Project Documentation: S&P 500 Stock Analysis and Forecasting

1. Introduction

The S&P 500 is a widely recognized stock market index that tracks the performance of 500 of the largest publicly traded companies in the U.S. This project aims to analyze historical stock price data of the S&P 500 and implement various forecasting models to predict future trends.

2. Objectives

- To explore the historical trends of the S&P 500 index.
- To apply different forecasting techniques, including:
  - Linear Regression
  - Naïve Forecasting
  - Exponential Smoothing (ETS)
  - ARIMA (AutoRegressive Integrated Moving Average)
- To compare the accuracy of these models and determine the best-performing method.

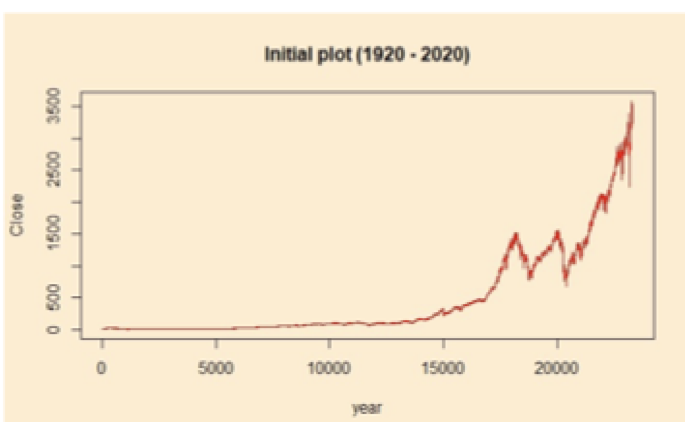
3. Data Source

- The dataset used for this project is obtained from **Kaggle** and contains daily stock prices of the S&P 500 from **December 1927 to November 2020**.
- The data excludes weekends and public holidays when the market is closed.

4. Methodology

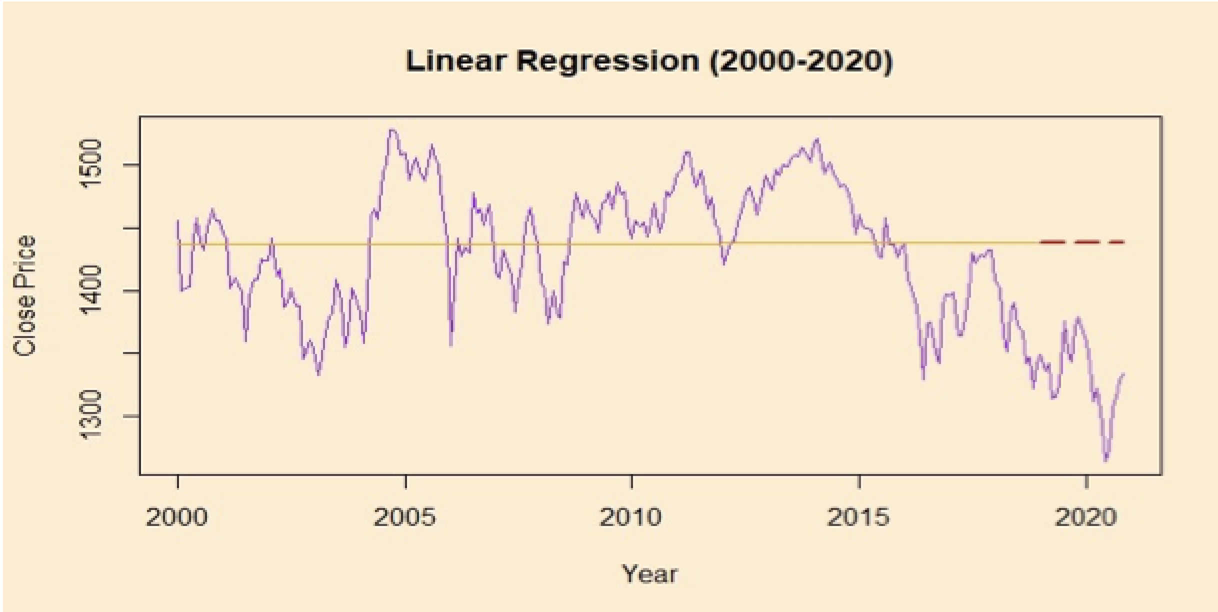
The project follows these steps:

1. **Data Cleaning and Preprocessing**
   - Load the dataset and format the `Date` column.
   - Convert the dataset into a **time series format** for analysis.
2. **Exploratory Data Analysis (EDA)**
   - Initial visualization of S&P 500 closing prices over time.

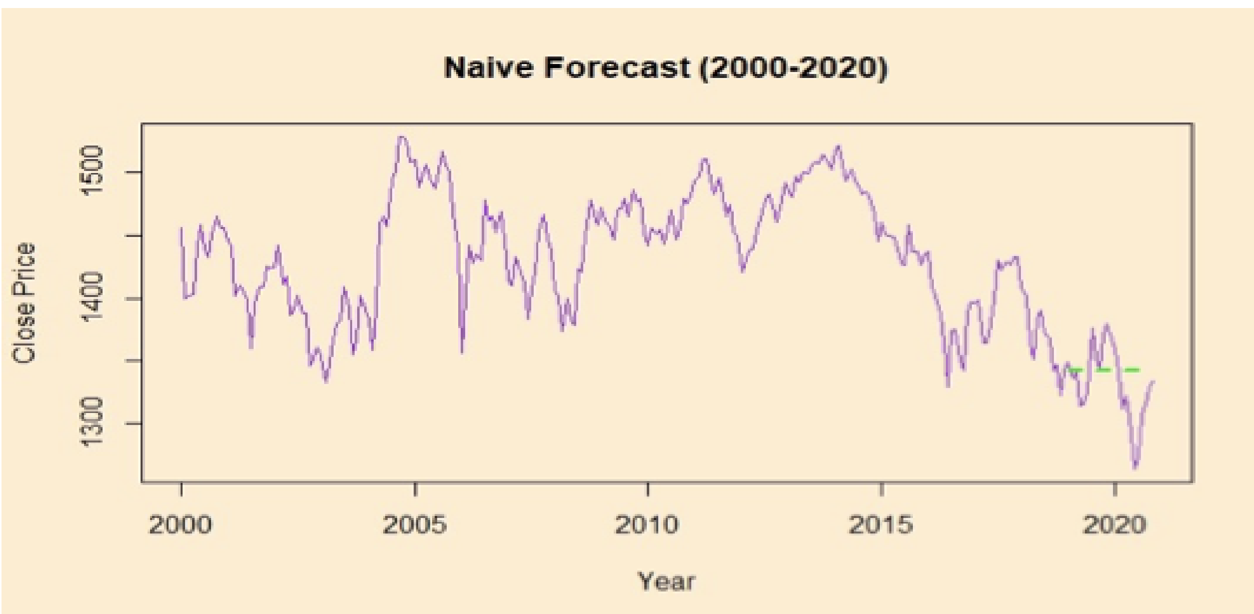


   - Filtering data from the year **2000 onwards** for more recent insights.

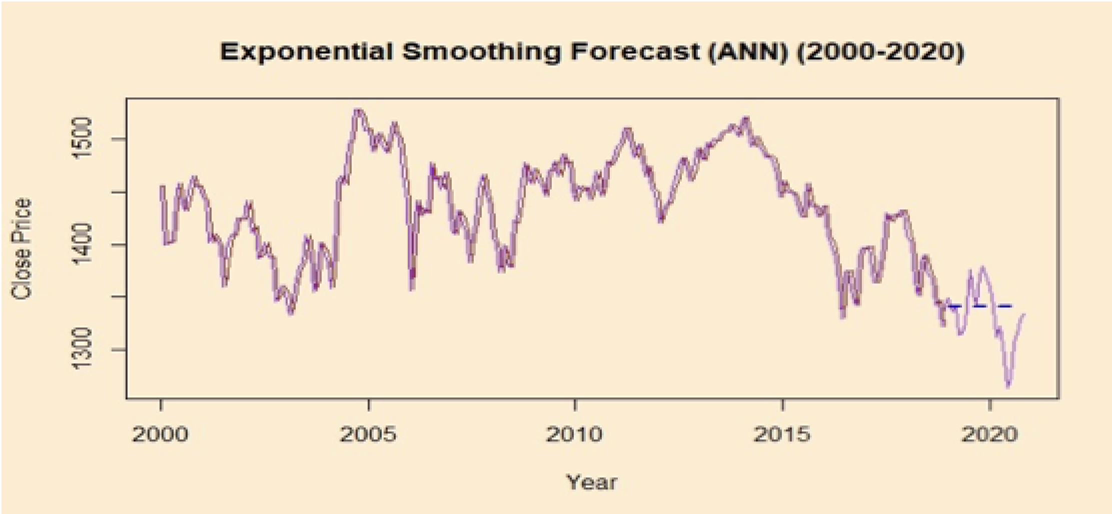
3. **Forecasting Models**
   - **Linear Regression:** Models the trend of stock prices as a function of time.



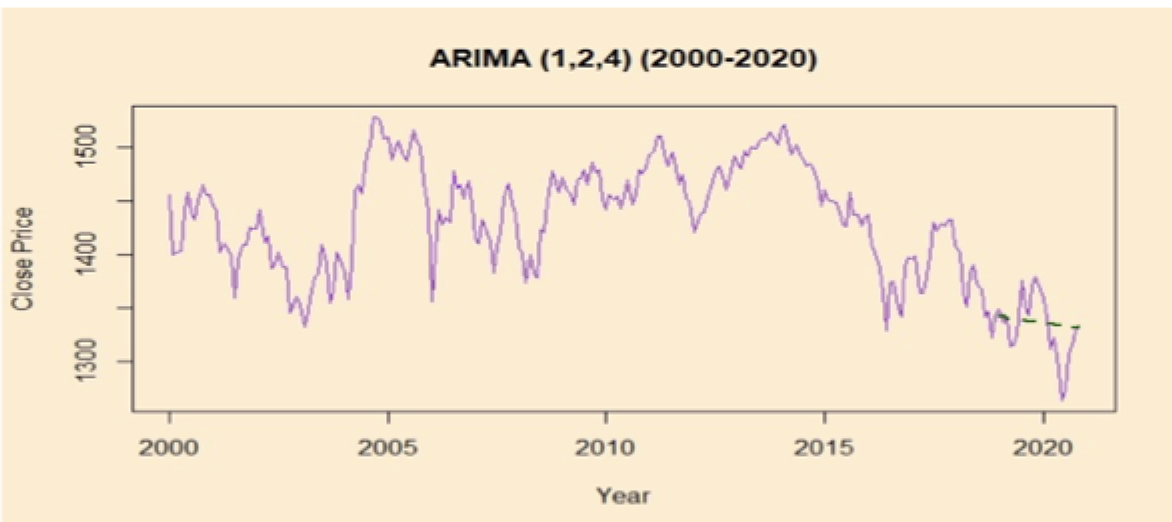
   - **Naïve Forecasting:** Assumes that future prices will equal the most recent observed value.

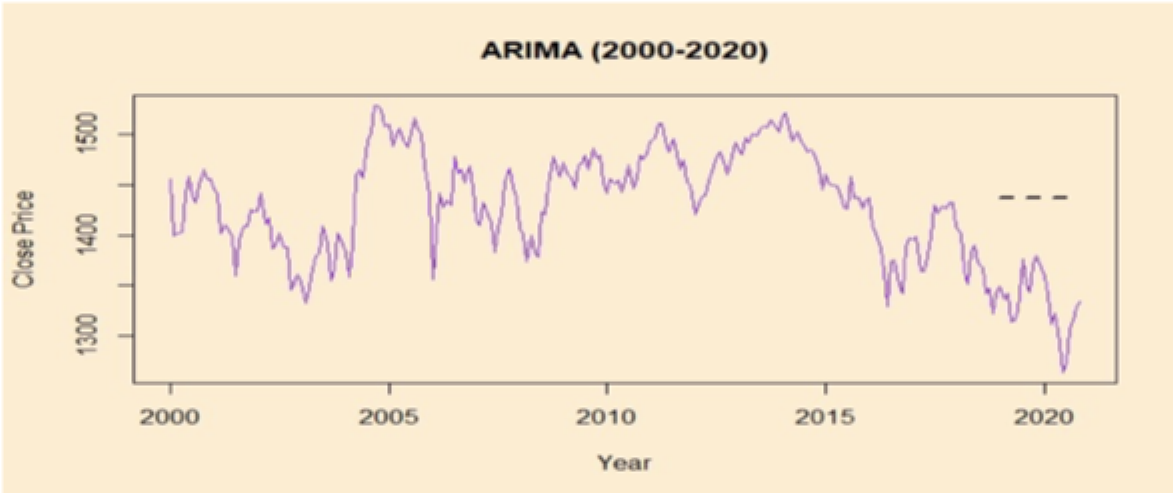


   - **Exponential Smoothing (ETS):** Accounts for short-term fluctuations and seasonality.



- **ARIMA Model:** Uses past data points to predict future values based on autoregression and moving average techniques.





4. **Accuracy Evaluation**
    - The forecasting models are evaluated using performance metrics like **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**.
    - The performance of each model is compared to determine the most accurate prediction technique.

Linear model forecast accuracy:

```
> accuracy(spx.lm.pred$mean, spx_valid)
                ME     RMSE      MAE       MPE     MAPE      ACF1 Theil's U
Test set -105.4953 109.4657 105.4953 -7.971338 7.971338 0.7375023  5.194347
```

Naïve Forecast accuracy:

```
> accuracy(spx.lm.pred$mean, spx_valid)
                ME     RMSE      MAE       MPE     MAPE      ACF1 Theil's U
Test set -105.4953 109.4657 105.4953 -7.971338 7.971338 0.7375023  5.194347
```

Exponential Smoothing forecast accuracy (model = ANN by computer, only alpha):

```
> accuracy(spx.es.pred$mean, spx_valid)
                ME     RMSE     MAE        MPE     MAPE      ACF1 Theil's U
Test set -9.348356 30.66522 24.1233 -0.7505845 1.828798 0.7373737  1.484776
```

ES-2 (model = "MAA", with alpha, beta and gamma):

```
> accuracy(spx.es2.pred$mean, spx_valid)
                ME     RMSE      MAE        MPE     MAPE      ACF1 Theil's U
Test set -1.023769 26.80269 21.47313 -0.1201741 1.619948 0.7470554  1.283005
```

ARIMA forecast accuracy:

```
> accuracy(spx.arima.pred$mean, spx_valid)
                ME    RMSE      MAE        MPE    MAPE      ACF1 Theil's U
Test set -4.620281 28.4734 22.49812 -0.3933777 1.69997 0.715505  1.371601
```

Arima 2(with computer chosen values):

```
> accuracy(spx.arima2.pred$mean, spx_valid)
                ME    RMSE      MAE      MPE    MAPE      ACF1 Theil's U
Test set -105.0923 109.075 105.0923 -7.94106 7.94106 0.7373737   5.17604
>
```

5. **Results and Conclusion**
    - The **Exponential Smoothing (ETS, MAA)** model was identified as the most effective for forecasting stock prices during the specified period.
    - The study highlights the importance of continuously updating forecasting models to adapt to market fluctuations.

5. Tools & Technologies Used

- **Programming Language:** R
- **Libraries:**
    - `dplyr` (Data Manipulation)
    - `ggplot2` (Data Visualization)
    - `lubridate` (Date Formatting)
    - `forecast` (Time Series Forecasting)

6. Future Scope

- Expand the dataset to include additional macroeconomic indicators (e.g., GDP, interest rates).
- Implement machine learning-based forecasting models (e.g., LSTM, Random Forest Regression).
- Incorporate real-time stock price updates for dynamic analysis.