CAPSTONE PROJECT

Walmart Sales Forecasting and Analysis for Retail Stores

- Chaitanya Keshav

2. Table of Contents

1. Cover Page

2. Table of Contents

3. Problem Statement

4. Project Objective

5. Data Description

6. Data Pre-processing Steps and Inspiration

7. Exploratory Data Analysis (EDA)

8. Methodology and Model Selection Process for forecasting model

9. Overview of Model Performance

10. Forecasted Sales

11. Implications and Next Steps

12. Conclusion

13. References

14. Appendices

---

3. Problem Statement

- **Overview of the Problem:**
    - A retail chain with multiple outlets across the country is encountering challenges in managing its inventory to effectively align demand with supply.
- **Key Objectives:**
    - The objective is to derive actionable insights from the provided dataset and develop predictive models to forecast sales over a specified time horizon (e.g., several months or years). Sales forecasting is a key problem to address in improving inventory management, so forecasting future sales for the next 12 weeks to help improve store operations.
- **Importance of the Project:**
    - This will enable the company to optimize inventory management, ensuring that supply meets demand and minimizing the risk of stockouts or overstocking across its stores.

---

4. Project Objective

- **Short-term Objectives:**
    - Analyse historical sales data to identify patterns and trends.
    - Identify key drivers (e.g., unemployment, CPI) affecting sales.
- **Long-term Objective:**
    - Build a predictive model to forecast future sales for each store over the next 12 weeks.
    - Ensure inventory management aligns with predicted demand.

---

5. Data Description

- **Dataset Overview:**
    - The dataset contains 6435 rows and 8 columns.
- **Feature List and Explanation:**
    - **Store:** Store number.
    - **Date:** Week of sales data.
    - **Weekly_Sales:** Sales for the given store in the week.
    - **Holiday_Flag:** Indicator of whether it's a holiday week.
    - **Temperature:** Temperature on the day of the sale.
    - **Fuel_Price:** Cost of fuel in the region.
    - **CPI:** Consumer Price Index.
    - **Unemployment:** Unemployment rate.
- **Data Types and Missing Values:**
    - **Store (int64):** Categorical variable representing store number.
    - **Date (object):** Stored as a string type, will be converted to datetime for time series analysis.
    - **Weekly_Sales (float64)**: Continuous numerical variable representing weekly sales.
    - **Holiday_Flag (int64):** Binary variable indicating holiday week (1) or not (0).
    - **Temperature (float64):** Continuous variable representing temperature on the day of sale.
    - **Fuel_Price (float64):** Cost of fuel during the sale week.
    - **CPI (float64):** Consumer Price Index for the week.
    - **Unemployment (float64):** Unemployment rate for the corresponding week.
    - **Missing Values:** Upon examination of the dataset, it was observed that there were **no missing values** in any of the columns. All columns, including **Store**, **Date**, **Weekly_Sales**, **Holiday_Flag**, **Temperature**, **Fuel_Price**, **CPI**, and **Unemployment**, contain **6435 non-null entries**. Therefore, no data imputation or handling of missing values was required for this dataset.

    This completeness of data is beneficial as it allows for direct analysis and model building without the need for additional data preprocessing steps for handling missing values.

- **Summary Statistics:**

Feature	Min	Max	Mean	Std Dev
Store	1	45	-	-
Weekly_Sales	2,09,986.20	38,18,686	10,46,965	5,64,366.60
Holiday_Flag	0	1	0.07	-
Temperature (°F)	-2.06	100.14	60.66	18.44
Fuel_Price	2.47	4.47	3.36	0.46
CPI	126.06	227.23	171.58	39.36
Unemployment	3.88%	14.31%	8.00%	1.88%

  - **Store:** A categorical variable with store numbers ranging from 1 to 45.
  - **Weekly_Sales:** Ranges from $209,986.20 to $3,818,686, with an average weekly sale of $1,046,965, showing significant variation across stores.
  - **Holiday_Flag**: A binary variable indicating holiday weeks (0 = non-holiday, 1 = holiday), with holidays being relatively infrequent (mean = 0.07).
  - **Temperature:** Ranges from -2.06°F to 100.14°F, with an average of 60.66°F, reflecting diverse climate conditions across store regions.
  - **Fuel_Price:** Spans from $2.47 to $4.47 per gallon, with an average of $3.36, indicating regional variations in fuel prices.
  - **CPI (Consumer Price Index):** Ranges from 126.06 to 227.23, with an average of 171.58, representing economic conditions over time.
  - **Unemployment:** Ranges from 3.88% to 14.31%, with an average of 8.00%, reflecting varying regional unemployment rates.

---

6. Data Pre-processing Steps and Inspiration

- **Data Cleaning:**

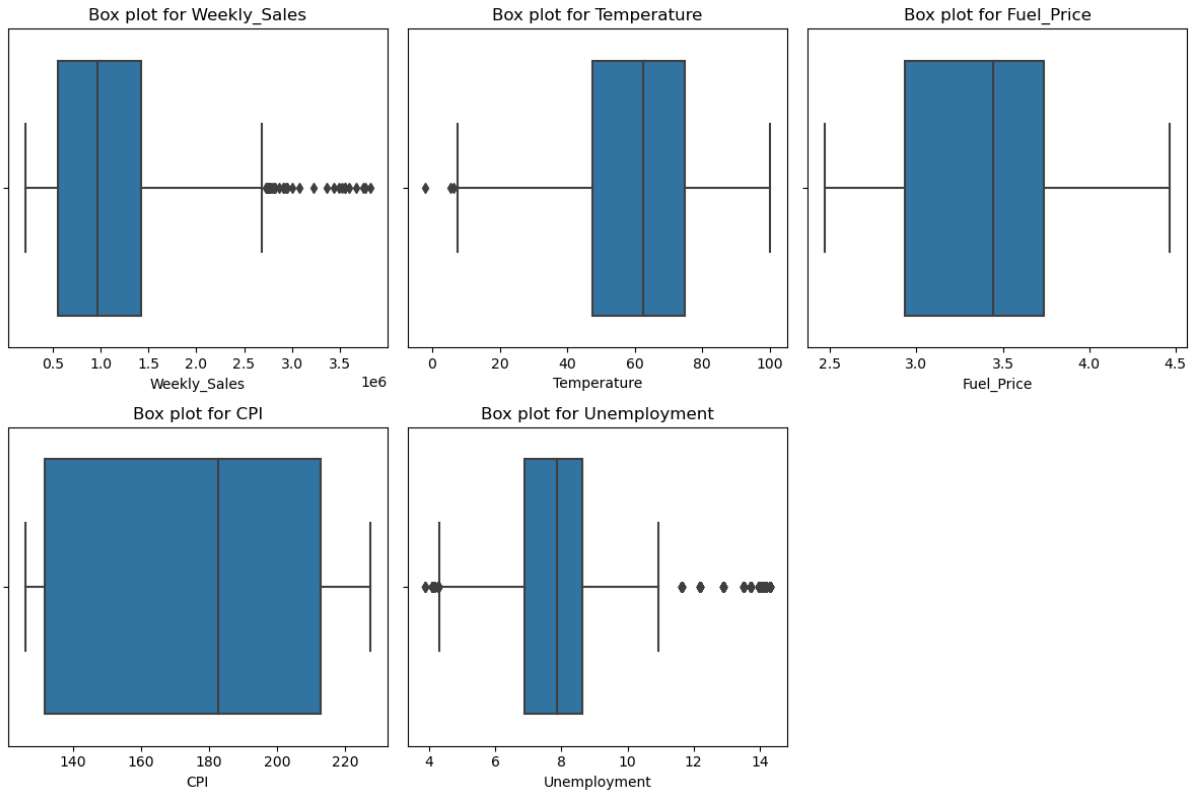
  As no missing or duplicated values/entries were encountered, therefore no need to perform any kind of data cleaning.

- **Outlier Detection and Handling:**

  - There is two ways to detect outliers:

    - Outliers in overall data.
    - Outliers in data related to specific store number.

  - **In the case of overall data** outliers were present for **Weekly_Sales**, **Temperature** and **Unemployment** variables.



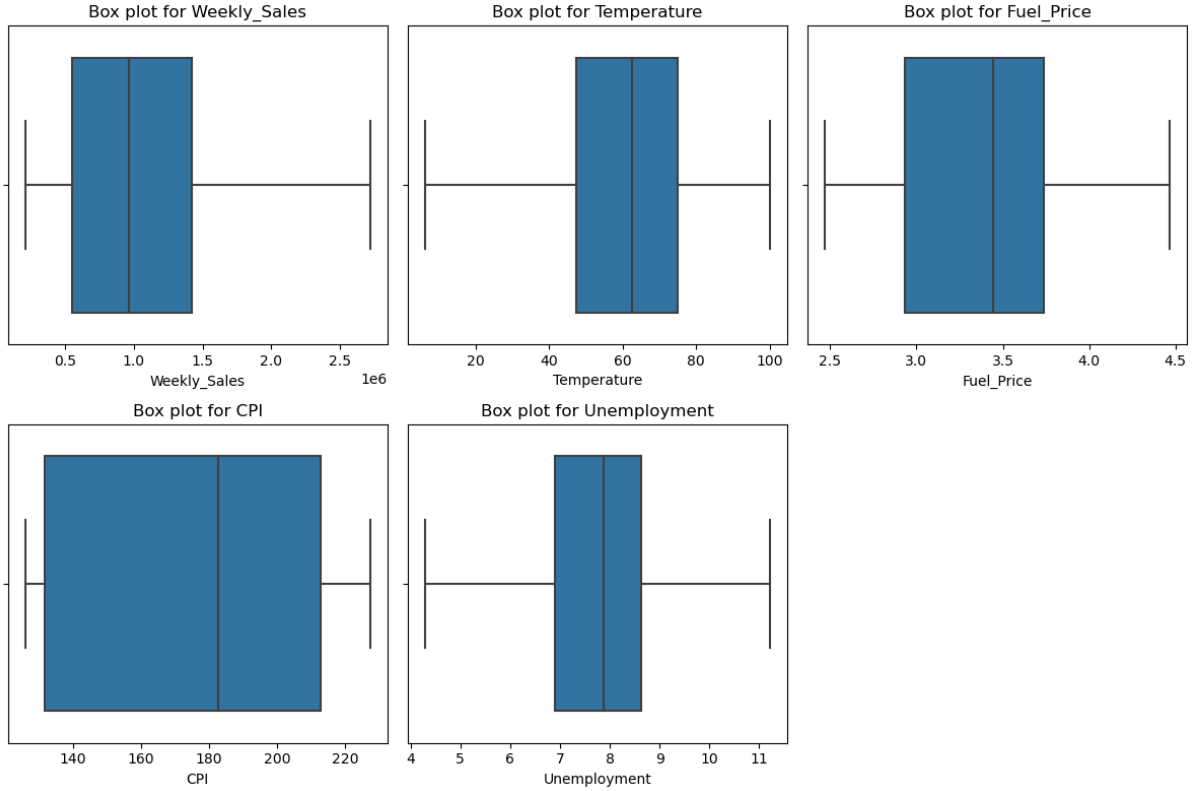
  - In the case of store specific subset of data, boxplot analysis revealed the presence of outliers predominantly in the **Weekly_Sales** variable. Which will be handled during store specific analysis.

  - **Handling:** To handle outliers, Winsorization was applied, capping extreme values at predefined bounds. This method preserves the dataset's structure and avoids data loss from dropping outliers, ensuring the integrity of temporal data and preventing distortions in statistical summaries or model predictions.

  - Outliers were not treated in certain cases, as the dataset contains no meaningless or impossible values, suggesting that these outliers may hold

meaningful information. Therefore, the raw data was retained to allow for comparative analysis in specific aspects of the study.

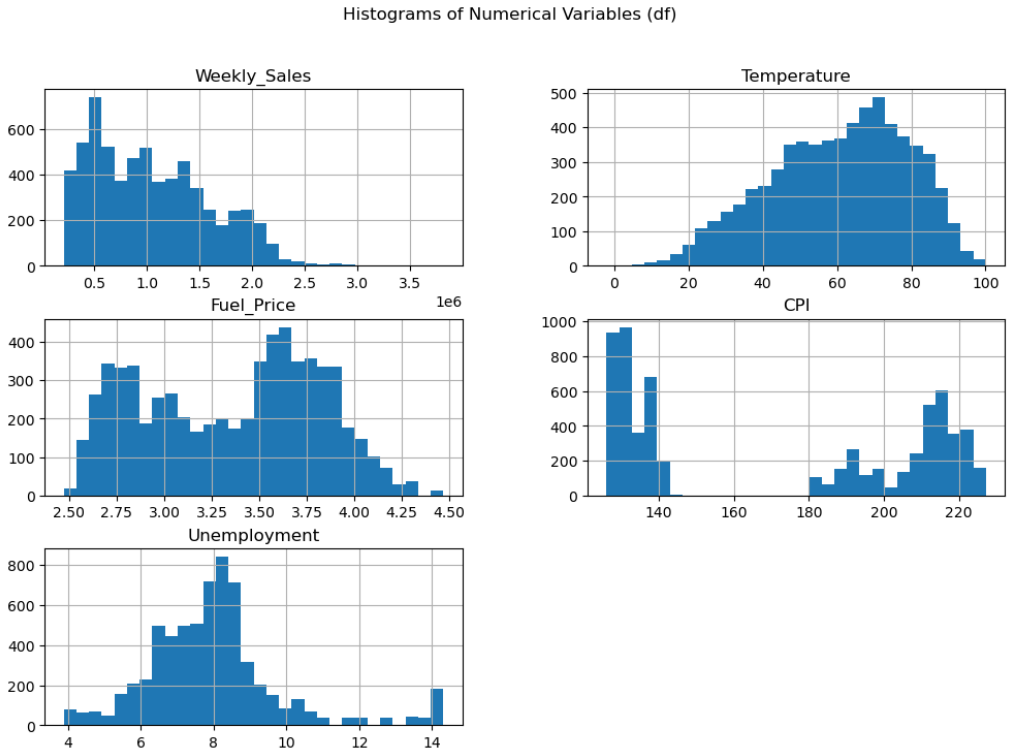
- Boxplot of overall data **(df)** after handling outliers **(df_clean):**



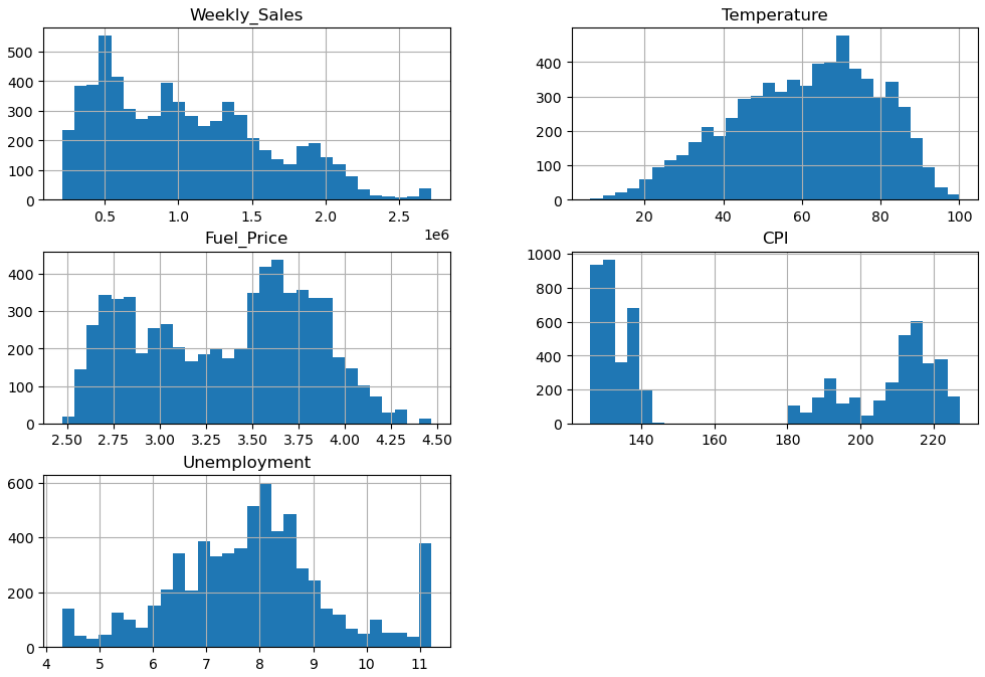
7. Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
  - Distribution of key variables:

Before handling outliers:



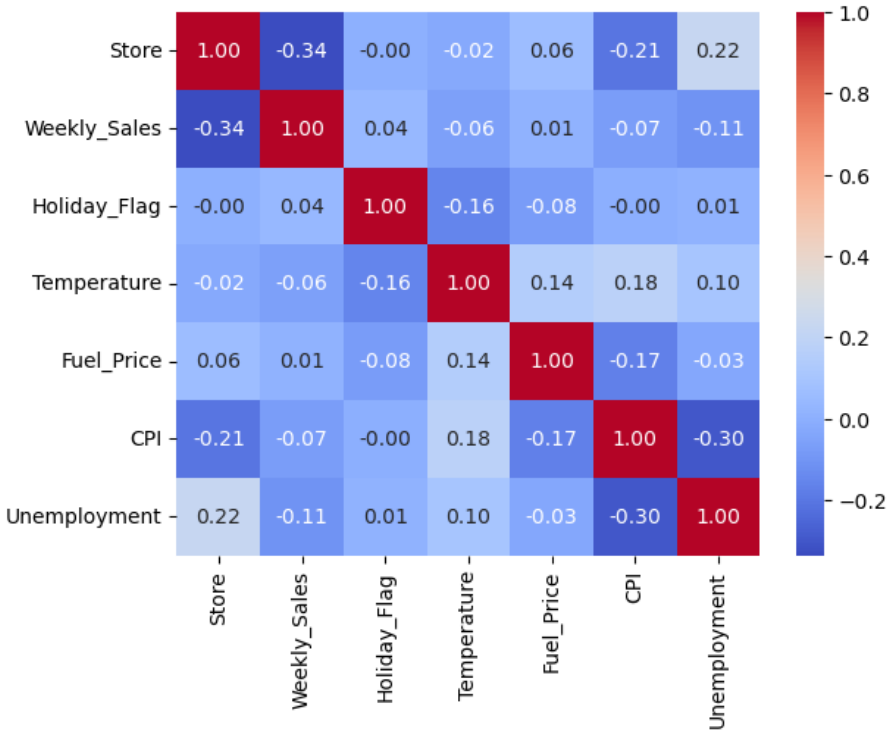
After handling outliers:



- **Interpretation:** The two sets of histograms provide a similar overall picture of the distributions of the numerical variables. Both sets show that Weekly_Sales, Fuel_Price, and Unemployment have right-skewed distributions, indicating a majority of data points concentrated towards lower values with a few outliers extending towards higher values. In contrast, Temperature and CPI exhibit approximately bell-shaped distributions, suggesting they follow a normal distribution pattern. While the overall shapes and centres of the distributions are consistent between the two sets, minor differences exist**.**

  The remainder of the exploratory data analysis (EDA) was conducted on both datasets, which further confirmed that the results were largely consistent across the two. **Consequently, the subsequent analyses will be based on the dataset containing the outliers (the raw data (df)).**

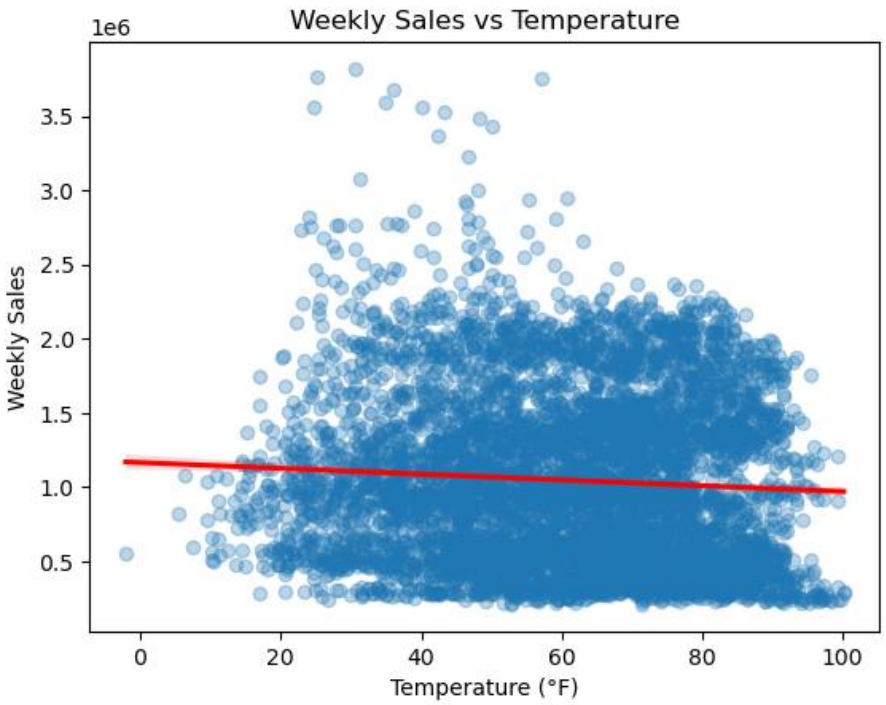
- **Bivariate Analysis:**
  - ***Correlation matrix (overall data):***



  - **Weekly Sales and Unemployment:** There is a **negative correlation** of -0.106 between weekly sales and the unemployment rate. While this indicates a slight inverse relationship, the correlation is weak, meaning that changes in the unemployment rate may have a minimal impact on weekly sales overall.

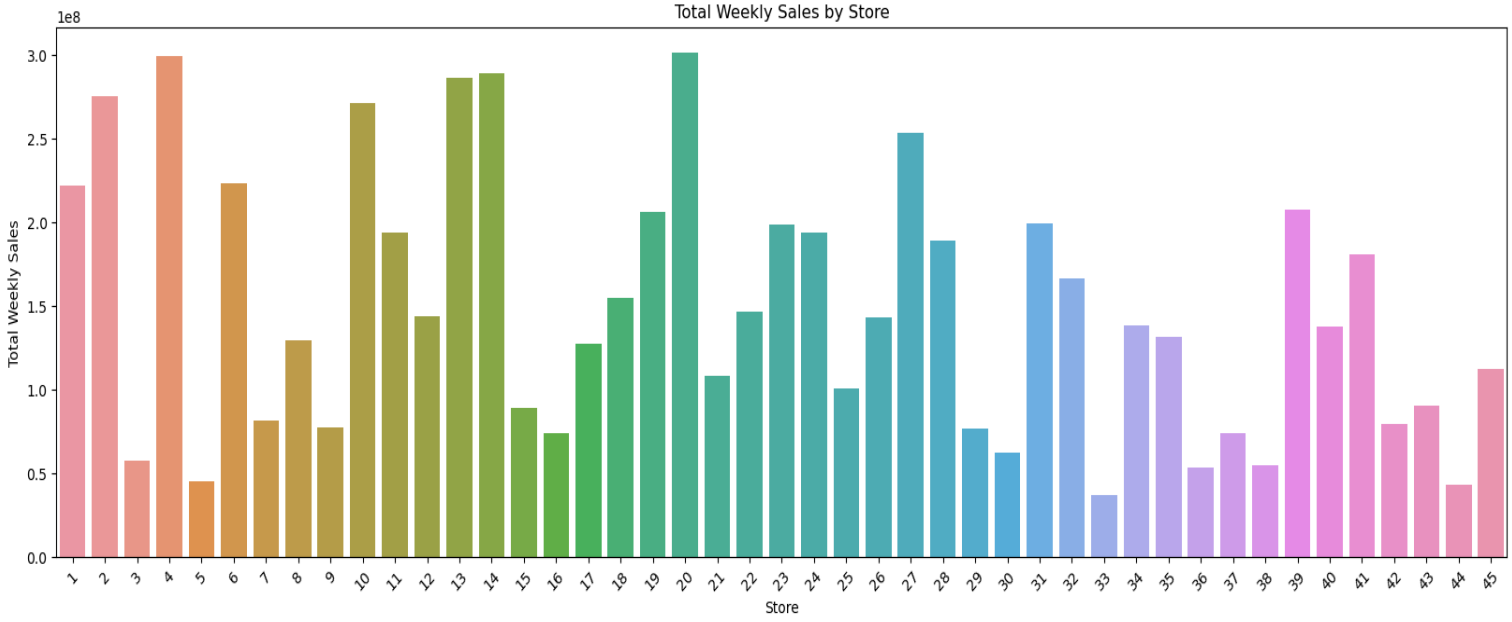
    Analysing store wise data **store number 38(correlation = -0.785)** and **44(correlation = -0.78)** were the **most affected stores.**

- **Weekly Sales and Temperature:** The correlation of **-0.0638** suggests a overall very weak negative relationship between temperature and weekly sales.



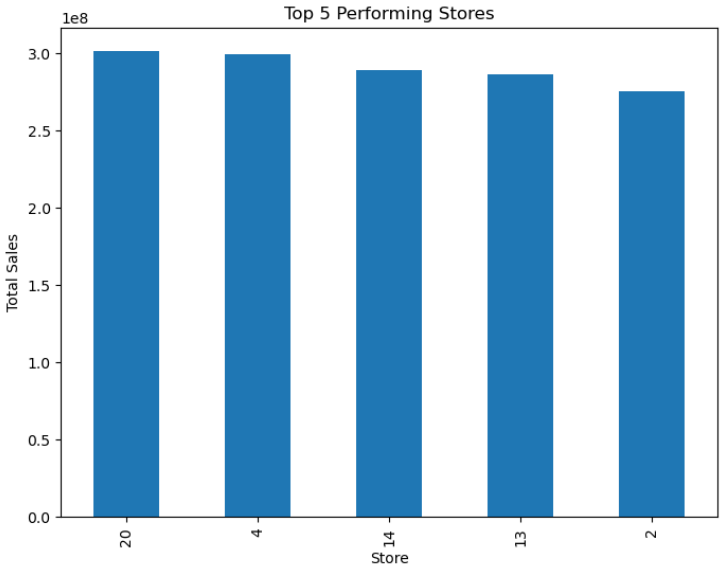
- **Weekly Sales and CPI (Consumer Price Index):** The correlation is **-0.0726** overall, indicating a very weak negative relationship between CPI and weekly sales. Store wise analysis revealed the **store 38** and **36** to be the highest positively and highest negatively correlated stores with the correlation values of **0.813** and **-0.915** respectively.

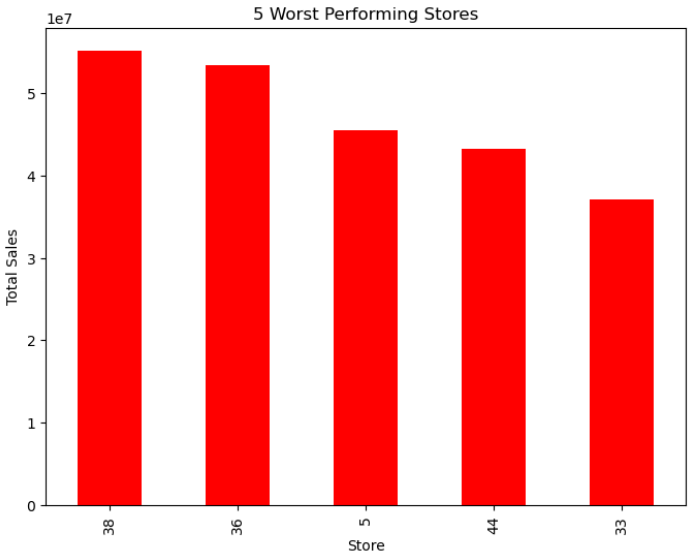
- **Store wise performance:**



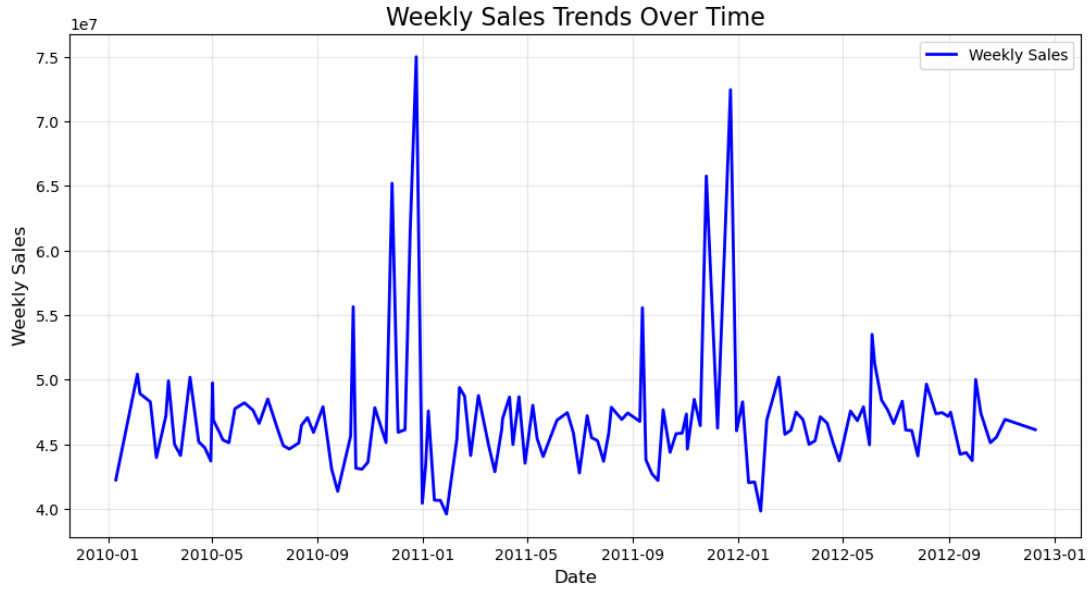
- To analyse store wise performance sum total of weekly sales for the available historical data was considered which revealed **Store 20 ($301.4 million), Store 4 ($299.5 million), Store 14 ($289.0 million), Store 13 ($286.5 million), and Store 2 ($275.4 million)** to be the best-performing stores. These stores exhibited consistently high sales, potentially due to favourable locations, strong consumer demand, or effective operational strategies.

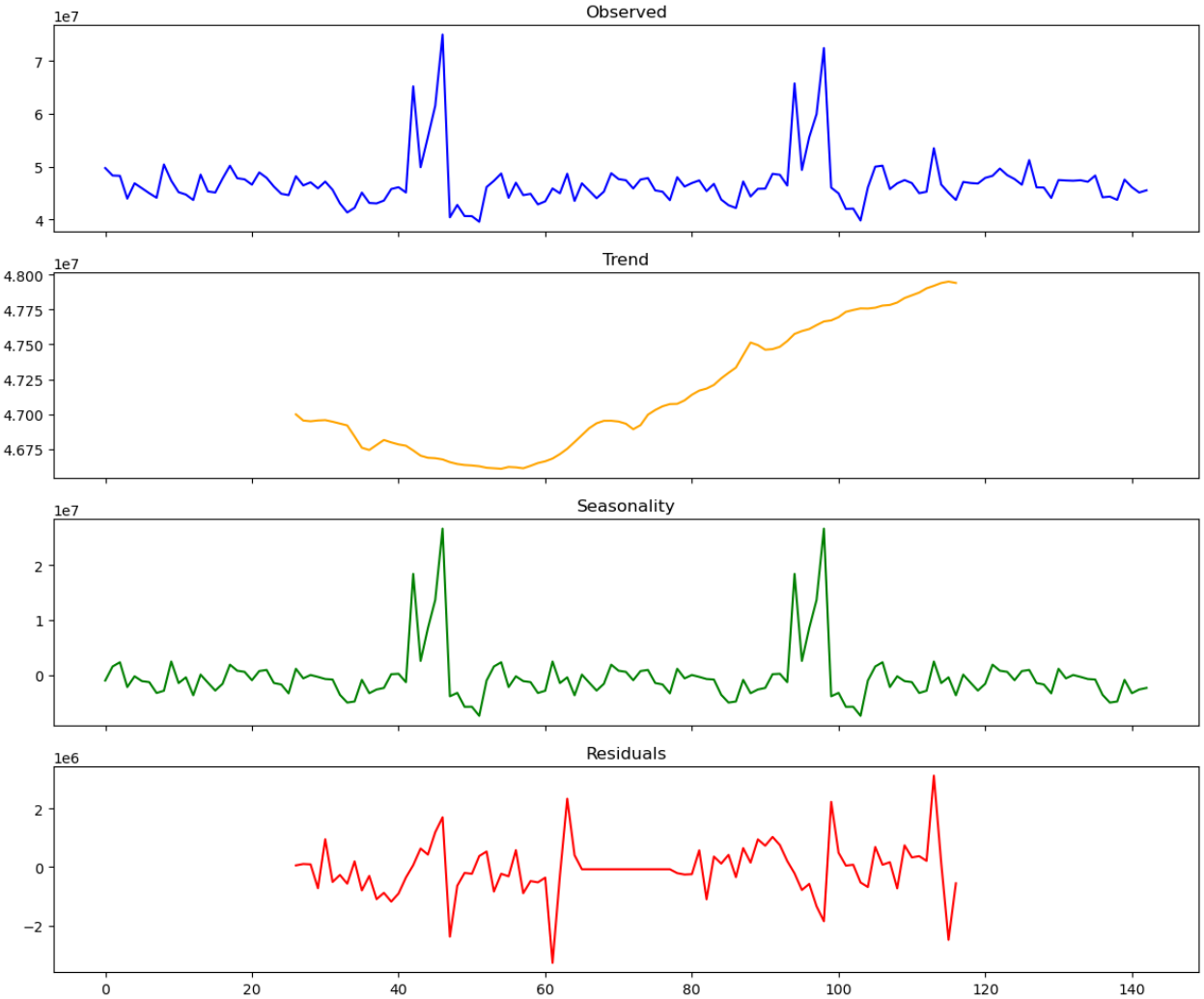
  On the other hand, the lowest-performing stores based on cumulative weekly sales were **Store 38 ($55.2 million), Store 36 ($53.4 million), Store 5 ($45.5 million), Store 44 ($43.3 million), and Store 33 ($37.2 million)**. The comparatively lower sales in these stores could be attributed to factors such as lower foot traffic, regional economic conditions, or variations in store size and product offerings. Further analysis of store-specific characteristics and external influences could provide deeper insights into the observed performance differences.



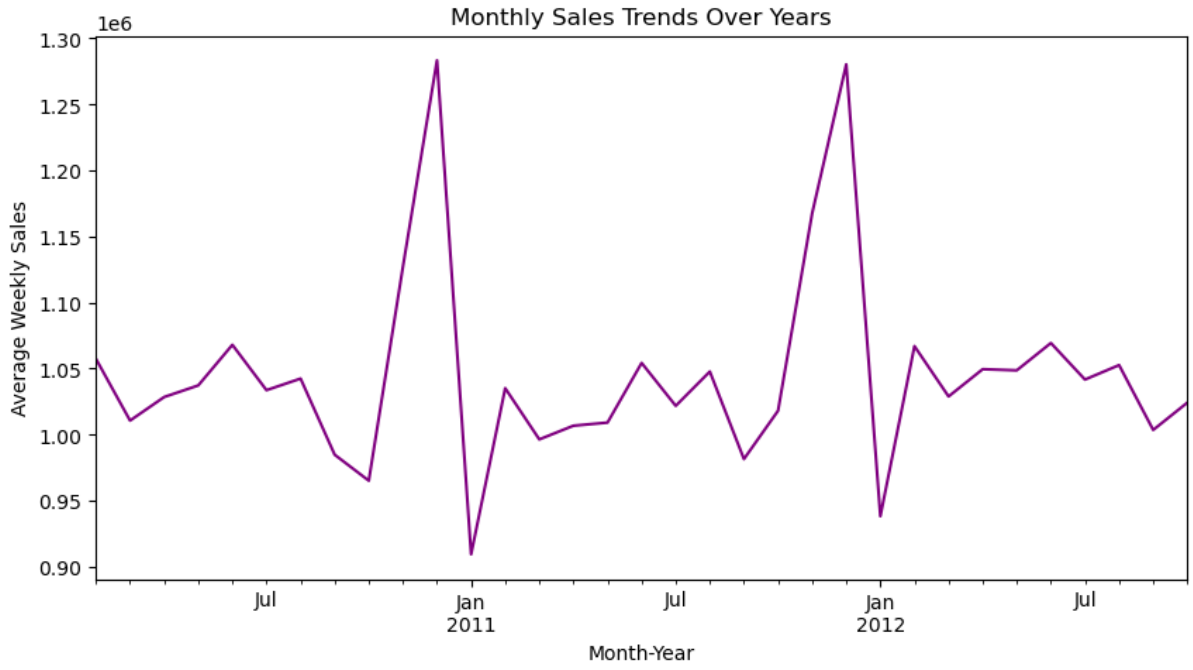


- **Seasonality Analysis:**
  - ***Overall (sum of weekly sales of all stores) weekly sales over time:***





- **Seasonal Spikes**: Noticeable peaks around specific times of the year, likely corresponding to holiday seasons or special events.
- **Downturns**: There are periods with sharp drops in sales, which might coincide with post-holiday slumps or external economic factors.
- **General Trend**: The overall sales seem to hover around a similar range but show fluctuations likely tied to seasonality.
- ***Monthly sales trend:***



- Sales spike during specific months, especially around major holidays like Thanksgiving, Christmas, or New Year. Lower sales could occur in off-seasons (e.g., post-holiday slumps).
- **Potential Reasons:**
- Consumer behaviour during festivals and holidays leads to higher spending.
- Seasonal products (e.g., summer/winter goods) contribute to variability.
- External factors like weather or marketing campaigns (discount seasons) also impact sales.

8. Methodology and Model Selection Process for forecasting model

- **Data Preprocessing:** The Date column was converted to datetime format and set as the index. The raw data (df) was used which had no missing values and the outliers present were assumed to be meaningful.
- **Understanding SARIMAX Model Parameters:**

  The SARIMAX model used in this study is based on AutoARIMA, which optimizes the selection of parameters:

- **(p, d, q): These parameters define the non-seasonal components:**
  - p: Autoregressive order (number of past observations used for forecasting).
  - d: Differencing order (number of times data is differenced to achieve stationarity).
  - q: Moving average order (number of lagged forecast errors used in the model).
- **(P, D, Q, m): These parameters define the seasonal components:**
  - P, D, Q: Analogous to p, d, q but for seasonal trends.
  - m: Seasonal period (52 weeks in this case, assuming yearly seasonality).
- **Model Development:**
  - **AutoARIMA** was used to identify the best (p, d, q) and (P, D, Q, m) values for each store, optimizing based on AIC.
  - The **SARIMAX** model was trained for each store using the identified parameters. (without any X component, i.e. any independent parameter predicting sales)
  - A **dynamic forecasting** approach was used for test set predictions.
  - The **forecast horizon was set to 12 weeks**, with one-step-ahead predictions dynamically updated based on prior forecasts.
- **Model Evaluation and Forecasting:**
  - Train-test split (80-20%) was used to evaluate model performance.
  - RMSE was calculated for each store to assess forecasting accuracy.
  - The best-performing models were selected based on the lowest RMSE values.
  - The next 12 weeks of sales were forecasted for each store.

9. Overview of Model Performance

The forecasting models were evaluated across the 45 stores. The dataset was split into an 80-20 train-test split, and performance was measured using Root Mean Square Error (RMSE).

Key Observations:

- The RMSE values vary significantly across stores, ranging from approximately 15,000 to 470,000, indicating differences in the complexity of sales patterns across locations.
- The best ARIMA orders vary, with some stores requiring more complex models (e.g., (5,0,3)) and others performing well with simpler models (e.g., (0,1,0)).
- The seasonal order is predominantly (1,0,0,52), which aligns with the assumption of yearly seasonality based on weekly data.
- Certain stores, such as Store 14 (RMSE ~ 470,760) and Store 23 (RMSE ~ 209,805), exhibit significantly higher forecasting errors, suggesting possible external influences or high variability in sales patterns.
- Conversely, some stores, such as Store 33 (RMSE ~ 14,814) and Store 30 (RMSE ~ 15,726), exhibit relatively low RMSE, indicating more stable and predictable sales trends.

---

10. Forecasted Sales

The SARIMAX [containing no X component (independent variable predicting sales)] models were used to generate sales forecasts for each store for the next 12 weeks.

Key Forecast Observations:

- The predicted sales values show fluctuations, with some stores exhibiting consistent trends while others display high volatility.
- Stores with high historical RMSE (e.g., Store 14 and Store 23) continue to exhibit large variations in forecasted sales, indicating that external factors may be influencing demand.
- Store 1 shows relatively stable sales trends, fluctuating between 1.5M and 2.3M, suggesting steady consumer demand.
- Smaller stores (e.g., Store 33 and Store 36) have lower sales volumes but maintain more predictable patterns.
- Some stores show seasonal peaks, particularly in the 4th and 8th weeks, aligning with previous observations of seasonality in the dataset.

- **Example Forecast Data (1st 5 stores):**

Store	Week 1 Sales	Week 2 Sales	Week 3 Sales	Week 4 Sales	Week 5 Sales
1	1,854,754	1,767,574	1,720,311	2,141,198	1,758,323
2	2,103,893	2,070,372	2,054,356	2,645,714	2,095,704
3	489,347	448,066	441,222	569,332	501,579
4	2,281,217	2,203,029	2,243,946	3,004,702	2,180,999
5	403,568	370,243	367,080	526,702	421,588

---

11. Implications and Next Steps

- **Business Implications**
- Stores with **high RMSE** require further investigation—external factors like regional economic shifts, promotional events, or supply chain disruptions may be affecting sales.
- Stores with low RMSE have more predictable sales trends, making them ideal candidates for automated inventory optimization.
- Seasonal patterns should be further explored to determine how holidays, promotions, or weather influence sales variability.

- **Potential Model Evaluation:**
- Feature Engineering: Introduce external factors such as holiday indicators, regional economic data, or competitor activities.
- Alternative Models: Test XGBoost, LSTM, and Prophet for comparative analysis.
- Hybrid Models: Combine SARIMAX with machine learning techniques to capture both short-term fluctuations and long-term trends.
- Fine-Tuning Seasonal Parameters: Re-evaluate whether all stores require m=52, as some may have shorter seasonal cycles.

---

12. Conclusion

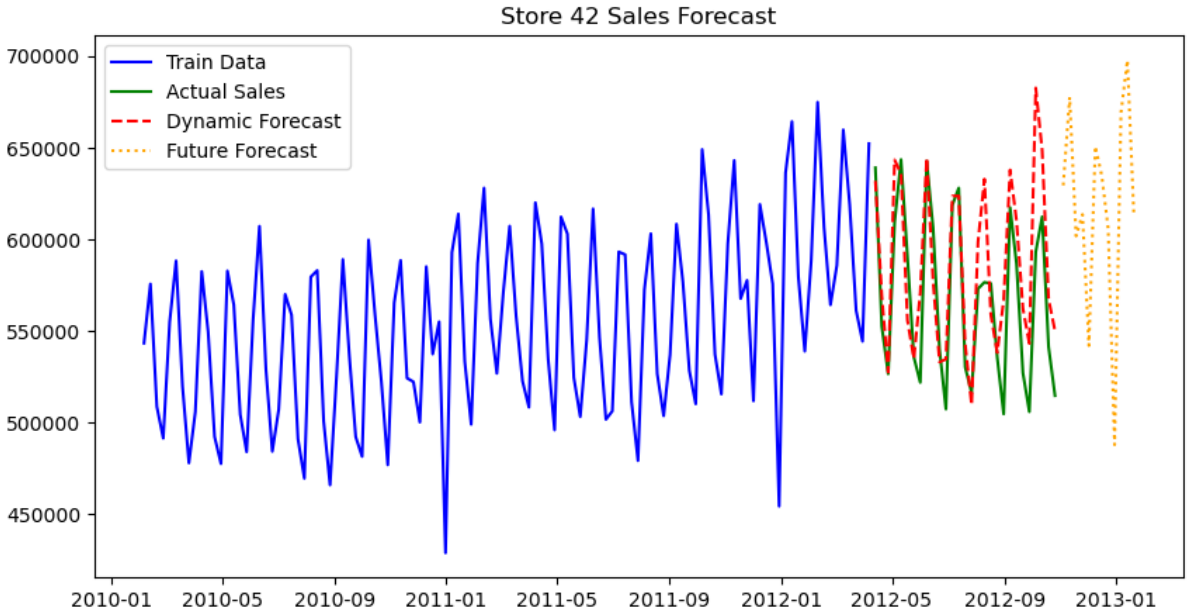
This study provides a data-driven approach to forecasting weekly sales for a retail chain with multiple outlets. While the models demonstrate reasonable predictive capability, improvements can be made by incorporating external features and testing alternative models. Future work should focus on refining model performance, particularly for stores with high forecasting errors, to enhance inventory planning and business decision-making.

13. References

- https://alkalineml.com/pmdarima/modules/generated/pmdarima.arima.ARIMA.html
- https://www.geeksforgeeks.org/winsorization/

14. Appendix

- Example graph of model training, testing and future forecasting for store 42:



- Excel files are also submitted for store wise model summary and future forecasting for 12 weeks.
- **Model summary table:**

Store	Train Size	Test Size	RMSE	Best Order (p,d,q)	Seasonal Order (P,D,Q,m)
1	114	29	132527.7	(1, 1, 1)	(1, 0, 0, 52)
2	114	29	118340.5	(2, 0, 2)	(1, 0, 0, 52)
3	114	29	15916.83	(5, 0, 3)	(1, 0, 0, 52)
4	114	29	109690.5	(0, 0, 1)	(0, 1, 0, 52)
5	114	29	37293.45	(0, 1, 1)	(1, 0, 0, 52)
6	114	29	170715.8	(2, 0, 2)	(1, 0, 0, 52)
7	114	29	36231.95	(0, 0, 1)	(0, 1, 0, 52)
8	114	29	45420.05	(2, 0, 2)	(1, 0, 0, 52)
9	114	29	69956.74	(2, 1, 1)	(0, 1, 0, 52)
10	114	29	87425.08	(2, 0, 2)	(1, 0, 0, 52)
11	114	29	139118	(2, 0, 2)	(1, 0, 0, 52)
12	114	29	81495.04	(2, 0, 2)	(1, 0, 0, 52)
13	114	29	189312.1	(2, 0, 2)	(1, 0, 0, 52)
14	114	29	470760.3	(2, 0, 2)	(1, 0, 0, 52)
15	114	29	50983.42	(2, 0, 2)	(1, 0, 0, 52)
16	114	29	24696.26	(2, 0, 2)	(1, 0, 0, 52)
17	114	29	56120.98	(1, 1, 1)	(1, 0, 0, 52)
18	114	29	124355.9	(4, 0, 2)	(1, 0, 0, 52)
19	114	29	137735.2	(2, 0, 2)	(1, 0, 0, 52)
20	114	29	147309.2	(2, 0, 2)	(1, 0, 0, 52)
21	114	29	42313.14	(1, 0, 2)	(1, 0, 0, 52)
22	114	29	95505.88	(2, 0, 2)	(1, 0, 0, 52)
23	114	29	209805.7	(2, 0, 2)	(1, 0, 0, 52)
24	114	29	117255.1	(2, 0, 2)	(1, 0, 0, 52)
25	114	29	56682.21	(2, 0, 2)	(1, 0, 0, 52)
26	114	29	63527.34	(2, 0, 2)	(1, 0, 0, 52)
27	114	29	99471.52	(2, 0, 2)	(1, 0, 0, 52)
28	114	29	216996.5	(2, 0, 2)	(1, 0, 1, 52)
29	114	29	47474.6	(2, 0, 2)	(1, 0, 0, 52)
30	114	29	15726.54	(1, 1, 0)	(1, 0, 0, 52)
31	114	29	58129.85	(1, 1, 1)	(1, 0, 0, 52)
32	114	29	65330.28	(2, 0, 2)	(1, 0, 0, 52)
33	114	29	14814.34	(0, 1, 0)	(1, 0, 0, 52)
34	114	29	33304.76	(2, 0, 2)	(1, 0, 0, 52)
35	114	29	58723.08	(1, 1, 1)	(1, 0, 0, 52)
36	114	29	23272.2	(2, 1, 3)	(2, 0, 0, 52)
37	114	29	19160.1	(1, 1, 0)	(1, 0, 0, 52)
38	114	29	58523.77	(0, 1, 0)	(0, 1, 1, 52)
39	114	29	156247.5	(4, 1, 0)	(1, 0, 0, 52)
40	114	29	78729.3	(2, 0, 2)	(1, 0, 0, 52)
41	114	29	62356.06	(4, 1, 0)	(1, 0, 0, 52)
42	114	29	32112.6	(0, 1, 0)	(0, 1, 1, 52)
43	114	29	18897.16	(0, 1, 0)	(0, 1, 1, 52)
44	114	29	16123.46	(1, 1, 0)	(1, 0, 0, 52)
45	114	29	44484.01	(2, 0, 2)	(1, 0, 0, 52)