Predicting Stock Market Trends with Financial Big Data Analytics

In this project, the aim is to get better insights upon big data analytics using PySpark and Python to predict stock market trends. By analyzing stock data of the “SP500\_data” Excel, which gathered data from 07-Jan-2020 until 05-Jan-2024, provided by Yahoo Finance, we aim to discover patterns and answer the following questions: "How can we calculate the change in price of the stock over time?", "In which way can we calculate the daily return of the stock?", "What is the moving average of the stock price?", "How can the standard deviation of the daily returns (risk) be calculated?" and "Predicting the future stock behavior using Linear Regression”.

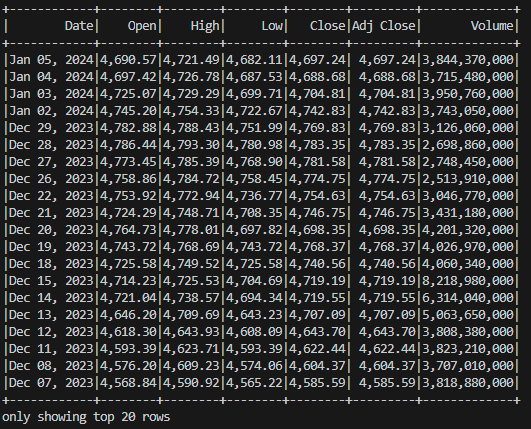


Fig. 1.

In Fig. 1. there are displayed the first 20 rows of the DataFrame. It provides us with data concerning features such as the value of stocks on the day they were open, the highest and lowest price, the closing price and the trading volume.

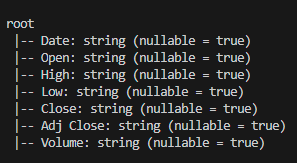


Fig. 2.

In Fig. 2. it can be observed the schema of the DataFrame, from which we can see the fact that all the data types are strings.

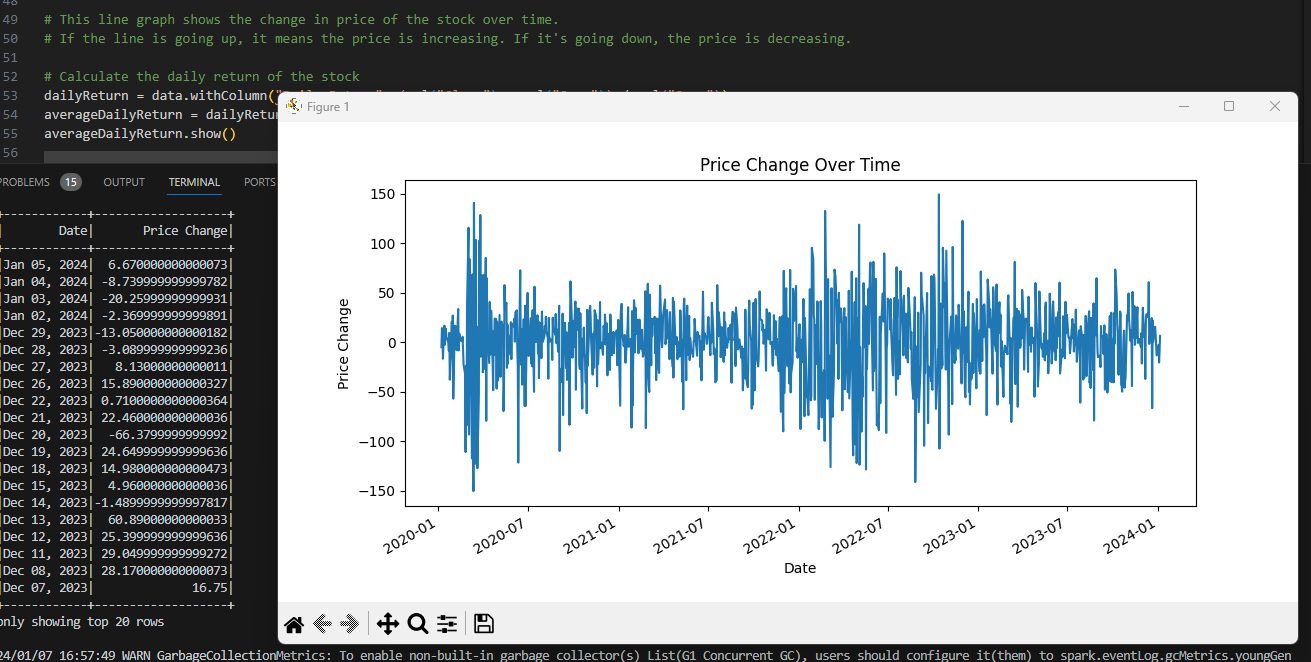
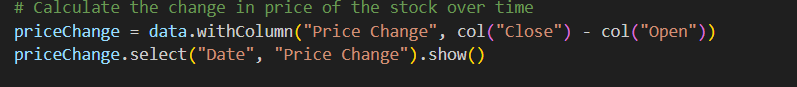


Fig. 3.

Fig. 3. provides us with insights of changes in the price of the stock over time. If the line goes up, it means the price is increasing, and whenever it goes down the price decreases. In order to get to the graphic’s values, “Open” column was subtracted from the “Close” column, as it can be seen below:



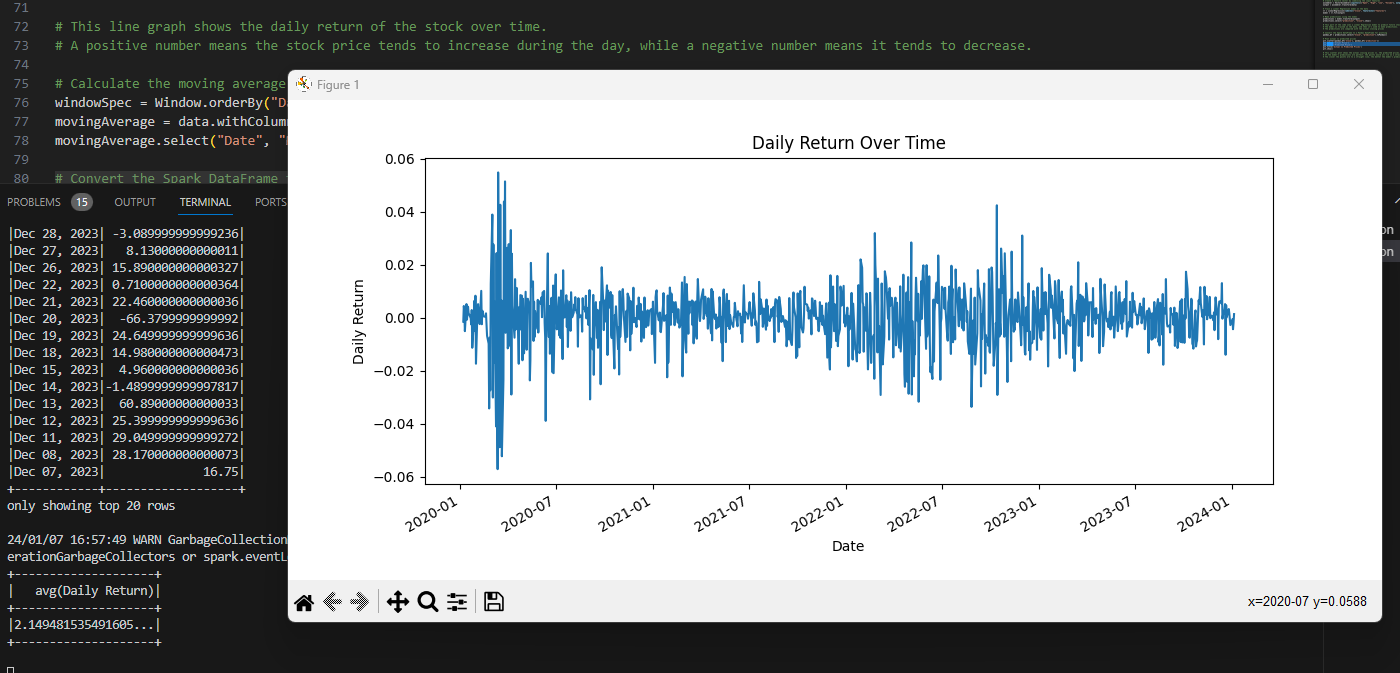
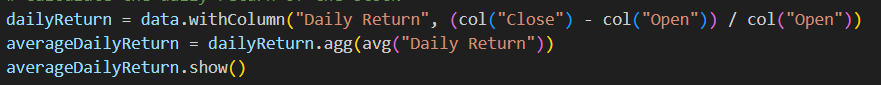


Fig. 4.

 This line graph (Fig. 4.) shows the daily return of the stock over time. A positive number means the stock price tends to increase during the day, while a negative number means it tends to decrease.



Above can be seen that “Daily Return” was obtained by dividing the subtraction between “Close” and “Open” columns with “Open” columns.

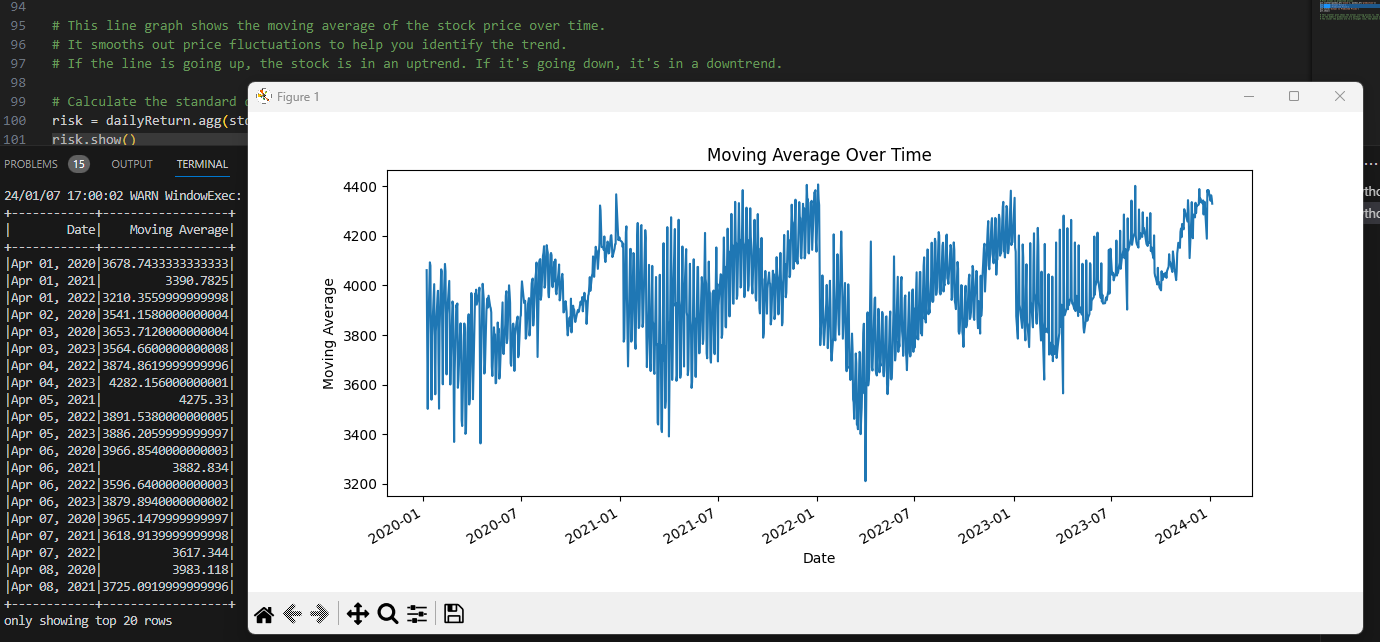
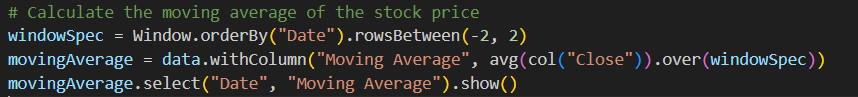


Fig. 5.

In the above graph it is shown the moving average of the stock price over time. It smooths out price fluctuations to help identify the trend. If the line is going up, the stock is in an uptrend. If it's going down, it's in a downtrend.



This code calculates the moving average of the "Close" column using the window specification and assigns the result to a new column called "Moving Average" in the "data" DataFrame.

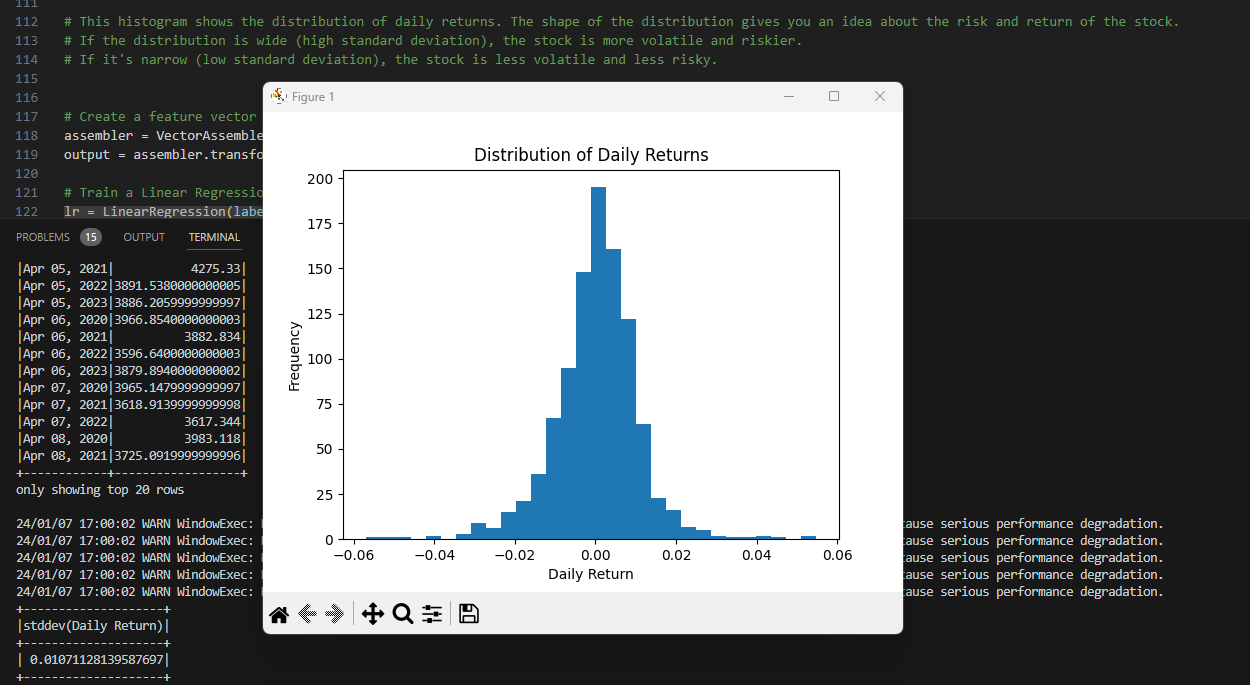
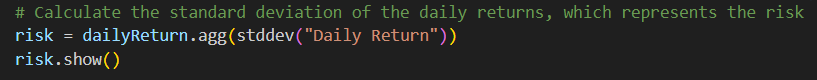


Fig. 6.

This histogram shows the distribution of daily returns. The shape of the distribution gives us an idea about the risk and return of the stock. If the distribution is wide, the stock is more volatile and riskier, but if it's narrow, the stock is less volatile and less risky.



The provided line of code calculates the standard deviation of the daily returns. The "agg" function is employed to aggregate the standard deviation, and the "stddev" function is applied to the "Daily Return" column within the DataFrame. The resulting "risk" variable holds the computed standard deviation value, providing a measure of the variability in the daily returns.

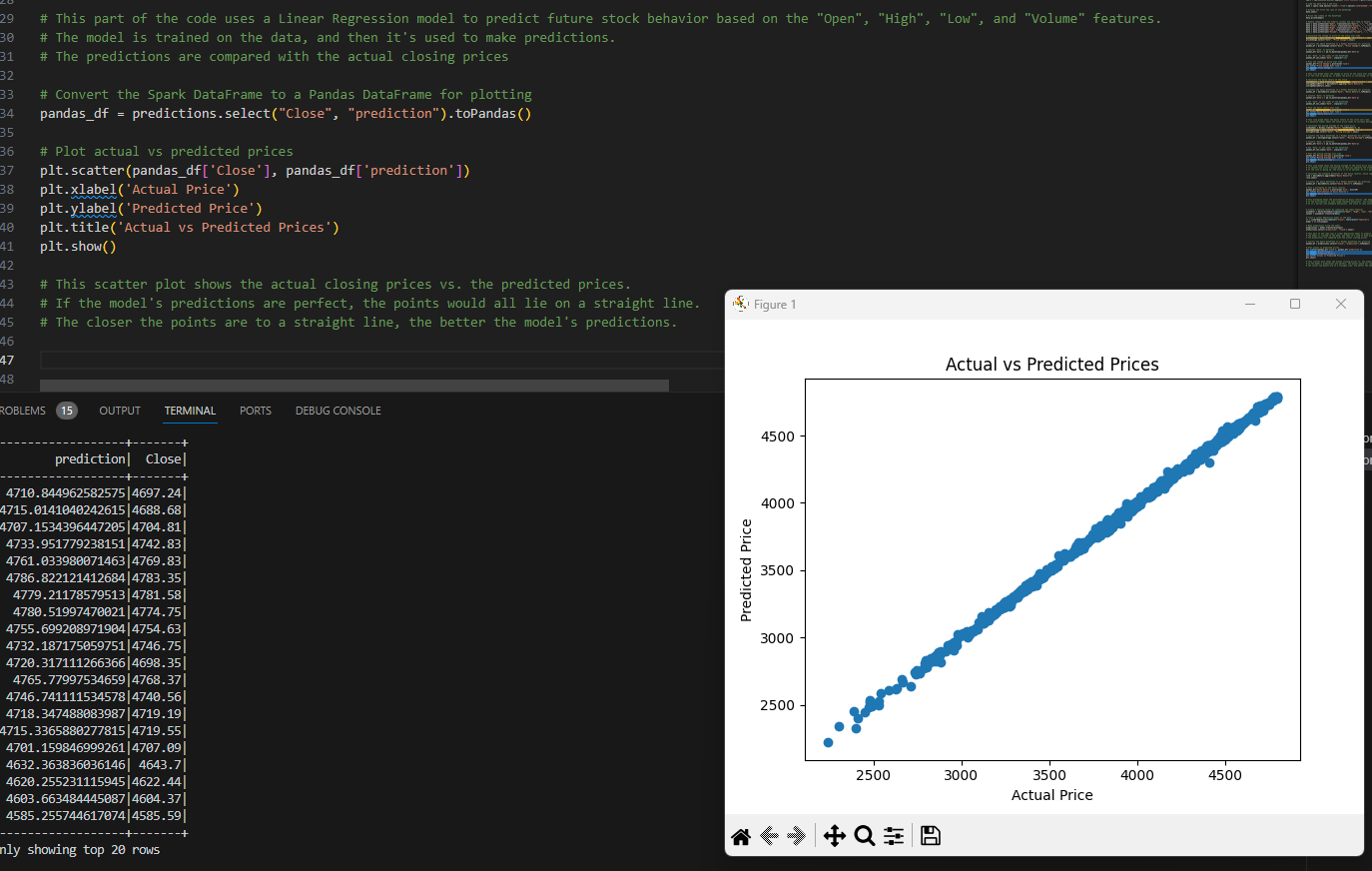


Fig. 7.

This part of the code uses a Linear Regression model to predict future stock behavior based on the "Open", "High", "Low", and "Volume" features.

Linear regression is a statistical method used to model the relationship between a dependent variable (also called the response or outcome variable) and one or more independent variables (also called predictors or features). The goal is to find the best-fitting linear relationship that describes the data.

The model is trained on the data, and then it's used to make predictions. The predictions are compared with the actual closing prices. This scatter plot shows the actual closing prices vs. the predicted prices. If the model's predictions are perfect, the points would all lie on a straight line. The closer the points are to a straight line, the better the model's predictions.

