



CAPSTONE PROJECT



PROJECT TITLE:

WALMART SALES ANALYSIS AND FORECASTING

Sekhar Chandra Padhi



Table of Contents:

Introduction

- 1.1 Project Objective
- 1.2 Dataset

Data Cleaning and Preprocessing

- 2.1 Data Loading
- 2.2 Data Type Conversion
- 2.3 Handling Missing Values
- 2.4 Handling Duplicates
- 2.5 Outlier Detection
- 2.6 Outlier Treatment

Exploratory Data Analysis (EDA) and Insights

- 3.1 Unemployment Rate Impact
- 3.2 Seasonal Trends
- 3.3 Temperature Impact
- 3.4 Consumer Price Index (CPI) Effect
- 3.5 Top Performing Stores
- 3.6 Worst Performing Stores
- 3.7 Performance Difference Significance

Sales Forecasting

- 4.1 Data Preparation
- 4.2 Stationarity Check
- 4.3 Model Selection
- 4.4 Model Training and Evaluation
- 4.5 Sales Prediction

Conclusion & FAQ



1. Introduction

1.1 Project Objective

This project aims to analyze historical sales data from Walmart stores to understand the factors influencing weekly sales and to develop a predictive model for forecasting sales for the next 12 weeks. The insights gained from this analysis will help Walmart optimize inventory management, pricing strategies, and promotional activities to improve business performance.

1.2 Dataset

The dataset used for this project is the "Walmart.csv" file. It contains historical weekly sales data for 45 Walmart stores across the United States. The dataset includes the following variables:

- Store: Store number.
- Date: Week of sales.
- Weekly_Sales: Sales for the given week.
- Holiday_Flag: Whether the week is a special holiday week (1 – Holiday week, 0 – Non-holiday week).
- Temperature: Average temperature in the region.
- Fuel_Price: Cost of fuel in the region.
- CPI: Consumer Price Index.
- Unemployment: Unemployment rate.

2. Data Cleaning and Preprocessing

2.1 Data Loading

- The dataset was loaded into a pandas DataFrame using the `pd.read_csv()` function.

2.2 Data Type Conversion

- The 'Date' column was converted to datetime format using `pd.to_datetime()` to facilitate time series analysis.

2.3 Handling Missing Values

- The dataset was checked for missing values using `df.isnull().sum()`. There were no missing values found.

2.4 Handling Duplicates

- The dataset was checked for duplicate rows using `df.duplicated().sum()`. No duplicates were found.

2.5 Outlier Detection

- Box plots were used to visually identify potential outliers in numerical columns like 'Weekly_Sales', 'Temperature', 'Fuel_Price', 'CPI', and 'Unemployment'.

2.6 Outlier Treatment

- Outliers were removed using the Interquartile Range (IQR) method to ensure data quality and robustness of the analysis.

3. Exploratory Data Analysis (EDA) and Insights

3.1 Unemployment Rate Impact

- Analysis revealed that weekly sales tend to be higher when the unemployment rate is between 6 and 10. Stores 7, 12, 14, 1, and 17 were found to be suffering the most due to high unemployment rates.

3.2 Seasonal Trends

- The analysis, supported by the seasonal decompose graphs, strongly suggests an inverse correlation between temperature and weekly sales. Lower temperatures are generally associated with higher sales, while warmer temperatures coincide with lower sales. This pattern is likely influenced by seasonal consumer behavior, where colder weather may drive increased demand for certain products or encourage more indoor shopping activities.

3.3 Temperature Impact

- A weak negative correlation (-0.042) was observed between temperature and weekly sales. While temperature does have an effect, it is likely minimal compared to other factors.

3.4 Consumer Price Index (CPI) Effect

- Analysis revealed that as the CPI increases, weekly sales tend to decrease, highlighting the impact of inflation on consumer spending.



3. Exploratory Data Analysis (EDA) and Insights

3.5 Top Performing Stores



- Stores 1, 2, 4, 10, 13, 14, 20, and 27 were identified as the top-performing stores based on their total weekly sales.

3.6 Worst Performing Stores



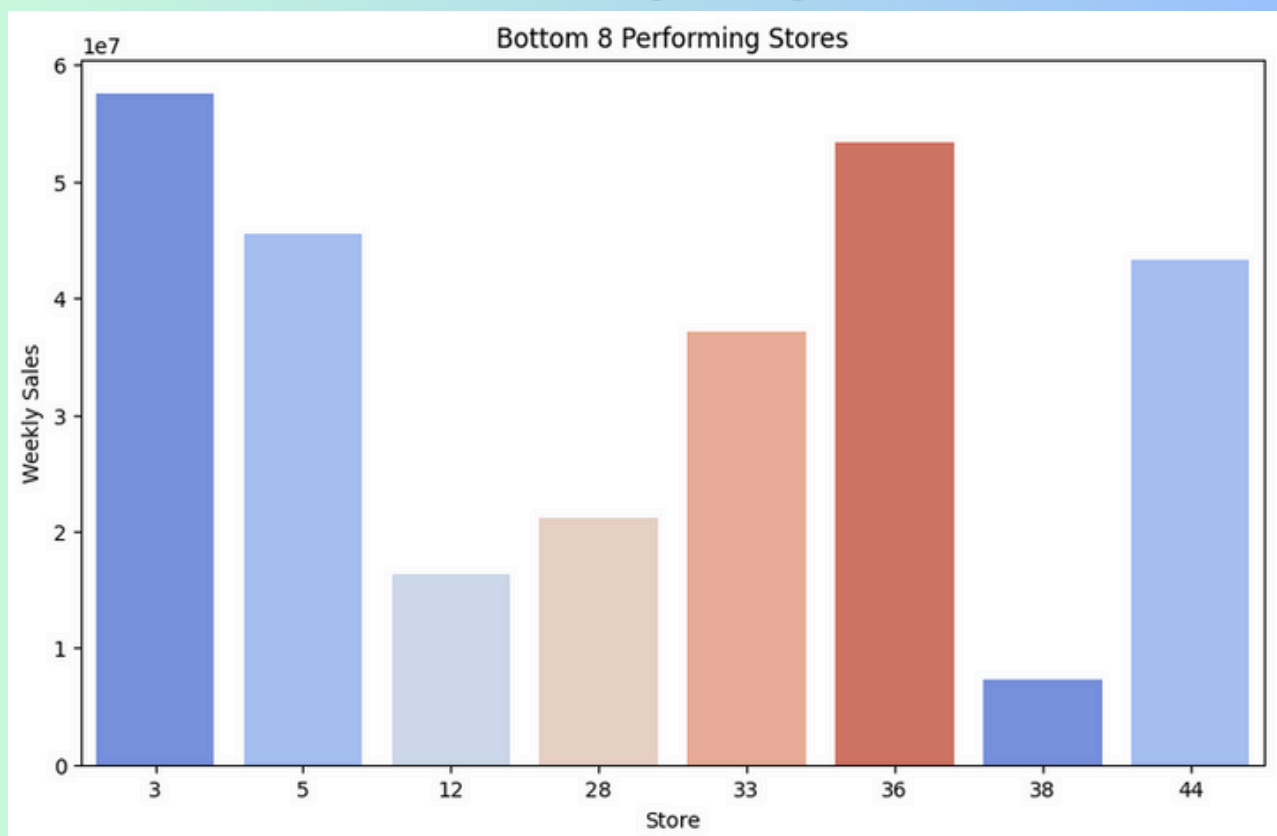
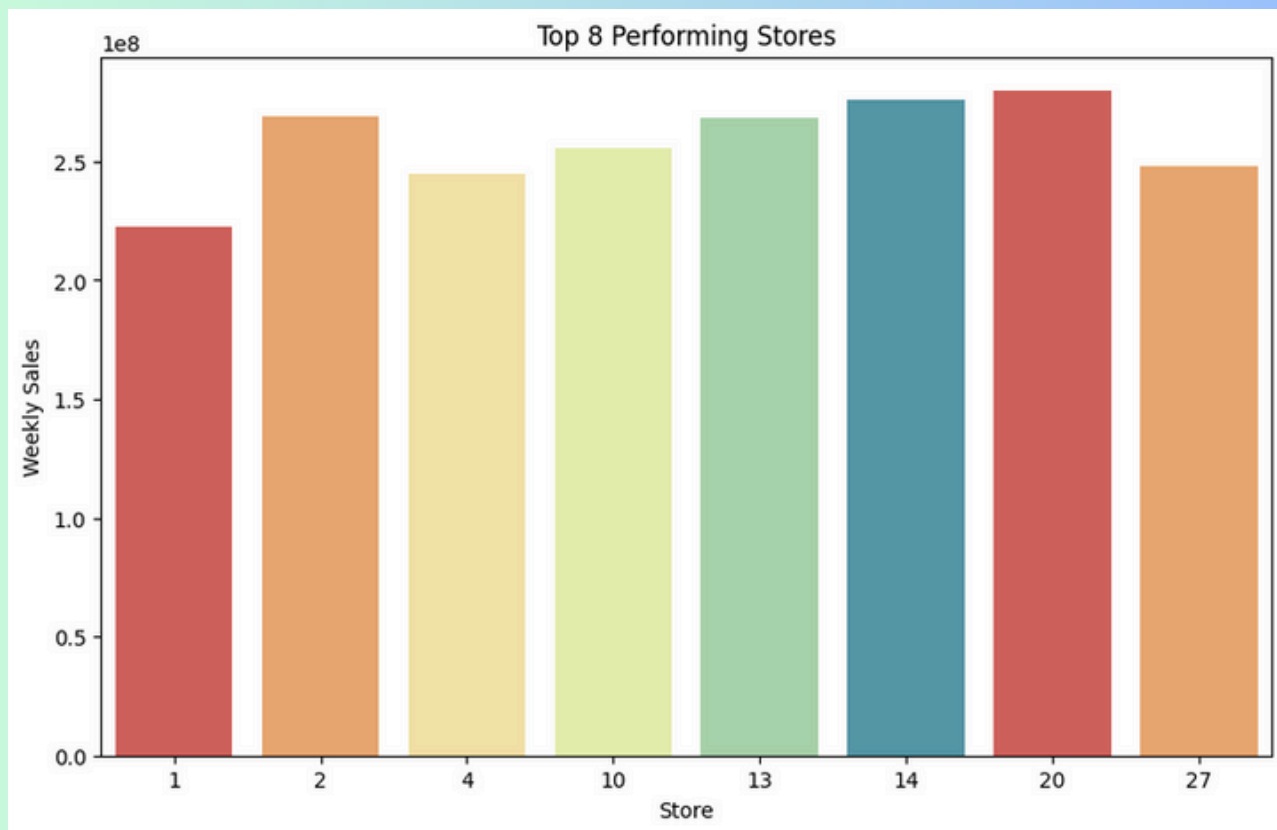
- Stores 3, 5, 12, 28, 33, 36, 3

3.7 Performance Difference Significance

- A t-test was conducted to compare the sales of the top-performing and worst-performing stores. The results indicated a statistically significant difference ($p\text{-value} < 0.05$) between the two groups, confirming that the performance gap is substantial.
- 
- 



visualization



4. Sales Forecasting

4.1 Data Preparation

- For sales forecasting, the dataset was grouped by 'Date' and the total weekly sales were calculated. The 'Date' column was then set as the index.

4.2 Stationarity Check

- The Augmented Dickey-Fuller (ADF) test was performed to check the stationarity of the time series data. The results indicated that the data is stationary, which means there is no need for differencing.

4.3 Model Selection

- ARIMA and SARIMAX models were chosen for sales forecasting due to their ability to handle time series data with trend and seasonality. The `auto_arima` function was used to automatically determine the optimal parameters (p, d, q) for the ARIMA model.






4. Sales Forecasting

4.4 Model Training and Evaluation

- The ARIMA and SARIMAX models were trained on the historical sales data for selected stores (Store 1, 33, 5, and 11). To evaluate the models on a portion of existing data, predictions were made for the last 50 data points, and their fit against the actual values was visualized.

4.5 Sales Prediction

- 
- The trained models were then used to forecast weekly sales for the next 12 weeks for the selected stores. Visualizations were created to display the predicted sales alongside the historical data. In Store 11's case, initial model predictions were found to be inaccurate and a second attempt with `auto_arima` using modified parameters yielded better results.
- 
- 

5. Conclusion

- This project successfully analyzed historical sales data from Walmart stores, identified key factors influencing weekly sales, and developed predictive models to forecast sales for the next 12 weeks.

Key Findings:

- Unemployment rate and CPI have a significant impact on weekly sales.
- There is an inverse relationship between temperature and weekly sales, indicating seasonal trends.
- There is a significant difference in performance between the top and worst-performing stores.

Recommendations:

- Walmart should closely monitor unemployment rates and CPI to adjust pricing and inventory strategies accordingly. Promotional activities should be aligned with seasonal trends to capitalize on sales opportunities.
- Strategies for improving the performance of the worst-performing stores should be explored.

5. Conclusion

Limitations:

- The models were trained on a limited dataset, and their accuracy may vary with new data. External factors not included in the dataset (e.g., competitor actions, economic conditions) could influence future sales.
- This analysis provides valuable insights for Walmart to optimize its operations and improve business performance. Further research and model refinement can enhance the accuracy of sales forecasts and enable more effective decision-making.

FAQ

Given that our data does not exhibit any obvious trend or pattern, why we have used the SARIMAX model?

Hidden Seasonality:

While initial analysis might not reveal obvious seasonality, there could be subtle seasonal patterns that SARIMA can capture. The weekly sales data shows some recurring fluctuations, particularly during holiday seasons or specific times of the year (refer to question (b) inferences), and temperature which is an external factor which usually causes seasonality in this case showing a weak correlation but overall with external factor data and date we had some seasonal pattern and trends. These patterns might not be prominent but can still influence the model's accuracy. SARIMA is designed to identify and model such hidden or complex seasonal patterns.