FINANCE AND RISK ANALYTICS PROJECT REPORT

Prepared By: Barkha Agarwal

PART-1

Credit Risk

Problem Statement

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Data that is available includes information from the financial statement of the companies for the previous year.

Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

Perform the following in given order:

1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach.

1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model.

1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach.

1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model.

1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve).

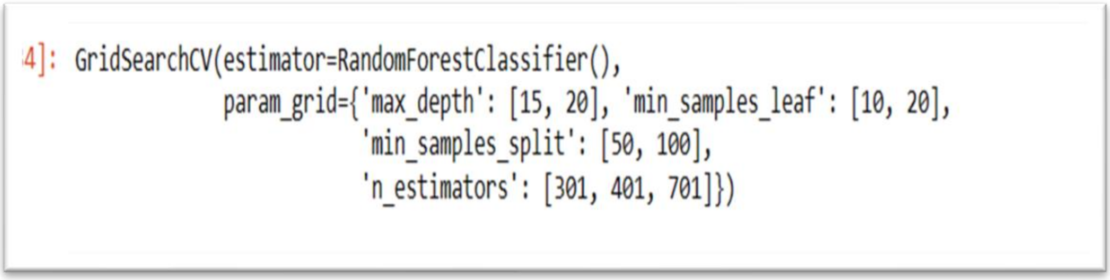
1.13 State Recommendations from the above models.

INTRODUCTION TO PROBLEM STATEMENT

The provided information indicates that the goal is to predict credit default for businesses or companies using their balance sheet data. The dataset includes information from the financial statements of companies for the previous year. The "Default" variable in the dataset will be considered as the dependent variable for the prediction.

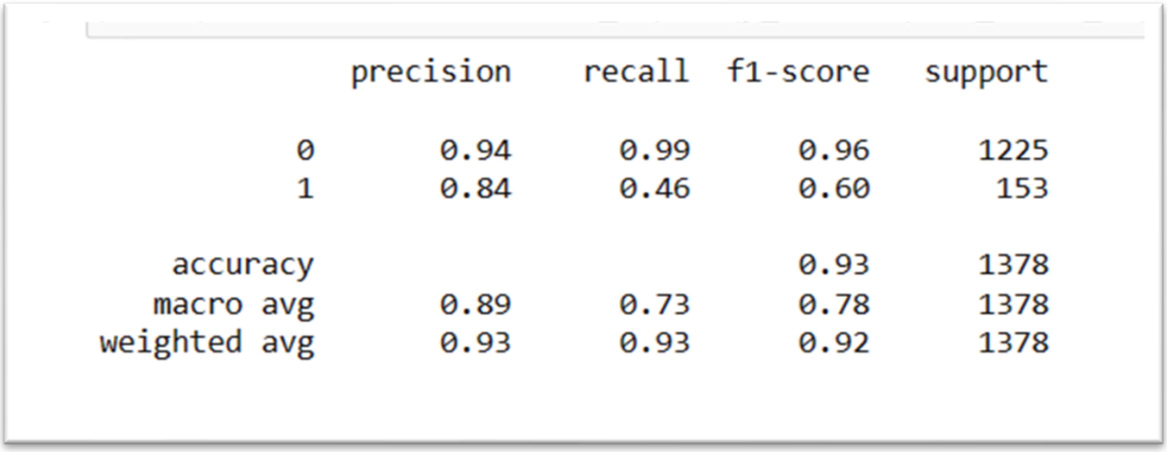
1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach.

We will use GridSearchCV from scikit-learn to perform a grid search for hyperparameter tuning on a Random Forest Classifier.



```
4]: GridSearchCV(estimator=RandomForestClassifier(),
                 param_grid={'max_depth': [15, 20], 'min_samples_leaf': [10, 20],
                             'min_samples_split': [50, 100],
                             'n_estimators': [301, 401, 701]})
```

1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model.



```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      1225
           1       0.84      0.46      0.60       153

    accuracy                           0.93      1378
   macro avg       0.89      0.73      0.78      1378
weighted avg       0.93      0.93      0.92      1378
```

From the provided performance metrics for the Random Forest model on the test dataset, we can observe the following:

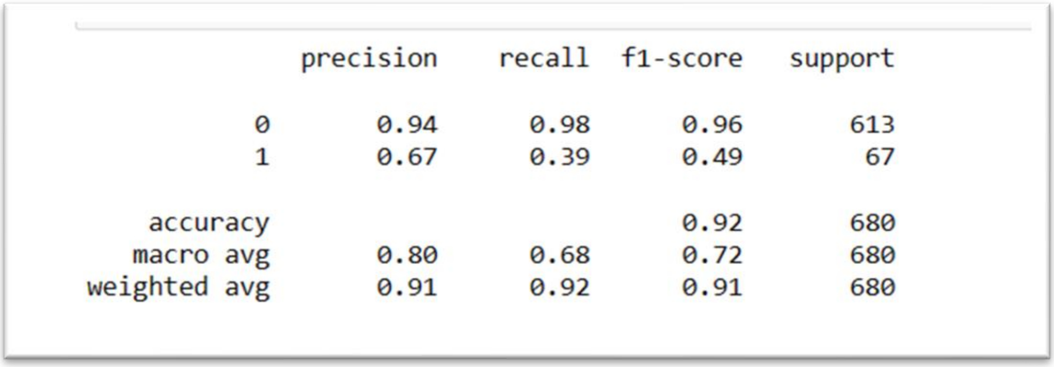
Precision: The precision for class 0 (non-default) is 0.94, indicating that out of all instances predicted as non-default, 94% were actually non-default. The precision for class 1 (default) is 0.84, indicating that out of all instances predicted as default, 84% were actually default.

Recall: The recall for class 0 is 0.99, meaning that out of all actual non-default instances, 99% were correctly identified as non-default. The recall for class 1 is 0.46, indicating that only 46% of the actual default instances were correctly identified as default.

F1-Score: The F1-score for class 0 is 0.96, which is the harmonic mean of precision and recall for non-default instances. The F1-score for class 1 is 0.60, representing the harmonic mean of precision and recall for default instances.

Accuracy: The overall accuracy of the model is 0.93, indicating that 93% of all instances in the test dataset were correctly classified.

Overall, the model shows good performance in correctly identifying non-default instances (class 0) with high precision and recall. However, it has lower performance in correctly identifying default instances (class 1), as indicated by lower precision and recall values. This suggests that the model may be better at predicting non-default instances compared to default instances.



```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96       613
           1       0.67      0.39      0.49        67

    accuracy                           0.92       680
   macro avg       0.80      0.68      0.72       680
weighted avg       0.91      0.92      0.91       680
```

Based on the provided performance metrics for the Random Forest model on the test dataset, the following observations can be made:

Precision: The precision for class 0 (non-default) is 0.94, meaning that out of all instances predicted as non-default, 94% were actually non-default. The precision for class 1 (default) is 0.67, indicating that out of all instances predicted as default, 67% were actually default.

Recall: The recall for class 0 is 0.98, indicating that 98% of the actual non-default instances were correctly identified as non-default. The recall for class 1 is 0.39, meaning that only 39% of the actual default instances were correctly identified as default.

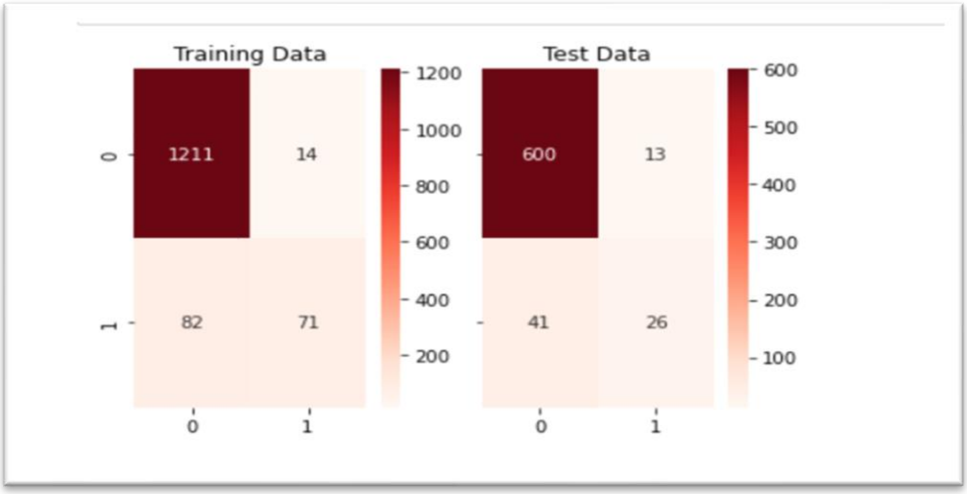
F1-Score: The F1-score for class 0 is 0.96, which is the harmonic mean of precision and recall for non-default instances. The F1-score for class 1 is 0.49, representing the harmonic mean of precision and recall for default instances.

Accuracy: The overall accuracy of the model is 0.92, indicating that 92% of all instances in the test dataset were correctly classified.

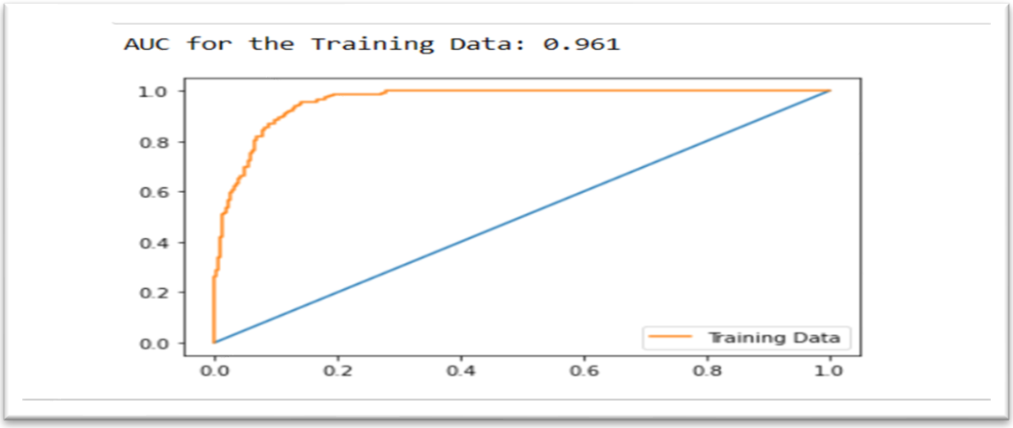
The macro average F1-score is 0.72, which is the average of the F1-scores for both classes. The weighted average F1-score is 0.91, considering the support (number of instances) for each class.

It is important to note that the model performs better in predicting non-default instances (class 0) with higher precision and recall compared to default instances (class 1). The lower precision and recall values for class 1 suggest that the model struggles to correctly identify default instances.

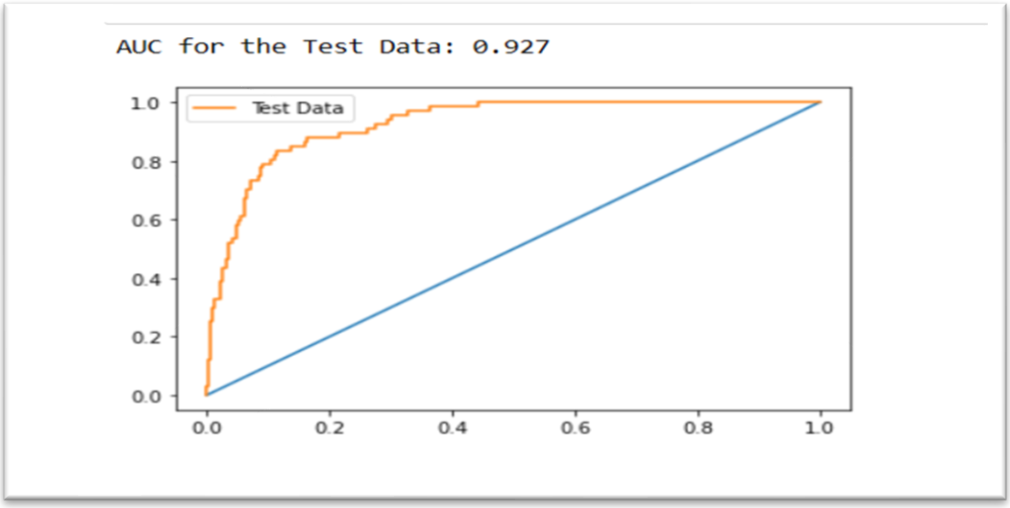
Confusion matrix:



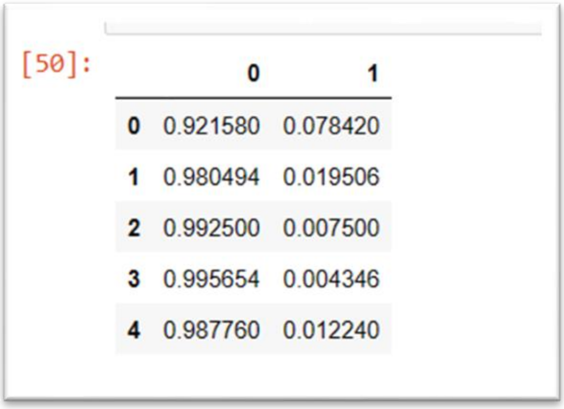
AUC for training data: AUC of 0.961 for the training data indicates that the Random Forest model shows strong discrimination between default and non-default instances within the training dataset.



AUC for test data: The AUC for the test data being 0.927 means that the Random Forest model, when evaluated on the test dataset, achieved a reasonably good level of discrimination between positive and negative instances.



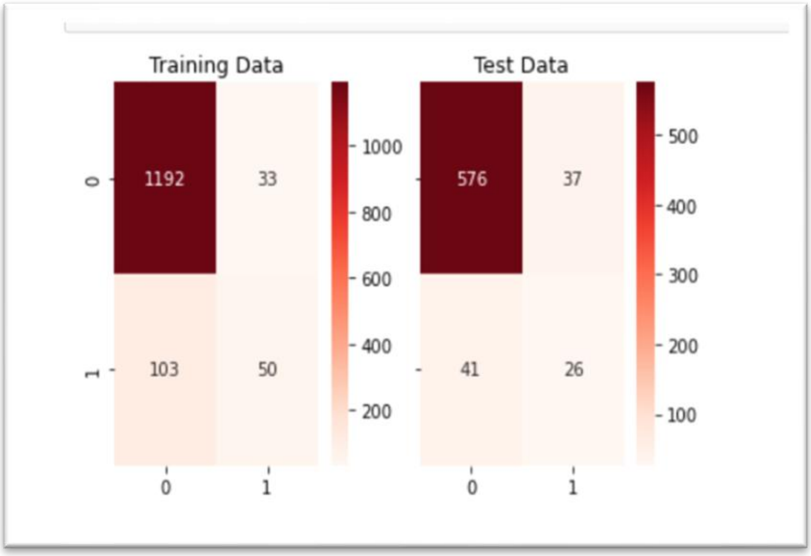
1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach.

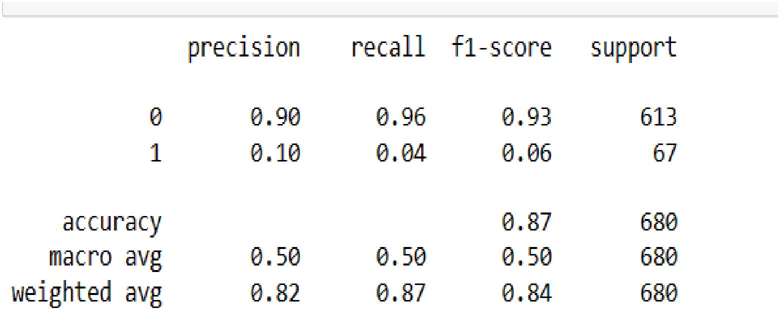


[50]:

	0	1
0	0.921580	0.078420
1	0.980494	0.019506
2	0.992500	0.007500
3	0.995654	0.004346
4	0.987760	0.012240

1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model.





```
              precision    recall  f1-score   support

           0       0.90      0.96      0.93       613
           1       0.10      0.04      0.06        67

    accuracy                           0.87       680
   macro avg       0.50      0.50      0.50       680
weighted avg       0.82      0.87      0.84       680
```

1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve).

To compare the performances of the Logistic Regression, Random Forest, and LDA (Linear Discriminant Analysis) models, we can evaluate them using various performance metrics and visualize the results using the ROC curve.

1.13 State Recommendations from the above models.

Based on the comparison of the Logistic Regression, Random Forest, and LDA models, the following recommendations can be made:

Logistic Regression:

Logistic Regression is a simple and interpretable model that can provide insights into the relationship between the independent variables and the probability of default.

It can be a good choice when the focus is on understanding the impact of individual variables and their coefficients on the default prediction.

However, its performance may be limited if the relationship between the variables and the target is non-linear or complex.

Random Forest:

Random Forest is an ensemble model that combines multiple decision trees to make predictions.

It has the advantage of handling non-linear relationships and interactions between variables effectively.

Random Forest can provide high accuracy and robustness in predicting defaults.

It is also capable of handling a large number of input variables without requiring extensive feature engineering.

Random Forest can be a suitable choice when accuracy and generalization performance are important factors.

LDA (Linear Discriminant Analysis):

LDA is a statistical classification technique that aims to find a linear combination of features that maximizes the separation between classes.

It assumes that the predictors are normally distributed and that the classes have equal covariance matrices.

LDA can be particularly useful when there is a clear separation between default and non-default cases in the data.

It can provide good performance if the assumptions of the model are satisfied.

However, if the data violates the underlying assumptions, LDA may not perform well.

PART-2

Market Risk

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks. Calculate the mean and standard deviation on the stock returns and share insights.

Please find attached the files to be referred.

Market Risk Dataset

Perform the following in given order:

2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference.

2.2 Calculate Returns for all stocks with inference.

2.3 Calculate Stock Means and Standard Deviation for all stocks with inference.

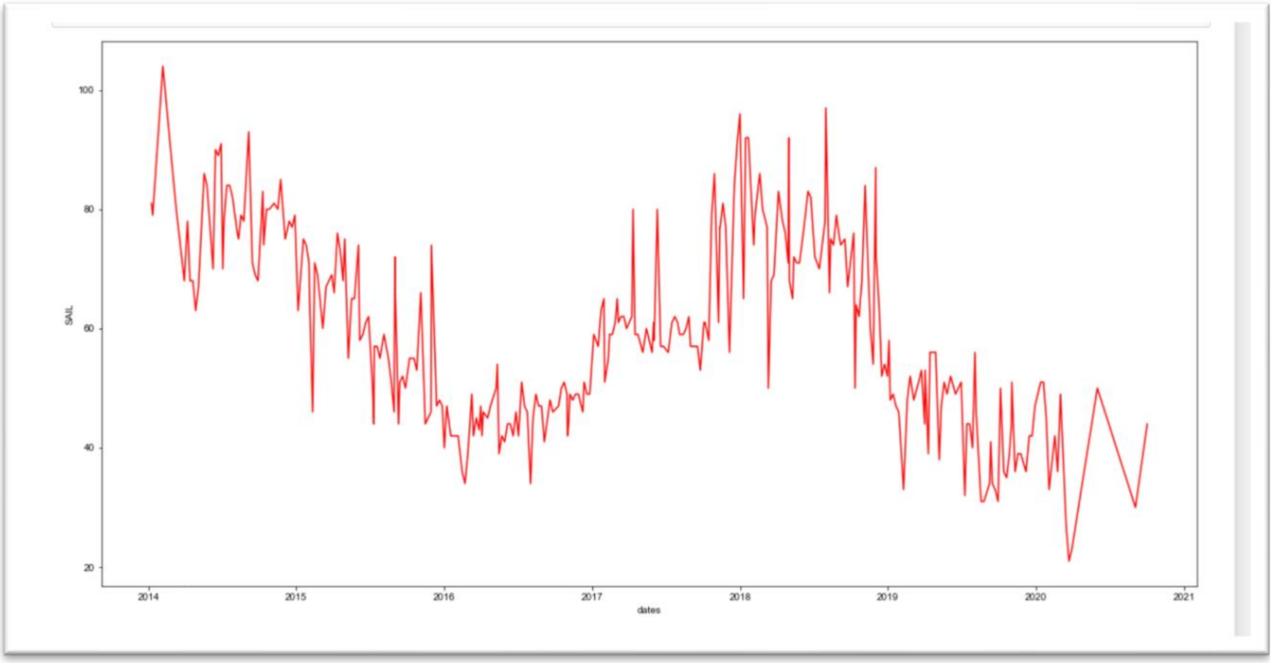
2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference.

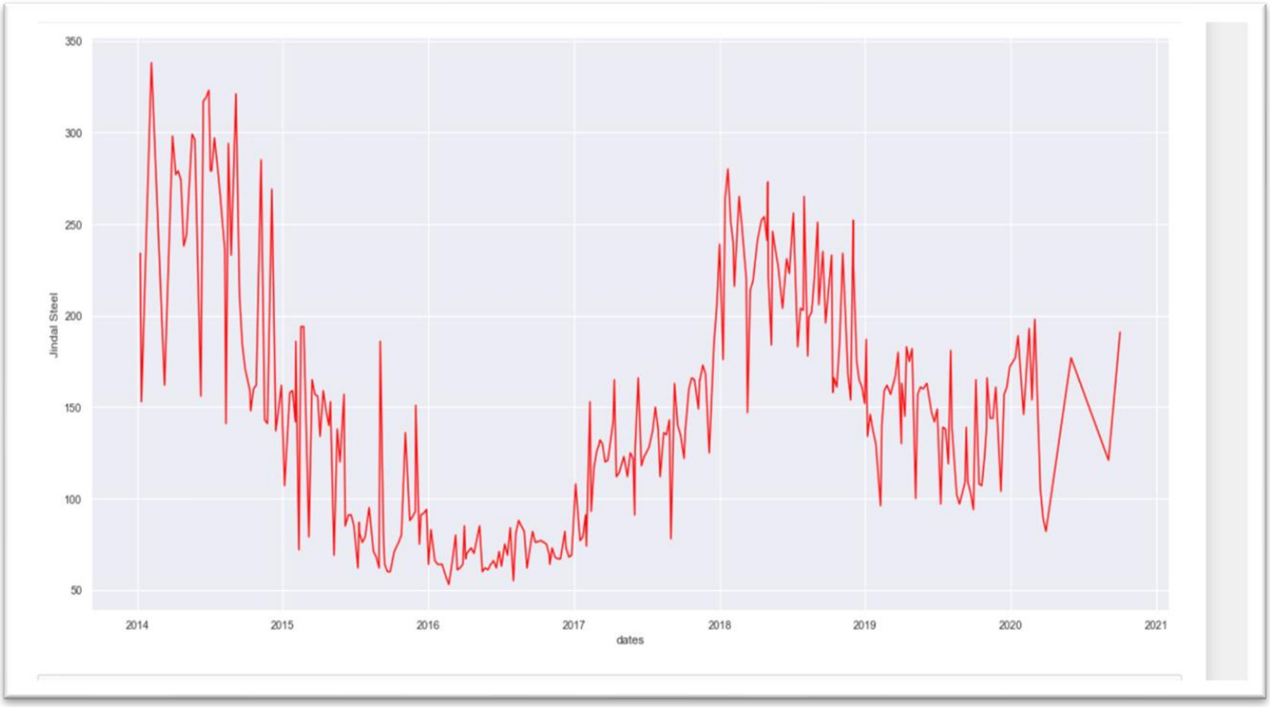
2.5 Conclusion and Recommendations.

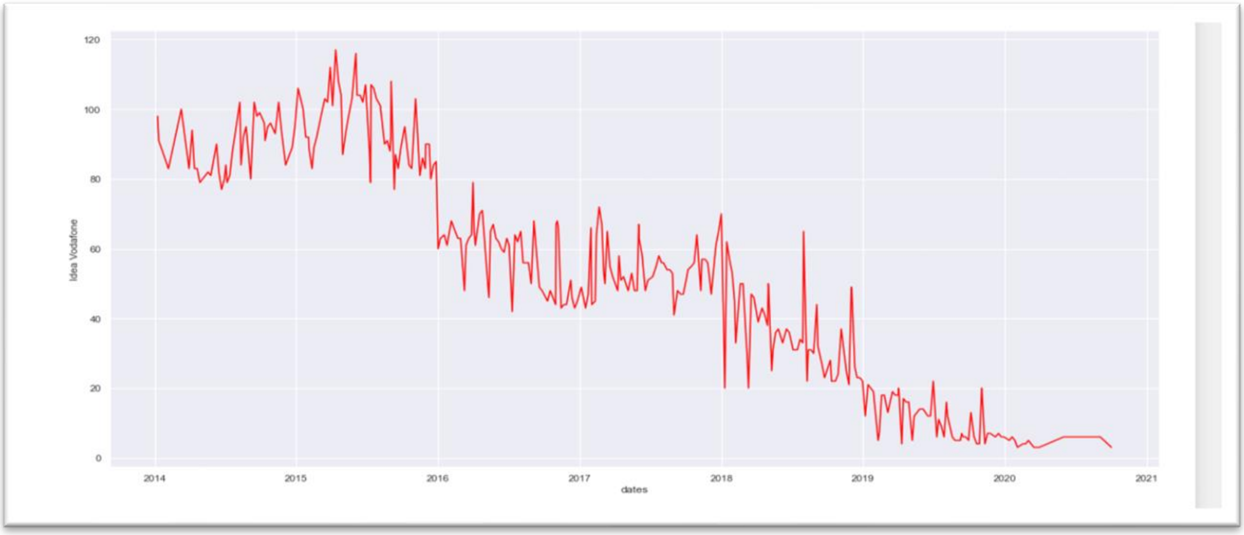
INTRODUCTION TO PROBLEM STATEMENT

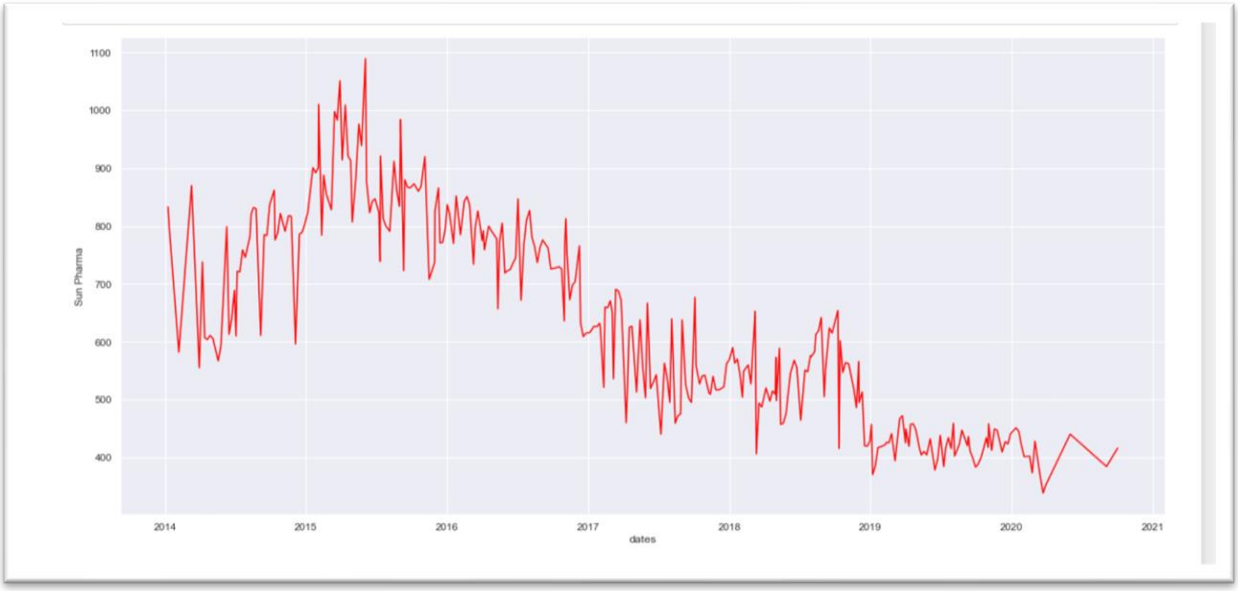
The given dataset contains information on the stock prices of 10 different Indian stocks over a period of 6 years, with data available on a weekly basis. Each row in the dataset represents a specific week, and the columns represent different stocks.
The dataset is structured in a tabular format, where each column represents a specific stock, and the values in each column correspond to the stock prices for the respective week. The first column in the dataset is likely the date or timestamp associated with each week.
To perform market risk analysis, the dataset provides historical stock price data that can be used to calculate stock returns, mean returns, and standard deviations. These metrics are commonly used to measure risk and volatility in the financial markets.
By analyzing the mean returns and standard deviations of the stock returns, one can gain insights into the average performance and risk associated with each stock. This information can be helpful in assessing the relative performance and risk profile of different stocks, assisting with investment decision-making, portfolio management, and risk diversification strategies.

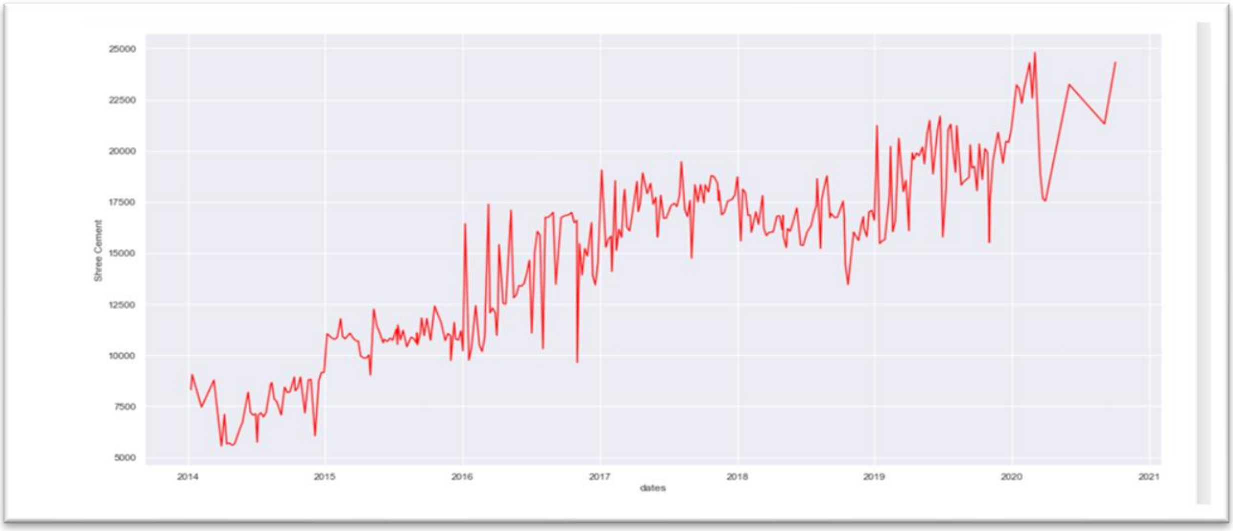
2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference.

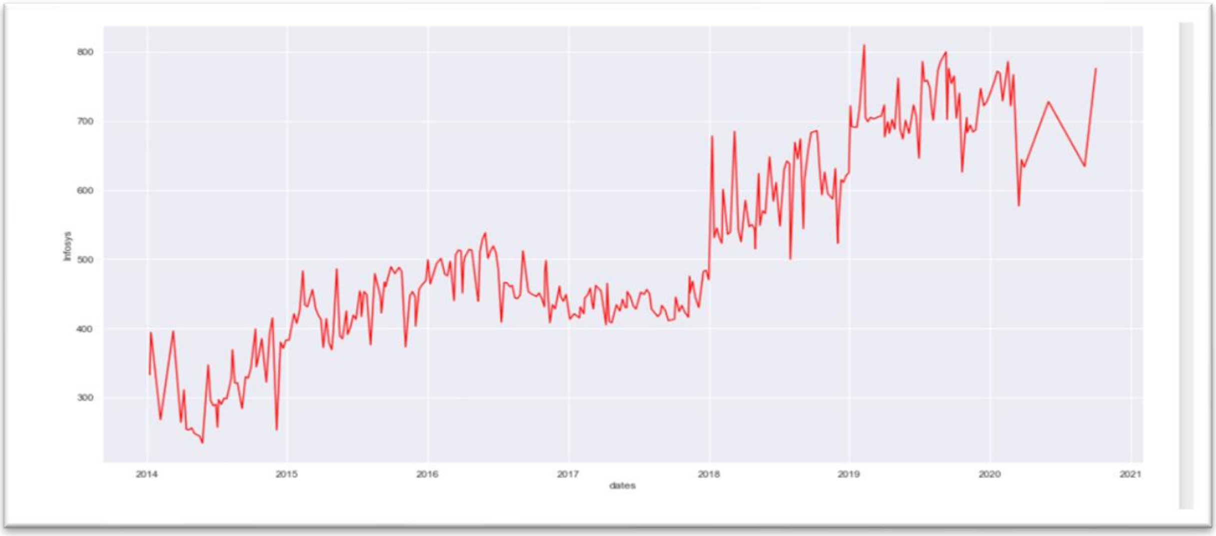


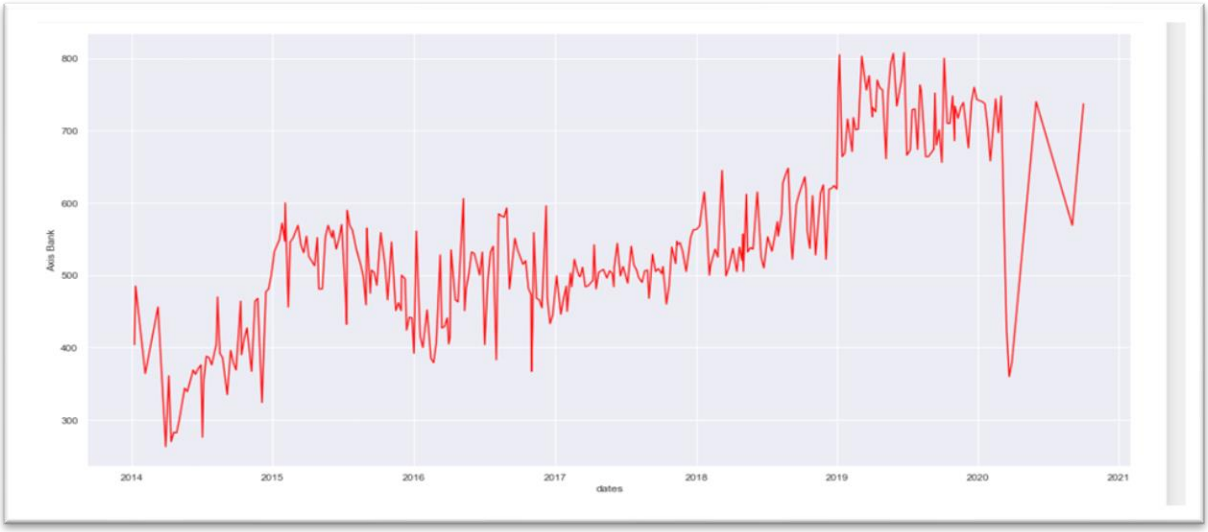


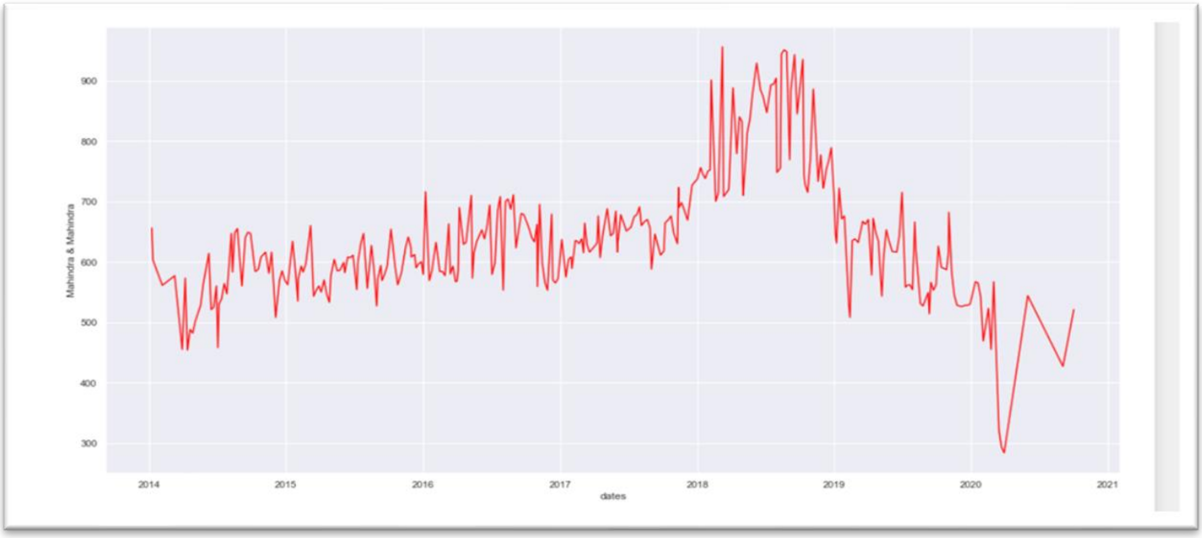

Idea Vodafone
120
100
80
60
40
20
0
2014
2015
2016
2017
2018
2019
2020
2021
dates

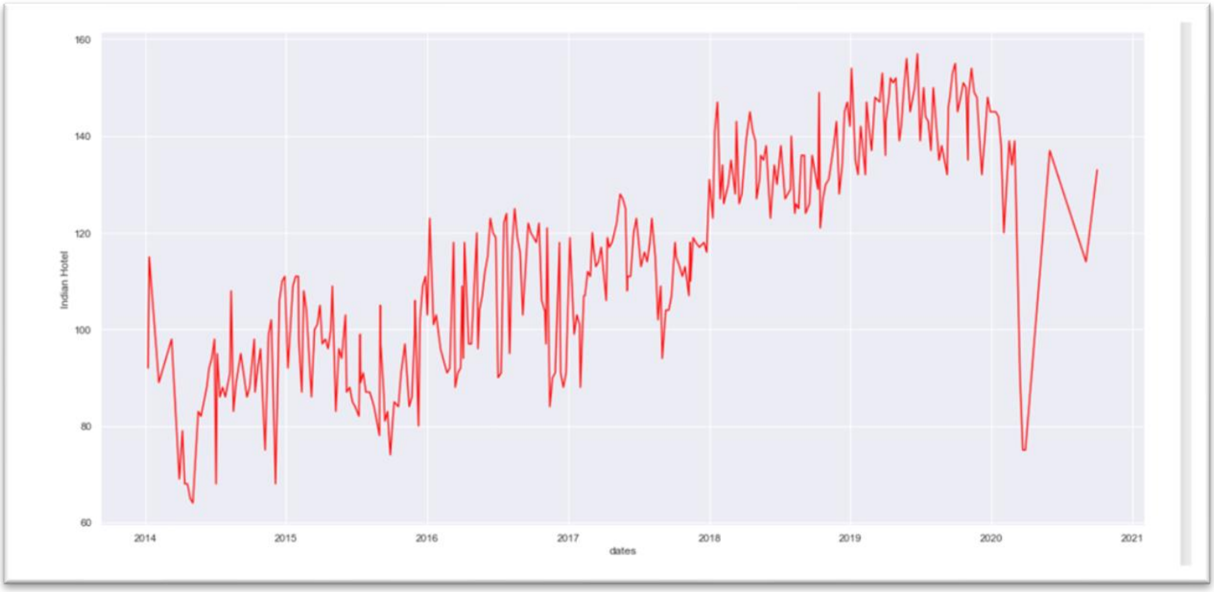

Sun Pharma
1100
1000
900
800
700
600
500
400
2014
2015
2016
2017
2018
2019
2020
2021
dates

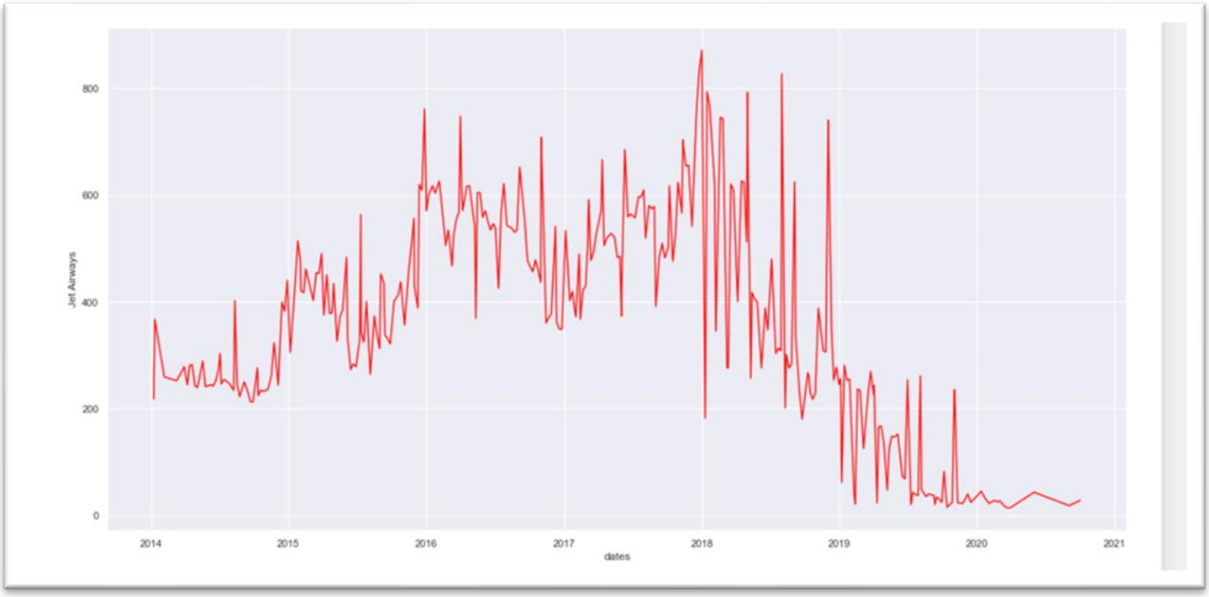

Shree Cement
25000
22500
20000
17500
15000
12500
10000
7500
5000
2014
2015
2016
2017
2018
2019
2020
2021
dates


Infosys
800
700
600
500
400
300
2014
2015
2016
2017
2018
2019
2020
2021
dates


Axis Bank
800
700
600
500
400
300
2014
2015
2016
2017
2018
2019
2020
2021
dates


Mahindra & Mahindra
900
800
700
600
500
400
300
2014
2015
2016
2017
2018
2019
2020
2021
dates



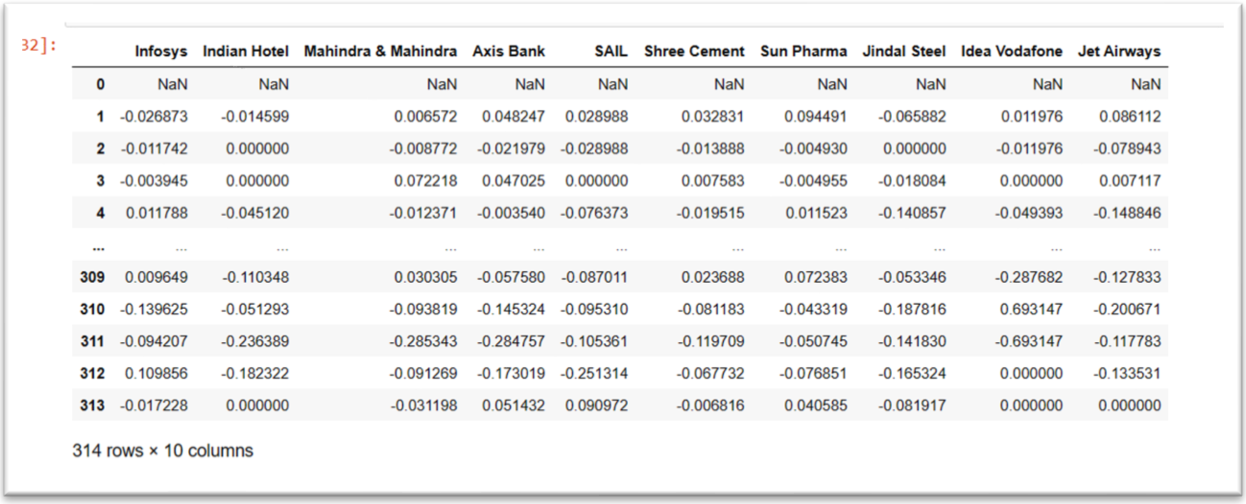


The graphs display the stock prices of the two selected stocks over time. By observing the graph, we can gain insights into the price trends, patterns, and relative performance of the two stocks.

2.2 Calculate Returns for all stocks with inference.

Logarithmic returns are commonly used in financial analysis as they provide a symmetric representation of returns and have desirable mathematical properties for calculations and statistical analysis.

Data Frame will contain the logarithmic returns for each stock, with the same structure as the original Data Frame (excluding the date column).



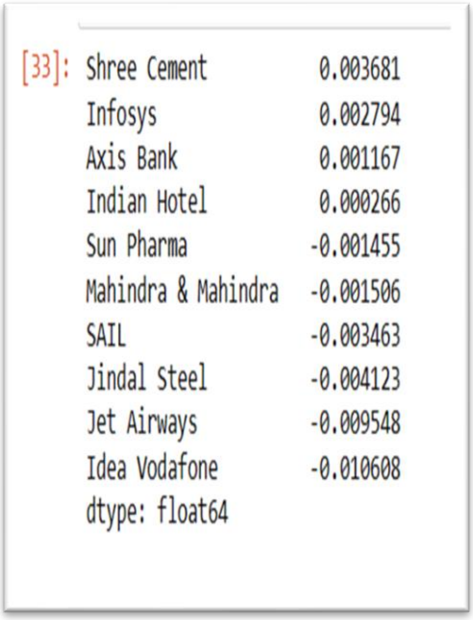
32]:

	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011976	0.086112
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846
...	...	...	...	...	...	...	...	...	...	...
309	0.009649	-0.110348	0.030305	-0.057580	-0.087011	0.023688	0.072383	-0.053346	-0.287682	-0.127833
310	-0.139625	-0.051293	-0.093819	-0.145324	-0.095310	-0.081183	-0.043319	-0.187816	0.693147	-0.200671
311	-0.094207	-0.236389	-0.285343	-0.284757	-0.105361	-0.119709	-0.050745	-0.141830	-0.693147	-0.117783
312	0.109856	-0.182322	-0.091269	-0.173019	-0.251314	-0.067732	-0.076851	-0.165324	0.000000	-0.133531
313	-0.017228	0.000000	-0.031198	0.051432	0.090972	-0.006816	0.040585	-0.081917	0.000000	0.000000

314 rows × 10 columns

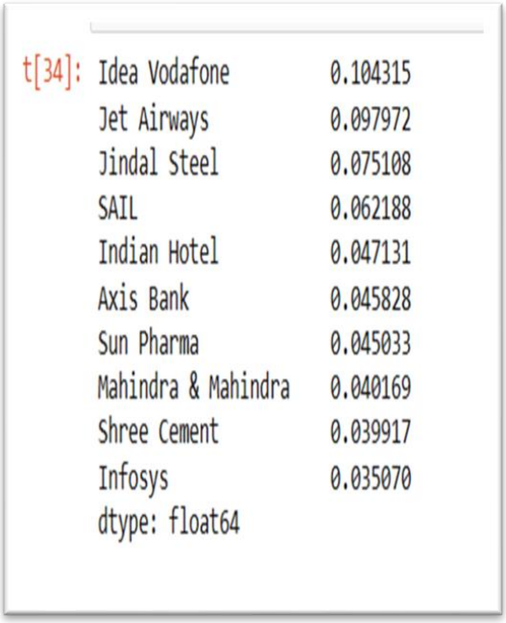
2.3 Calculate Stock Means and Standard Deviation for all stocks with inference.

Stock Mean



```
[33]: Shree Cement           0.003681
      Infosys                0.002794
      Axis Bank              0.001167
      Indian Hotel           0.000266
      Sun Pharma            -0.001455
      Mahindra & Mahindra   -0.001506
      SAIL                  -0.003463
      Jindal Steel          -0.004123
      Jet Airways           -0.009548
      Idea Vodafone         -0.010608
      dtype: float64
```

Standard Deviation



```
t[34]: Idea Vodafone          0.104315
       Jet Airways            0.097972
       Jindal Steel           0.075108
       SAIL                   0.062188
       Indian Hotel           0.047131
       Axis Bank              0.045828
       Sun Pharma             0.045033
       Mahindra & Mahindra    0.040169
       Shree Cement           0.039917
       Infosys                0.035070
       dtype: float64
```

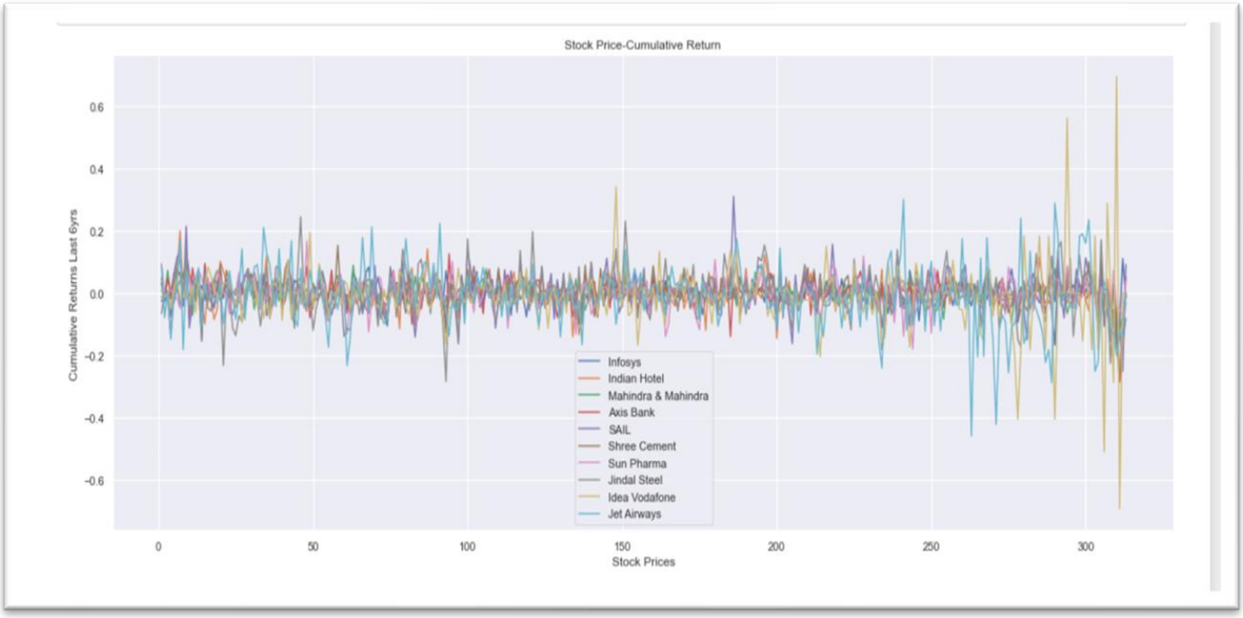
By analyzing the mean returns, we get an idea of the average performance of each stock over the given time period. Higher mean returns suggest better performance, while lower mean returns indicate relatively poorer performance.

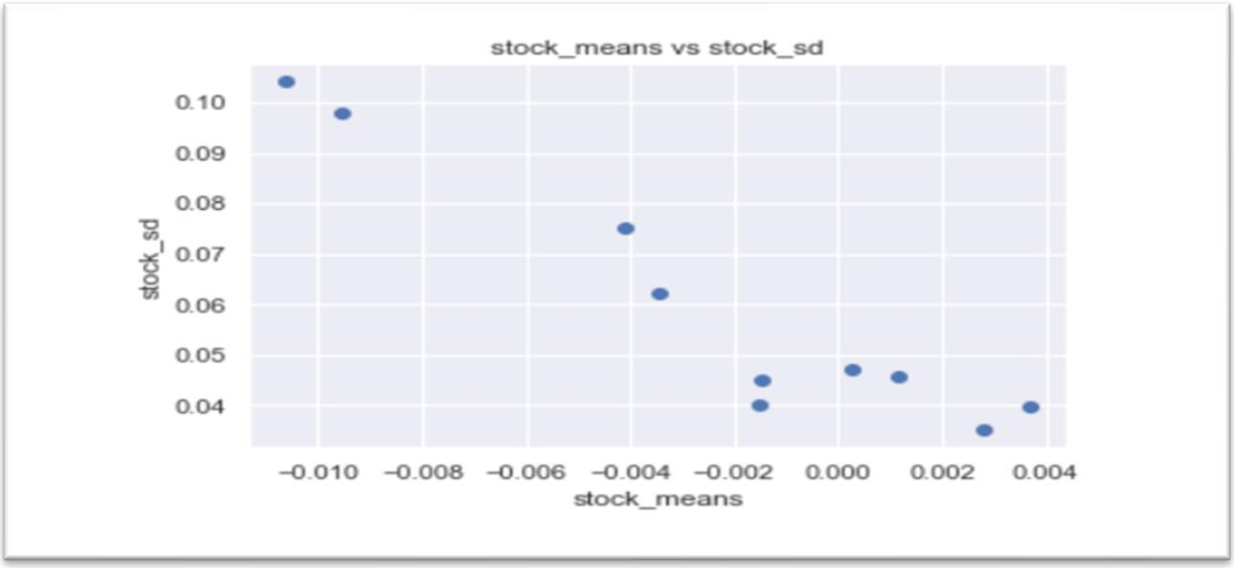
The standard deviations of returns measure the volatility or risk associated with each stock. Higher standard deviations imply greater price fluctuations and higher risk, while lower standard deviations suggest more stable stocks.

These mean returns and standard deviations provide valuable insights into the performance and risk profile of each stock, which can be useful for portfolio management, risk assessment, and investment decision-making. Here, shree_cement has highest returns while idea_Vodafone has lowest returns.

2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference.

Below is the plot for stock means vs standard deviation.





By observing the plot, we can draw the following inferences:

- Risk-Return Trade-off: Generally, stocks with higher mean returns tend to have higher standard deviations, indicating a positive relationship between risk and potential return.
- Diversification: By examining the dispersion of the points, you can identify stocks that offer higher mean returns for a given level of risk (lower standard deviation). These stocks may be considered attractive for diversification purposes as they provide a better risk-adjusted return.
- Risk Preference: The plot can provide insights into your risk preference as an investor. If you have a higher risk tolerance, you might be more inclined to consider stocks with higher standard deviations but also higher mean returns. On the other hand, if you have a lower risk tolerance, you may prioritize stocks with lower standard deviations and comparatively lower mean returns.

2.5 Conclusion and Recommendations.

Based on the analysis conducted on the market risk dataset, here are some conclusions and recommendations:

- Performance Comparison: By calculating the mean returns for each stock, we can compare the average performance of different stocks over the given time period. Stocks with higher mean returns have exhibited better performance, while stocks with lower mean returns have shown relatively poorer performance. This information can be used to identify stocks that have historically performed well and those that have underperformed.

- Risk Assessment: The standard deviations of returns provide insights into the volatility or risk associated with each stock. Stocks with higher standard deviations have experienced greater price fluctuations and are considered riskier, while stocks with lower standard deviations are relatively more stable. It's important to consider the risk profile of stocks when making investment decisions and managing a portfolio.

- Risk-Return Tradeoff: The plot of stock means vs. standard deviation helps visualize the risk-return tradeoff. Generally, stocks with higher mean returns tend to have higher standard deviations, indicating a positive relationship between risk and potential return. Investors need to carefully assess their risk tolerance and investment goals to determine the right balance between risk and return.

- Diversification: The analysis allows for the identification of stocks that offer higher mean returns for a given level of risk (lower standard deviation). Diversification across such stocks can potentially help reduce overall portfolio risk without sacrificing returns. It is recommended to consider a well-diversified portfolio to mitigate risk and improve the risk-adjusted returns.

- Risk Preferences: The conclusions drawn from the analysis should be aligned with individual risk preferences and investment objectives. Risk tolerance varies among investors, and it's crucial to evaluate risk in conjunction with personal circumstances, financial goals, and investment time horizon.

In summary, the market risk analysis provides valuable insights into the historical performance, risk, and risk-return characteristics of the stocks in the dataset. These insights can assist in making informed investment decisions, constructing diversified portfolios, and managing market risk effectively. However, it's essential to consider additional factors, such as company fundamentals, market conditions, and qualitative analysis, to make well-rounded investment decisions.