also smaller peaks around 30 days, hinting at possible monthly patterns, likely influenced by factors like paydays or scheduled promotions.

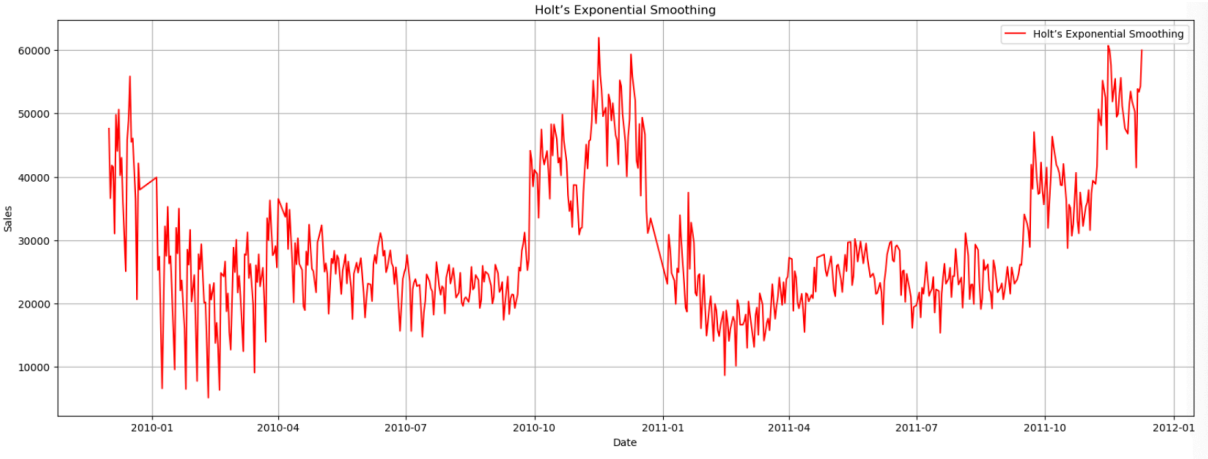
Exponential Smoothing

In the data, we have applied Exponential Smoothing with alpha around 0.3.



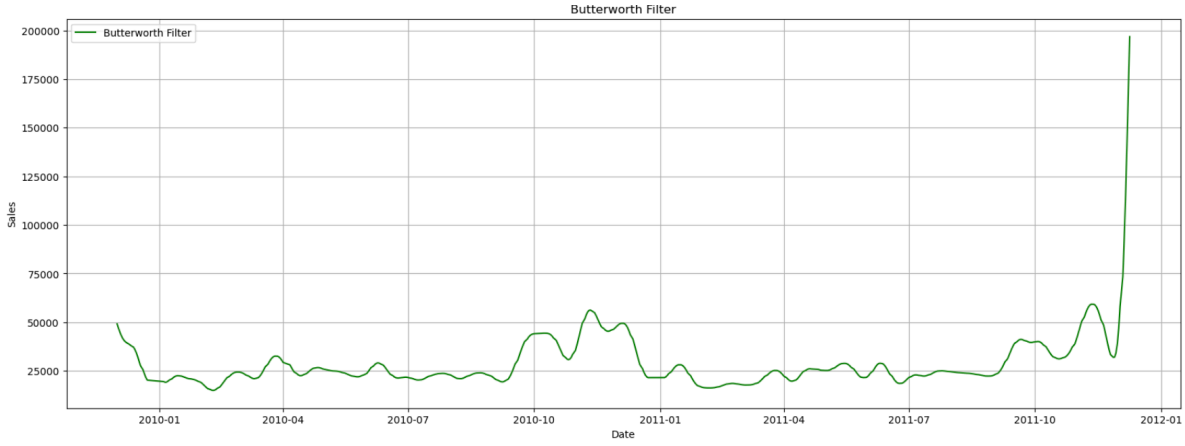
The code smooths the sales data using exponential smoothing with a factor of 0.3. We have experimented with different values of alpha ranging from 0.1 to 0.8. When we set alpha = 0.1, the smoothing is less sensitive to the most recent changes, leading to a smoother, less reactive curve. The past data will have a greater influence on the smoothed values. When we set alpha as 0.8, the smoothing becomes much more sensitive to recent data. This means that the smoothed line will follow the fluctuations and trends of the most recent observations more closely and be less smooth.

Holt Winters Exponential Smoothing



We have applied Holt Winters Exponential Smoothing model to the daily sales data with additive trend and seasonality features. It effectively captures fluctuations in the data. Towards the beginning of 2010, there were significant fluctuations in sales data which reduced towards the end of the year as sales numbers picked up and gave a trend. It is giving an idea of how the future sales can be by showing the trend and cycles.

Butterworth Low Pass Filter



The sales data are passed through Butterworth Low Pass filter with cutoff = 0.1. The graph presents a smoothed version of the sales trend by filtering out high-frequency noise while retaining the overall structure of the data. The filter effectively highlights long-term patterns and suppresses short-term fluctuations.

Unlike simple MA or exponential smoothing, the Butterworth filter preserves significant underlying trends while removing short-term irregularities, making it a useful tool for detecting broad sales patterns.

Training & Testing Split

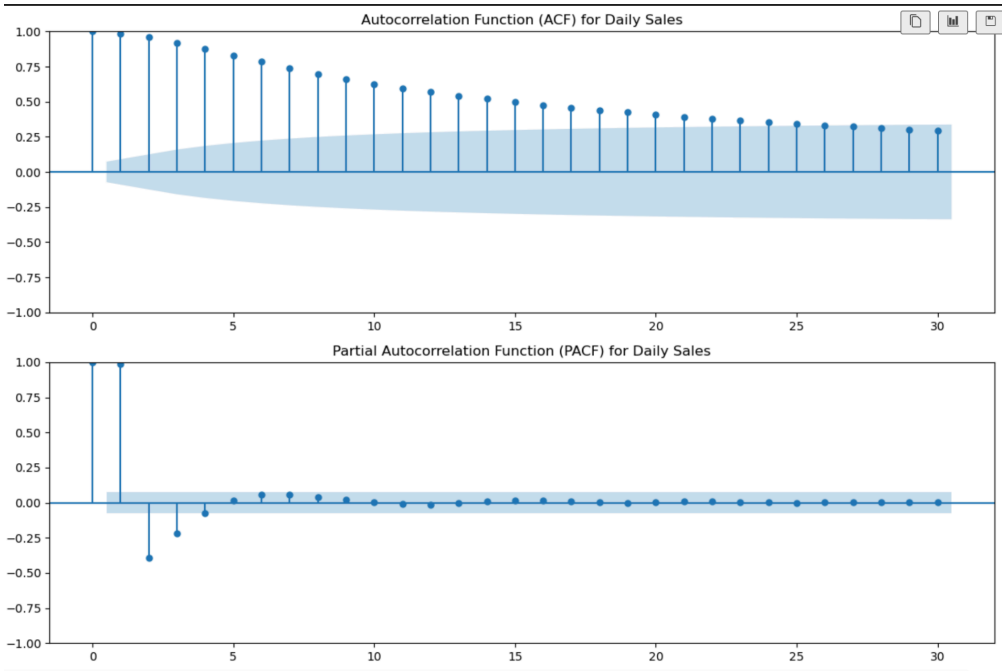
Data was split into training and test data with a ratio of 80% to 20%

Modelling

1. ARIMA 2. FBProphet 3. LSTM

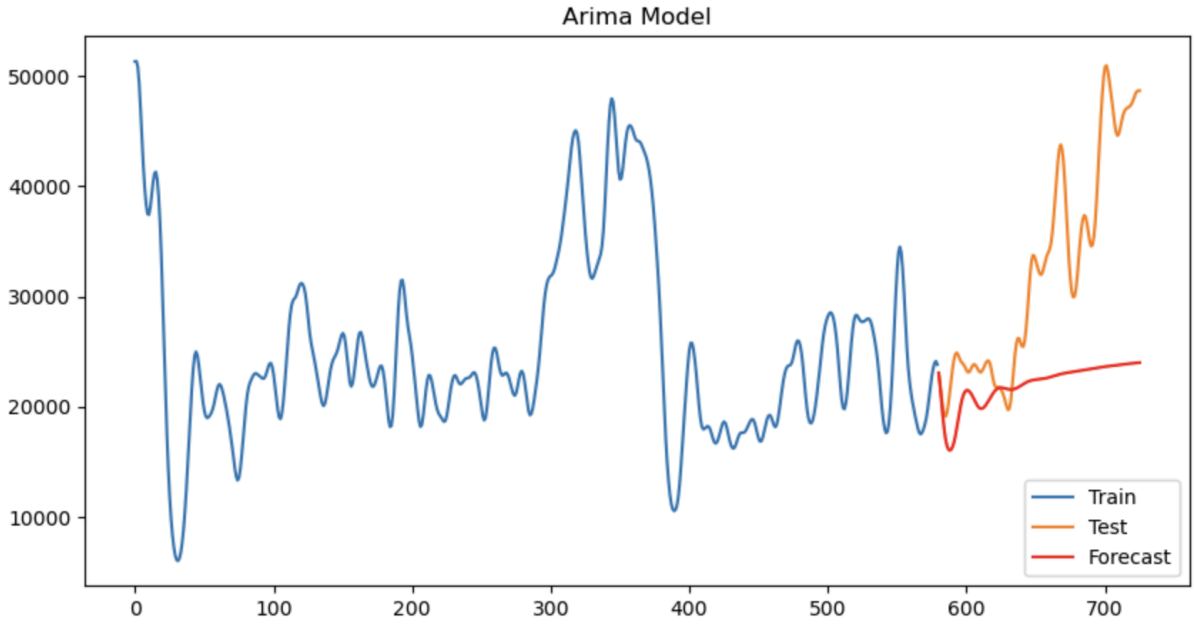
ACF and PACF Plots for ARIMA:

The ACF and PACF plots will provide a guide on what parameters to use for our ARIMA model. In the ACF plot, we notice that there is a decline in the values, indicating that the data might be non-stationary and will need to be differenced. The PACF graph shows a sudden change at 1, suggesting that the time series follows an AR(1) process, where past values influence future values. Given this, the parameters that should be used are (1,1,0), where the (1,1,0) parameters indicate an AR(1) process, with one differencing step for stationarity and no moving average component.



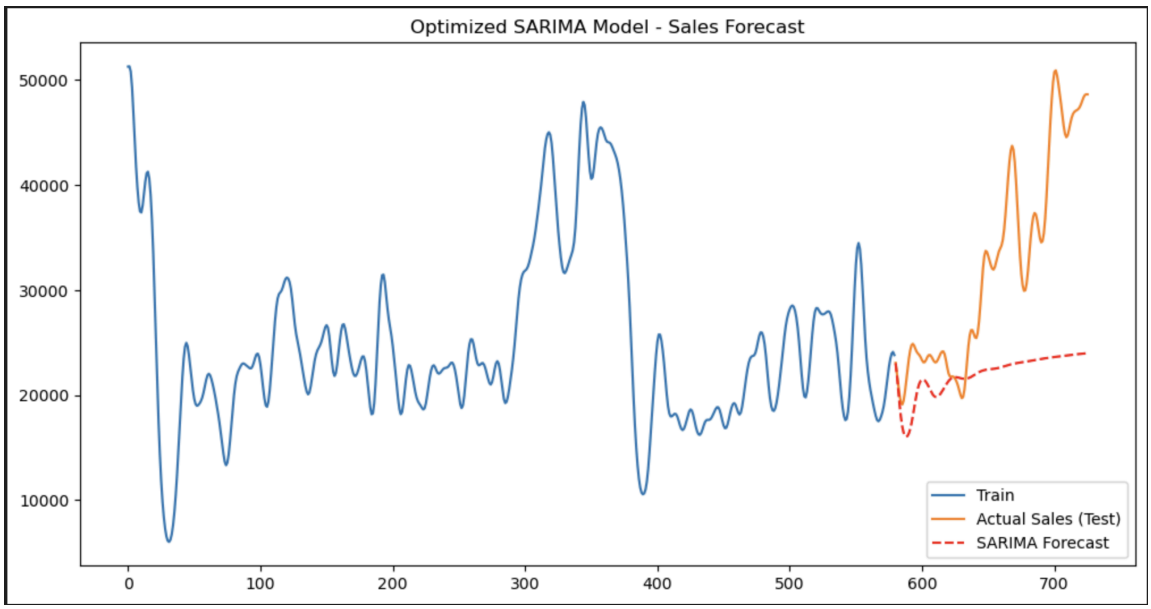
ARIMA

The ARIMA model (AutoRegressive Integrated Moving Average) is a forecasting model that is used to predict trends in our data. It utilizes three components, Autoregressive, Integrated, and Moving Average to forecast futures sales. Autoregressive allows using past values to predict future values, integrating differences in the data to make it stationary, and Moving Average uses past errors to correct and predict future values. Using ARIMA, we can predict the future sales for this dataset. However, ARIMA does not account for seasonality in the trends, therefore later we will apply a SARIMA to account for this. When plotting this data, we specified the holidays in the United Kingdom to account for any abrupt changes and seasonality of the data. We found the mean absolute percentage error (MAPE) to be 37.89%, which needs to be improved.



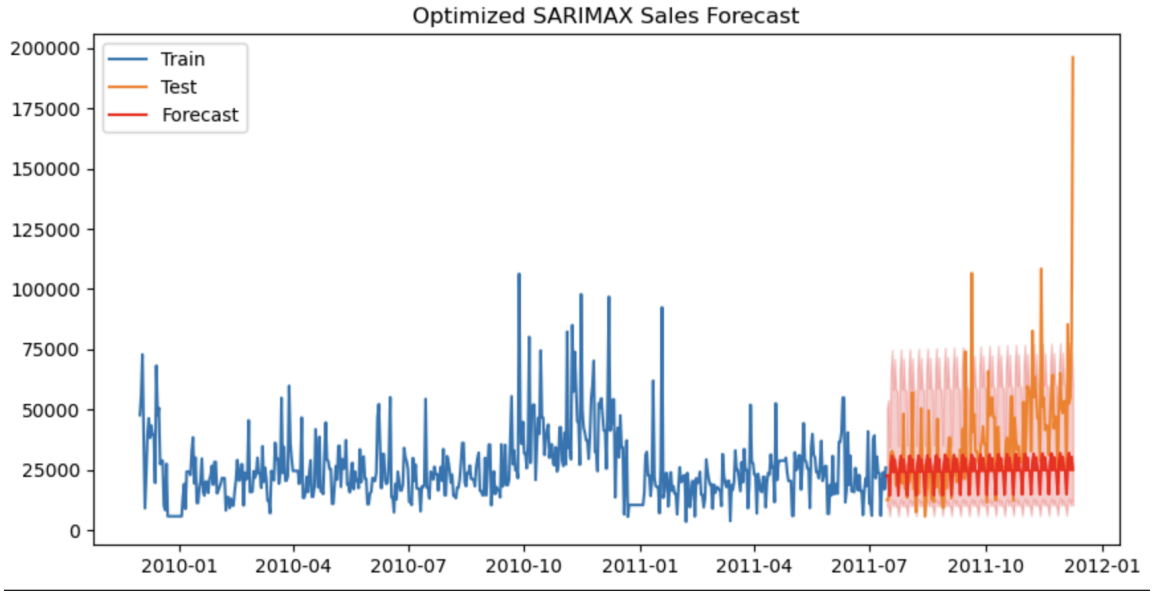
SARIMA Model:

Because the data showed signs of seasonality, we have also implemented the SARIMA model. We can see here that the model fails to forecast properly to the test data, with a MAPE score 27.68%. While we see that the forecast is following the trend at first, it fails to catch the spike in sales, indicating that there may be outliers in the data or noise that is disrupting the data.



SARIMAX Model:

Here, we have optimized the SARIMAX model for our dataset to account for seasonality. We can see that the forecast fails to adapt to the fluctuating sales in the test set. This could be due to the external factors or noise in the data. There are also confidence intervals in the forecast to show possible ranges. We also accounted for holidays in the United Kingdom as an exogenous variable.

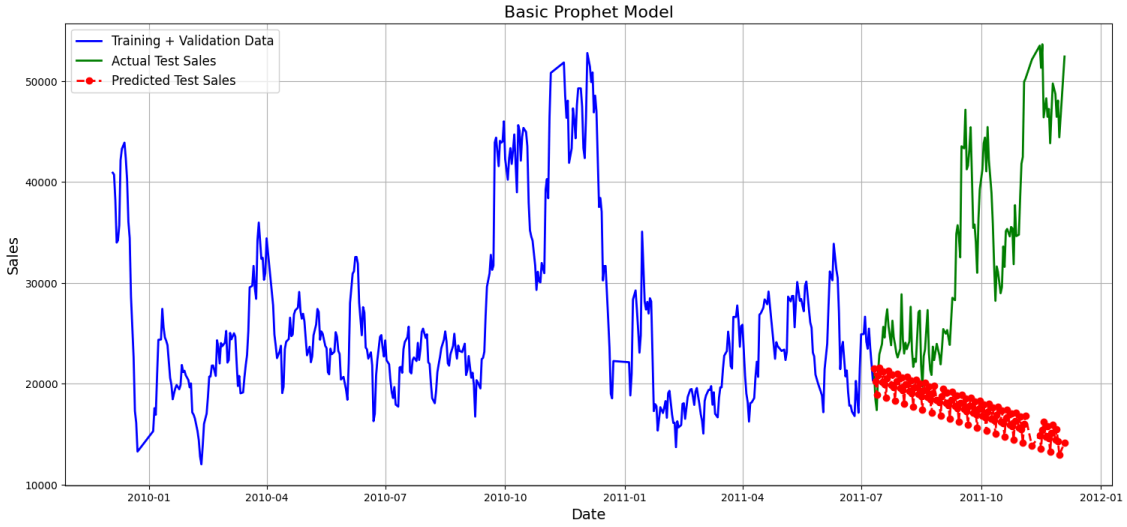


FbProphet

Facebook Prophet is an additive regression model specifically designed for time series forecasting. It is useful for data with strong seasonal patterns, trends, and external influences such as holidays. Prophet is built to be robust to missing data and outliers and is particularly effective for business forecasting.

1. **Base Model**

   This model was created with no tuning and no additional parameters, intended to be the baseline model.

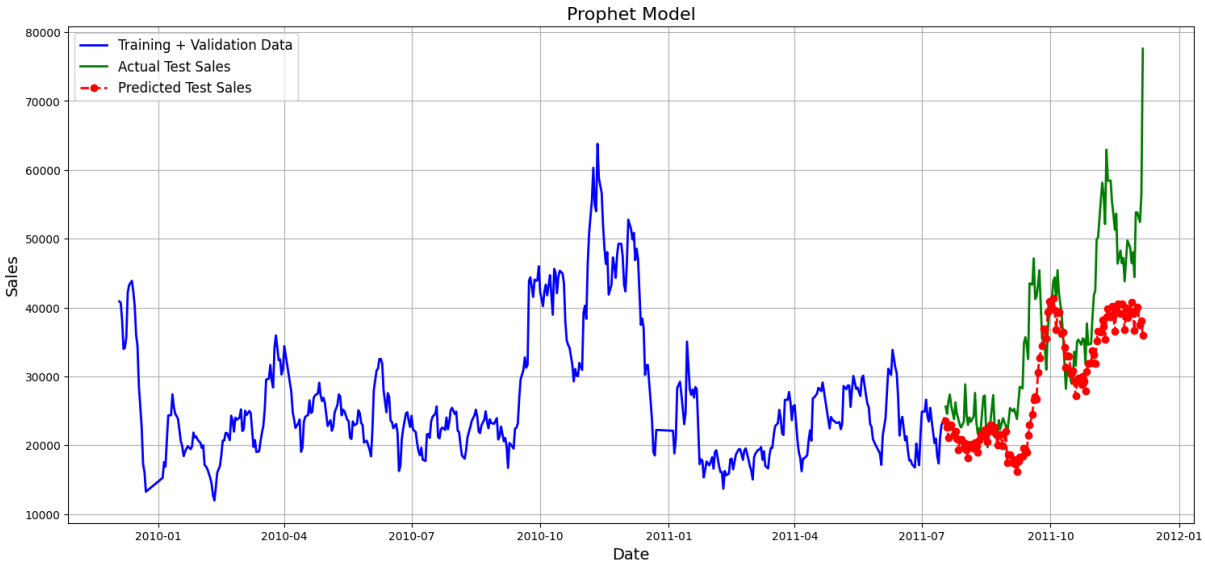


   MAPE: 44.71%

2. **Improved Model**

For this iteration of the Prophet model, several improvements were made to enhance forecasting accuracy. First, a log transformation was applied to the target variable to stabilize variance and prevent negative predictions, ensuring more reliable forecasts. A weekday regressor was introduced to capture weekly seasonality effects, accounting for potential fluctuations in sales based on different days of the week. UK public holidays were also incorporated to allow the model to adjust for demand spikes or drops around special events. The changepoint_prior_scale was manually set to 0.2, providing a balance between flexibility and overfitting when detecting trend changes. Additionally the parameters daily seasonality, weekly seasonality, yearly seasonality were marked as true.

The results indicate an improved model performance, as evidenced by a more aligned forecast with actual sales, minimizing error.

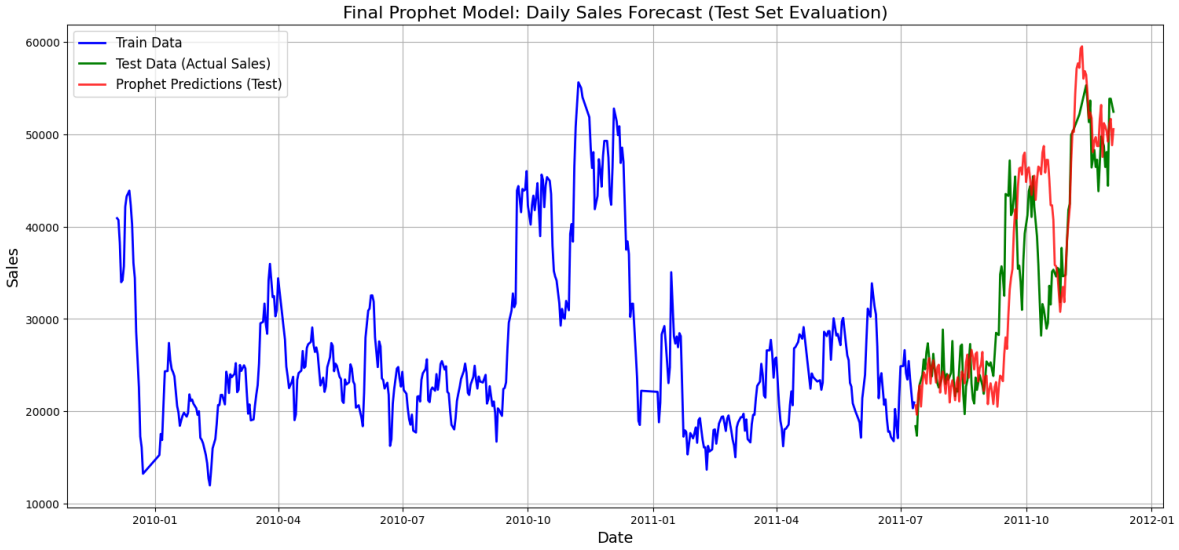


MAPE: 17.73%

3. Logistic Growth Model

This final iteration of the Prophet model incorporates several enhancements to improve forecasting accuracy and account for real world complexities in sales trends. First, extreme outliers (top 1% highest sales) were removed to prevent skewing the model, ensuring it captures realistic demand patterns rather than rare spikes. A logistic growth model was applied to account for natural constraints in sales, introducing a carrying capacity (cap) to prevent unrealistic. Increased yearly and weekly seasonalities (20 and 10 components, respectively) were included to improve sensitivity to both long-term and short-term trends.

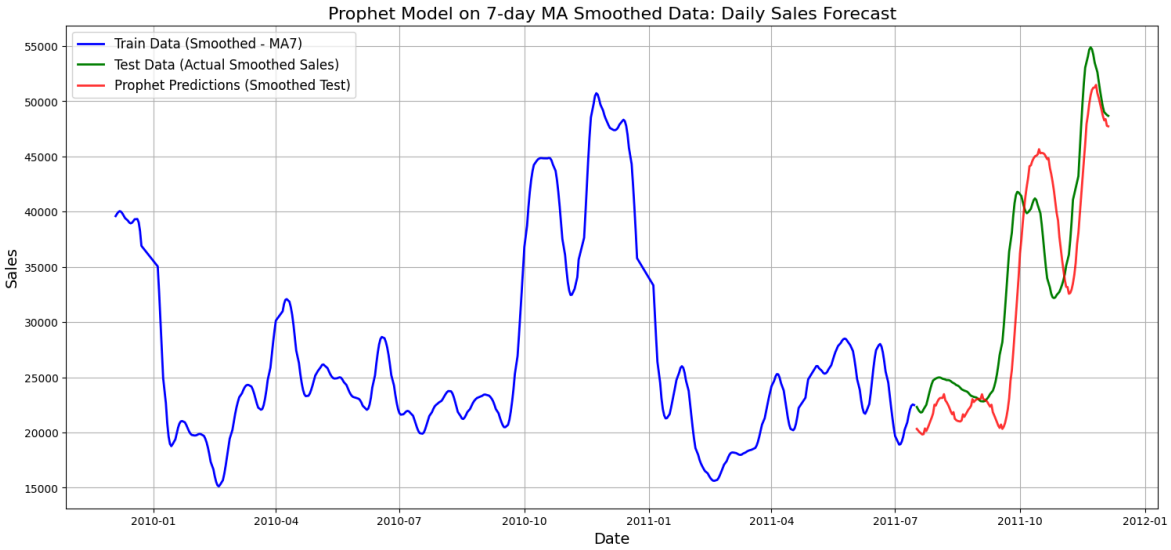
The test set predictions show improved alignment with actual sales trends, reducing errors and capturing seasonal peaks.



MAPE: 13.06%

4. Smoothed MA + Prophet Model

In this updated FBProphet model, a 7-day moving average (MA7) smoothing was applied to the sales data before training. This reduces short-term fluctuations, allowing the model to focus on overall trends rather than daily noise.



MAPE: 12.29%

Long Short-Term Memory (LSTM)

LSTMs are a type of recurrent neural network designed to handle long-term dependencies by overcoming the vanishing gradient problem, making them effective for sequential data. They regulate information flow using three key mechanisms:

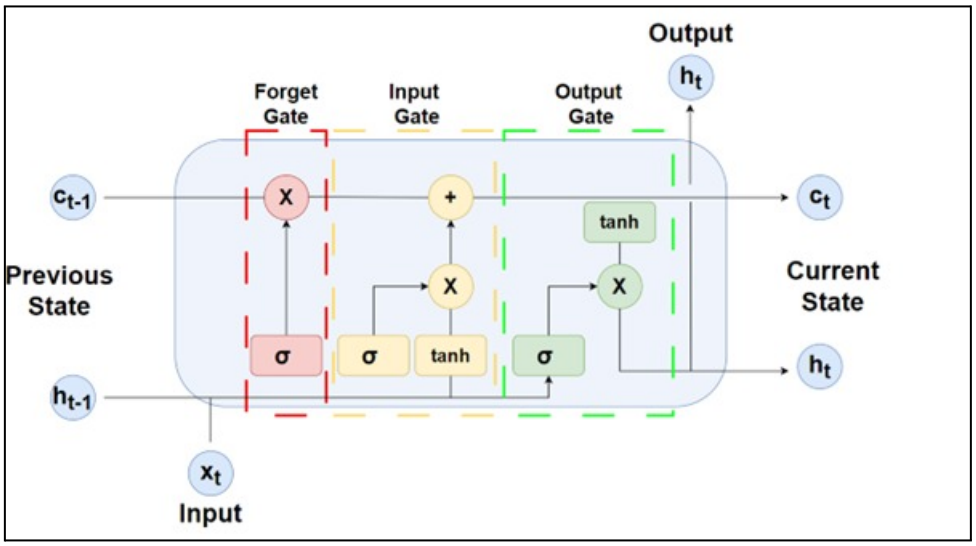
- **Forget Gate:** Determines which past information to discard, with values closer to 0 meaning "forget" and closer to 1 meaning "keep."
- **Input Gate:** Decides what new information to store using a combination of a sigmoid layer and a tanh layer.
- **Output Gate:** Controls what part of the current state is used as output, ensuring only relevant information is passed forward.

These mechanisms allow LSTMs to retain important patterns over long sequences, making them well-suited for time-series forecasting.

Bidirectional LSTM (BiLSTM)

BiLSTM improves standard LSTM by processing data in both forward and backward directions, allowing it to capture patterns from both the past and future. It consists of

- **Forward Layer:** Reads the sequence from start to finish, picking up patterns as they appear naturally.
- **Backward Layer:** Works in reverse, helping to catch relationships that might be clearer when looking from the end back to the beginning.
- **Output Layer:** Combines insights from both directions, giving a fuller picture of the data.



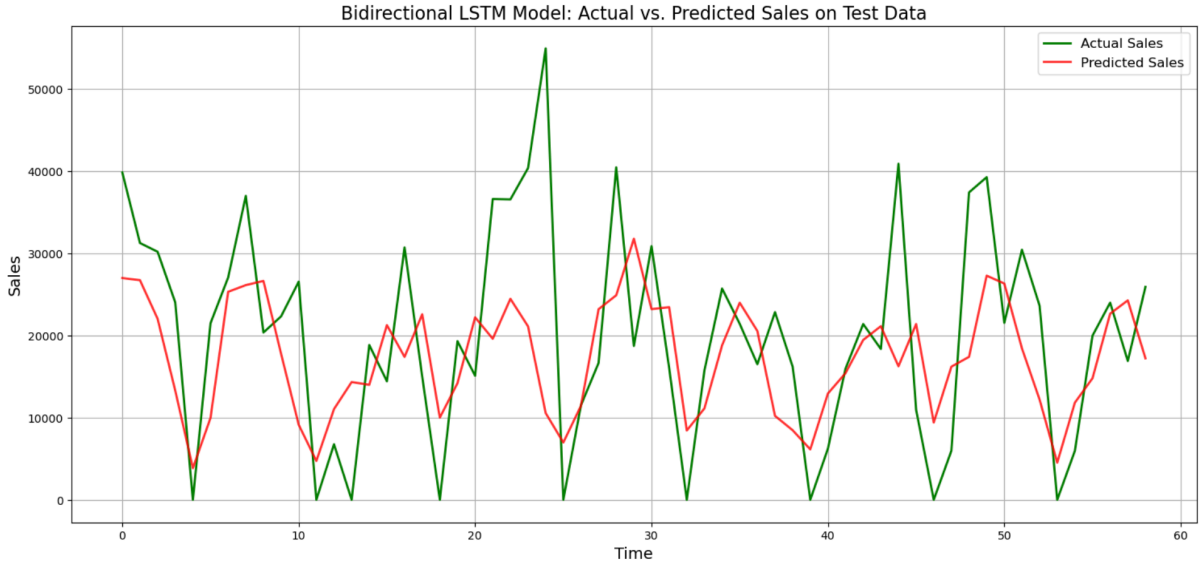
We have used the Bidirectional LSTM model for doing the time series analysis. The below picture represents the Bidirectional LSTM we have built.

Model: "sequential"

Layer (type)	Output Shape	Param #
bidirectional (Bidirectional)	(None, 30, 200)	81,600
bidirectional_1 (Bidirectional)	(None, 30, 200)	240,800
bidirectional_2 (Bidirectional)	(None, 30, 200)	240,800
bidirectional_3 (Bidirectional)	(None, 30, 200)	240,800
bidirectional_4 (Bidirectional)	(None, 200)	240,800
dense (Dense)	(None, 100)	20,100
dense_1 (Dense)	(None, 1)	101

Total params: 1,065,001 (4.06 MB)
Trainable params: 1,065,001 (4.06 MB)
Non-trainable params: 0 (0.00 B)

This model consists of five Bidirectional LSTM layers followed by two Dense layers. It takes input sequences of shape (None, 30, 200), where 30 is the sequence length and 200 represents the features. The first four Bidirectional LSTM layers keep this shape, with the first having 81,600 parameters and the next three having 240,800 each. The fifth Bidirectional LSTM layer reduces the sequence to (None, 200) while keeping 240,800 parameters. After that, a Dense layer with 100 units brings the output to (None, 100) with 20,100 parameters. Finally, the last Dense layer has one unit, producing (None, 1) with 101 parameters. In total, the model has 1,065,001 trainable parameters and is well-suited for tasks like sequence classification or regression.



The LSTM model demonstrates some ability to predict sales, though it is not entirely accurate. With a MAPE score of approximately 39%, it serves as a baseline model, providing a reference point for evaluating the performance of Prophet and other forecasting methods.

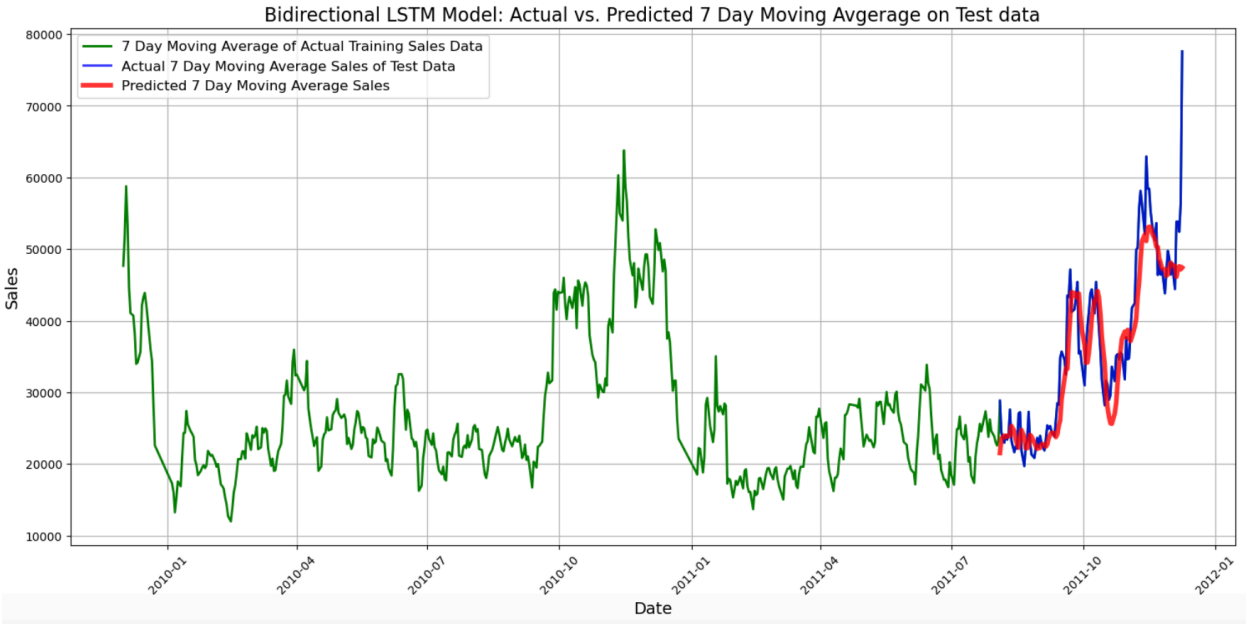
Improved Bidirectional LSTM Model

Model: "sequential_11"

Layer (type)	Output Shape	Param #
bidirectional_51 (Bidirectional)	(None, 60, 512)	528,384
dropout_46 (Dropout)	(None, 60, 512)	0
bidirectional_52 (Bidirectional)	(None, 60, 256)	656,384
dropout_47 (Dropout)	(None, 60, 256)	0
bidirectional_53 (Bidirectional)	(None, 60, 128)	164,352
dropout_48 (Dropout)	(None, 60, 128)	0
bidirectional_54 (Bidirectional)	(None, 128)	98,816
dropout_49 (Dropout)	(None, 128)	0
dense_34 (Dense)	(None, 64)	8,256
dense_35 (Dense)	(None, 32)	2,080
dense_36 (Dense)	(None, 16)	528
dense_37 (Dense)	(None, 1)	17

Total params: 1,458,817 (5.56 MB)
Trainable params: 1,458,817 (5.56 MB)
Non-trainable params: 0 (0.00 B)

First, the input to the model is changed from daily sales data to 7-day moving average of the sales data. This leads to less fluctuations in the data which helps the model to learn patterns easily. In this model, the number of bidirectional LSTM layers is 3 and the number of dense layers is also 3. There is mix and match in the activation functions in each of these layers with 'tanh' and 'relu' being used. There are 3 fully connected dense layers with a progressively reduced number of neurons. This helps the model to reach close to the actual values intuitively and it produces better results with MAPE 9.54%.



MAPE: 9.54%

Results

Model	MAPE (%)
ARIMA	37.89
SARIMA	27.68
SARIMAX	39.29
FbProphet	12.29
Bi-LSTM	39.81
Improved Bi-LSTM	9.54

Evaluation Metric

The metric chosen for model evaluation is MAPE (Mean Absolute Percentage Error), which was calculated on the test data.

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error
n = number of times the summation iteration happens
A_t = actual value
F_t = forecast value

Future Scope

We plan to implement a clustering algorithm in order to group products by their product descriptions or their stock codes.

First, we experimented in creating clusters with similar stock codes. When looking at the data, we noticed that products in similar categories had similar stock codes. Therefore we utilized a MiniBatch model to cluster the products into 5 categories and created a new column in our data frame to assign each product. Using this altered dataframe, we plan to create forecasts on each of the clusters to understand the trends with similar products.

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Sales	Cluster	Stock_Numeric	Stock_Cluster	Cluster_Label	Dates
0	489434	85048	15cm christmas glass ball 20 lights	12	2009-12-01 07:45:00	6.95	13085	United Kingdom	83.4	1	85048.0	1	vintage, zinc, star, white, hanging, tree, woo...	2009-12-01
1	489434	79323P	pink cherry lights	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0	4	79323.0	1	vintage, zinc, star, white, hanging, tree, woo...	2009-12-01
2	489434	79323W	white cherry lights	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0	0	79323.0	1	vintage, zinc, star, white, hanging, tree, woo...	2009-12-01
3	489434	22041	record frame 7" single size	48	2009-12-01 07:45:00	2.10	13085	United Kingdom	100.8	0	22041.0	0	white, tin, assorted, feltcraft, vintage, hot,...	2009-12-01
4	489434	21232	strawberry ceramic trinket box	24	2009-12-01 07:45:00	1.25	13085	United Kingdom	30.0	19	21232.0	0	white, tin, assorted, feltcraft, vintage, hot,...	2009-12-01

Another approach we tried was to categorize the products by their similarities in their product descriptions using Natural Language Processing (NLP) techniques. First, we removed words from the descriptions such as colors, shapes, and sizes manually. Additionally we used spaCY to filter out adjectives. Then, it applies TF-IDF vectorization to convert the cleaned text into numerical representations. We created 10 clusters using K-means clustering to group the products and assign them labels. Doing this we examined the output and also printed the top terms in each cluster. We can use these to assign labels manually or computationally to the clusters.

```
Cluster 0:
 - PINK CHERRY LIGHTS
 -  WHITE CHERRY LIGHTS
 - RECORD FRAME 7" SINGLE SIZE
 - STRAWBERRY CERAMIC TRINKET BOX
 - PINK DOUGHNUT TRINKET POT
 - SAVE THE PLANET MUG
 - CAT BOWL
 - LUNCHBOX WITH CUTLERY FAIRY CAKES
 - DOOR MAT BLACK FLOCK
 - ASSORTED COLOUR BIRD ORNAMENT
 - CHRISTMAS CRAFT WHITE FAIRY
 - FULL ENGLISH BREAKFAST PLATE
 - PIZZA PLATE IN BOX

Cluster 1:
 - 15CM CHRISTMAS GLASS BALL 20 LIGHTS
 - RETRO SPOT TEA SET CERAMIC 11 PC
 - SET/2 RED SPOTTY TEA TOWELS
 - VICTORIAN GLASS HANGING T-LIGHT
 - HOT WATER BOTTLE TEA AND SYMPATHY
 - GOLD APERITIF GLASS
 - GOLD WINE GLASS
 - SILVER APERITIF GLASS
 - ANTIQUE SILVER TEA GLASS ETCHED
 - ANTIQUE SILVER TEA GLASS ETCHED
 - RETRO SPOT TEA SET CERAMIC 11 PC
 - PICCADILLY TEA SET

Cluster 2:
 - ASSORTED COLOUR MINI CASES
 - RED SPOTTY ROUND CAKE TINS
 - PACK OF 60 PINK PAISLEY CAKE CASES
 - 60 TEATIME FAIRY CAKE CASES
 - PACK OF 72 RETRO SPOT CAKE CASES
 - 60 TEATIME FAIRY CAKE CASES
 - FAIRY CAKE CANDLES
 - CERAMIC CAKE BOWL + HANGING CAKES
 - DOOR MAT FAIRY CAKE
 - VINTAGE CREAM 3 BASKET CAKE STAND
 - DOOR MAT FAIRY CAKE
 - CERAMIC CAKE STAND + HANGING CAKES
 - MINI CAKE STAND WITH HANGING CAKES

Cluster 3:
 - DOG BOWL , CHASING BALL DESIGN
 - STRIPES DESIGN MONKEY DOLL
 -  VINTAGE DESIGN GIFT TAGS
 - DOG BOWL , CHASING BALL DESIGN
 - LUNCH BAG RED SPOTTY
 - LUNCH BAG CARS BLUE
 - LUNCH BAG WOODLAND
 - BIRDS MOBILE VINTAGE DESIGN
 - CERAMIC BOWL WITH STRAWBERRY DESIGN
 - COFFEE MUG CAT + BIRD DESIGN
 - COFFEE MUG DOG + BALL DESIGN
 -  VINTAGE DESIGN GIFT TAGS
 - RIBBON REEL SPOTS DESIGN
```

```
Cluster 4:
 - CHARLIE AND LOLA CHARLOTTE BAG
 - JUMBO BAG CHARLIE AND LOLA TOYS
 - JUMBO BAG TOYS
 - RETRO SPORT PARTY BAG + STICKER SET
 - JUMBO BAG PINK VINTAGE PAISLEY
 - JUMBO BAG SCANDINAVIAN PAISLEY
 - JUMBO BAG PINK VINTAGE PAISLEY
 - JUMBO BAG RED WHITE SPOTTY
 - CHARLOTTE BAG , PINK/WHITE SPOTS
 - RED SPOTTY CHARLOTTE BAG
 - GREY FLORAL FELTCRAFT SHOULDER BAG
 - PINK FLORAL FELTCRAFT SHOULDER BAG
 - CHARLIE AND LOLA CHARLOTTE BAG
```

```
Cluster 5:
 - BAKING SET 9 PIECE RETROSPOT
 - LUNCHBOX WITH CUTLERY RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - MILK PAN RED RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - LUNCHBOX WITH CUTLERY RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - LUNCHBOX WITH CUTLERY RETROSPOT
```

```
Cluster 6:
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
 - BLACK LOVE BIRD T-LIGHT HOLDER
 - LOVE HEART POCKET WARMER
 - FOOD CONTAINER SET 3 LOVE HEART
 - LADLE LOVE HEART RED
 - RED LOVE HEART SHAPE CUP
 - 6 CHOCOLATE LOVE HEART T-LIGHTS
 - LOVE HEART POCKET WARMER
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
```

```
Cluster 7:
 - FANCY FONT HOME SWEET HOME DOORMAT
 - HOME BUILDING BLOCK WORD
 -  PEACE WOODEN BLOCK LETTERS
 - BATH BUILDING BLOCK WORD
 - PEACE SMALL WOOD LETTERS
 - JOY LARGE WOOD LETTERS
 - WOOD S/3 CABINET ANT WHITE FINISH
 - WOOD 2 DRAWER CABINET WHITE FINISH
 - WOOD 2 DRAWER CABINET WHITE FINISH
 - FANCY FONT HOME SWEET HOME DOORMAT
 - HOME BUILDING BLOCK WORD
 - 12 EGG HOUSE PAINTED WOOD
 - NOEL WOODEN BLOCK LETTERS
 - WOOD S/3 CABINET ANT WHITE FINISH
```

```
Cluster 8:
 - HEART MEASURING SPOONS LARGE
 - HEART IVORY TRELLIS LARGE
 - HEART FILIGREE DOVE LARGE
 - CHRISTMAS CRAFT HEART DECORATIONS
 - CHRISTMAS CRAFT HEART STOCKING
 - HANGING HEART ZINC T-LIGHT HOLDER
 - GINGHAM HEART  DOORSTOP RED
 - RED WOOLLY HOTTIE WHITE HEART.
 - HEART IVORY TRELLIS LARGE
 - WHITE HANGING HEART T-LIGHT HOLDER
 - PINK FELT HANGING HEART W FLOWER
 - BLUE FELT HANGING HEART W FLOWER
```

```
Cluster 9:
 - AREA PATROLLED METAL SIGN
 - PLEASE ONE PERSON  METAL SIGN
 - BATHROOM METAL SIGN
 - LADIES & GENTLEMEN METAL SIGN
 - AREA PATROLLED METAL SIGN
 - PLEASE ONE PERSON  METAL SIGN
 - LADIES & GENTLEMEN METAL SIGN
 - LAUNDRY 15C METAL SIGN
 - AIRLINE LOUNGE,METAL SIGN
 - METAL SIGN CUPCAKE SINGLE HOOK
 - NO JUNK MAIL METAL SIGN
```

Top words in each cluster:

```
Cluster 0 top terms: ['box', 'paper', 'christmas', 'water', 'bottle']
Cluster 1 top terms: ['glass', 'tea', 'tlight', 'time', 'holder']
Cluster 2 top terms: ['cake', 'cases', 'fairy', 'stand', 'retro']
Cluster 3 top terms: ['design', 'lunch', 'bag', 'spaceboy', 'box']
Cluster 4 top terms: ['bag', 'paisley', 'vintage', 'charlotte', 'strawberry']
Cluster 5 top terms: ['retrospot', 'bag', 'lunch', 'piece', 'baking']
Cluster 6 top terms: ['love', 'heart', 'london', 'word', 'building']
Cluster 7 top terms: ['wood', 'home', 'block', 'letters', 'finish']
Cluster 8 top terms: ['heart', 'holder', 'tlight', 'hanging', 'decoration']
Cluster 9 top terms: ['sign', 'metal', 'bathroom', 'chocolate', 'person']
```

From these results we plan to apply hyperparameter tuning to our models to optimize and improve the performances. Using cross validation we can improve the robustness of our model and find weaknesses in our model. We will also perform time series analysis using the clusters to analyze the trends in their sales. There are other optimization techniques that we can use to improve our models and improve our clustering performance and improve the overall models. We will also address other external factors such as economic factors, holidays that were not in the Holidays python package, and discounts offered by the retail location. Finally, we will look into other models that we can apply such as BERTopic, XGBoost, or Density-Scan Clustering (DBSCAN) that can address the fluctuating trends in the sales.

Conclusion:

This project aimed to identify the best model for forecasting daily sales in the United Kingdom by exploring both traditional statistical methods and deep learning approaches. The goal was to find a model that could accurately predict future sales while considering seasonal trends, holiday effects, and underlying patterns in the data.

Among the statistical models, ARIMA and SARIMA struggled with the complexity of the sales data. Their reliance on assumptions about stationarity and linearity made it difficult to fully capture sales fluctuations. While SARIMA (MAPE: 27.68%) performed better than ARIMA (MAPE: 37.89%), neither model proved to be an ideal fit for this dataset.

Facebook Prophet (MAPE: 12.29%) stood out as a strong alternative. By automatically detecting trend shifts and incorporating UK holidays, it handled seasonality and external factors well. Its ability to manage missing data and model complex trends made it a powerful forecasting tool.

On the deep learning side, Bi-LSTM (MAPE: 39.81%) underperformed. However, after refining the approach, the improved LSTM model (MAPE: 9.54%) delivered the most accurate forecasts. This suggests that, with proper tuning, deep learning models can effectively capture complex, non-linear sales patterns.

Overall, the Improved LSTM model provided the most precise predictions, but Facebook Prophet remains a strong choice, particularly for businesses that prioritize interpretability and flexibility.