# **FINAL PROJECT SUMMARY**

# **Project Title:** Hand-picked Giftware

# **Team members:**

# Madhuri Patibandla, Madhumathi Sekar, Thy Kieu

# **Purpose**

# The purpose of this project is to forecast UK retail sales using a time series analysis for 2012 by analyzing actual sales observed from 2009-2011.

# The project highlights the importance of adaptive forecasts in retail environments characterized by changing sales patterns and irregularities.

# The analysis involves comparing the forecasting performance of the Exponential Smoothing (ETS) and Autoregressive Integrated Moving Average (ARIMA) models.

# **How Did You Build Your Project (Tools/Database Technologies)**

# The project is implemented using the R programming language. Key R libraries used include:

# fpp3 for time series analysis and forecasting.

# readxl for reading Excel files.

# ggplot2 for data visualization.

# forecast for additional forecasting tools.

# scales for plot scaling.

# tsibble for time series data structures.

# seasonal for seasonal decomposition.

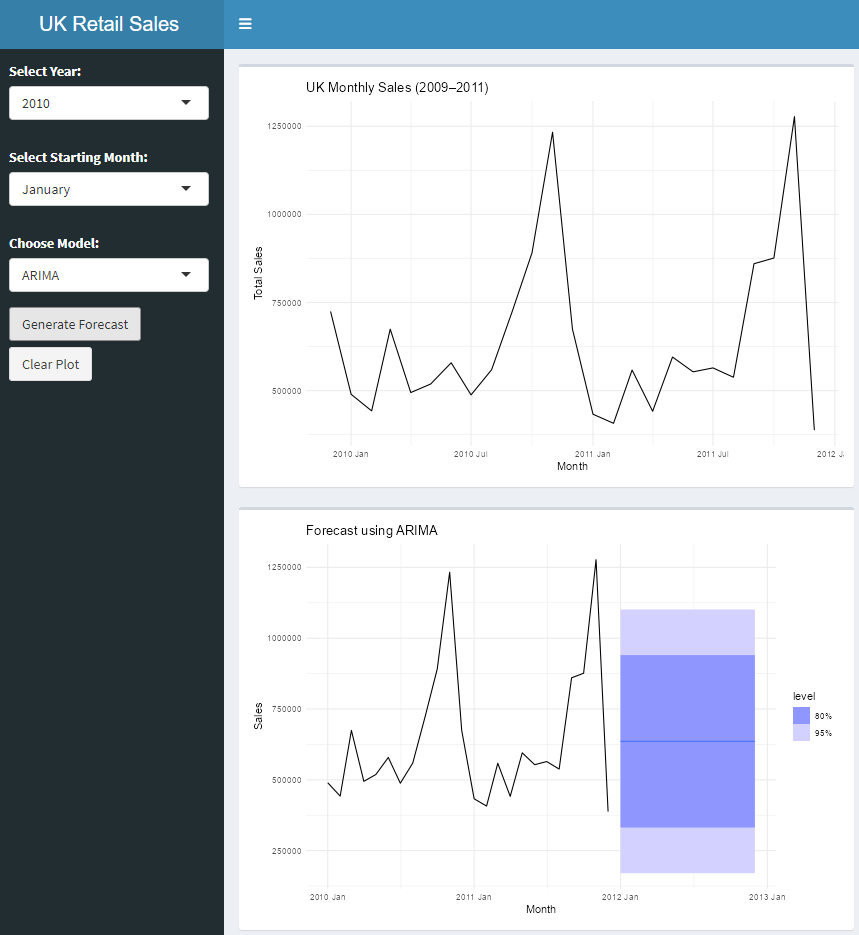
**Data**

* The project uses the "Online Retail" data set, which contains UK retail sales data.
* The dataset is in Excel format (xlsx). It appears to consist of two sheets, one for the period 2009-2010 and another for 2010-2011.
* Data cleaning steps included handling missing values in CustomerID, removing duplicates, and filtering records for the UK. Transformations involved calculating total prices and creating monthly and yearly aggregated time series.
* The data is characterized by changing sales patterns and irregularities, making it suitable for demonstrating the strengths of adaptive forecasting methods.

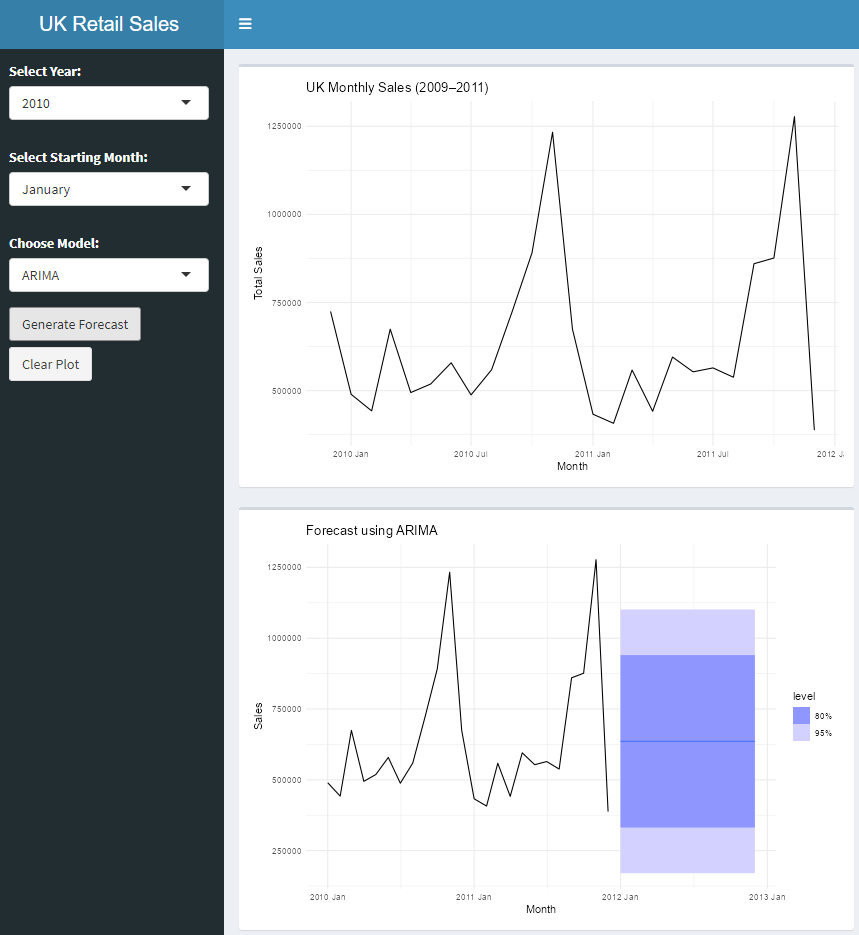
**Functionalities**

* **Data Loading and Preprocessing:** The project loads the Excel datasets, combines them, handles missing values (specifically for CustomerID), and removes duplicates. It also performs data transformations, such as calculating total prices and extracting month and year information from dates.
* **Exploratory Data Analysis (EDA):** The project includes EDA to understand the data's characteristics. This involves:
  + Calculating and visualizing daily, monthly, and yearly sales trends.
  + Analyzing summary statistics and data structures.
* **Time Series Analysis:** The core functionalities revolve around time series analysis:
  + **Time Series Visualization:** Creating seasonal plots, ACF (Autocorrelation Function) plots, PACF (Partial Autocorrelation Function) plots, and bar plots to visualize sales patterns.
  + **Time Series Decomposition:** Performing classical (additive) and STL (Seasonal-Trend decomposition using Loess) decomposition to separate the time series into trend, seasonal, and remainder components.
  + **Time Series Modeling**: Fitting ETS (Exponential Smoothing) and ARIMA models to the sales data. A Neural Network model (NNAR) is also used.
  + **Forecasting:** Generating forecasts for future sales using the fitted models.
  + **Model Comparison and Evaluation:** Comparing the forecasts from different models (ETS, ARIMA) and evaluating their accuracy using metrics like MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error).
  + **Predictions:** All models predicted an overall increasing trend in sales with regular seasonal peaks, particularly around the year-end (likely due to holidays). NNAR offered the most flexible and adaptive forecasts, especially suited for UK sales retail environments where sales patterns change, and irregularities exist. ETS and ARIMA provided solid, interpretable baselines and can be useful in more stable forecasting environments.

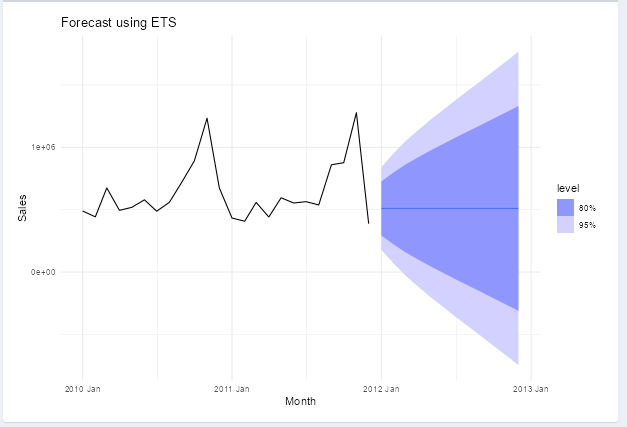
**Deployment**



Using the Shiny server, the initial dashboard shows the UK Monthly Sales from 2009 to 2011, and we added filters for selecting the year, starting month, and model (shown above).



When the ‘generate forecast’ button is clicked, the generated plot with the specific filters will be shown below the plot of the UK Total Monthly Sales. As shown above, the forecast using ARIMA was generated starting from January 2010.



The graph of the forecast using ETS was generated (shown above) using the filter starting in the month of January, ETS model, and starting in the year 2010.

**Issues**

* **Data Quality and Missing Values**
* **Issue**: The dataset had a significant number of missing CustomerID values, which posed a challenge for customer-level analysis.
* **Solution**: We replaced missing CustomerIDs with a placeholder value "Unknown" to ensure consistency in the dataset while still allowing aggregate analysis.
* **Data Duplication**
* **Issue:** There were duplicate rows in the combined dataset which could lead to inaccurate analysis.
* **Solution:** We used the distinct() function in R to remove duplicates and ensure the integrity of the dataset.
* **Inconsistent Date Formats**
* **Issue:** The InvoiceDate field had inconsistent formatting, which made time-based analysis difficult initially.
* **Solution**: We converted the dates using as\_datetime() and as.Date() for compatibility with tsibble and time series packages.
* **Revenue Calculation Mistake**
* **Issue:** While calculating Total\_Price, a sum over the entire dataset was mistakenly applied to each row.
* **Solution:** We corrected it by computing Total\_Price = Quantity \* Price row-wise before summarizing at the desired time intervals.
* **Model Selection Confusion**
* **Issue**: Choosing between ETS, ARIMA, and NNAR models required experimentation due to similar performance.
* **Solution:** We used cross-validation and accuracy metrics like MAE and RMSE to evaluate each model's performance and selected NNAR for its flexibility and robustness in handling irregular seasonality.
* **Forecast Visualization Clarity**
* **Issue:** Overlapping forecast and actual data in plots caused confusion.
* **Solution:** We adjusted transparency and used distinct colors and shading for forecasts to clearly differentiate them from historical data.
* **Online Deployment**
* **Issue:** Shiny app couldn’t be deployed online.
* **Solution:** Can be deployed and used locally.

**Contributions**

|  |  |
| --- | --- |
| **Name** | **Average Time for Final Project** |
| Madhuri Patibandla | 8 hours |
| Madhumathi Sekar | 8 hours |
| Thy Kieu | 8 hours |

**Reflection**

* We had a challenge in managing the data cleaning process, particularly dealing with missing values and ensuring the time series data was correctly formatted for modeling.
* Getting the data into a tsibble structure and troubleshooting errors during modeling required a strong understanding of R's time series analysis structure.
* One of the most challenging aspects of this project was selecting the right forecasting model. While ETS and ARIMA provided strong baselines, their performance was very close, and it took careful experimentation and evaluation to understand their differences. Understanding when and why to use models like NNAR, especially in a retail context with non-linear and changing seasonal patterns was both a technical and conceptual learning experience.
* It was also our first time working with deployment approaches such as Shiny, and it working with it helped develop more experience and what could be done in the application.
* The most interesting part was seeing how various models interpreted the same dataset differently. Comparing their forecasts side-by-side helped us visualize predictions better. We also enjoyed using visual storytelling to explain trends like holiday spikes in sales which made the data feel more connected to real-world behavior.

**Sharing:** Github : <https://github.com/madhuripatibandla/Timeseries-Project.git>

**Link to the deployed application:** was deployed locally (screenshots of local deployment in both report section above and presentation slides)