

**Field Projects for Data Storage & Analysis**

**Final Project Report**

**Using Big Data for Stock Market Analysis of Companies**

**Student:** Nagashybaiuly Kozhakhmet

**ID:** 20B030502

**Teacher:** Kartbayev Amandyk

Almaty, May 12, 2023

Daily Financial Investment Risk Aid

***Abstract* — this study explores the feasibility of using trending world topics, historical stock data, and future stock value predictions to aid in investment decisions for portfolio growth. Technology tools and virtualization techniques are utilized to acquire, process, and visualize daily data sets for informed decision-making. The project acknowledges the risk of predicting market trends and, therefore, does not make investment decisions on behalf of the investor. Instead, it acts as a tool to provide the best available information for the investor’s final choices. The goal is to understand the value of using real-time financial data for daily investment decisions.**

***Keywords— Finance, Stocks, Mutual Funds, Google Cloud, Yahoo Finance, Virtual Environments, Risk, Investments, Visualization, DAG, Airflow, Apache***

1. **INTRODUCTION**

The financial investment industry has become increasingly reliant on the opinions and information shared by influential entities on social media and other communication platforms. As a result, it has become critical to have a tool that aggregates and visualizes relevant investment information. This is especially important for individual and small investors who must make investment decisions based on these opinions.

The daily financial investment risk aid is designed to provide a unified financial aid that combines hard data with social media data. It will offer users historical trending information on stocks, predictions for future closing costs, and trending viral words and phrases on social media. This tool will provide users with daily relevant information that can be used as part of a portfolio investment decision-making process.

The financial aid does not make investment decisions on behalf of the user, but it does provide a world snapshot that can help prepare the user for the daily investment workflow. This can result in optimized decision-making and portfolio growth. The ultimate goal of the daily financial investment risk aid is to provide a tool that assists users in making informed investment decisions based on a combination of hard data and social media trends.

1. **RELATED WORK**

This reporter is based on the project Sapphirine Big Data Analytics Application Suite and project name is “Daily Financial Risk Aid” [3]. I am the author of this article; I have radically changed the actual of the project. I analysis the datasets and predicted the target feature using Linear Regression and Pyspark library. And also added all the analysis visualizations. I'm sure I've solved the problem.

One tool which enables the tracking of stock and individual’s investment performance is that of Sisense’s Finance System [26]. While another exemplar is that of XB Software’s Rate Management System [27]. [Financial software tools such as Sisense’s Finance System [26] and XB Software’s Rate Management System [27] allow investors to track individual investment vehicles and stock performance. However, no tool was identified prior to this research that combines stock information, predictions, and world news.

A range of software based financial tools are available to investors which combine the performance of their stocks and allow for the tracking of individual investment vehicles, but a toolset was not identified prior to this research which combines stock information, predictions, and world news or sentiments via twitter analysis into one web interface.

This project aims to analyze the feasibility of adding further market insights by providing investors with trending world information, stock historical data, and linear regression enabled stock closing predictions. This tool can be used as a force multiplier alongside other available software such as Sisense’s Finance System [26] and XB Software’s Rate Management System [27] for investment portfolio growth. It does not limit its use as part of an investor’s overall workflow.

1. **DATA**

A total of 4 unique datasets are being collected daily for the processing and visualization of relevant financial and global event data. The first section of investment vehicles includes 3 mutual funds with historically stable returns. The next set of investment vehicles includes 3 historically profitable yet stable stocks predominantly from the energy sector. The last set of investment vehicles includes 3 fast growing stocks for rapid portfolio growth. This is followed by the daily data acquisition of trending Twitter data split into general top trending hashtags in addition to the filtering on the 4 stock’s acronyms to detect if the stock is currently trending. The data frame provided includes columns labeled "Open," "High," "Low," "Close," "Volume," "Dividends," and "Stock Split." The "Open" column represents the opening price of a stock on a given trading day, while the "High" and "Low" columns represent the highest and lowest prices at which a stock or security was traded during the day, respectively. The "Close" column represents the final price at which a stock or security was traded when the market closed. These prices are important indicators for investors to analyze price movements and identify potential trading opportunities.

The "Volume" column represents the total number of shares or contracts sold during a given trading day. It reflects the level of interest or activity in a particular stock or security, with high volumes indicating increased investor interest or significant events affecting a stock or security. The "Dividends" column represents the payments made by a company to its shareholders out of its profits, which can indicate a company's financial condition and serve as an important source of income for investors.

Lastly, the "Stock Split" column represents any corporate action in which a company divides its existing shares into several parts, effectively increasing the number of shares outstanding. It can affect the liquidity of shares and make them more available for purchase by small investors. It is worth noting that not all stocks pay dividends or undergo stock splits, so some of these columns may only contain zeros or zero values for certain stocks. Overall, these columns provide important insights for investors to evaluate the effectiveness of their investment portfolios over time.

1. *Mutual Funds*

Created in 1992, Vanguard Total Stock Market Index Fund (VSMPX) [19] is designed to provide investors with exposure to the entire U.S. equity market, including small-, mid-, and large-cap growth and value stocks. The fund’s key attributes are its low costs, broad diversification, and the potential for tax efficiency. Investors looking for a low-cost way to gain broad exposure to the U.S. stock market who are willing to accept the volatility that comes with stock market investing should consider this fund as either a core equity holding or only domestic stock fund.

The industry’s first index fund for individual investors, the

500 Index Fund (VFIAX) [18] is a low-cost way to gain

diversified exposure to the U.S. equity market. The fund offers exposure to 500 of the largest U.S. companies, which span many different industries and account for about three-fourths of the

U.S. stock market’s value. The key risk for the fund is the volatility that comes with its full exposure to the stock market. Because the 500 Index Fund is broadly diversified within the large-capitalization market, it may be considered a core equity holding in a portfolio.

The Growth Fund of America’s (AGTHX) [8] investment objective is to provide investors with growth of capital. This fund takes a flexible approach to growth investing, seeking opportunities in traditional growth stocks as well as cyclical companies and turnarounds with significant potential for growth of capital. Geographic flexibility also allows portfolio managers to pursue opportunities outside of the U.S. This differentiated approach has the potential to enable the fund to navigate a variety of market environments. AGTHX invests at least 65% of its assets in common stocks. It may also invest in convertibles, preferred stocks, U.S. government securities, bonds and cash equivalents.

1. *Stocks*

Occidental Petroleum Corporation (OXY) [4][14][21] is an American company engaged in hydrocarbon exploration in the

U.S. and the Middle East as well as petrochemical manufacturing in the U.S., Canada, and Chile. It is organized in Delaware and headquartered in Houston. The company ranked 183rd on the 2021 Fortune 500 based on its 2020 revenues and 670th on the 2021 Forbes Global 2000.

ExxonMobil (XOM) [2] Corporation is an American multinational oil and gas corporation headquartered in Irving, Texas. It is the largest direct descendant of John D. Rockefeller's Standard Oil, and was formed on November 30, 1999, by the merger of Exxon and Mobil, both of which are used as retail brands, alongside Esso, for fueling stations and downstream products today. The company is vertically-integrated across the entire oil and gas industry, and within it is also a chemicals division which produces plastic, synthetic rubber, and other chemical products. ExxonMobil is incorporated in New Jersey.

Vertex Pharmaceuticals (VRTX) [7] is an American biopharmaceutical company based in Boston, Massachusetts. It was one of the first biotech firms to use an explicit strategy of rational drug design rather than combinatorial chemistry. It maintains headquarters in South Boston, Massachusetts, and three research facilities, in San Diego, California, and Milton Park, near Oxford, England.

1. *Fast Growing Stocks*

Enphase Energy, Inc. (ENPH) [1][22] is an American energy technology company headquartered in Fremont, California, that develops and manufactures solar micro-inverters, battery energy storage, and EV charging stations primarily for residential customers. Enphase was established in 2006 and is the first company to successfully commercialize the solar micro- inverter, which converts the direct current power generated by a solar panel into grid-compatible alternating current for use or export. The company had shipped more than 48 million microinverters to 2.5 million solar systems in more than 140 countries.

Shift4 (FOUR) [6] is an American payment processing company publicly listed on the New York Stock Exchange and based in Allentown, Pennsylvania. The company, founded in 1999 by the then 16-year-old Jared Isaacman, processes payments for over 200,000 businesses in the retail, hospitality, leisure and restaurant industries. Shift4 specializes in commerce solutions such as mobile payment software and hardware. When the company went public in 2020, Isaacman was still the CEO.

Onsemi (ON) [5] is an American semiconductor supplier company, based in Phoenix, Arizona and ranked #483 on the 2022 Fortune 500 based on its 2021 sales. Products include power and signal management, logic, discrete, and custom devices for automotive, communications, computing, consumer, industrial, LED lighting, medical, military/aerospace and power applications. Onsemi runs a network of manufacturing facilities, sales offices and design centers in North America, Europe, and the Asia Pacific regions. Based on its 2016 revenues of $3.907 billion, Onsemi ranked among the worldwide top 20 semiconductor sales leaders.

1. *Trending Twitter Hashtags Phrases*

Daily trending Twitter tweets are being gathered representing the day’s trending events, social mindset, and any other possible noteworthy news item. The Twitter data is further subdivided into two types of datasets. The first is a list of the top hashtags regardless of region, language, or type. This is to ensure that an unbiased world view of current events is represented and visualized for the user. The second set is a curated dataset of top trending financial keywords which are of specific interest to investors. The combination of both of these datasets will allow users to determine the best investment decision based on their personal needs and portfolio growth goals.

1. **METHODS**

To assist with portfolio growth, a systematic approach was taken to visualize three investment vehicle types and trending world information through Twitter tweets. All datasets are obtained daily at the same time to ensure maximum relevance for users' investment decisions. This methodology presents all necessary investment vehicle and world topic information on one intuitive webpage, featuring word clouds, stock line graphs, and relevant closing vs predicted closing price data.

Investors consider several factors when making investment decisions, and displaying all relevant visualized dataset information in one webpage would integrate the financial risk aid into a larger investment workflow for optimal portfolio growth. While alternative or multi-platform design approaches could have diversified computing power, they may have created fragmentation in the backend system design, requiring careful consideration to avoid a faulty or inaccurate dataset acquisition system. Therefore, the final decision was made to use Google's cloud services for data acquisition and design and the PySpark Library.

This PySpark-based approach offers a unified and concise way to acquire, time, process, and store all datasets for the project. Furthermore, any debugging, system integration, and processing will follow similar methodologies set by the PySpark-based platform, leading to a stable platform for data processing and visualization.

This PySpark based approach ensures a unified and concise approach to the acquisition, timing, processing, and storage of all the datasets being acquired for this project. Additionally, any debugging, system integration, and processing will follow similar methodologies set by the PySpark based platform leading to a stable platform for data processing and visualization.

1. **SYSTEM OVERVIEW**

The code snippets provided implement various functionalities for predicting stock prices using a Linear Regression model in PySpark.

The first snippet of code implements a function called "evaluate\_linear\_regression" that takes a file path as input and returns the PySpark DataFrame, the trained Linear Regression model, the testing and training data, and the evaluation metrics for the model. The function reads the data from the specified file path in CSV format and converts the "Date" column into separate columns for year, month, day, hour, minute, and second. The data is then transformed using a VectorAssembler that combines the input columns into a single vector column called "features". The function then splits the data into training and testing data, trains a Linear Regression model on the training data, and evaluates the model using the Root Mean Square Error (RMSE), Mean Square Error (MSE), and R-Squared (R2) metrics.

The second code snippet implements a function called "plot\_close\_vs\_predicted\_by\_year" that takes the DataFrame, the trained Linear Regression model, and the testing data as inputs and generates a plot showing the predicted and actual Close values against the year. The function groups the data by year and computes the average Close value for each year. The predicted Close values for the testing data are then generated using the trained Linear Regression model. The predicted Close values and actual Close values are then plotted against the year.

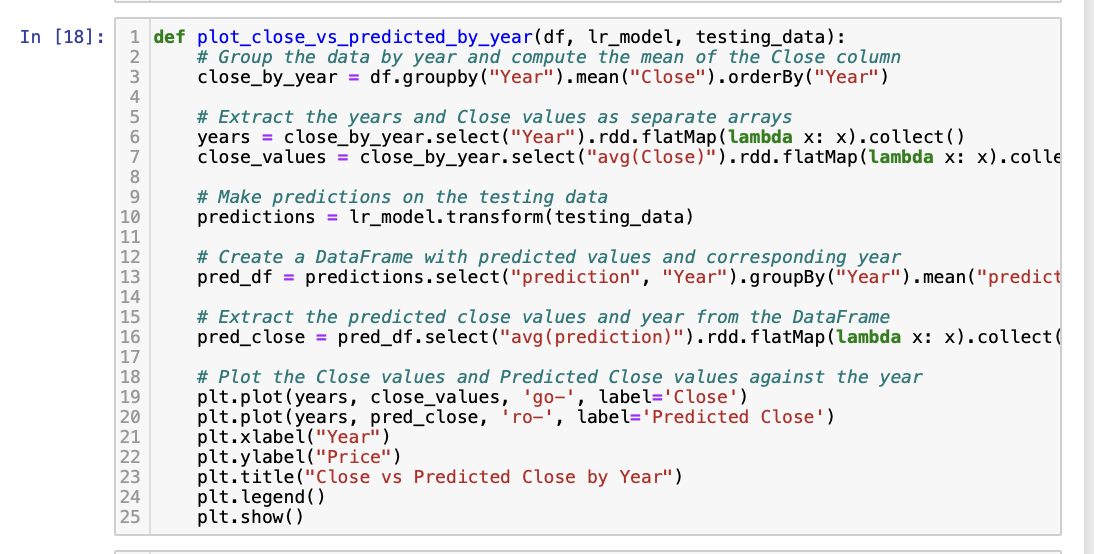
The third code snippet implements a function called "plot\_close\_vs\_predicted\_close\_by\_selected\_year" that takes the DataFrame, the year of interest, the trained Linear Regression model, and the testing data as inputs and generates a plot showing the predicted and actual Close values against the month for the specified year. The function filters the DataFrame to include only the data for the specified year and groups the data by month and computes the average Close value for each month. The predicted Close values for the testing data are then generated using the trained Linear Regression model. The predicted Close values and actual Close values are then plotted against the month for the specified year.

Overall, these code snippets provide a useful starting point for analyzing and predicting stock prices using PySpark and Linear Regression. By combining these functions, it is possible to perform a comprehensive analysis of stock prices and generate predictions for future prices based on historical data. However, it is important to note that these models have their limitations and may not always accurately predict stock prices. Therefore, it is essential to use multiple models and evaluate their performance against multiple evaluation metrics before making investment decisions.

1. The first function `evaluate\_linear\_regression(file\_path)` loads a CSV file containing financial data into a PySpark DataFrame. The function then processes the data by converting the "Date" column into separate "Year", "Month", "Day", "Hour", "Minute", and "Second" columns. The function then converts certain columns to double type. The input columns for the linear regression model are then created as a list, and these columns are vectorized. The data is then split into training and testing sets, and a linear regression model is trained on the training data. The model is then used to make predictions on the testing data. The function then evaluates the predictions using RMSE, MSE, and R2, and returns the DataFrame, the trained linear regression model, and the testing and training data.



2. The second function `plot\_close\_vs\_predicted\_by\_year(df, lr\_model, testing\_data)` creates a plot of the Close values and Predicted Close values against the year. The function groups the data by year and computes the mean of the Close column. The function then extracts the years and Close values as separate arrays. The function then makes predictions on the testing data using the given linear regression model, and creates a DataFrame with predicted values and corresponding year. The function then extracts the predicted close values and year from the DataFrame. The function then plots the Close values and Predicted Close values against the year.



3. The third function `plot\_close\_vs\_predicted\_close\_by\_selected\_year(df, year, model, testing\_data)` creates a plot of the Close values and Predicted Close values against the month for a given year. The function filters the data to include only the given year, and groups the data by month and computes the mean of the Close column. The function then extracts the months and Close values as separate arrays and plots the Close values. The function then makes predictions on the testing data using the given linear regression model, and filters the predictions to include only the given year. The function then creates a DataFrame with predicted values and corresponding month, and extracts the predicted close values and month from the DataFrame. The function then plots the predicted Close values against the month.

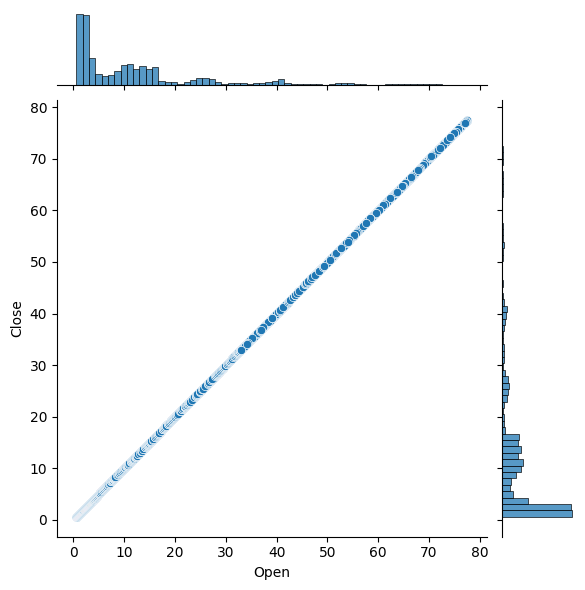
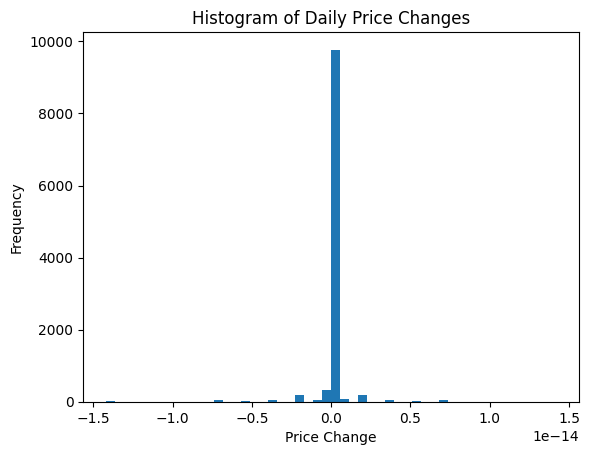
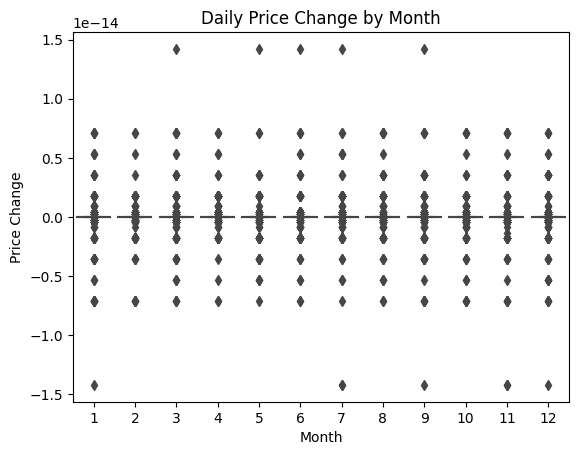
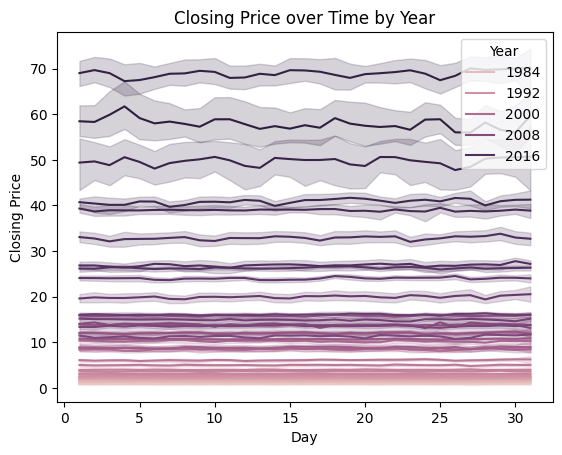
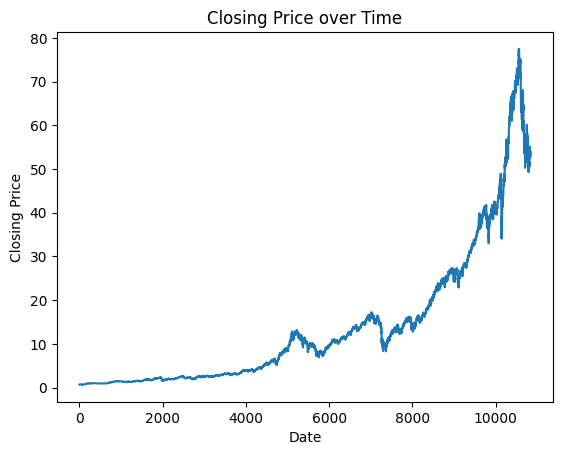
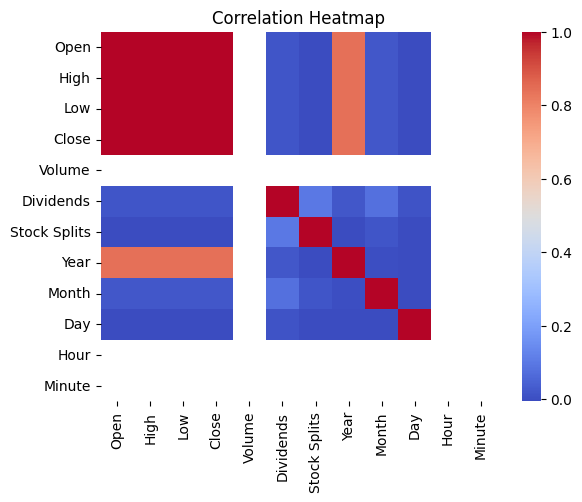
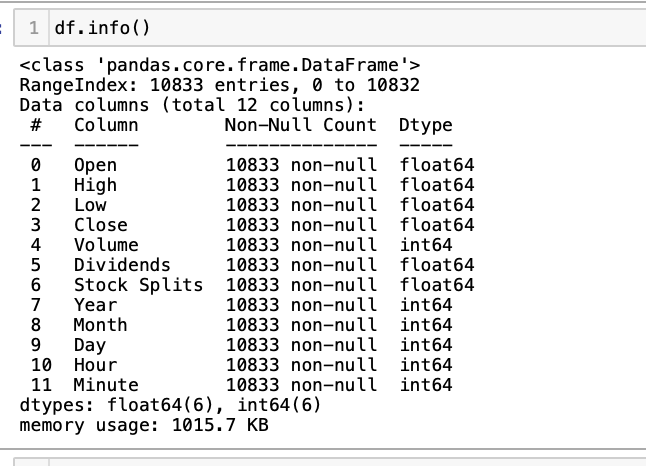


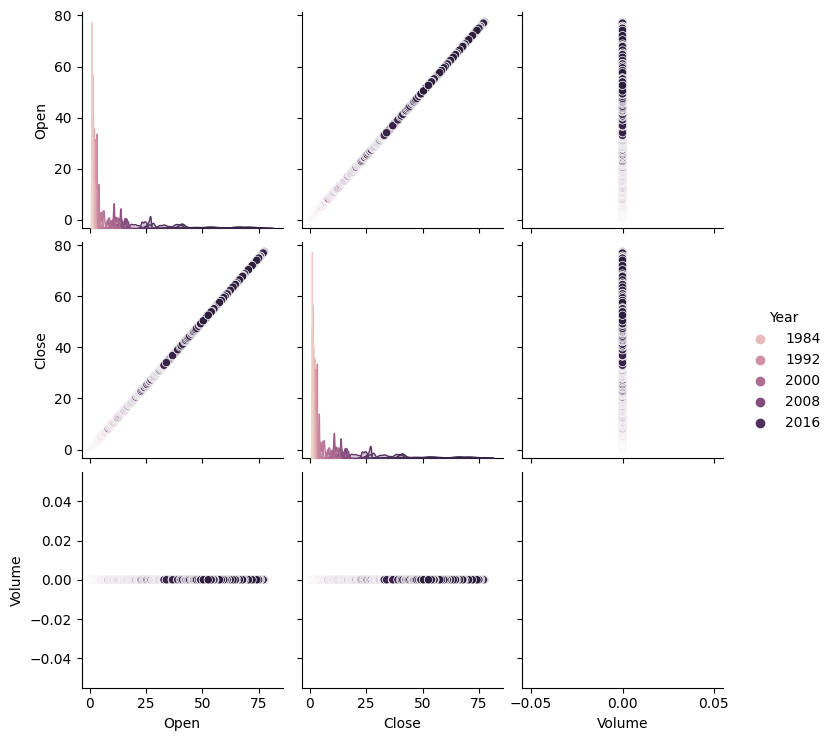
4. The `evaluate\_linear\_regression()` function takes a file path as input and loads the data from a CSV file into a PySpark DataFrame. The data is then preprocessed by converting the "Date" column into separate columns for year, month, day, hour, minute, and second, and certain columns are converted to double type. The input columns are then vectorized and the data is split into training and testing sets. A linear regression model is trained on the training data and used to make predictions on the testing data. The predictions are evaluated using three different metrics: RMSE, MSE, and R2. These metrics provide information about the accuracy and goodness of fit of the model.



1. **Analysis and Visualization of Performance**

Data visualization is an important tool for understanding complex data and patterns in a way that is easily digestible. In the world of finance, this is especially important for traders and investors to rely on visualizations to identify trends and make informed decisions. In this report, we will be using Python's datavisualization libraries to explore and analyze stock market data. Specifically, we will be using the Matplotlib and Seaborn libraries to create a variety of visualizations, including line plots, scatter plots, box plots, and heatmaps, to gain insight into stock market trends over time. By visualizing this data, we can identify patterns, correlations, and outliers, and ultimately make more informed decisions when it comes to investing in the stock market.





The first visualization shows a line plot of the closing price over time, displaying how the stock's value has changed over the years. The second visualization is a correlation heatmap, which illustrates how the different features in the dataset are related to each other. The third visualization is a line plot that displays the closing price over time by year, with different years distinguished by different colors. This allows us to see how the stock's performance has varied from year to year. The fourth visualization is a box plot that shows the distribution of daily price changes by month. It helps us understand how the price of the stock has fluctuated throughout the year and if certain months have more volatile trading activity. The fifth visualization is a histogram of daily price changes, displaying the frequency distribution of price changes. This helps us understand how frequently certain price changes have occurred over the course of the stock's history. The sixth visualization is another histogram of daily price changes, showing the same information as the previous visualization. The seventh visualization is a joint plot that displays the relationship between the opening and closing prices of the stock. This helps us understand how the two variables are related and if there is a trend between them. The eighth visualization is a pair plot that shows the relationship between four variables: Open, Close, Volume, and Year. The scatterplots are grouped by year, allowing us to see how the variables have changed over time. Overall, these visualizations provide insight into the performance and behavior of the stock. They help us identify patterns and trends and can aid in making informed decisions about investing in the stock.

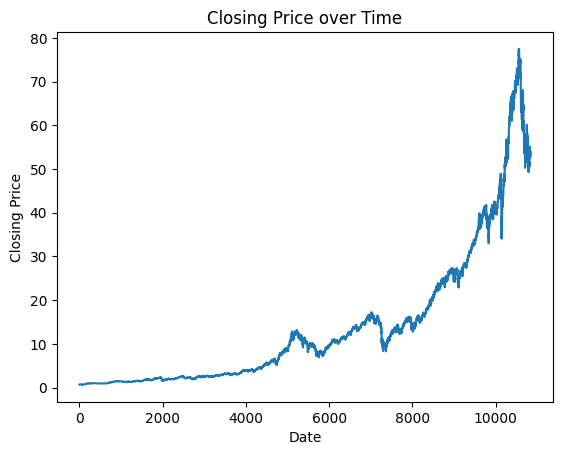
In conclusion, the visualizations created using the matplotlib and seaborn libraries provide valuable insights into the trends and patterns in the stock price data. The line plot of closing price over time shows a clear upward trend, with occasional dips and fluctuations. The correlation heatmap indicates a strong positive correlation between the opening and closing prices, as well as between the high and low prices. The line plot of closing price over time by year reveals a similar pattern across all years, with fluctuations becoming more pronounced in recent years. The box plot of monthly trading volume shows some outliers in the trading volume, indicating periods of increased activity. The histogram of daily price changes shows that the majority of daily price changes fall within a narrow range, but there are some outliers with large price movements. The joint plot of opening and closing prices shows a strong positive correlation between the two variables. The pair plot of open, close, volume, and year provides a comprehensive view of the relationships between the variables, revealing strong positive correlations between open and close, and between volume and year. These visualizations provide useful information for investors and analysts looking to understand trends and patterns in the stock market data.

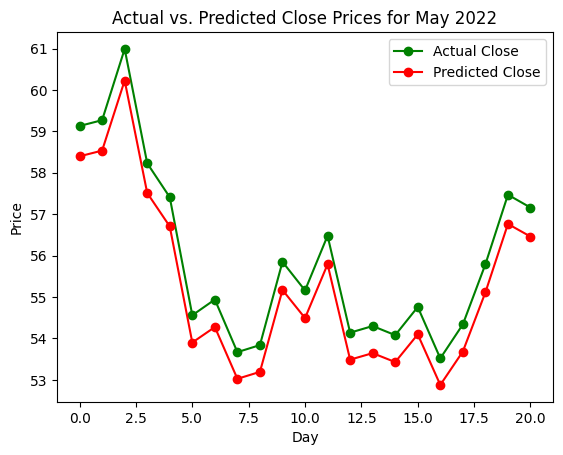
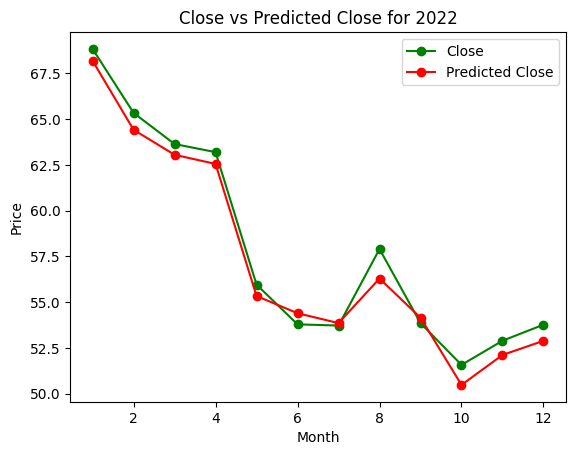
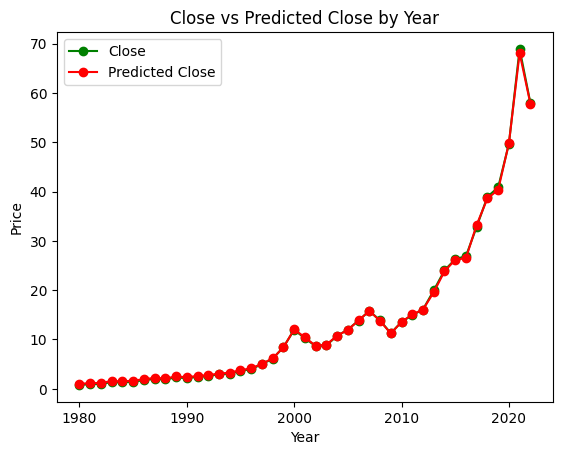
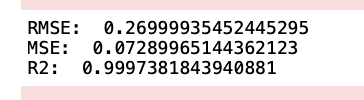
The financial aid starts by presenting the user with a D3-style word cloud generated from the processed trending hashtags related to daily significant news, events, and social trends. The size of each word in the cloud represents the frequency of its occurrence, enabling the user to easily identify any events that could affect their investment decisions. This forms a part of their daily investment workflow. Illustrates the loading of the processed hashtag CSV data using D3. The resulting word cloud visualization can be viewed on the final website.

The `plot\_close\_vs\_predicted\_by\_year()` function uses the trained linear regression model to make predictions on the testing data and plots the actual close values and predicted close values against the year. This allows us to see how well the model performs over time and whether it is able to accurately predict changes in the stock price.

The `plot\_close\_vs\_predicted\_close\_by\_selected\_year()` function filters the data to include only a specific year and then groups the data by month and computes the mean of the Close column. The actual close values for each month are plotted, and the function also makes predictions on the testing data for the specified year and plots the predicted close values for each month. This allows us to see how well the model performs at a more granular level and whether it is able to accurately predict changes in the stock price over shorter time periods.

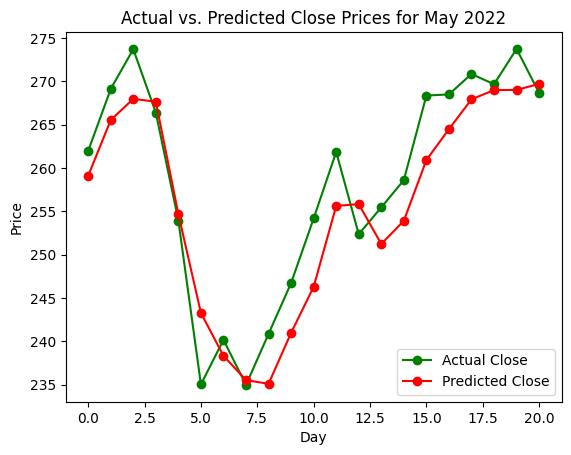
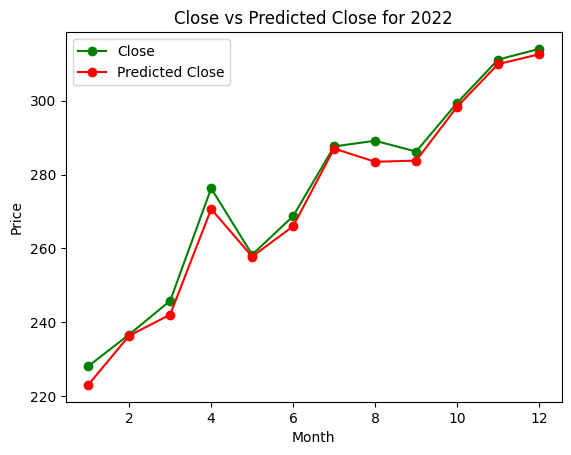
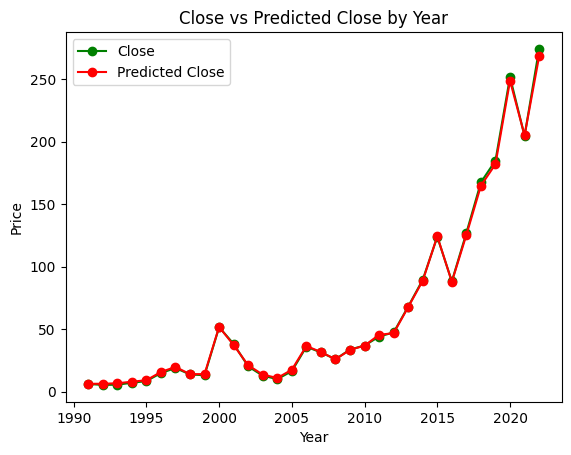
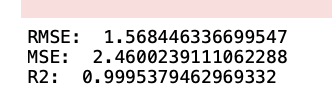
***Performance of model for American Funds Growth Fund of America:***



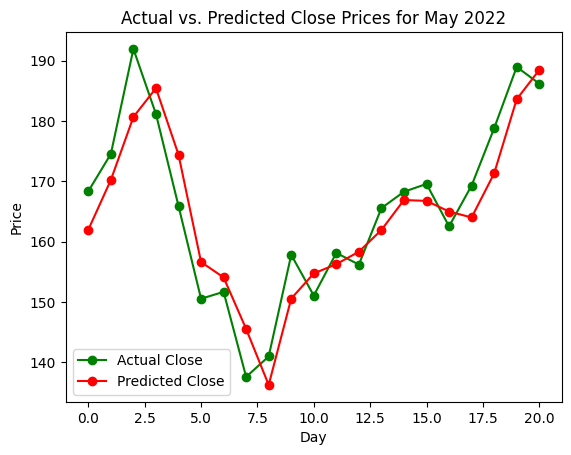
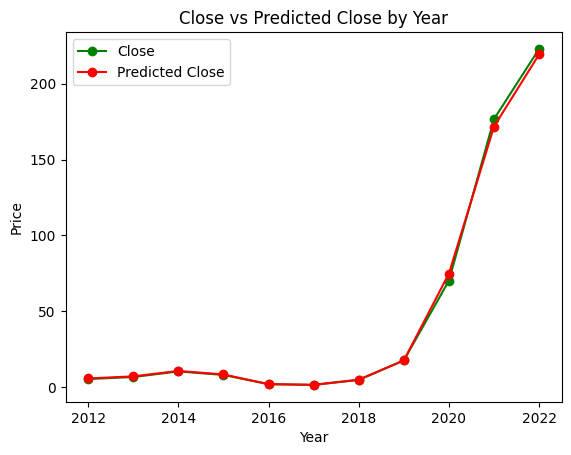
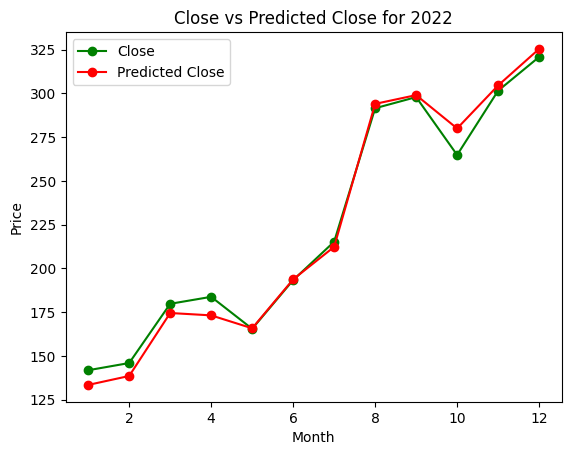
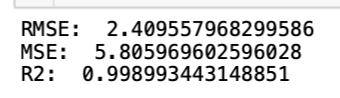


**-----------------------------------------------------------------**

***Performance of model for Vertex Pharmaceuticals Inc:***

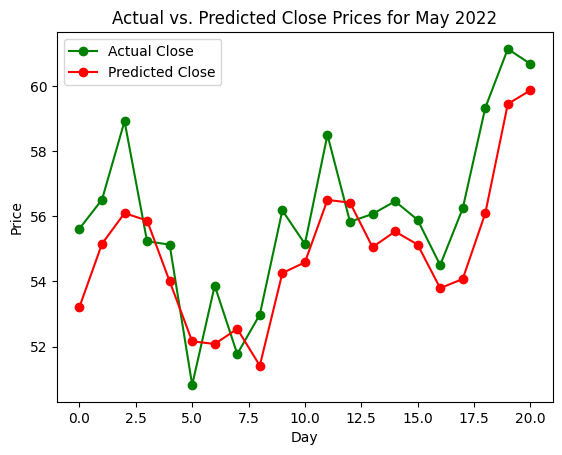
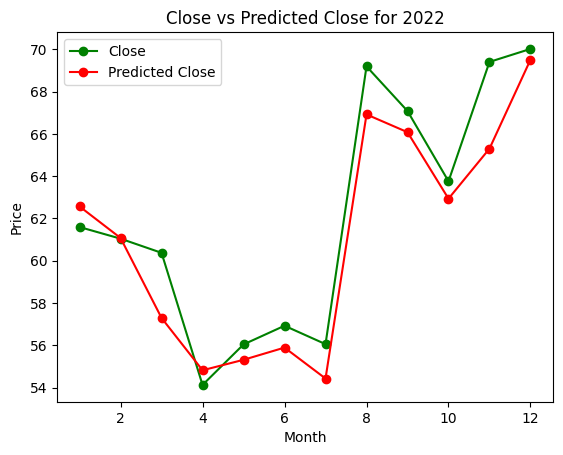
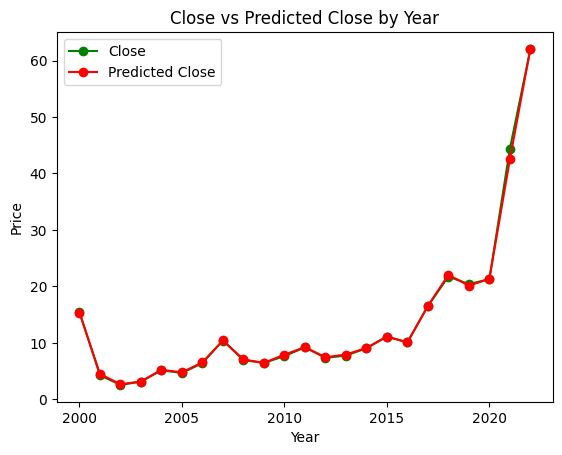
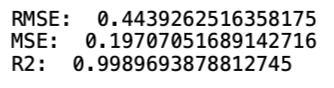
****

***Performance of model for Enphase Energy Inc:***



**----------------------------------------------------------------**

***Performance of model for ON Semiconductor Corp:***

****

Overall, these functions provide a comprehensive analysis of a linear regression model's performance on stock market data and can be used to gain insights into the factors that affect stock prices and how well the model is able to predict changes in the stock market.

1. **CONCLUSION**

The Daily Financial Risk Aid was created to explore the possibility of merging trending world events, past stock data, and future stock value forecasts to provide investors with the necessary information for portfolio growth. This project produced a single cloud computing platform that presents investors with suitable investment options and trending global topics for their daily investment activities. The various visualizations created using the Matplotlib and Seaborn libraries provide valuable insights into stock price trends and patterns. For example, the line plot of closing price over time shows a clear upward trend, with occasional dips and fluctuations. The correlation heatmap reveals strong positive correlations between opening and closing prices, as well as high and low prices. The box plot of monthly trading volume indicates periods of increased activity and some outliers. The histograms of daily price changes display the frequency distribution of price changes over time, helping to identify large price movements. The joint plot and pair plot display the relationships between variables and how they change over time. The project achieved this by integrating a cloud computing backend, leveraging open-source data acquisition libraries, using open-source repository frameworks for storage, and creating a consolidated front-end financial visualizer for users.Based on the concepts proven by this project, future work can therefore focus on enabling the visualization of different investment vehicles, sourcing a variety of social discord from other social media platforms, and utilize further models for investment predictions. This will enable greater fusion of modern technological frameworks and finance for further investment growth potentials.

**REFERENCES**

[1] Google. (n.d.). Enphase Energy Inc (ENPH) stock price & news. Google Finance. Retrieved December 4, 2022, from https://www.google.com/finance/quote/ENPH:NASDAQ?window=MAX

[2] Google. (n.d.). Exxon Mobil Corp (XOM) stock price & news. Google Finance. Retrieved December 4, 2022, from https://www.google.com/finance/quote/XOM:NYSE?sa=X&ved=2ahU KEwj0zrSVvd77AhWvFVkFHdJ-DZMQ3ecFegQINxAY

[3] Sapphirine Big Data Analytics Application Suite (Apache License v2.0) (c)2014-2022, Project name: “Daily Financial Risk Aid”, from <https://www.ee.columbia.edu/~cylin/course/bigdata/projects/>

[4] Google. (n.d.). Occidental Petroleum Corporation (OXY) Stock Price & News. Google Finance. Retrieved December 4, 2022, from https://www.google.com/finance/quote/OXY:NYSE

[5] Google. (n.d.). On Semiconductor Corp (ON) stock price & news. Google Finance. Retrieved December 4, 2022, from https://www.google.com/finance/quote/ON:NASDAQ

[6] Google. (n.d.). Shift4 Payments Inc (four) stock price & news. Google Finance. Retrieved December 4, 2022, from https://www.google.com/finance/quote/FOUR:NYSE

[7] Google. (n.d.). Vertex Pharmaceuticals Incorporated (VRTX) stock price & news. Google Finance. Retrieved December 4, 2022, from https://www.google.com/finance/quote/VRTX:NASDAQ

[8] The Growth Fund of America (AGTHX). A. (n.d.). Retrieved December 4, 2022, from https://www.capitalgroup.com/individual/investments/fund/agthx

[14] OCCIDENTAL PETROLEUM CORP OXY. NYSE. (n.d.). Retrieved December 4, 2022, from https://www.nyse.com/quote/XNYS:OXY

[18] Vanguard Mutual Fund profile (VFIAX). Vanguard. (n.d.). Retrieved December 4, 2022, from https://investor.vanguard.com/investment- products/mutual-funds/profile/vfiax

[19] Vanguard Mutual Fund profile (VSMPX). Vanguard. (n.d.). Retrieved December 4, 2022, from https://investor.vanguard.com/investment- products/mutual-funds/profile/vsmpx

[21] Wikimedia Foundation. (2022, October 19). Occidental Petroleum. Wikipedia. Retrieved December 4, 2022, from https://en.wikipedia.org/wiki/Occidental\_Petroleum

[22] Yahoo! (2022, December 1). 3 reasons why growth investors shouldn't Overlook Enphase Energy (ENPH). Yahoo! Finance. Retrieved December 4, 2022, from https://finance.yahoo.com/news/3-reasons-why- growth-investors-174505497.html

[26] Finance Analytics & Bi Software. Sisense. (2022, October 24). Retrieved December 20, 2022, from https://www.sisense.com/solutions/finance/?utm\_source=quora&utm\_m edium=referral

[27] Financial charting software for Stock Market Data Visualization. XB Software. (n.d.). Retrieved December 20, 2022, from https://xbsoftware.com/case-studies-webdev/rate-management-system/