CAPSTONE PROJECT
WALMART SALES FORECASTING

Table of Contents

1. Problem Statement

Walmart, being one of the largest retail chains globally, generates massive volumes of sales data. This project focuses on forecasting Walmart's weekly sales, considering various factors that influence sales trends, such as holiday effects and seasonal variations. Accurate sales forecasting is crucial for inventory management, financial planning, and ensuring product availability while minimizing overstock scenarios.

Key Goals:

- Improve inventory accuracy and reduce stockouts to please customers.
- Streamline product replenishment to minimize excess inventory and lower costs.
- Enhance demand forecasting to better anticipate customer needs and avoid stockouts or overstocking.
- Develop strategies to handle supply chain disruptions and ensure continuous product availability.
- Use data-driven insights to guide inventory management decisions and foster continuous improvement.

2. Project Objectives

The primary objectives of this project are:

- To understand and preprocess the Walmart sales data.
- To perform exploratory data analysis (EDA) to identify patterns, trends, and outliers.
- To analyse correlations among variables and assess seasonal trends.
- To evaluate the models and recommend the best approach for forecasting future sales.

The important objectives of this project are:

1. You are provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:
    a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
    b. If the weekly sales show a seasonal trend, when and what could be the reason?
    c. Does temperature affect the weekly sales in any manner?
    d. How is the Consumer Price index affecting the weekly sales of various stores?
    e. Top performing stores according to the historical data.
    f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.
2. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

3. Data Description

The dataset comprises weekly sales data from 45 Walmart stores across the United States. The key features include:

Feature Name	Description
Store	Store number
Date	Week of sales
Weekly_Sales	Sales for the given store in that week
Holiday_Flag	If it is a holiday week
Temperature	Temperature on the day of the sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

The dataset spans 33 weeks, offering a broad view of sales patterns over time.

4. Data Pre-processing Steps and Inspiration

Data Preprocessing:

Key preprocessing steps included:

- Converting the Date column to a datetime format.
- Analysing and handling outliers in the Weekly Sales using Z-scores and IQR methods.
- Visualizing sales trends over time to identify any evident patterns.

1. **Loading the Dataset:** Loading the dataset using tools like Pandas is the standard initial step in any data analysis or forecasting project. Ensuring the dataset is accessible for further analysis is crucial.

2. **Checking for Data Types:** It's essential to check data types (e.g., numeric, categorical, datetime) to ensure appropriate handling of each column during analysis. Incorrect data types can lead to issues when performing operations or applying models.

3. **Converting Date Column:** For time series data, converting the date column to a datetime type is important for performing time-based operations, visualizations, and analysis. This is a standard and necessary step in time series preprocessing.

4. **Handling Outliers:** Outliers can distort the results of your analysis, especially in forecasting models. Identifying and handling them properly, either through visualization or statistical methods (like z-scores or IQR), is key to maintaining the integrity of the dataset.

5. **Correlation Analysis:** Utilizing tools like heatmaps aids in understanding the interrelationships between various features within the dataset, providing insights into potential dependencies and guiding further exploration.

6. **Exploring Relationships:** Beyond correlation analysis, exploring relationships between columns through visualizations and statistical methods unveils additional patterns and dependencies, enriching the understanding of the dataset's dynamics.

7. **Time Series Analysis:** Time series analysis, including checking for stationarity (using rolling statistics, ADF test, etc.), is crucial. Many forecasting models, like ARIMA, assume the data is stationary, so confirming and achieving stationarity through differencing or other transformations is essential.

8. **Forecasting Models Selection:** Selecting appropriate forecasting models based on data characteristics is critical for accurate predictions. Each method (e.g., Moving Average, ARIMA) has strengths depending on the data's nature (stationarity, seasonality, etc.). This ensures better alignment between the model and the data.

These steps form a comprehensive and logical data preprocessing pipeline for time series analysis and forecasting projects. Each step builds upon the last, ensuring the dataset is clean, well-understood, and ready for accurate forecasting and analysis.

5. Choosing the Algorithm for the Project

For this project, the ARIMA (AutoRegressive Integrated Moving Average) model was chosen. ARIMA is a popular and widely used forecasting technique for time series data, particularly when there is evidence of non-stationarity in the data. Given that we are analyzing Walmart's weekly sales data, which is time-dependent, ARIMA was an appropriate choice as it models the relationship between an observation and its past values.

6. Motivation and Reasons for Choosing the Algorithm

The primary reasons for selecting the ARIMA model are:

- **Handling Non-Stationary Data**: Weekly sales data often show trends, seasonality, or other patterns over time, making ARIMA a suitable option for capturing these dynamics by differencing the data to make it stationary.

- **Simplicity and Effectiveness**: ARIMA is relatively simple to implement and effective in capturing autocorrelation within time series data, which is crucial for generating accurate forecasts.

- **Forecasting Capabilities**: ARIMA models are excellent for short- to medium-term forecasting, which aligns with the need to predict Walmart's weekly sales for the coming weeks.

- **Modeling Flexibility**: ARIMA allows for fine-tuning parameters (p, d, q) based on autocorrelation and partial autocorrelation plots, enabling it to adapt well to the data structure.

7. Assumptions

The ARIMA model relies on several key assumptions:

- **Linearity:** The model assumes that the time series follows a linear relationship, where future values are a linear function of past values and errors.
- **Stationarity**: Though ARIMA can handle non-stationary data by differencing, it assumes that after differencing, the data becomes stationary (constant mean and variance over time).
- **No Seasonality**: In this specific case, a seasonal component wasn't modeled. If seasonality existed, SARIMA (Seasonal ARIMA) would have been more appropriate.
- **Independence of Residuals**: The model assumes that the residuals (errors) are uncorrelated and normally distributed over time.

8. Model Evaluation and Techniques

To evaluate the performance of the ARIMA model:

- **Residual Analysis**: The residuals (difference between predicted and actual values) were plotted to check for any patterns. Ideally, residuals should be randomly distributed around zero, indicating that the model has captured all underlying patterns in the data. The residuals plot shows no obvious patterns, indicating the model's adequacy.
- **Autocorrelation and Partial Autocorrelation Plots**: Before fitting the model, we examined the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to help determine the appropriate order for the ARIMA model. The order (1, 1, 1) was selected based on these plots, which indicated significant autocorrelations.
- **Summary Statistics**: The ARIMA model's statistical summary was reviewed to ensure parameters were significant and to evaluate model fit.

9. Inferences from the Same

Based on the model output and evaluation:

- **Short-Term Forecast**: The ARIMA model provided a forecast for the next 12 weeks of Walmart's sales. The forecasted values suggest a continuation of the existing sales trend, with no significant changes or spikes predicted in the near term.

- **Residuals Behaviour**: The residual plot showed no significant autocorrelation, indicating that the model has effectively captured the underlying patterns in the time series data. This suggests that the ARIMA model provides a reliable forecast.

- **Model Performance**: While ARIMA performed well for the current dataset, further refinement, such as incorporating seasonality (SARIMA) or testing other model orders, might improve the forecasts if more detailed patterns, like holiday effects or promotions, are present.

10. Forecasting Results

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?

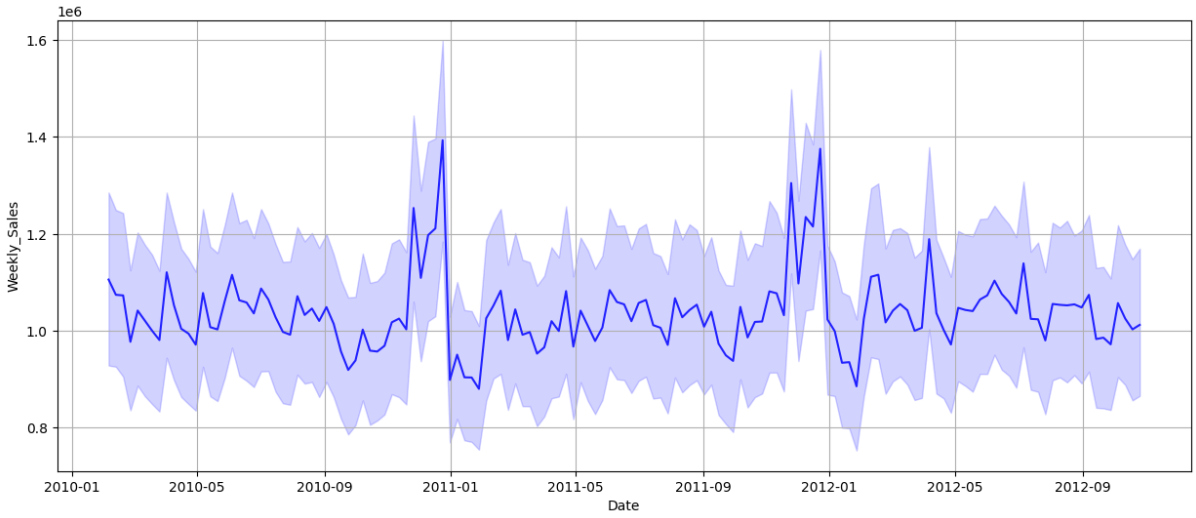
Ans:

Yes**,** Stores suffering most due to unemployment rate:

	Store	correlation
35	36	0.833734
37	38	0.638697
43	44	-0.780076

b. If the weekly sales show a seasonal trend, when and what could be the reason?

Ans:



- Weekly sales show a seasonality in its dataset near the month of December in both years.
- Reason behind this seasonality can be the Christmas festival.
- As per Wikipedia, the economics of Christmas are significant because Christmas is typically a high-volume selling season for goods suppliers around the world. Sales increase dramatically as people purchase gifts, decorations, and supplies to celebrate.

c. Does temperature affect the weekly sales in any manner?

Ans:

- If we consider correlation between weekly sales and temperature there is moderate correlation between them and strong positive or strong negative correlation is not observed.
- while weekly sales are higher when temperature is between 20 and 80.
- Below temperature of 20 weekly sales are dropping significantly.

d. How is the Consumer Price index affecting the weekly sales of various stores?

Ans:

- In above graph we can see three different clusters of weekly sales with specific Consumer Price Index.

- But no correlation is observed between Consumer Price Index and Weekly Sales.

e. Top performing stores according to the historical data.

Ans:

Top performing stores: Store

4 288579000

20 286749000

14 279970600

13 273966900

2 270643600

f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

Ans:

Worst performing store: Store

33 37160221.96

44 43293087.84

5 45475688.90

36 53412214.97

38 55159626.42

- Difference between highest and lowest performing stores: **251418790.3**

Q 2. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

- **12 Weeks Forecast for Top 5 Stores**

Store 4

```
Forecasted values:
 2012-11-02   2137037
2012-11-09   2134985
2012-11-16   2134649
2012-11-23   2134594
2012-11-30   2134585
2012-12-07   2134584
2012-12-14   2134583
2012-12-21   2134583
2012-12-28   2134583
2013-01-04   2134583
2013-01-11   2134583
2013-01-18   2134583
```

Store 20

```
Forecasted values:
 2012-11-02   2039693
2012-11-09   2041900
2012-11-16   2042505
2012-11-23   2042671
2012-11-30   2042717
2012-12-07   2042729
2012-12-14   2042733
2012-12-21   2042734
2012-12-28   2042734
2013-01-04   2042734
2013-01-11   2042734
2013-01-18   2042734
```

Store 14

```
Forecasted values:
 2012-11-02   1647107
2012-11-09   1644110
2012-11-16   1643953
2012-11-23   1643945
2012-11-30   1643944
2012-12-07   1643944
2012-12-14   1643944
2012-12-21   1643944
2012-12-28   1643944
2013-01-04   1643944
2013-01-11   1643944
2013-01-18   1643944
```

Store 13

```
Forecasted values:
 2012-11-02   2026058
2012-11-09   2023823
2012-11-16   2023276
2012-11-23   2023142
2012-11-30   2023109
2012-12-07   2023101
2012-12-14   2023099
2012-12-21   2023099
2012-12-28   2023099
2013-01-04   2023099
2013-01-11   2023099
2013-01-18   2023099
```

Store 2

```
Forecasted values:
 2012-11-02   1866074
2012-11-09   1873410
2012-11-16   1875112
2012-11-23   1875507
2012-11-30   1875598
2012-12-07   1875620
2012-12-14   1875625
2012-12-21   1875626
2012-12-28   1875626
2013-01-04   1875626
2013-01-11   1875626
2013-01-18   1875626
```

➢ Weeks Forecast for Bottom 5 Stores

Store 33

```
Forecasted values:
 2012-11-02    254072
2012-11-09    254297
2012-11-16    254446
2012-11-23    254544
2012-11-30    254609
2012-12-07    254651
2012-12-14    254680
2012-12-21    254698
2012-12-28    254711
2013-01-04    254719
2013-01-11    254724
2013-01-18    254728
```

Store 44

```
Forecasted values:
 2012-11-02    349051
2012-11-09    348042
2012-11-16    347958
2012-11-23    347950
2012-11-30    347950
2012-12-07    347950
2012-12-14    347950
2012-12-21    347950
2012-12-28    347950
2013-01-04    347950
2013-01-11    347950
2013-01-18    347950
```

Store 5

```
Forecasted values:
 2012-11-02    321882
2012-11-09    322758
2012-11-16    323088
2012-11-23    323211
2012-11-30    323258
2012-12-07    323275
2012-12-14    323282
2012-12-21    323284
2012-12-28    323285
2013-01-04    323285
2013-01-11    323286
2013-01-18    323286
```

Store 36

```
Forecasted values:
 2012-11-02    274275
2012-11-09    275255
2012-11-16    275792
2012-11-23    276087
2012-11-30    276249
2012-12-07    276338
2012-12-14    276386
2012-12-21    276413
2012-12-28    276427
2013-01-04    276436
2013-01-11    276440
2013-01-18    276442
```

Store 38

```
Forecasted values:
 2012-11-02    426577
2012-11-09    427962
2012-11-16    428169
2012-11-23    428199
2012-11-30    428204
2012-12-07    428205
2012-12-14    428205
2012-12-21    428205
2012-12-28    428205
2013-01-04    428205
2013-01-11    428205
2013-01-18    428205
```

11. Future Possibilities of Project

There are several potential areas for expanding and enhancing this project:

- **Incorporating Seasonality**: By extending the current ARIMA model to a SARIMA (Seasonal ARIMA) model, we can account for any seasonal patterns that might exist in the weekly sales data, such as increased sales during holidays or promotional periods.

  Exploring Exogenous Variables: Adding external variables, such as marketing spend, promotions, holidays, or economic factors, through an ARIMAX (ARIMA with Exogenous Variables) model, could improve forecast accuracy by including these additional drivers of sales.

- **Longer-Term Forecasting**: The current ARIMA model focuses on short-term forecasts. With more data and further analysis, extending forecasts over a longer horizon could provide strategic value for Walmart's supply chain and inventory management decisions.

- **Advanced Time Series Techniques**: Other machine learning techniques such as LSTM (Long Short-Term Memory networks) or Prophet by Facebook could be explored to compare forecasting performance and accuracy, especially in the presence of more complex patterns or when scalability is required.

- **Real-Time Forecasting**: By implementing a real-time forecasting system, updated with the most recent data, Walmart could make dynamic sales predictions and immediately respond to changing market conditions.

12. Conclusions

In conclusion, this project successfully applied the ARIMA model to forecast Walmart's weekly sales over a short-term period. The model was chosen due to its simplicity and effectiveness in handling time series data with non-stationarity. Based on the analysis of residuals and model fit, ARIMA provided reliable sales forecasts, which can be valuable for operational decision-making.

However, there is still potential for improving the model, such as by incorporating seasonality or external factors influencing sales. Future work could focus on enhancing forecast precision and extending its scope. Overall, the insights gained from this project can serve as a foundation for further development of more sophisticated forecasting tools to help Walmart optimize its sales strategies and inventory planning.

13. References

- Dataset: Walmart Sales Data (provided as part of the project)
- Statistical models: Box, G.E.P., Jenkins, G.M., & Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control*.
- Python libraries: Pandas, NumPy, Scikit-learn, Statsmodels, Matplotlib, Seaborn.