# A blue and white flyer with a blue circle and white text Description automatically generated

# Table Of Content

Introduction………………………………………………………………………… 3

About dataset………………………………………………………………………. 4

Data Preprocessing ………………………………………………………………… 5

Exploratory Data Analysis …………………………………………………………. 6

Stationarity Check …………………………………………………………………. 11

Detail Analysis of models …………………………………………………………. 12

Model Comparison…………………………………………………………………. 23

Monte Carlo Simulations…………………………………………………………… 25

Optimizations………………………………………………………………………. 26

Decision Tree ………………………………………………………………………. 30

Conclusion …………………………………………………………………………. 31

# Introduction

The objective of this project is to analyse and predict Walmart's sales on a weekly basis using advanced time series methodologies. A dynamic retail environment like Walmart's is full of competitions; therefore, it requires very accurate forecasting. Proper prediction will help in optimum inventory management and strategic planning to achieve operational efficiency.

In this project, we will review Walmart's historical sales data using advanced techniques of time series analysis. We will forecast only for store number 7. Time series analysis is one of the procedures essential to decode the underlying structure and trend of any kind of data recorded over time. The main objective will be to develop predictive models that not only capture but also forecast the future trends more accurately using historical sales information.

The key objective of this project will be the application of time series analysis techniques to understand the sales dynamics at Walmart and to build models that provide reliable forecasts. This involves understanding and analyzing trend, seasonality, and irregular fluctuations in a time series. This approach enables actionable insights that might drive critical strategic decisions and enhance operational effectiveness.

In summary, this project is developing robust predictive models of Walmart sales using advanced methodologies in time series to improve inventory management, strategic planning, and business operations in the competitive retail environment.

# Dataset Description:

The dataset contains weekly sales records from 45 Walmart stores, covering weekly dates. It includes several factors that may influence sales, such as whether a holiday occurred, temperature, fuel prices, Consumer Price Index (CPI), and unemployment rates. This dataset enables a time series analysis to explore sales patterns and understand the impact of these factors on Walmart's performance over time.

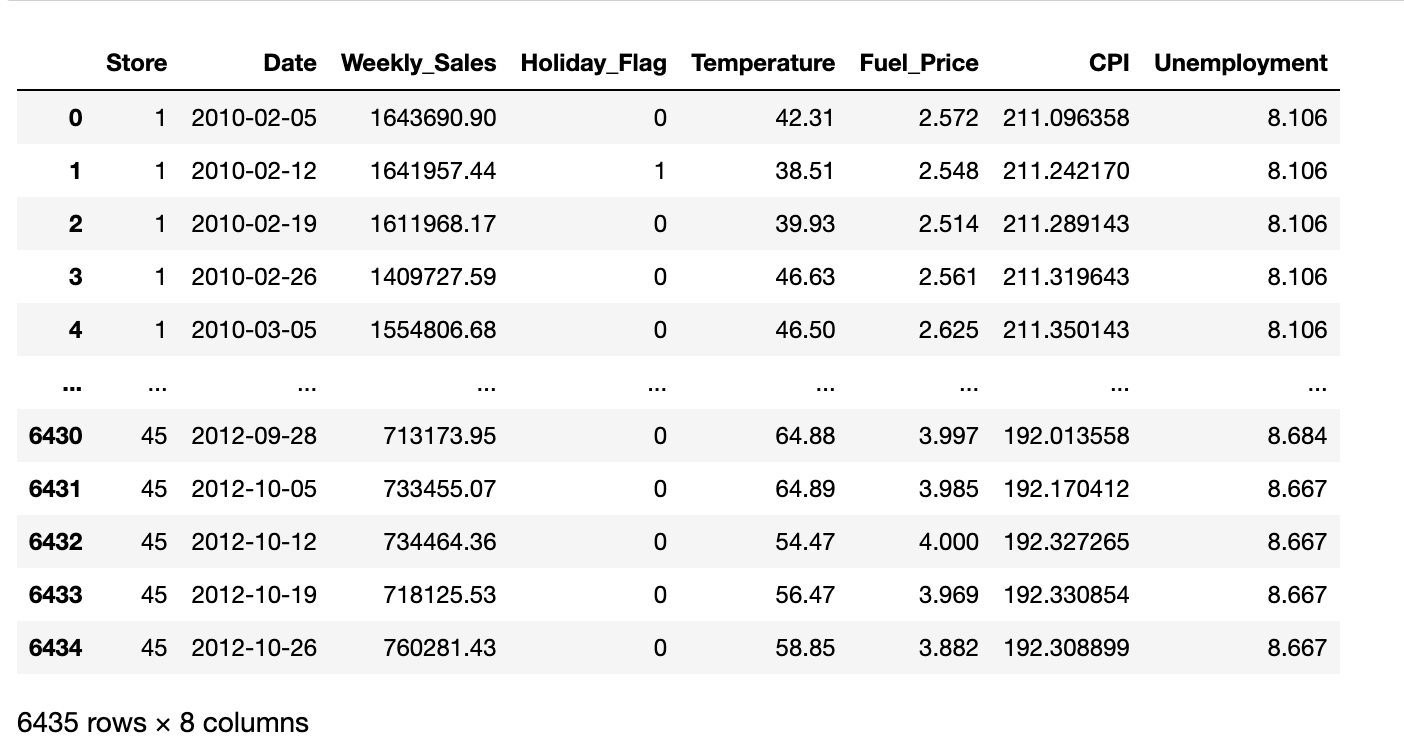
**About Variables:**

|  |
| --- |
| **Response Variable:** |
| **Weekly Sales:** The sales amount for the given week. |

|  |
| --- |
| **Predictors:** |
| **Store:** The store number. |
| **Date:** The date of the sales entry. |
| **Holiday Flag:** Indicates whether the week includes a special holiday. |
| **Temperature:** The temperature on that week. |
| **Fuel Price:** The fuel price on that week. |
| **CPI:** The Consumer Price Index on that week. |
| **Unemployment:** The unemployment rate on that week. |

**About Dataset:**

It includes 6,435 rows and 8 columns, which holds the weekly sales data from the years 2010 to 2012 for different stores. In total, it contains data from 45 different stores. However, we will only forecast Sales Growth for Store: 7.



# Data Pre-processing:

In the data preprocessing phase for our time series analysis, we performed several key steps to clean and prepare the dataset using pandas and dplyr library.

1. **Data Inspection**:

* Used df.head() to view the first few rows of the dataset.
* Checked for missing values with df.isnull().sum().
* Summarized the dataset with df.describe() to get basic statistics.

1. **Data Inspection:**

* Used df.head() to view the first few rows of the dataset.
* Checked for missing values with df.isnull().sum().
* Summarized the dataset with df.describe() to get basic statistics.

1. **Date Handling:**

* Ensured the 'Date' column was in the correct datetime format using pd.to\_datetime().
* Set the 'Date' column as the index of the dataframe with df.set\_index('Date', inplace=True).

1. **Data Filtering:**

* Filtered the data to include only records from store number 7 using df = df[df['Store'] == 7].

1. **Verification:**

* Checked the remaining columns with print(df.columns) and verified the index with print(df.index).

A screenshot of a graph

Description automatically generated

**Exploratory Data Analysis (EDA):**

**Correlation Matrix:**

The correlation matrix helps in understanding the linear relationship between different variables. A higher absolute value of correlation indicates a stronger relationship.

A chart with different colored squares

Description automatically generated

#### Key Observations:

* There is a very high negative correlation between Consumer Price Index with unemployment -0.95, which essentially means that if the Consumer Price Index is high, then the rate of unemployment falls drastically.
* The fuel price is strongly positively correlated 0.79 with Consumer Price Index, indicating that a high gasoline price goes with a higher Consumer Price Index.
* The fuel price, however, has a very high negative relationship -0.79 with unemployment, which means that high fuel prices go in conjunction with lower unemployment rates.
* The temperature is at a reasonable correlation with Fuel\_Price at 0.24 and -0.26 with Unemployment.
* Weekly Sales has weak positive correlations with Fuel Price and Temperature, 0.11 and 0.015 respectively.
* Temperature and CPI are weakly positively correlated at 0.15.

**Boxplot for Holiday Effect:**

**Holiday Impact:** A boxplot is used to compare sales distributions during holidays versus non-holidays (Holiday Flag = 1) and non-holidays (Holiday Flag = 0). It shows how sales differ when a holiday is flagged, revealing any notable differences in sales patterns.

A chart of a graph

Description automatically generated with medium confidence

**Holiday Sales vs. Non-Holiday Sales :**

* The median sales figures during holidays are higher compared to during non-holidays; this means that generally, the weekly sales figures are higher during the holiday periods.
* The IQR and range for sales holidays are much bigger. This implies that although sales are expected to be much higher with holidays, these will also experience more variability.
* There are a few weeks of non-holiday periods with far greater sales than the range that might be an indication of occasional peaks in sales outside of holidays. (outliers in non- holiday sales)

**Rolling Statistics:**

**Rolling Mean and Standard Deviation:** Calculations of rolling mean and standard deviation over a time help in understanding trends and variability in sales data. The rolling mean provides a smoothed view of sales over time, while the standard deviation shows the variability around this mean. By highlighting fluctuations, the rolling standard deviation helps us understand the underlying patterns in the data. This measure shows how much the sales figures vary over time, making it easier to spot trends and make better decisions based on the observed changes.

A graph showing a line of sales

Description automatically generated with medium confidence

Key Observation:

* Visible spikes in weekly sales towards the ends of the years 2010 and 2011 indicate more sales during holiday seasons.
* The orange line shows the moving average, smoothing out the weekly sales data to give the overall trend and periodic highs and lows corresponding to the seasonal troughs and peaks in sales.
* The shaded area represents the rolling standard deviation, indicating the variability in weekly sales. The variability of sales increases during peak sales periods; the fluctuation of sales around holidays has higher variation.

**Time series plot and Outlier Detection:**

A time series plot displays data points over time, showing how a variable changes at different intervals. It helps visualize trends, seasonal patterns, and fluctuations in the data, making it easier to identify underlying patterns and predict future values.

**Outlier Visualization:** A plot with outliers highlighted helps in visualizing and understanding unusual sales spikes or drops.

* **Z-Score Calculation:** The Z-score is used to identify outliers in the Weekly\_Sales data. Outliers are flagged where the absolute Z-score exceeds 3, indicating unusually high or low sales compared to the mean.

A graph of blue lines with red dots

Description automatically generated

**Key Observations:**

The sales have been highly variable over the full period. These peaks and troughs show periods when sales are either high or low.

* The first outlier is somewhere towards the end of 2010, with a sales value of approximately 1 million.
* Another outlier is at the end of 2011 or beginning of 2012, peaking close to 1 million in sales.
* The presence of both the significant spikes in sales at certain times—maybe due to promotions, holidays, or other factors—and periods with more stable sales levels.
* There may be some seasonality or periodicity, for there seem to be recurrent peaks around the same time of every year. For example, at the end of 2010 and at the end of 2011. Moreover, we will get into a more accurate seasonal pattern in the Decomposition part.

**Time Series Decomposition:**

**STL Decomposition:** In this step, the time series data will be decomposed into trend, seasonal, and residual components using the STL decomposition. This will help in separating and showing possible trends, seasonal patterns, and residual noise underlying the data. It uses locally weighted regression Loess to handle complex, nonlinear trends and seasonality that is varying in an uneven manner. It is the flexibility that empowers it to model a wide array of seasonal effects, and its robustness makes it less sensitive to outliers and irregularities in the data; adaptability assures handling of changing seasonalities, making the method very versatile and suitable for general time series analysis.

A graph of a graph

Description automatically generated with medium confidence

**Key Points:**

* **Original Series**:
  + This is the raw time series data.
  + The data appears to fluctuate significantly, indicating high variability.
* **Trend Component**:
  + The trend component: This is the long-term progression of the series.
  + First, from the beginning of 2010, it shows a downward trend up to early 2011.
  + The trend stabilizes and comes up a bit more following early 2011, then heads down again towards the end of the period.
  + The trend, if anything, is downward over the whole period.
* **Seasonal Component**:
  + The seasonal component captures the repeating short-term cycles within the data.
  + There is a regular pattern observed, which indicates seasonality in the data.
  + The seasonal pattern seems to be relatively consistent in its amplitude and frequency throughout the observed period.
  + Peaks and troughs are regularly spaced, indicating a strong seasonal effect that repeats over time.
* **Residual Component**:
  + The residual component holds the random noise in the data after netting off the trend and seasonal components.
  + The residuals, which are normally distributed around zero, prove that the trend and seasonal components have been suitably isolated through decomposition.
  + The residuals show extreme variance, thus indicating other factors or random noise in the data that are not accounted for within the trend or seasonal components.
  + There are some large spikes, particularly around early 2012, which could indicate outliers or anomalies in the data.

**Autocorrelation Analysis:**

**ACF and PACF Plots:** Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are created to examine the relationships between lagged values of the time series. These plots help in identifying appropriate parameters for ARIMA or SARIMA models. A comparison of graphs with numbers

Description automatically generated with medium confidence

**Key Observations:**

* **ACF:** The gradual decline in the ACF plot suggests the presence of a Moving Average (MA) component in the time series. Specifically, the significant lags in the ACF plot indicate the order of the MA component. The plot suggests an MA component of order 4 or 5.
* **PACF:** The significant spike at lag 1 in the PACF plot suggests the presence of an AutoRegressive (AR) component. The plot suggests an AR component of order 1 or 2.

**Stationarity Check**

**ADF Test:**

The ADF statistic is used for testing for a unit root in a time series to ascertain whether it is stationary or not. This is a right-tailed test where the null hypothesis is that the series has a unit root, with the alternative hypothesis being that it is stationary. Critical values compare with the ADF statistic to allow for making a decision on rejecting or otherwise of the null hypothesis. If the statistic is less than the critical value, one decides that it is a stationary time series and does not require further differencing. • Stationarity: When the ADF statistic and p-value confirm, it is a stationarity in the series. This means that the variance and mean of the series are constant, and hence the series is appropriate for many time series forecasting models.

* ADF Statistic: -4.476254338547259
* p-value: 0.0002170071890711737
* Critical Values: {'1%': -3.479742586699182, '5%': -2.88319822181578, '10%': -2.578319684499314}
* The time series is likely stationary.It means if adf of weekly sales is lower than the 0.05 then the time series is likely to stationary.

Here, p-value is 0.000217 which is less than the 0.05, it means, it is stationary.

### 8. Detailed Analysis of Models:

1. **ETS Model:**

ETS (Error-Trend-Seasonality) models are a class of time series forecasting methods that decompose data into three key components: error, trend, and seasonality. These models are particularly useful for handling time series data with distinct trend and seasonal patterns. The error component captures the random noise or irregular fluctuations that cannot be attributed to the trend or seasonality.

ETS models are favored for their ability to handle complex patterns by adapting to changing trends and seasonal effects over time. They simplify the forecasting process by breaking down the time series into its fundamental components, allowing for intuitive and accurate forecasts. There are two main types of ETS models: additive, which is suitable for data with constant trend and seasonal effects, and multiplicative, which works well for data where these components vary proportionally. Forecasts are generated by combining these components, either through addition in additive models or multiplication in multiplicative models.

ETS model generated by combining Level, Trend and Seasonality:

Additive ETS model: Multiplicative ETS Model:

Forecast = Level + Trend + Seasonality Forecast = Level \* Trend \* Seasonality

In the exploration of Exponential Smoothing State Space Models (ETS), various parameter combinations were tested to identify the most suitable model for forecasting Walmart's weekly sales. The tested combinations included both additive and multiplicative trends and seasonal components, with seasonal periods set to 52 weeks to capture annual patterns. By evaluating multiple parameter configurations, the model with the lowest Root Mean Squared Error (RMSE) was selected, ensuring the best fit for the historical sales data. This rigorous process allowed us to identify the optimal ETS model for accurate and reliable sales forecasting.

Basically, in time series forecasting based on the ETS model, various combinations of error, trend, and seasonality are fitted against the series to land the best combination. Major combinations would be additive trend with additive seasonality, additive trend with multiplicative seasonality, multiplicative trend with additive seasonality, and multiplicative trend with multiplicative seasonality. The best option gives forecast accuracy and performance to understand the underlying pattern of data better for more reliable forecasting.

|  |  |  |
| --- | --- | --- |
| Combinations | RMSE | MAE |
| Add, Add | 25299.56 | 15525.52 |
| Add, Multiple | 24790.32 | 15560.12 |
| Multiple, Add | 25337.87 | 15557.20 |
| Multiple, Multiple | 24831.22 | 15575.32 |

Best combination we get out of all is (Add, Multiple). Because it has lower MAE and RMSE. Below is the graph of forecasting with best combination of (Add, Multiple).

A graph of a graph

Description automatically generated with medium confidence

**Key Observations:**

* The fitted values closely follow the observed data, indicating that the model captures the underlying trend and seasonality quite well.
* The forecast shows a significant increase in sales, with a large spike towards the end of 2012 and into 2013.
* Peaks typically occur around the same time each year, possibly indicating a seasonal demand or event-driven sales pattern.
* However, the forecasted data shows more volatility, indicating that future sales could be subject to higher uncertainty or a potential shift in the trend.
* The sales trend appears relatively stable with some fluctuations, but there is a noticeable spike in the forecasted period, suggesting either an expected event or a significant change in market conditions.

**Summary of (Add, Multiple) combinations:**

**A screenshot of a computer

Description automatically generated**

**Limitations:**

* **Limited Handling of Complex Patterns:** ETS models might struggle with complex patterns or non-linear relationships in the data that aren't easily captured by simple additive or multiplicative trends and seasonality.
* **Seasonal Component Limitation:** While ETS can handle seasonality, it might not be suitable for very irregular or highly variable seasonal patterns.

1. **ARIMA (Auto Regressive Integrated Moving Average)**

* ARIMA—Autoregressive Integrated Moving Average—a statistical model used for the forecasting of time series data. This technique puts together three basic constituents: autoregression, which makes use of past observations in predicting future values; differencing, a method by which trends and seasonality are removed to stationarize the data; and moving averages, which models the relationship relating an observation to past forecast errors.
* This finds quite useful applications in sales forecasting and stock market analysis, among others, since it deals with non-stationary data and can accommodate different time series patterns. It therefore entails the identification of parameters for the AR, I, and MA components, fitting of the model, and then making a prediction for the future.
* While ARIMA models complex relationships and non-stationary data, they can also be more difficult to tune and expand to include seasonality. ETS is simple and does well with time series containing plain seasonal patterns; thus, it is easier to implement in the presence of seasonality and trends.
* Now, for the performance of the Arima checking, stationarity is a very important one. Check whether the time series is stationary or not; that is, its statistical properties don't change over time.
* Either with the help of tools using ACF and PACF plots explaining how it changes, or by a statistical test like the Dickey-Fuller test, check whether it is not stationary. Apply differencing to make it stationary if it is not.
* Then analyze the ACF and PACF plots to select the appropriate values for 'p' and 'q'. These plots are useful in checking the extent of autocorrelation and partial autocorrelation in the data.
  + **AR (p):** Number of lag observations included in the model.
  + **I (d):** Number of times the data is differenced to achieve stationarity.
  + **MA (q):** Size of the moving average window used to model errors.

We perform some combinations based on observation and select best combination.

|  |  |  |
| --- | --- | --- |
| Combinations (p, d, q) | RMSE | MAE |
| (3, 0, 5) | 81906.21 | 51686.1 |
| (2, 0, 5) | 82100.90 | 51825.99 |
| (3, 1, 5) | 101009.10 | 60430.45 |
| (2, 1, 5) | 101677.72 | 61951.01 |
| (2, 0, 4) | 82097.46 | 51820.92 |
| (3, 1, 4) | 103140.95 | 63103.85 |

We made a forecast using the best combination in the ARIMA model:

A graph of a graph

Description automatically generated with medium confidence

**Key Observations:**

* The model used is an ARIMA(3, 0, 5), indicating an autoregressive component of order 3, no differencing, and a moving average component of order 5.
* The fitted values generally stay very close to the observed data; there are some discrepancies, particularly around sharp peaks and troughs.
* The forecast, relative to the previous data, is quite flat. That means the model believes in a stable, though slightly growing, sales trend without the peaks and troughs of the past.However, the model appears to struggle with the extreme spikes observed in the sales data, often underestimating or smoothing these sharp changes.
* Unlike the previous model (with trend=add and seasonal=mul), the ARIMA model does not explicitly model seasonality, which may explain why it struggles with some of the periodic spikes.
* The differences between the observed and fitted lines indicate that there may be some patterns in the residuals, particularly where the model fails to capture the peaks and troughs.

**Summary of (3, 0, 5) Combination:**

A white paper with black text and numbers

Description automatically generated

**Limitations:**

* **Requires Stationarity**: ARIMA models require the time series to be stationary, which means the statistical properties of the series (mean, variance) should be constant over time. Non-stationary series often require differencing or other transformations, which can complicate modeling.
* **Limited Handling of Seasonality**: Standard ARIMA models do not account for seasonal effects directly. Handling seasonality requires extending the model to SARIMA, which adds complexity.

**Seasonal ACF and PACF plots:**

A seasonal ACF plot is essential for analyzing time series data with seasonality. It detects repeating cycles, identifies significant autocorrelation lags (e.g., annual or quarterly), and helps in specifying and fitting models like SARIMA. It also aids in diagnosing seasonality issues and informs data preprocessing, leading to improved forecasts and analyses. By highlighting these seasonal effects, it ensures that the models capture the true underlying patterns, which enhances the accuracy and reliability of predictive insights. Additionally, it provides a visual tool for assessing the effectiveness of seasonal adjustments made to the data.

A graph showing a line graph

Description automatically generated with medium confidence

**Key Observations:**

* Seasonality: The ACF plot reveals a clear seasonal pattern, with significant positive autocorrelation at specific lags and damped oscillations as the lag increases.
* Initial Lags: The first few lags show strong positive correlation, indicating that the time series is highly influenced by its recent past values.
* Cyclical Behavior: The consistent peaks and troughs at regular intervals confirm the presence of strong seasonal cycles in the data.
* Damping Effect: The decreasing magnitude of autocorrelation with increasing lag suggests that the seasonal effect becomes less pronounced over time but remains detectable.

**Seasonal PACF:**

A seasonal PACF (Partial Autocorrelation Function) plot is useful for understanding seasonal patterns in a time series. It shows the direct relationship between a variable and its lags, while ignoring the influence of other lags. This helps identify key seasonal lags, set up seasonal models like SARIMA, and pinpoint true seasonal effects. By doing so, it helps fix seasonality problems and improve model accuracy, leading to better forecasts and analyses. Additionally, it clarifies which seasonal terms are most relevant for the model and helps avoid overfitting by focusing on direct seasonal influences.

A graph with blue dots and numbers

Description automatically generated

**Key Observations:**

* Strong Initial Correlation: The first lag shows a very high partial autocorrelation, indicating a strong direct influence of the most recent past value on the current value.
* Damping Effect: The significant correlations diminish rapidly after the first few lags, suggesting that more distant past values have less direct influence on the current value.
* **Occasional Significant Points: There are a few sporadic significant partial autocorrelations beyond the initial lags, which might indicate occasional direct relationships that could be of interest for further analysis.**

**3.SARIMA (Seasonal Auto Regressive Integrated Moving Average)**

**SARIMA—Seasonal AutoRegressive Integrated Moving Average—is an extension of the ARIMA model in order to account for seasonality in time series data. While ARIMA is efficient with non-seasonal data by modeling trends and patterns, SARIMA combines additional seasonal components: seasonal autoregressive, seasonal differencing, and seasonal moving average, along with a parameter that defines the length of the seasonal cycle. This enhancement enables SARIMA to precisely pick up and carry out periodic trends, like spikes in sales during certain months or economic shifts every quarter. Compared to ARIMA, which may require additional preprocessing steps to become resilient against these effects, SARIMA makes it easier to model data with a strong seasonal component. Hence, SARIMA is especially suitable for time series data that has noticeable seasonal cycles, increasing the accuracy of the forecast in various applications with regular cycles.**

To use SARIMA, first identify the parameters: ARIMA parameters (p, d, q) and seasonal parameters (P, D, Q, s), where s represents the seasonal cycle length. Fit the model by ensuring stationarity of the series, using ACF and PACF plots of Walmart sale for non-seasonal parameters, and seasonal plots for seasonal parameters. Finally, use the fitted SARIMA model to forecast future values, accounting for both seasonal and non-seasonal components.

**Non - seasonal Parameter:**

* **AR (p):** The number of lagged observations included in the model for non-seasonal effects.
* **I (d):** The number of times the data is differenced to make it stationary, removing trends.
* **MA (q):** The size of the moving average window for non-seasonal residuals.

**Seasonal Parameter:**

* **SAR (P):** The seasonal autoregressive order, capturing the effect of past values at seasonal lags.
* **SI (D):** The seasonal differencing order, which differences the series at seasonal intervals to handle seasonal non-stationarity.
* **SMA (Q):** The seasonal moving average order, modeling the seasonal residual errors.
* **s (Seasonal Period):** The length of the seasonal cycle (e.g., 12 for monthly data with yearly seasonality).

**Identify Parameters by ACF and PACF plot:**

* The parameters for the nonseasonal AR, I, and MA are determined using ACF and PACF plots.
* Seasonal plot and seasonal ACF/PACF plots can be used to specify the seasonal AR, I, and MA parameters and determine the seasonal period.

We perform some combinations based on observation and select best combinations out of that.

|  |  |  |  |
| --- | --- | --- | --- |
| ARIMA parameter | SARIMA parameter | RMSE | MAE |
| (2, 0, 5) | (1, 0, 2, 52) | 84859.76 | 46937.28 |
| (2, 0, 5) | (2, 0, 2, 52) | 84854.17 | 47056.99 |
| (2, 1, 5) | (1, 0, 2, 52) | 92543.05 | 54540.29 |
| (2, 1, 5) | (2, 0, 2, 52) | 92543.67 | 54541.47 |

We made a forecast using the best combination in the SARIMA model:

A graph of a graph

Description automatically generated with medium confidence

**Key Observations:**

* The model used is SARIMA (Seasonal ARIMA) with non-seasonal parameters (2, 0, 5) and seasonal parameters (1, 0, 2, 52).
* The fitted values generally align well with the observed data, particularly capturing both the trend and some seasonal variations.
* The periodic peaks and troughs in the data are reflected in both the fitted and forecasted values, aligning with the yearly seasonality.
* The forecast indicates a moderate and steady pattern, reflecting both the underlying trend and seasonal factors without the sharp spikes observed in the historical data.
* However, some of the sharp spikes in the observed data are still not fully captured, implying that while the SARIMA model is a good fit, there might still be some unexplained variability in the data.
* The forecast is more reliable, incorporating seasonal variations and suggesting that future sales will follow a similar seasonal trend with moderate fluctuations.
* Despite its strengths, the model still shows some limitations in capturing extreme spikes, indicating potential areas for further refinement or consideration of additional factors in the model.

**Limitations:**

* More complex to implement and requires more computational resources.
* **Overfitting Risk**: The increased number of parameters in SARIMA models can lead to overfitting, particularly if the model is too complex relative to the amount of data available.

1. **Linear Regression**

In linear regression, the relationship between the dependent variable and one or more independent variables is modelled. Such independent variables in the case of this problem include temperature, fuel price, CPI, and unemployment. It assumes that there is a straight-line relationship between the predictors and the outcome. This model would thus help to know how changes in these independent variables impact sales. Although useful in understanding the influence of different factors on sales, this may be better done by using time-series models. Forecasts made with linear regression rest on such linear relationships, which provide valuable insights into how various factors impact sales. Linear regression in time series analysis can help learn and quantify the various factors affecting the time series data. It also provides a naïve way to make out-of-sample forecasts from the historical relationship established. Here is a step-by-step breakdown of how it works:

**Feature Engineering**: Preparing the dataset Create features of the independent variables and transformations, if needed. For example, you can make time series features using lag values.

Equation for linear regression that includes more than one independent variable to predict a dependent variable on a time series

yt= β0+ β1X1,t + β2X2,t + β3X3,t +⋯+ βnXn,t + εt

Where, β0 is intercept, β1, β2, …, βn are coefficients

ϵt is the error term.

Yt: Dependent variable at time ttt (for example, weekly sales)

X1,t, X2,t, X3,t, …, Xn,t : Independent variables at time t (like temperature, fuel

price, CPI, unemployment rate).

We got some values from this model:

|  |  |
| --- | --- |
| Mean Absolute Error **(MAE):** | 74507.77 |
| Root Mean Squared Error **(RMSE):** | 99118.12 |
| R-squared: | 0.343 |
| Adj. R-squared: | 0.308 |

A graph of a graph showing the growth of the stock market

Description automatically generated with medium confidence

**Key Observations:**

* The sales data has numerous sharp peaks and troughs that call for highly volatile periods of high and low sales.
* The model seems to capture the general trend of the training data, but it does not quite fit perfectly with peaks and troughs because of its linearity.
* There is also high volatility in the test data as observed in the training data.
* The forecast would seem to hold for the general trend of the test data but fails to account for the volatility of the observed test data.
* The model is doing a reasonable job of capturing the overall trend. The linear regression model was having problems coping with the extreme peaks and troughs seen in the sales data.
* It might be the case that linearity is too simple for a model to display all the intricate patterns in the sales data, and more complex models—say, nonlinear or seasonal ones—may show better performance.

**Summary of Linear Regression:**

**A screenshot of a computer screen

Description automatically generated**

**Limitations:**

* **Complex Patterns:** Linear regression does not perform as well in modeling complex, time-dependent patterns or seasonality as it would in more specialized time-series models, like ARIMA or SARIMA.
* **Assumption:** Actually, the accuracy of the forecasts depends on how appropriate these assumptions made by the model are. If these assumptions are violated, then the forecasts may be less reliable.
* May not capture complex patterns and interactions.

### Model Comparison

This comparison of models is done to determine the one that works best for the data and gives an accurate prediction. All the models are going to have their strengths and weaknesses. We use several metrics, including RMSE—Root Mean Squared Error—and MAE—Mean Absolute Error—as a way to show which one is the best. RMSE gives a higher weight to larger errors, while MAE is simply the average magnitude of errors without considering their direction.

By checking the RMSE and MAE values for each model, we can identify which model has the lowest values for these metrics. The model with the smallest RMSE and MAE is typically selected as it indicates the best fit and most reliable predictions. This approach helps ensure that we choose a model that is both accurate and easy to use, while considering the strengths and weaknesses of each model.

## Comparison of Models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Best Combination | RMSE | MAE |
| ETS | (Add, Multiple) | 24790.32 | 15560.12 |
| ARIMA | (3, 0, 5) | 84047.84 | 52336.94 |
| SARIMA | (2, 0, 5), (1, 0, 2, 52) | 84859.76 | 46937.28 |
| Linear Regression |  | 99118.12 | 74507.77 |

**Model Suitability:**

* **ETS** is particularly effective when the data has clear trends and seasonal patterns, which it captures well with its additive and multiplicative components.
* **ARIMA** is useful for data without strong seasonality but struggles when seasonal patterns are present.
* **SARIMA** improves upon ARIMA by incorporating seasonality, making it more suitable for data with seasonal fluctuations, though it may still miss some complexities.
* **Linear Regression** is not well-suited for this time series data, as it likely fails to account for non-linear trends and seasonality.

**Results:**

Based on the performance metrics, the **ETS model** is the best option for forecasting, consistently outperforming other models in terms of both RMSE and MAE. However, here, seasonality is a critical component that needs to be explicitly modeled, **SARIMA** would be the next best choice, as it is specifically designed for such tasks. Despite SARIMA's theoretical advantage in handling seasonality, the ETS model might still be preferred if it captures seasonal patterns sufficiently and performs significantly better in practice.

Ultimately, the decision could involve testing both models on out-of-sample data to determine which one generalizes better and provides more accurate forecasts in the long run. **ARIMA** could be considered for datasets with less pronounced seasonality, while **Linear Regression** should be avoided due to its poor performance.

**Monte Carlo Simulations:**

Monte Carlo methods are computational techniques that use random sampling to estimate numerical results, inspired by the randomness of the Monte Carlo Casino in Monaco. They are useful for solving complex problems that are difficult to address analytically and are applied in finance, engineering, physics, and statistics.

These methods work by generating numerous random samples to explore possible outcomes. The process involves defining the problem, producing random samples, simulating outcomes, and analyzing the results to estimate quantities such as function values or probabilities. By aggregating these results and refining the model, Monte Carlo methods improve accuracy.

In data analytics, Monte Carlo methods are used to model and understand the uncertainty in data-driven predictions and decisions. By running simulations with random inputs, analysts can explore a wide range of potential scenarios and outcomes. This helps in assessing risks, estimating probabilities, and making more informed decisions based on the variability and distribution of the data. For instance, Monte Carlo simulations can be employed to forecast future trends, optimize resource allocation, and evaluate the impact of different strategies, thereby providing a robust framework for dealing with uncertainty in complex data analyses.

**Example:**

Product Adoption and Sales Bonus. Bianca Peterson is a marketing engineer for Hexagon Composites, a company which sells carbon composite storage tanks. To gain product adoptions from customers, Bianca goes on sales trips (often to foreign countries). For each of 120 previous sales trips, the file Sales Trips lists (1) whether the trip resulted in the visited customer adopting the product, and (2) the revenue generated by the adoption.

Sales Trips:

* 1. Bianca has six sales trips planned over the next couple of months. What is the average revenue that Bianca expects to generate from these six trips? What is the probability that she generates $200,000 or less from these six trips?
  2. Bianca receives a sales bonus if she gains three more product adoptions before the end of the year. The number of sales trips that Bianca will need to make to earn her bonus is uncertain. What is its distribution? If Bianca only has time to make 10 more sales trips before the end of the year, what is the likelihood that she earns her bonus?

**Solution:**

Average revenue expected from 6 trips: $295198.65

Probability of generating $200,000 or less: 0.0000

Probability of earning the bonus within 10 more trips: 1.0000

**Optimization:**

Optimization comes in to select the best solution from the set of solutions while considering the constraints. Optimization could either maximize or minimize an objective function that typically stands for profit or cost-related goals. Optimisation can be broadly segmented into four components: the objective function, which is to be either maximized or minimized, decision variables that can be adjusted, constraints that define the possible set of solutions, and finally, the feasible region within which the solution must lie. Thus, linear, nonlinear, and integer programming techniques are applied depending on the nature of the problem at hand.

Optimization is simply the process of choosing the best element from a set with respect to specified criteria and subject to a given set of constraints. Put differently, optimization either maximizes or minimizes an objective function that can be any quantity of interest, such as profit or cost. The major constituents of an optimization problem include the objective function, the decision variables which are adjustable, a set of defined constraints within which falls the feasible solution space, and the feasible region in which the optimal solution should lie. The techniques in use are linear, nonlinear, and integer programming, based on the nature of the problem. The concept of optimization applies to a great variety of fields: business when talking about maximizing profits and efficiently allocating resources; engineering in improving the design of systems and performance; finance when it is about portfolio optimization and risk minimization; logistics when talking about route planning and scheduling.

Optimization in data analytics helps deliver actionable insights from it. Optimization helps in tuning a model to come up with better predictions and makes it easier to make informed decisions because of the possibility of finding optimal parameters or strategies. Optimization techniques help in uncovering complicated trends in data for better decision-making.

**Mix -Integer Linear Problem:**

One of the problems of optimization that considers integer variables with continuous variables under linear constraints and a linear objective function is an MILP problem.

**Examples:**

Walmart should work out the inventory that will maximize profit for Store A. The demand for Store A is 180 units of Product 1, 120 units of Product 2, and 80 units of Product 3. Its inventory-purchasing budget is $10,000, and the storage capacity is 300 units. Each product has the following limitations: Product 1 can be ordered in lot sizes between 50 and 200; Product 2 can be ordered in lot sizes between 40 and 150; Product 3 can be ordered in lot sizes between 30 and 100. The costs for each product include a unit cost, a holding cost, and a fixed ordering cost. The problem is to determine how many units of each product should be ordered such that costs associated with this are balanced against the budget and storage constraints in pursuit of maximum possible profit. It means maximizing the total revenue from selling all of the products while considering the cost of buying and holding the inventory, and conforming to the fixed ordering costs without exceeding the constraints on budget or storage capacity.

**Solution:**

Maximize total Profit:

profit = (50*x1 + 40*x2 + 30*x3) - (30*x1 + 25*x2 + 20*x3) - (2 max([x1 - 180, 0]) *+* 1.5 max ([x2 - 120, 0]) + 1\*max ([x3 - 80, 0])) - (100 + 80 + 60)

Constraints:

50 <= x1 <= 200

40 <= x2 <= 150 order quantity constraints  
30 <= x3 <= 100

30*x1 + 25*x1 + 20\*x1 + 100 + 80 + 60 <= 10,000 budget constraints

x1 + x2+ x3 <= 300 storage capacity

**Result:**

Status: Success

Optimal Order Quantity for Product 1: 200.0

Optimal Order Quantity for Product 2: 70.0

Optimal Order Quantity for Product 3: 30.0

Maximum Profit: 4815.0

**Non- Linear Problem:**

One nonlinear problem is one in which there is no straight-line relationship between the input variables, or features, and the output variable, or the target. In other words, changing the input variables does not produce proportional, straightforward changes in the output variable. It can get quite complex—involving curves, interactions, or even more intricate patterns in the relationship.

**Example:**

A manufacturing company produces three kinds of products: A, B, and C. The profit of the company is maximized under restrictions on resources. Each product requires some amount of resources and brings profit according to the nonlinear function modeling diminishing returns when the quantity produced grows.

The profit function for Product A is defined as Pa (Xa) = 10\* Xa – 0.1\* Xa^2, where Xa denotes the quantity produced. For Product B, it is a Pb (Xb) = 15\* Xb – 0.05\* Xb^2, and for Product C, it is Pc (Xc) = 20\* Xc – 0.03\* Xc^2, where Xb and Xc are the quantities of Products B and C, respectively.

The company has 1,000 resource units in all. Each unit of A requires 2 resource units, B requires 3, and C requires 4. Finally, the production of each of the products is to remain within the following ranges: 0 to 100 units for Product A, 0 to 80 units for Product B, and 0 to 60 units for Product C.

The problem will be how much to produce of Products A, B, and C so that total profit is maximized, subject to the fact that the resource constraint is not violated and the given bounds on each product. This problem can be defined simply as a nonlinear optimization problem because it has an objective function where the nature of the constraints imposed by the problem is nonlinear.

**Solutions:**

**Objective Function:**

Maximize profit: (10*Xa* - 0.1 Xa^2) + (15Xb - 0.05 Xb^2) + (20Xc - 0.03 Xc^2)

**Constraints:**

2*Xa + 3*Xb + 4\*Xc <= 1000 resource usage

0 <= Xa <= 100

0 <= Xb <= 80

0 <= Xc <= 60

**Result:**

Status: Success

Optimal Production Quantity for Product A: 50.0

Optimal Production Quantity for Product B: 80.0

Optimal Production Quantity for Product C: 60.0

Maximum Profit: 2222.0

**Binary Problem Statement:**

A Binary Problem Statement is a specific type of optimization problem where the decision variables are restricted to binary values, meaning they can only be 0 or 1. This type of problem is a subset of integer programming and is often encountered in combinatorial optimization and operations research.

**Example:**

Walmart is considering opening new distribution centres in six potential locations. Each location has a fixed opening cost and offers different profit contributions. The fixed costs and profit contributions for each location are as follows:

Location 1 (300,000𝑐𝑜𝑠𝑡, 300,000cost, 900,000 profit),

Location 2 (250,000𝑐𝑜𝑠𝑡,250,000cost, 750,000 profit),

Location 3: (280,000𝑐𝑜𝑠𝑡, 280,000cost, 800,000 profit),

Location 4 (350,000𝑐𝑜𝑠𝑡, 350,000cost, 1,000,000 profit),

Location 5 (270,000𝑐𝑜𝑠𝑡, 270,000cost, 780,000 profit), and

Location 6 (320,000𝑐𝑜𝑠𝑡, 320,000cost, 850,000 profit).

With a total budget of $1,200,000, Walmart needs to decide which centres to open to maximize total profit. Each decision is binary: a centre is either opened (1) or not (0). The goal is to find the optimal combination of canters to open while staying within the budget.

**Solutions:**

**Objective Function:**

Maximize Z = 900,000x1​+750,000x2​+800,000x3​+1,000,000x4​+780,000x5​+850,000x6​

**Constraints:**

300,000x1​+250,000x2​+280,000x3​+350,000x4​+270,000x5​+320,000x6 ​≤ 1,200,000

xi​ ∈ {0,1}, for i =1,2,3,4,5,6

Optimal solution found:

Objective value (Total Profit): 3480000.0

Variable values (0 means not open, 1 means open):

Location 1: 1

Location 2: 0

Location 3: 1

Location 4: 1

Location 5: 0

Location 6: 0

**Decision Tree:**

A decision tree is a diagram that helps you make decisions by showing different options and their potential outcomes. It starts with a main decision, branches out to possible choices or chance events, and ends with results. This helps in evaluating which option might be the best based on the expected outcomes.

**Example:**

There is a donut company that manufactures three types of donuts: Jelly-Filled, Strawberry, and Chocolate. Each one of them can be sold either in 6-packs or 9-packs. The problem here is how to best model the distribution strategy in such a way that customer demand is met at the least cost while maximizing sales.

**Objective:** Develop a decision-making model that will help the company decide on the exact number of packs each type of donuts to produce, satisfying customers while maximizing profit.

**Solution:**

A diagram of a diagram

Description automatically generated

**Key Observation:**

* If the probability for any pack size is higher in relation to that donut type, the company should produce more of that pack size. For example, in relation to chocolate donuts, there will be more probability of purchasing a pack of 9 with a probability of 0.27 in comparison with a pack of 6 with a probability of 0.13. Therefore, the company should be producing more packs of 9 for the Chocolate donuts.
* Similarly, for Jelly-Filled donuts, though the probabilities are closer, more packs of 9 should be produced.
* The probabilities are even closer for the Strawberry Donuts: 0.10 to the pack of 6 versus 0.17 to the pack of 9, so the company should make slightly more packs of 9 for this item.

### Conclusion:

### This project is targeted at predicting weekly sales with Walmart using different time series models, namely, ETS, ARIMA, SARIMA, and Linear Regression, so a better method for its forecasting can be found. Of all the results, the best model was one of the techniques of ETS with additive trend and multiplicative seasonal components, since this method gave the lowest RMSE and MAE values by capturing seasonal and trend patterns in the series. Although the ARIMA model of order (5, 0, 2) was quite competitive in handling trends and autocorrelations, it did not turn out so well on seasonal effects. It is expected that SARIMA, which includes seasonal components, will outperform ARIMA for the seasonal pattern; however, no metrics are given. Hence, linear regression was useful in ascertaining the effect of the exogenous variables on temperature and fuel prices, but resulted in a higher forecast error since it is limited in its ability to capture complex temporal dynamics and seasonality.

### The analysis confirmed distinct seasonal patterns in Walmart’s sales, with peaks at the end of each year, which the ETS and SARIMA models effectively accounted for. While Linear Regression offered insights into external factors, it was less effective than time series models in capturing the data's temporal complexity. SARIMA’s complexity and computational demands necessitate careful consideration to avoid overfitting, and model accuracy relies on high-quality data. Future work should involve incorporating additional features, using advanced models, and regularly updating forecasts with new data to ensure ongoing accuracy. Overall, the project underscores the superior performance of time series models, particularly those addressing seasonal effects, with the ETS model emerging as the most effective for forecasting Walmart’s weekly sales.

Monte Carlo simulations are a versatile computational tool that may be used in the obtaining of numerical results through large random samples to assess related uncertainty. This is, therefore, most helpful in the solution of complex problems where it becomes hard to get analytical solutions. These simulation-based algorithms use Monte Carlo methods to provide insight into the variability and risks associated with data-driven predictions, simulating a wide range of scenarios to arrive at results in an aggregated manner. In these respects, it enables better decision-making in financial, engineering, and data analytics domains.

Optimization involves finding the most effective solution from a set of possible choices while adhering to defined constraints. It focuses on maximizing or minimizing an objective function, which could represent goals such as profit or cost. Optimization techniques, including linear, nonlinear, and integer programming, are employed to handle different types of problems. These methods are crucial in various domains, enabling businesses and researchers to make informed decisions, allocate resources efficiently, and enhance performance by identifying optimal strategies and solutions.