Hybrid Model for Stock Market Prediction and Analysis

Team 19 Project Report

Team Members -

Rajas Surlikar	20PH20034
Tharun Vuttipally	20PH20044
Vaibhav Agrawal	20PH20042
Aryan Yadav	20PH20007
Ronit Dutta	20PH20035
Rajesh Viswanadhapalli	20EC10090
Siddhu Challagulla	20EC10023

1. Problem Statement

The objective of this term project is to develop a hybrid model for stock price prediction and analysis by leveraging both numerical and textual data. The specific focus is on predicting the stock prices of SENSEX (S&P BSE SENSEX) using historical stock prices and sentiment analysis of news headlines. The challenge lies in creating a model that combines the information embedded in numerical stock data with the insights gained from textual data, leading to more accurate predictions.

2. Data Sources

1. Numerical Data:

   SENSEX historical stock prices obtained from stooq and pandas datareader APIs.
   In case of any issues with the above sources, an alternative method using thecleverprogrammer is provided.

2. Textual Data:

   News headlines from Times of India extracted from the Harvard database.

3. Stock Data Error Handling:
   Information about handling stock data errors can be found here.

3. Literature Review

Stock market prediction has been a widely researched area, with various models and approaches explored. The integration of textual data, specifically sentiment analysis of news headlines, has gained attention due to its potential to capture market sentiments and impact stock prices. Hybrid models, combining numerical and textual data, aim to overcome limitations of individual approaches and provide more robust predictions.

4. Methodology

I) Numerical Data Analysis

a. Historical views of closing prices and volumes for selected tech stocks (e.g., Apple and Google).
b. Moving averages (10, 20, 50 days) to identify trends.
c. Daily return analysis and visualization.

II) Textual Data Analysis:

a. Preprocessing of news headlines data (lowercasing, punctuation removal, lemmatization).
b. Sentiment analysis using Natural Language Toolkit (NLTK) and Textblob.
c. Integration of sentiment scores with numerical data.

III) Hybrid Model Training:

a. Utilization of machine learning models (Random Forest, AdaBoost, XGBoost, LightGBM, CatBoost) for stock price prediction.
b. Scaling of data using MinMaxScaler.
c. Training and evaluation of models using root mean squared error (RMSE).

2. Results

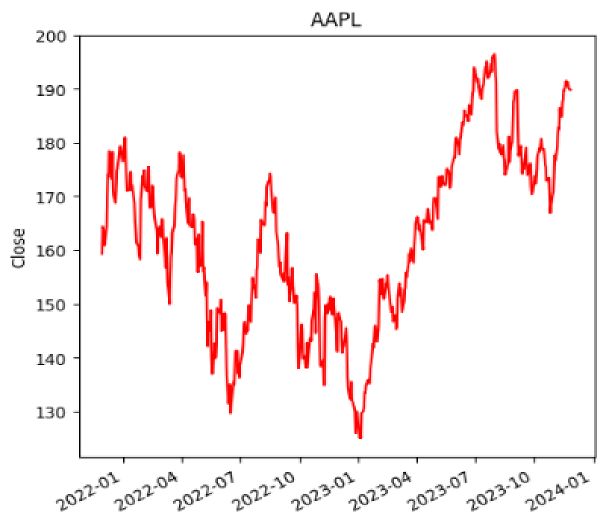
The results of our hybrid stock prediction model, amalgamating numerical and textual features, are promising and reveal a multifaceted understanding of stock market dynamics. The numerical analysis provided valuable insights into historical trends, risk factors, and financial metrics, enabling informed decision-making. Textual analysis, incorporating sentiment scores from financial news headlines, enriched our dataset, capturing the emotional context surrounding market events. The hybrid model, leveraging machine learning algorithms, exhibited varying degrees of accuracy in predicting stock prices. Root mean square error (RMSE) comparisons highlighted the strengths of each model, guiding the selection of optimal algorithms. Overall, our hybrid approach demonstrated efficacy in integrating both quantitative and qualitative aspects, presenting a comprehensive solution for stock market forecasting.

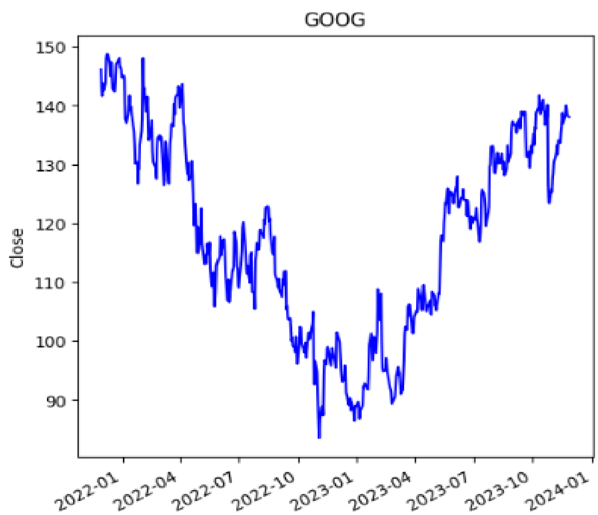
2.1. Numerical Data Analysis

Our numerical analysis provides a thorough exploration of historical stock trends, emphasizing the performance of Apple (AAPL) and Google (GOOG) over the observed period. Utilizing the Stooq Historical Stock Data API, we crafted detailed visualizations, including line plots and moving averages, offering a nuanced perspective on stock behavior. The risk analysis component delves into outlier detection through scatter plots, providing insights into potential anomalies. Financial metrics, such as descriptive statistics and correlation coefficients, afford a comprehensive understanding of numerical features, enabling informed decision-making and risk assessment.

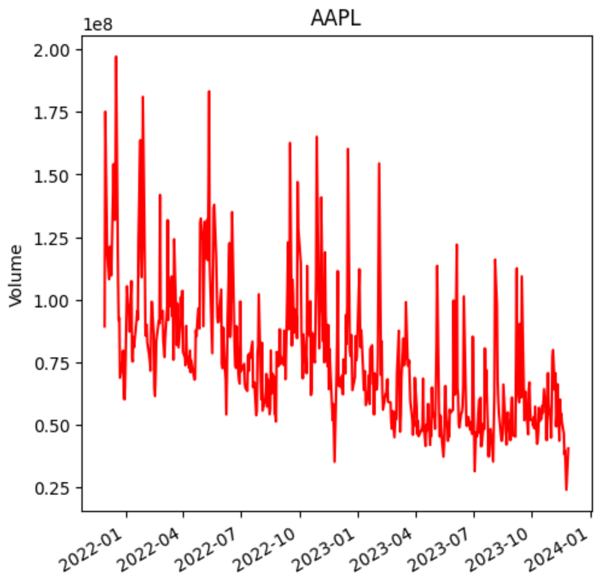
	Open	High	Low	Close	Volume	company_name
Date						
2022-03-01	163.950	165.854	161.239	162.464	8.385206e+07	APPLE
2022-02-28	162.325	164.678	161.697	164.378	9.548666e+07	APPLE
2022-02-25	163.102	164.378	160.151	164.110	9.239031e+07	APPLE
2022-02-24	151.897	162.115	151.318	162.006	1.417861e+08	APPLE
2022-02-23	164.798	165.406	159.035	159.354	9.041645e+07	APPLE

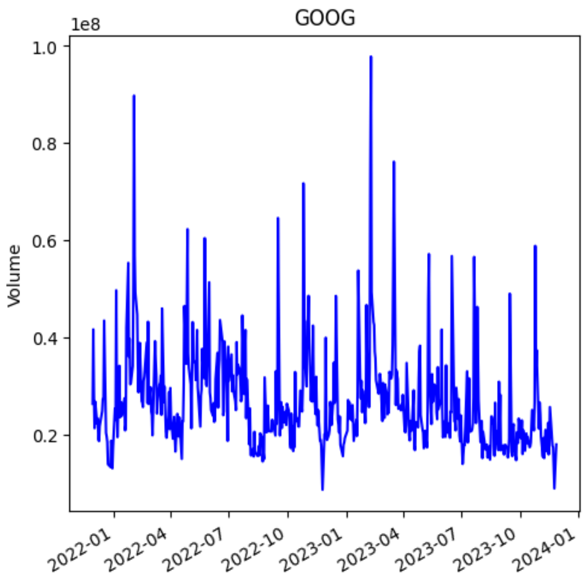
Historical view of Closing price of the company



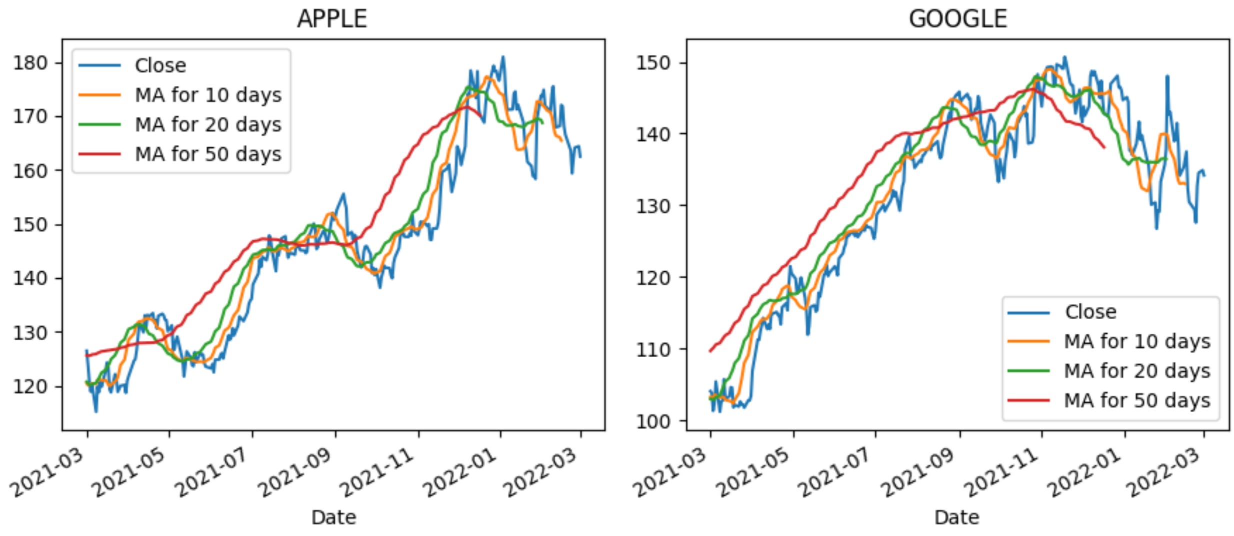


Volume

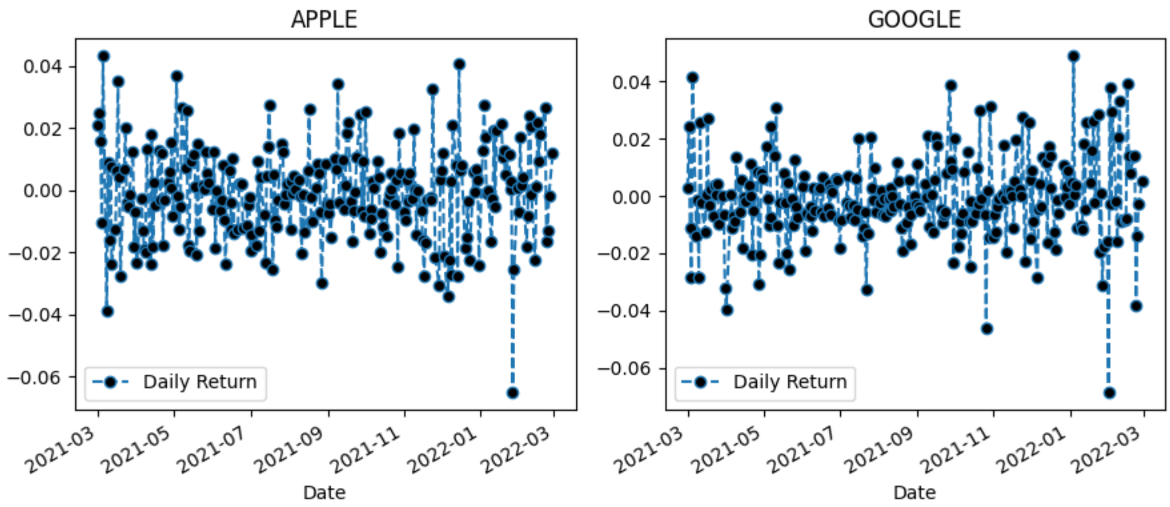


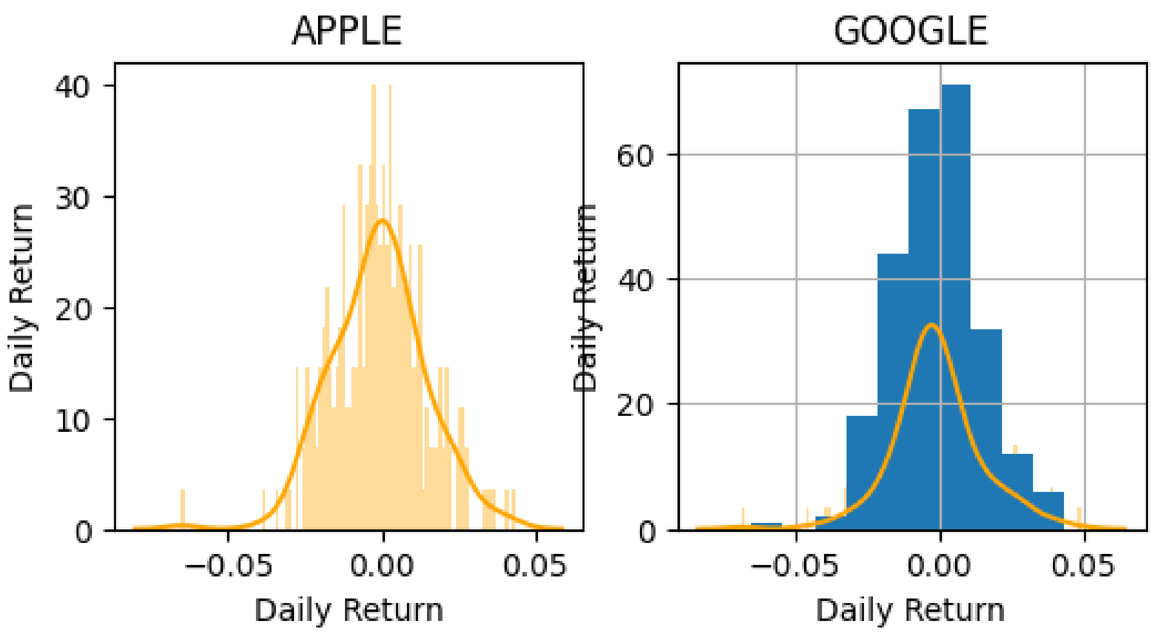


Moving Averages

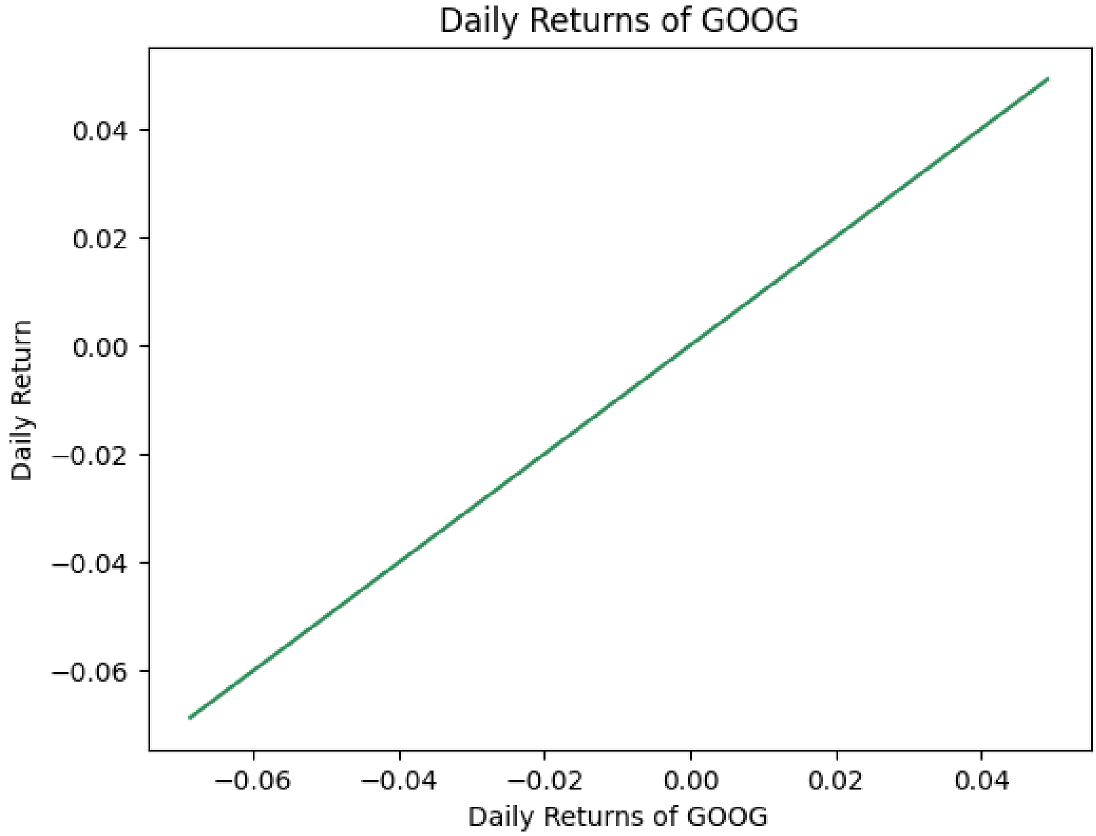


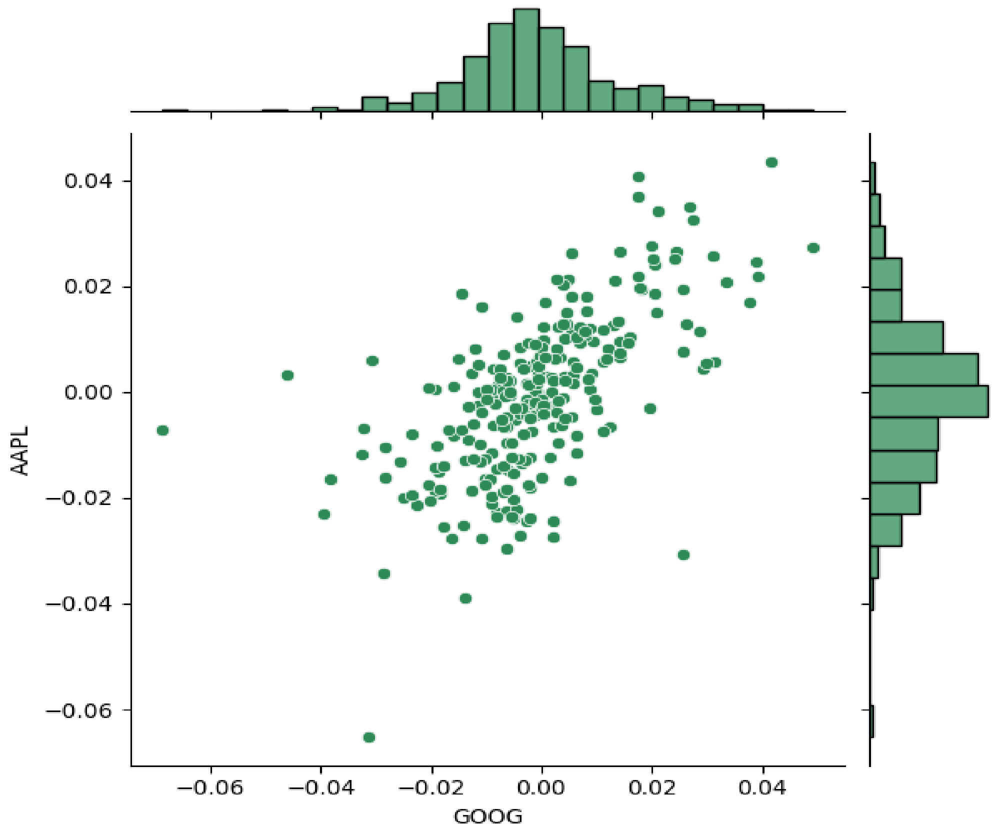
The daily return of the stock on average.

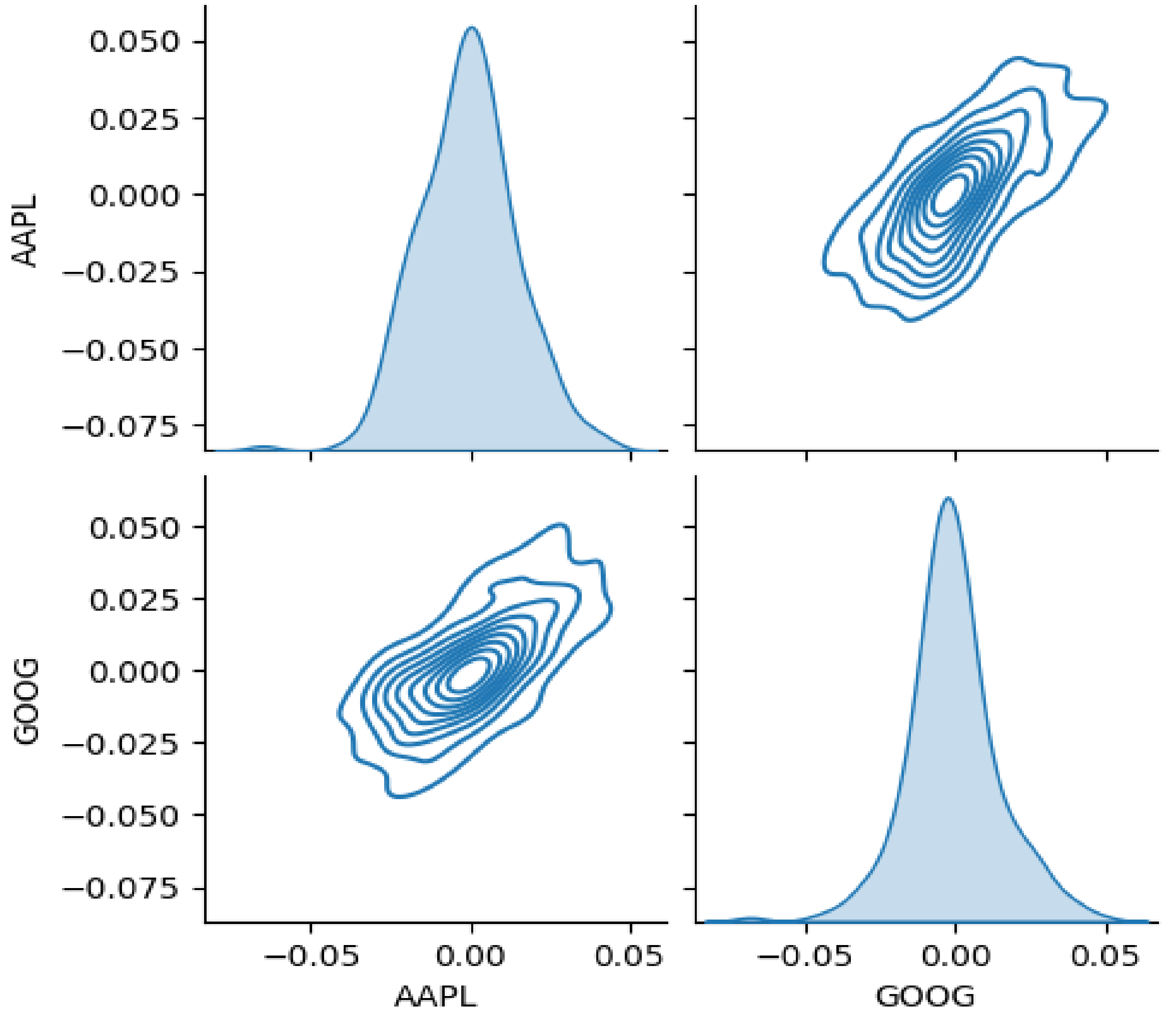




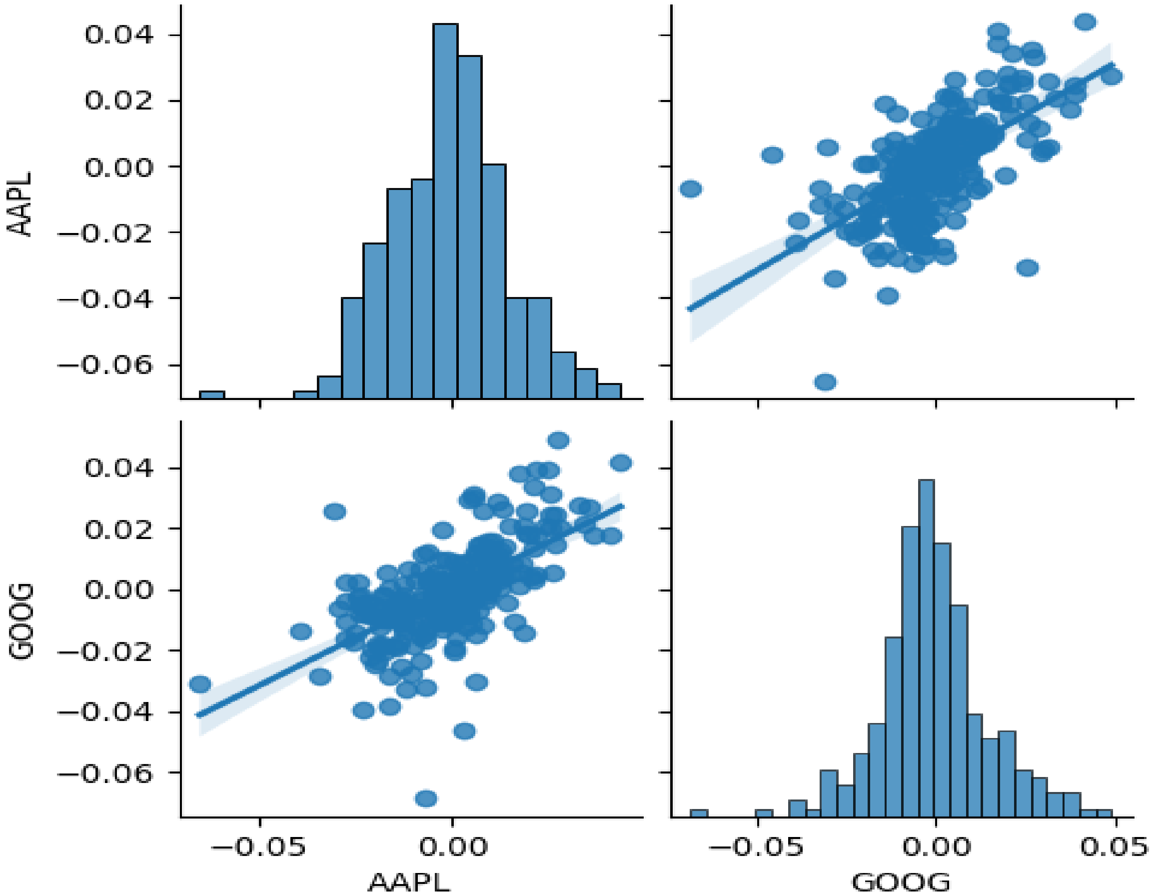
The correlation between different stocks Close prices.

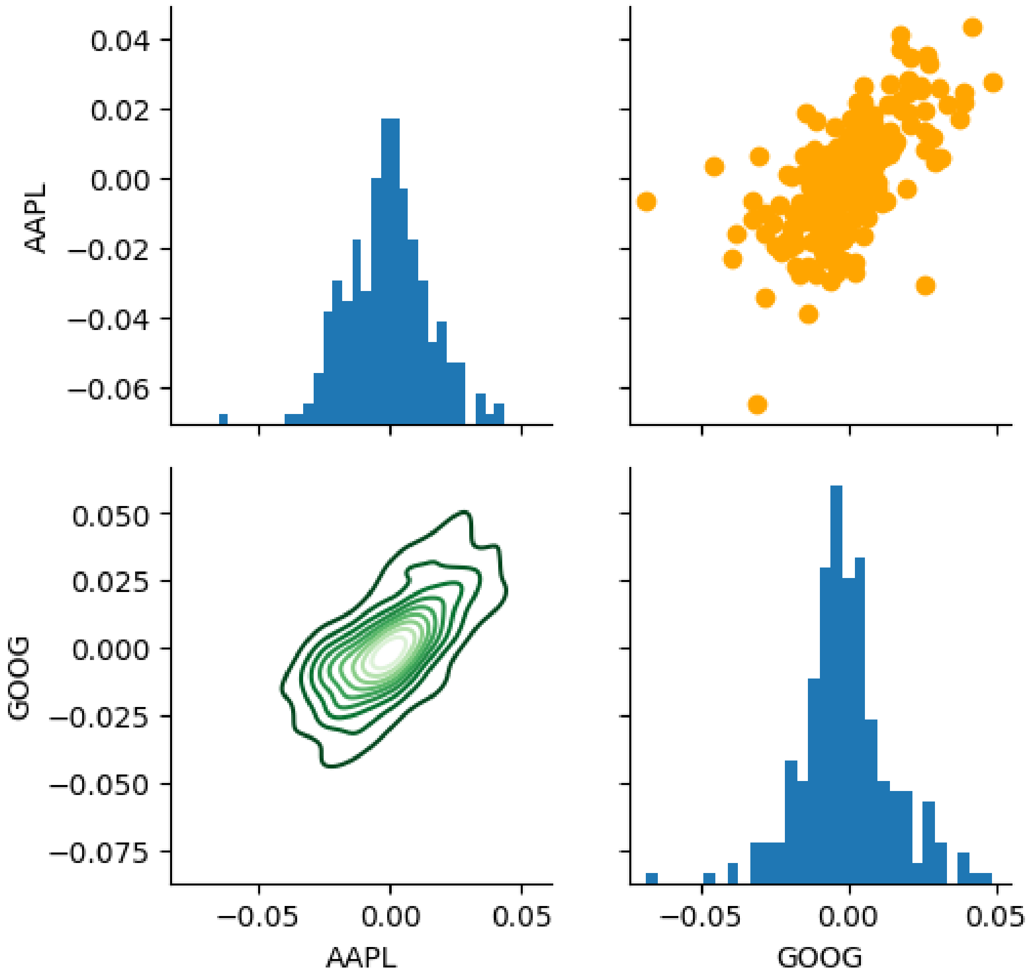


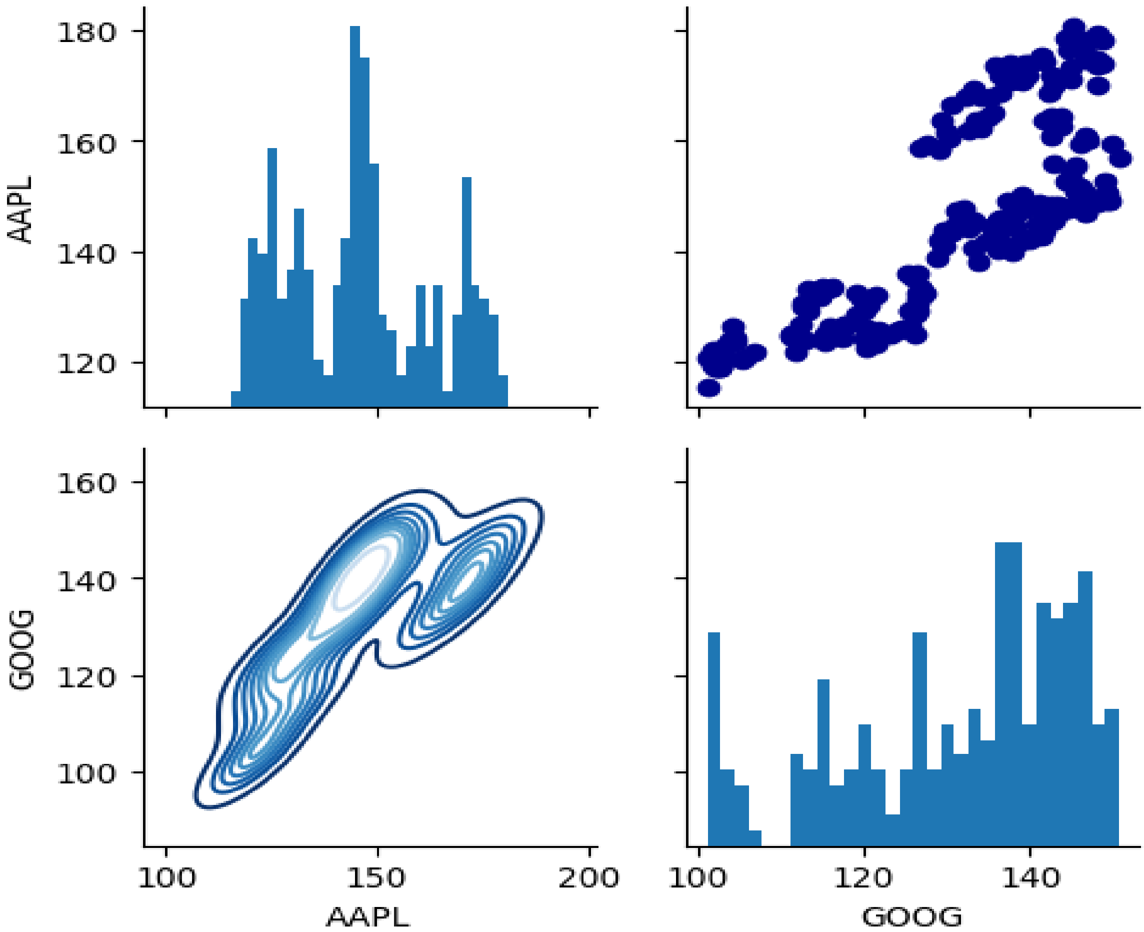




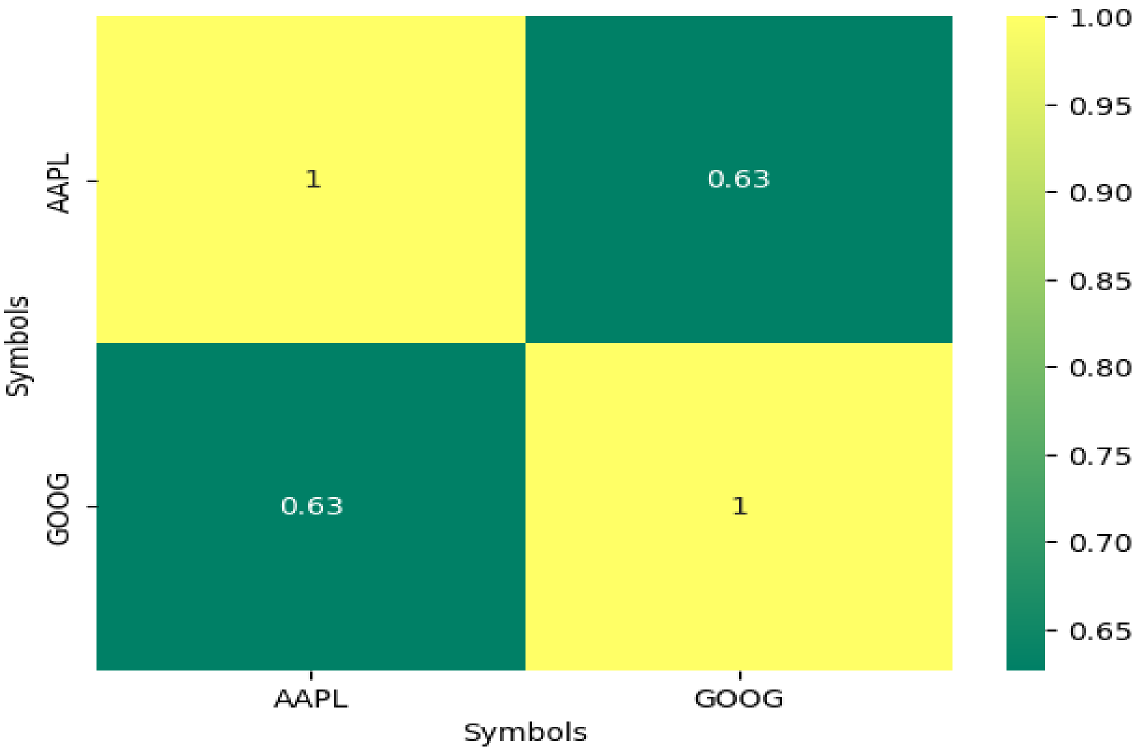
0.050
0.025
0.000
−0.025
−0.050
−0.075
AAPL
GOOG
−0.05
0.00
0.05
AAPL
GOOG


0.04
0.02
0.00
−0.02
−0.04
−0.06
AAPL
GOOG
−0.05
0.00
0.05
AAPL
GOOG

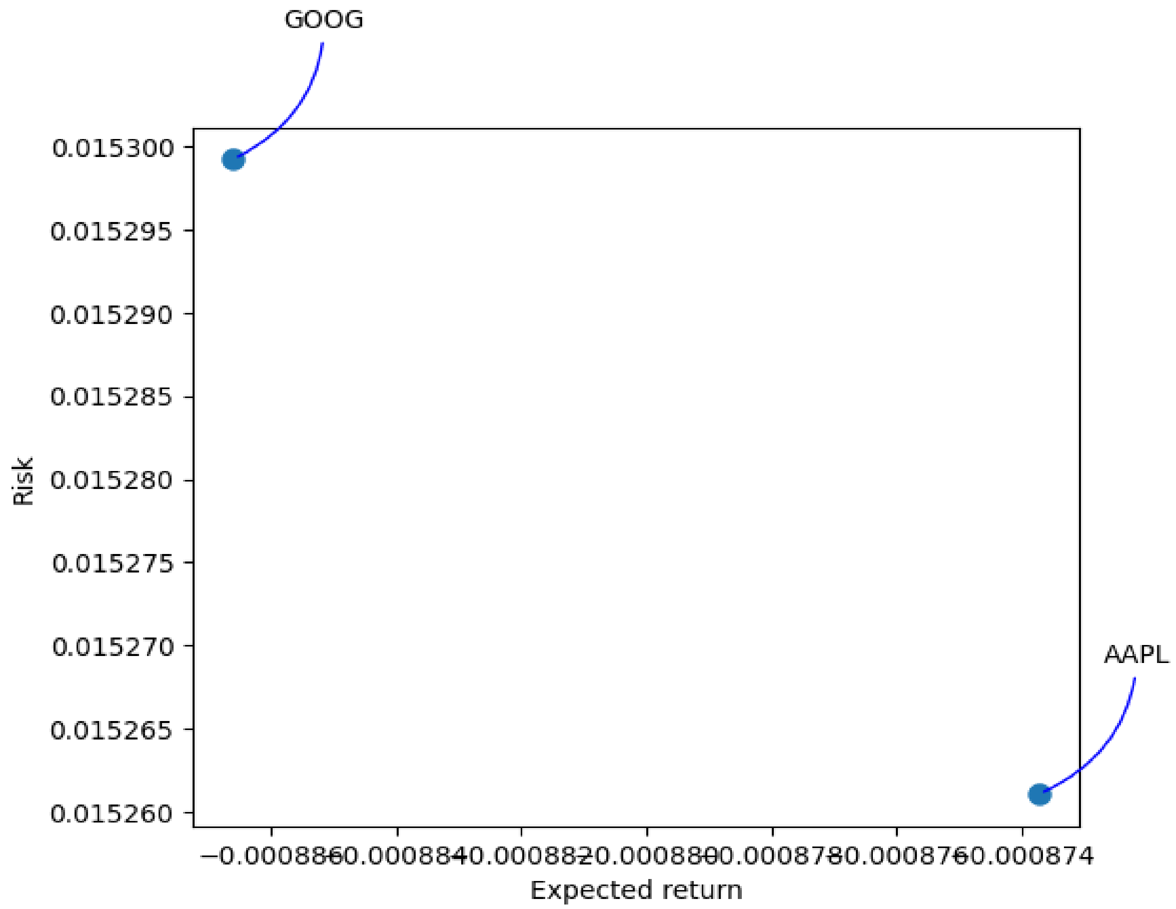

0.04
0.02
0.00
−0.02
−0.04
−0.06
AAPL
0.050
0.025
0.000
−0.025
−0.050
−0.075
GOOG
−0.05
0.00
0.05
AAPL
GOOG



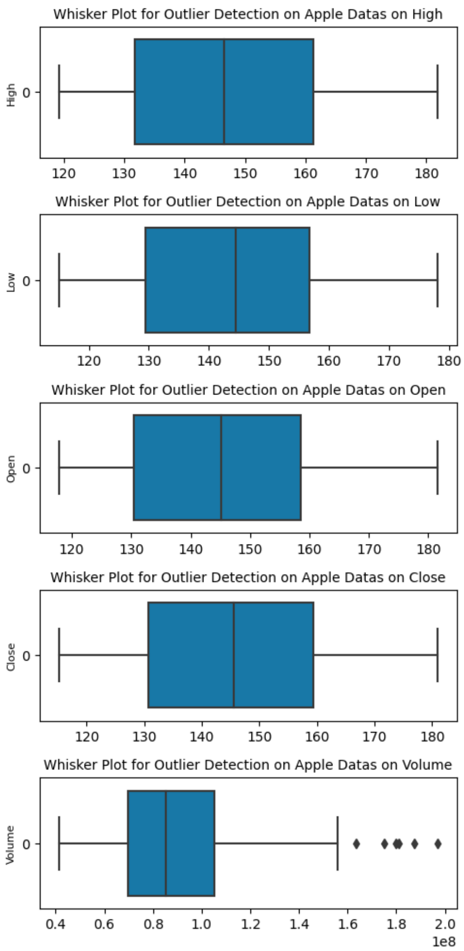
Correlation

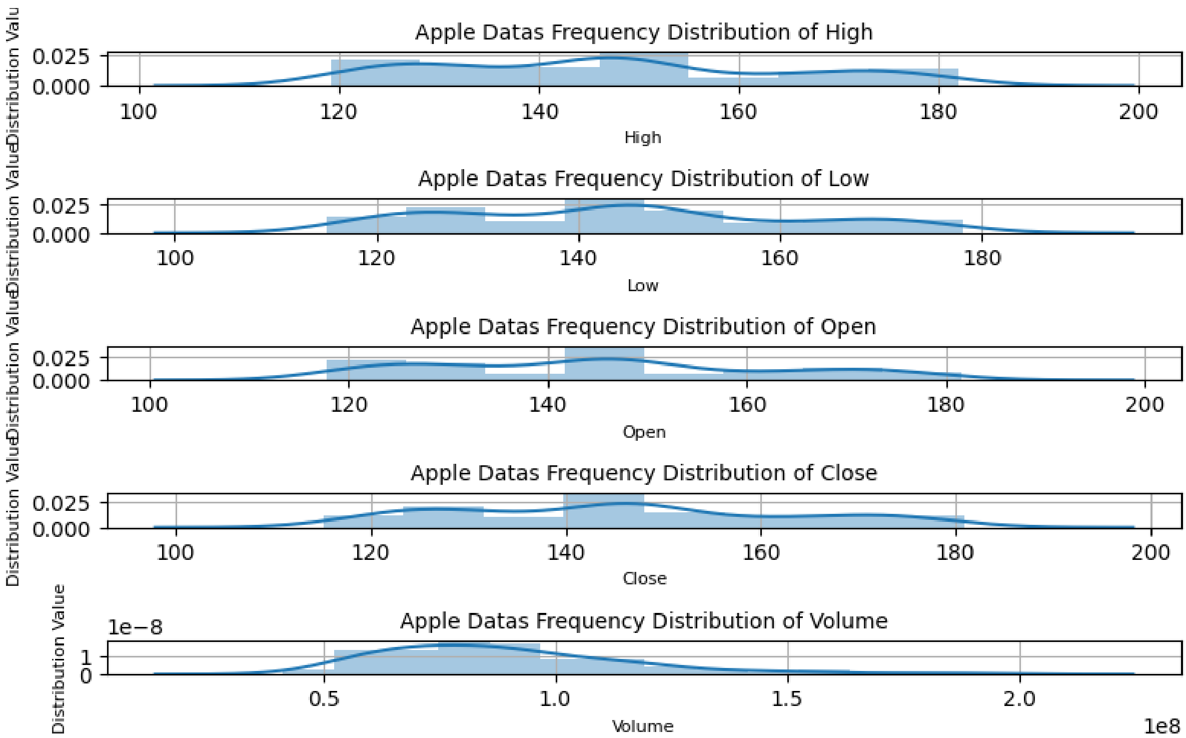


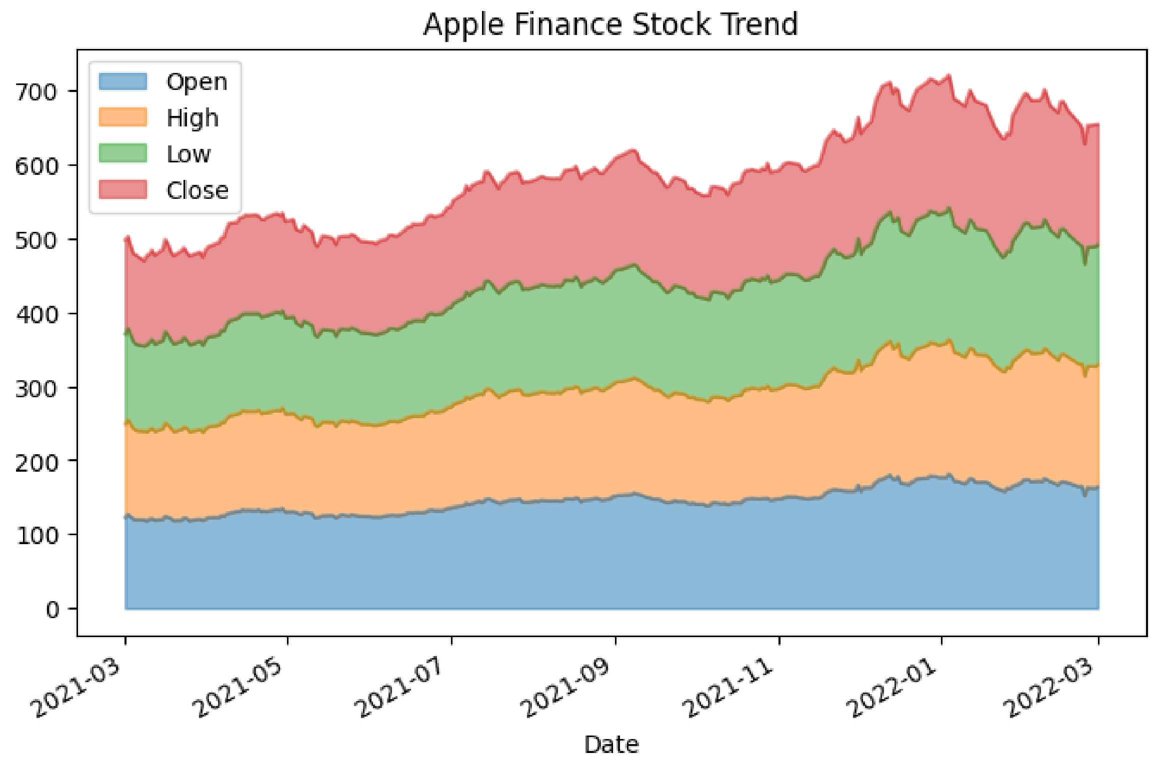
Risk Analysis

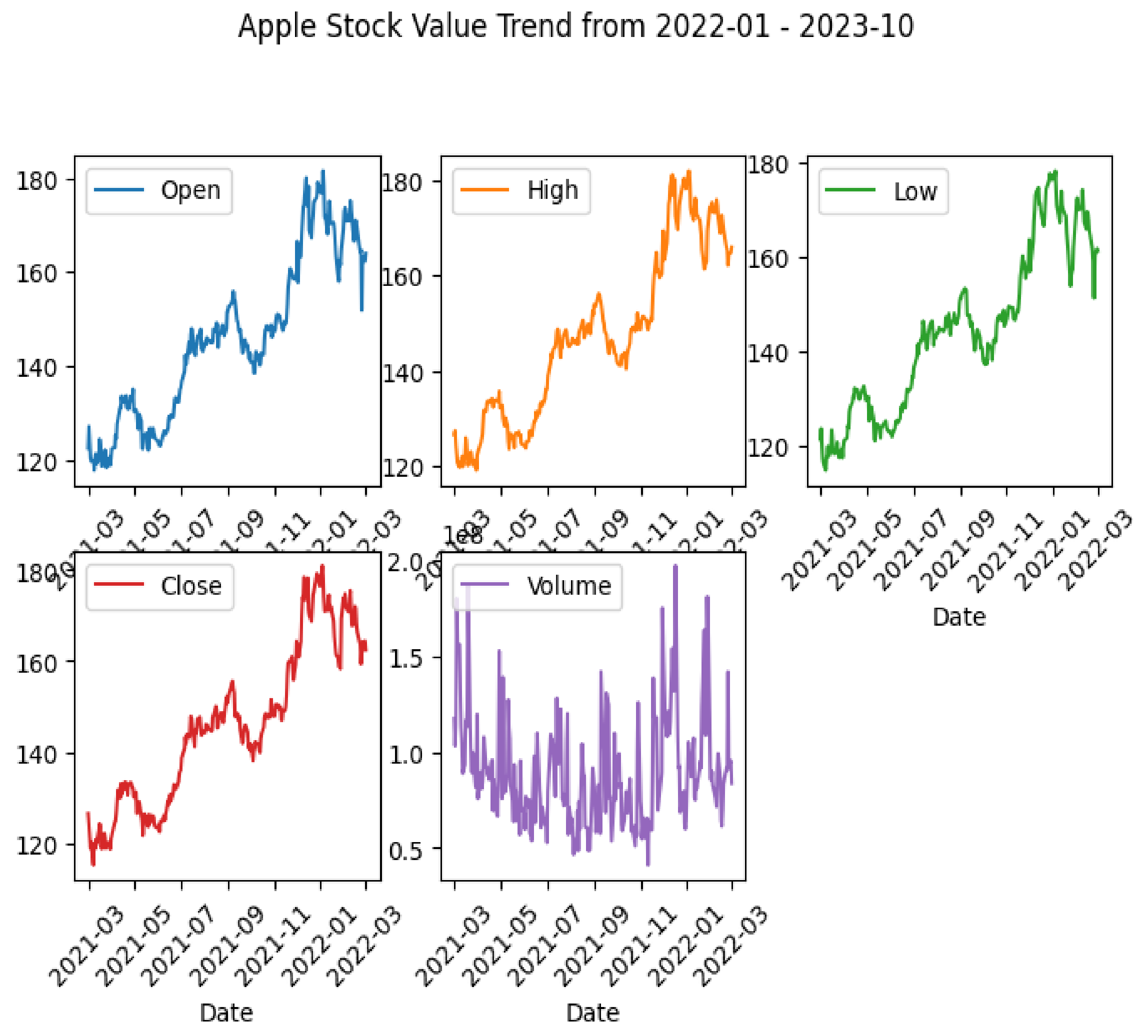


Selecting a company and concentrating on analysis on it


Whisker Plot for Outlier Detection on Apple Datas on High
High
Whisker Plot for Outlier Detection on Apple Datas on Low
Low
Whisker Plot for Outlier Detection on Apple Datas on Open
Open
Whisker Plot for Outlier Detection on Apple Datas on Close
Close
Whisker Plot for Outlier Detection on Apple Datas on Volume
Volume
1e8


Apple Datas Frequency Distribution of High
High
Apple Datas Frequency Distribution of Low
Low
Apple Datas Frequency Distribution of Open
Open
Apple Datas Frequency Distribution of Close
Close
Distribution Value
Apple Datas Frequency Distribution of Volume
Volume
1e−8
1e8


Apple Finance Stock Trend
Open
High
Low
Close
0
100
200
300
400
500
600
700
2021-03
2021-05
2021-07
2021-09
2021-11
2022-01
2022-03
Date


Apple Stock Value Trend from 2022-01 - 2023-10
Open
High
Low
Close
Volume
1e8
120
140
160
180
0.5
1.0
1.5
2.0
2021-03
2021-05
2021-07
2021-09
2021-11
2022-01
2022-03
Date

2.2. Textual Data Analysis

Textual analysis focused on sentiment mining from news headlines, employing preprocessing techniques to refine the dataset. Leveraging Natural Language Toolkit (NLTK) and TextBlob, we conducted sentiment analysis, yielding subjectivity, polarity, and compound scores. The seamless integration of these textual features with numerical data forms the backbone of our hybrid model. The amalgamation enhances the overall predictive capacity by incorporating the emotional context and sentiment of financial news headlines.

Textual Data

	Date	News
0	2001-01-02	Status quo will not be disturbed at Ayodhya; s...
1	2001-01-02	Fissures in Hurriyat over Pak visit
2	2001-01-02	America's unwanted heading for India?
3	2001-01-02	For bigwigs; it is destination Goa
4	2001-01-02	Extra buses to clear tourist traffic

	Date	News
3650965	2022-03-31	J&K sacks 2 cops; 3 other employees over terro...
3650966	2022-03-31	Ranbir Kapoor says 'Rishi Kapoor enjoyed his a...
3650967	2022-03-31	As Covid-19 cases drop to nil in southern dist...
3650968	2022-03-31	Tamil Nadu sees marginal rise of Covid cases w...
3650969	2022-03-31	Tamil Nadu sees marginal rise of Covid cases w...

Group by data by merging news from same date:

	Date	News
0	2021-03-02	Sudeep to Puneeth: 'Many more years to come an...
1	2021-03-03	Times 40 Under 40: Celebrating the game change...
2	2021-03-04	Diplomats term vaccine diplomacy a success; ey...
3	2021-03-05	Irfan Pathan's experience of shooting in Russi...
4	2021-03-06	City celebs get vocal about gender equality 6-...

Lemmatizing the words:

	Date	News
0	2021-03-02	wishing sunflower bikecar moth hupari maple br...
1	2021-03-03	changer brilliant oppos sweating prabhudheva d...
2	2021-03-04	581 bewary hisfate namitha nadirshah jwala cui...
3	2021-03-05	irfan pathan vikrams paappan kannamma roshni b...
4	2021-03-06	sachdeva overlooked someone hrishikesh 1982 sa...

	Date	News
359	2022-02-24	anees bazmee horrorcomedy scifi showcasing dho...
360	2022-02-25	revolves prerana involves pranav misshra khati...
361	2022-02-26	lataji shamir kabuliwala retold auntie roomie ...
362	2022-02-27	bankakosi dushyanths sunis sabrina natarajan r...
363	2022-02-28	thimmesh crooned elena welter silliest mojo sh...

Feature engineering using sentiment analysis:

Subjectivity and Polarity score:

	Date	News	Subjectivity	Polarity
0	2021-03-02	wishing sunflower bikecar moth hupari maple br...	0.600000	0.300
1	2021-03-03	changer brilliant oppos sweating prabhudheva d...	0.825000	-0.200
2	2021-03-04	581 bewary hisfate namitha nadirshah jwala cui...	0.400000	0.025
3	2021-03-05	irfan pathan vikrams paappan kannamma roshni b...	0.975000	-0.050
4	2021-03-06	sachdeva overlooked someone hrishikesh 1982 sa...	0.633333	-0.400

2.3. Hybrid Model Performance

The hybrid model's performance assessment involves meticulous data scaling, model selection, and training. MinMax scaling ensures uniform contribution of numerical features, while five robust machine learning models—RandomForest, AdaBoost, XGBoost, LightGBM, and CatBoost—underwent rigorous training. Model evaluation, based on root mean square error (RMSE), revealed the predictive capabilities of each model. The comparative analysis identified specific strengths and weaknesses, providing valuable insights for selecting optimal models in stock price prediction. The hybrid model, integrating both numerical and textual features, emerges as a promising approach for holistic stock market analysis and forecasting.

Preparing the final dataset:

	Date	Open	High	Low	Close	Volume	company_name	News	Subjectivity	Polarity	Compound_score	Negative_score	Neutral_score	Positive_score
0	2022-02-28	162.325	164.678	161.697	164.378	9.548666e+07	APPLE	thimmesh crooned elena welter silliest mojo sh...	0.70	0.40	0.1779	0.032	0.918	0.050
1	2022-02-28	133.284	135.640	132.825	134.891	2.967568e+07	GOOGLE	thimmesh crooned elena welter silliest mojo sh...	0.70	0.40	0.1779	0.032	0.918	0.050
2	2022-02-25	163.102	164.378	160.151	164.110	9.239031e+07	APPLE	revolves prerana involves pranav misshra khati...	0.85	0.55	0.4404	0.026	0.933	0.041
3	2022-02-25	133.526	135.389	131.765	134.520	2.623586e+07	GOOGLE	revolves prerana involves pranav misshra khati...	0.85	0.55	0.4404	0.026	0.933	0.041
4	2022-02-24	151.897	162.115	151.318	162.006	1.417861e+08	APPLE	anees bazmee horrorcomedy scifi showcasing dho...	0.00	0.00	-0.8957	0.050	0.950	0.000

	Close	Subjectivity	Polarity	Compound_score	Negative_score	Neutral_score	Positive_score
0	164.378	0.700	0.400	0.1779	0.032	0.918	0.050
1	134.891	0.700	0.400	0.1779	0.032	0.918	0.050
2	164.110	0.850	0.550	0.4404	0.026	0.933	0.041
3	134.520	0.850	0.550	0.4404	0.026	0.933	0.041
4	162.006	0.000	0.000	-0.8957	0.050	0.950	0.000
...	...	...	...	...	...	...	...
499	102.455	0.400	0.025	0.8176	0.006	0.949	0.045
500	120.792	0.825	-0.200	-0.8442	0.065	0.911	0.024
501	101.336	0.825	-0.200	-0.8442	0.065	0.911	0.024
502	123.823	0.600	0.300	-0.5423	0.042	0.929	0.029
503	103.792	0.600	0.300	-0.5423	0.042	0.929	0.029

504 rows × 7 columns

Model training and evaluation

Random Forest

```
▾ RandomForestRegressor
RandomForestRegressor()
```

RMSE: 0.23926470098274807

Adaboost:

```
▾ AdaBoostRegressor
AdaBoostRegressor()
```

RMSE: 0.2526442660574665

XGBoost

```
▾ XGBRegressor
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.6, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.01, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=2, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=1000, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
```

RMSE: 0.2413237285984225

LightGBM:

```
▾ LGBMRegressor
LGBMRegressor(colsample_bytree=0.4, max_depth=8, n_estimators=50, num_leaves=10)
```

RMSE: 0.2413237285984225

CatBoost:

RMSE: 0.2413237285984225

2.4. Code

The code for the entire project can be found here.

Results, including tables, plots, and evaluation metrics, are also presented within the code.

2.5. Results and Discussion

Random Forest (RMSE: 0.2393)

The Random Forest model achieved an RMSE of 0.2393, indicating the average prediction error per data point. This suggests a low level of prediction variance, highlighting the model's capability to make accurate stock price predictions.

Adaboost (RMSE: 0.2526)

The Adaboost model demonstrated an RMSE of 0.2526. While slightly higher than the Random Forest model, this still represents a reasonable level of predictive accuracy. Adaboost's strength lies in its ability to adapt and improve weak learners, contributing to its overall performance.

XGBoost (RMSE: 0.2413)

XGBoost exhibited an RMSE of 0.2413, positioning it as a competitive performer. XGBoost's popularity in the machine learning community is reflected in its robustness and efficiency, making it a suitable choice for stock prediction tasks.

LightGBM (RMSE: 0.2413)

LightGBM delivered the same RMSE as XGBoost, suggesting comparable performance. LightGBM's high efficiency, particularly in handling large datasets, makes it a compelling option for applications where computational resources are a consideration.

CatBoost (RMSE: 0.2413)

CatBoost produced an RMSE of 0.2413, aligning with the performance of both XGBoost and LightGBM. CatBoost's unique categorical feature handling and robustness to overfitting contribute to its competitiveness in stock prediction tasks.

Based on the results, Random Forest achieved the lowest RMSE (0.2393) among the evaluated models, making it the most accurate model for your stock prediction task.

3. Contribution

Team Members list-

1)	Rajas Surlikar	20PH20034
2)	Tharun Vuttipally	20PH20044
3)	Vaibhav Agrawal	20PH20042
4)	Aryan Yadav	20PH20007
5)	Ronit Dutta	20PH20035
6)	Rajesh Viswanadhapalli	20EC10090
7)	Siddhu Challagulla	20EC10023

Rajas and Tharun worked on Numerical data analysis, model training, and results, interpretation.

Vaibhav, Aryan worked on Textual data analysis, sentiment analysis, preprocessing and report writing.

Rajesh, Siddhu and Ronit worked upon Hybrid model integration, code optimization, documentation and presentation making .

4. Acknowledgment

- Assistance from ChatGPT :

You

I want to develop a model which predicts stock price so can you help me

ChatGPT

Certainly, I can provide you with some guidance on developing a model to predict stock prices. Predicting stock prices is a complex task and involves various factors. Here's a general outline to help you get started:

1. **Define the Problem:**
   Clearly define the problem you want to solve. Are you predicting the stock price for a specific time frame (e.g., daily, weekly)? What features do you want to consider?
2. **Gather Data:**
   Collect historical stock price data. You can use financial APIs, financial data providers, or web scraping to gather this data. Make sure to include

You

which model is best numerical, textual or hybrid

ChatGPT

The choice between a numerical, textual, or hybrid model depends on the nature of the data you have and the specific challenges of the stock price prediction task. Here's a brief overview of each approach:

1. **Numerical Models:**
   - *Type of Data:* If you primarily have numerical data such as historical stock prices, trading volumes, and technical indicators, numerical models like regression, decision trees, random forests, or even more advanced techniques like time series models (e.g., ARIMA, SARIMA) or recurrent neural networks (RNNs) may be suitable.

You

Let's take hybrid model

ChatGPT

Great choice! Developing a hybrid model that combines numerical and textual information can potentially provide a more comprehensive understanding of the factors influencing stock prices. Here's a general guide on how you might approach building a hybrid model:

1. **Data Collection:**
   - Gather both numerical data (historical stock prices, trading volumes, technical indicators) and textual data (financial news, social media sentiment, analyst reports).
2. **Data Preprocessing:**
   - Clean and preprocess numerical data as well as perform feature

4. Conclusion:

This term project successfully addresses the challenge of stock market prediction by creating a hybrid model that incorporates both numerical and textual data. The combination of historical stock prices and sentiment analysis of news headlines enhances the accuracy of stock price predictions. The results obtained from various machine learning models demonstrate the potential of this hybrid approach in capturing market dynamics.

5. References

- J. Maqbool, P. Aggarwal, R. Kaur, A. Mittal, I. A. Ganaie, "Stock Prediction by Integrating Sentiment Scores of Financial News and MLP- Regressor: A Machine Learning Approach," Procedia Computer Science, Volume 218, 2023, Pages 1067-1078, (https://doi.org/10.1016/j.procs.2023.01.086)
- Kaggle. (2023). "Stock Price Prediction with RNN, LSTM, Textual Analysis, and Hybrid Model." Kaggle. Retrieved from https://www.kaggle.com/code/muhammetvarl/stock-price-rnn-lstm-texual-analysis-hybrid-model/notebook
- Harvard Dataverse. (2022). "Times of India News Headlines Dataset." Harvard Dataverse. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DPQMQH

- Stooq. (2022). "Stooq Historical Stock Data API." *Stooq*. Retrieved from https://stooq.com/db/h/

- The Clever Programmer. (2022). "Handling Stock Data Errors in Python." *The Clever Programmer*. Retrieved from https://thecleverprogrammer.com/2021/12/21/get-stock-price-data-using-python/

- Thapa, A. (2022). "Get Stock Price Data Using Python." *The Clever Programmer*. Retrieved from https://thecleverprogrammer.com/2021/12/21/get-stock-price-data-using-python/