Thesis:

1. **Cover Page**
2. **Title Page**
3. **Certificate**
4. **Dedication**
5. **Declarations**
6. **Acknowledgment**
7. List of abbreviations
8. List of symbols
9. **Abstract**
10. **Contents**
11. **List of Figures**
12. **List of tables**
13. **Introduction and Literature review**
14. **Background and motivation/ motivation**
15. Literature review
16. **Problem statement/ Problem description and objective**
17. **Contribution/ summary of the contribution**
18. Thesis structure/ organization of the report
19. **Description of data**
20. **Analysis of simulated Data:**

**Data simulation steps and details.**

Relation between avgDD and avgAD.

Explanation of each concept with the proper equation.

1. Use of avgDD, avgDIST, maxW, and regularity.
2. Correlation matrix
3. Partial correlation
4. Pair plot
5. Considering the different combinations of variables
6. Ridge regression analysis
7. Lasso regression analysis

Discuss all observations

1. Correlation matrix(correlation between departure delay and graph invariant)
2. Univariate time series analysis(AR, MA, ARMA, ARIMA)
3. Two-layered model:

**Layer-1** Multivariate time series analysis (select the best model and use suitable p,d,q values or Some deep learning model for multivariate time series analysis)

**Layer-2** Regression analysis using different ML models. (Select the best model and use a suitable hyperparameter)

1. All of the above analyses on other Fragile datasets (Flight/Railways dataset)
2. Observations
3. Results
4. Conclusion and Future work
5. Appendices (if needed)
6. References

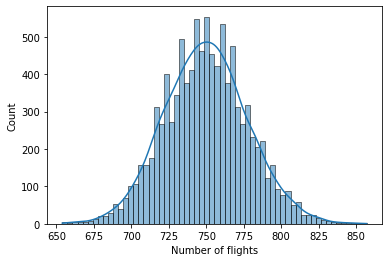
**Chapter 3  
Stochastic data modelling and analysis:**

The aviation industry is one of the most dynamic and challenging industries, where flights are often delayed or canceled due to various reasons, such as weather conditions, technical problems, or air traffic congestion. In this context, flight delay data analysis has become a crucial aspect of the industry's operations. In this project, we will create a stochastic model to simulate flight delay data and analyze its performance using several variables. It has been observed that the distribution of number of flights on daily basis follows poisson’s distribution. The idea is to simulate flight delay data using the Poisson distribution to get dialy basis information and perform an analysis of its performance using various variables.

The following steps were taken to create the stochastic model:

1. Simulate data using Poisson distribution:

We generated a Poisson distribution with a lambda value of 750, which is the average number of flights per day. Then, we performed an "iter" number of iterations, where we selected the nth day’s number of flights and randomly select that many flights (data points) from the original data. This process resulted in a new dataset with simulated flight delay data.



1. Compute the following variables from the simulated data: We computed the following variables from the simulated data:

* Average departure delay: The average difference between the scheduled and actual departure times of all flights.
* Average distance: The average distance between the origin and destination airports of all flights.
* Maximum Indegree of graph: The maximum number of incoming flights to a particular airport.
* Regularity of graph: It is a graph invariant that encapulates the graph structural property.
* Average arrival delay: The average difference between the scheduled and actual arrival delay time of all flights. This is the response variable.

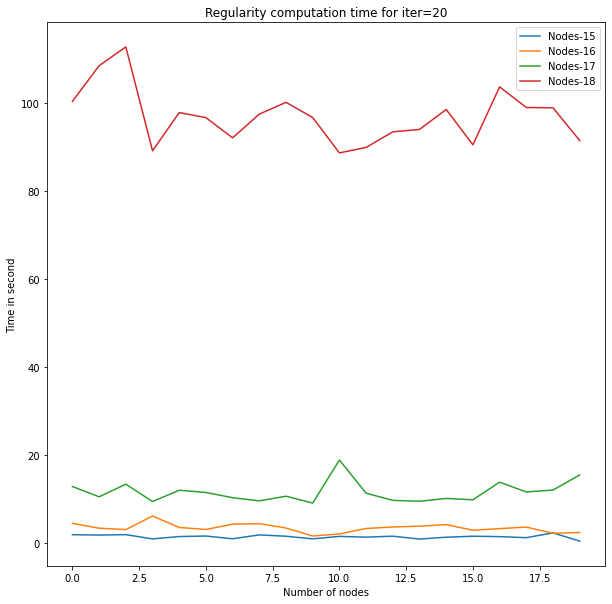
(c) Fit the Machine learning regression model: We then fit a machine learning regression model to the simulated data to analyze the relationship between the variables and flight delays. The model was trained using a subset of the simulated data, and the remaining data was used for testing. The model was evaluated based on its accuracy in predicting flight delays.

(d) Analyze performance: We analyzed the performance of the stochastic model using the variables computed in step (b) and the machine learning regression model fitted in step (c). We evaluated the accuracy of the model in predicting flight delays, as well as its ability to capture the relationships between the variables and flight delays.

In this analysis, we created a stochastic model to simulate flight delay data and analyzed its performance using various variables. We demonstrated that the multilinear regression model strongly fitted to the simulated data.

Here the Macaulay 2 software is needed for the computation of regularity. Macaulay 2 is a software that help in research in the field of commutative algebra and algebraic geometry. It has core algorithsm for computing multigraded free resolutions of modules and Grobner bases.

Regularity calculation are computationally very expensive.



Hence we will do analysis by considering only 24 nodes i.e the airports in our data.



7Var\_Quantiles\_corr\_pearson

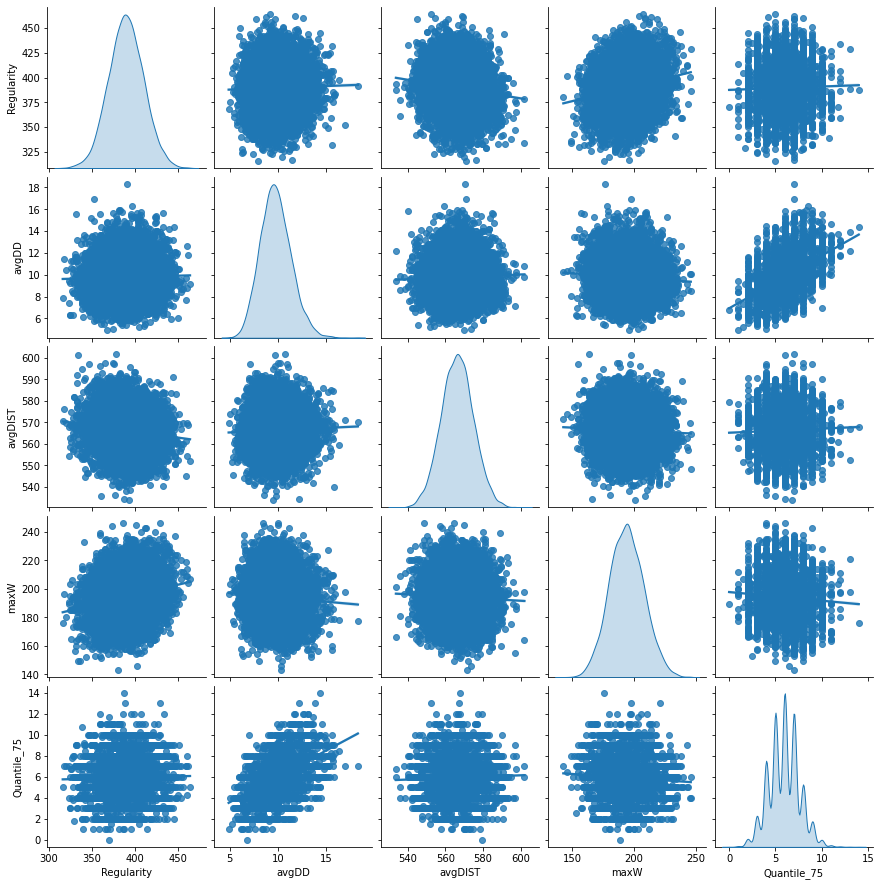


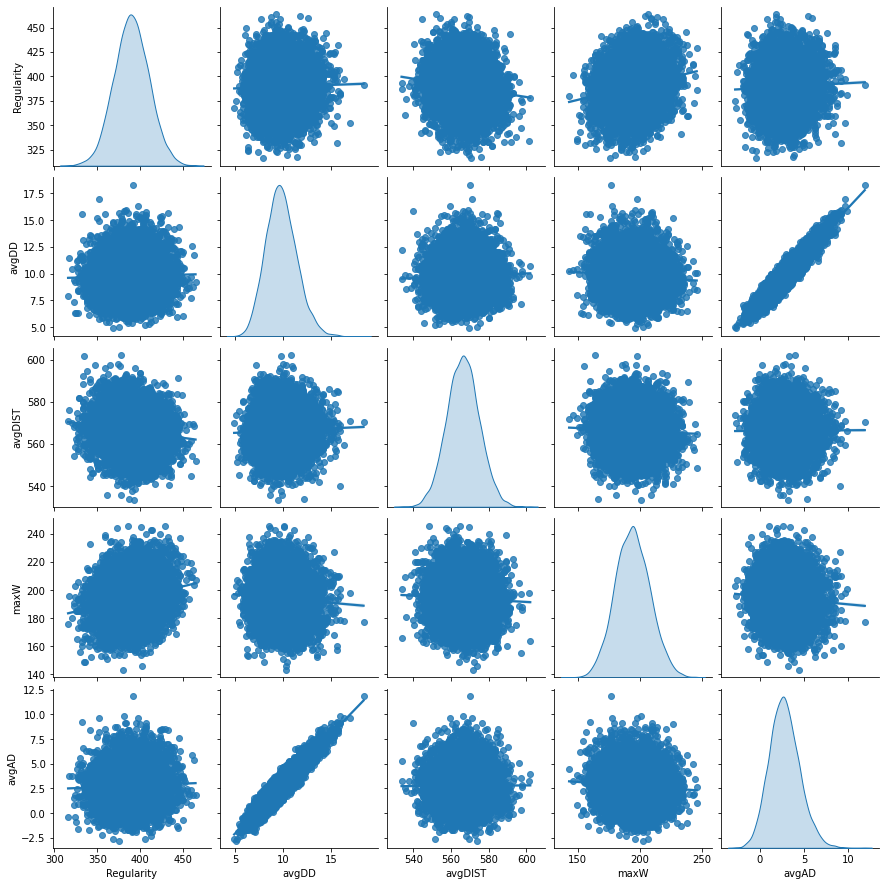
7Var\_Quantiles\_corr\_spearman

Partial coefficient of correlation:

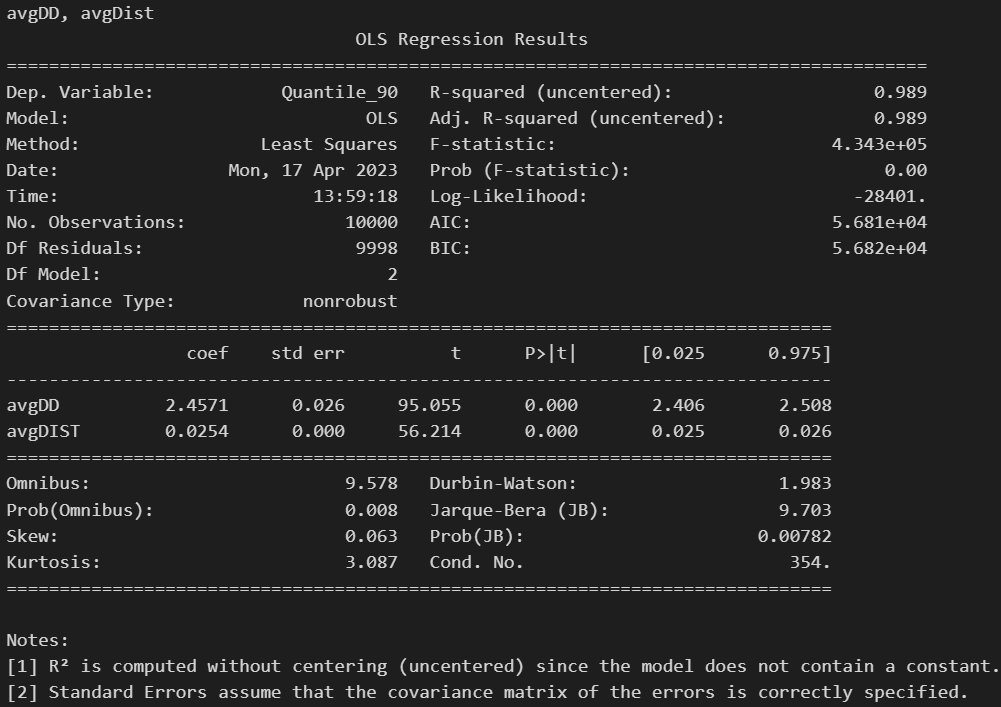
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **n** | **var** | **r** | **CI95%** | **p-val** |
| 500 | avgDD + avgAD | 0.94953 | [0.94, 0.96] | 2.13E-251 |
| 500 | avgDIST + avgAD | -0.1469 | [-0.23, -0.06] | 0.001025 |
| 500 | maxW + avgAD | 0.06001 | [-0.03, 0.15] | 0.181648 |
| 500 | Regularity + avgAD | -0.045 | [-0.13, 0.04] | 0.317228 |

The difference between Pearson correlation coefficient and partial correlation coefficient is that Pearson correlation coefficient measures the strength and direction of a linear relationship between two variables, whereas partial correlation coefficient measures only the strength of the relationship. It is a positive integer between -1 and 1. Partial correlation coefficient, on the other hand, gauges the intensity of a relationship between two variables while controlling for one or more extra variables. After controlling for one or more variables, it is used to derive correlation coefficients.





Pairplot

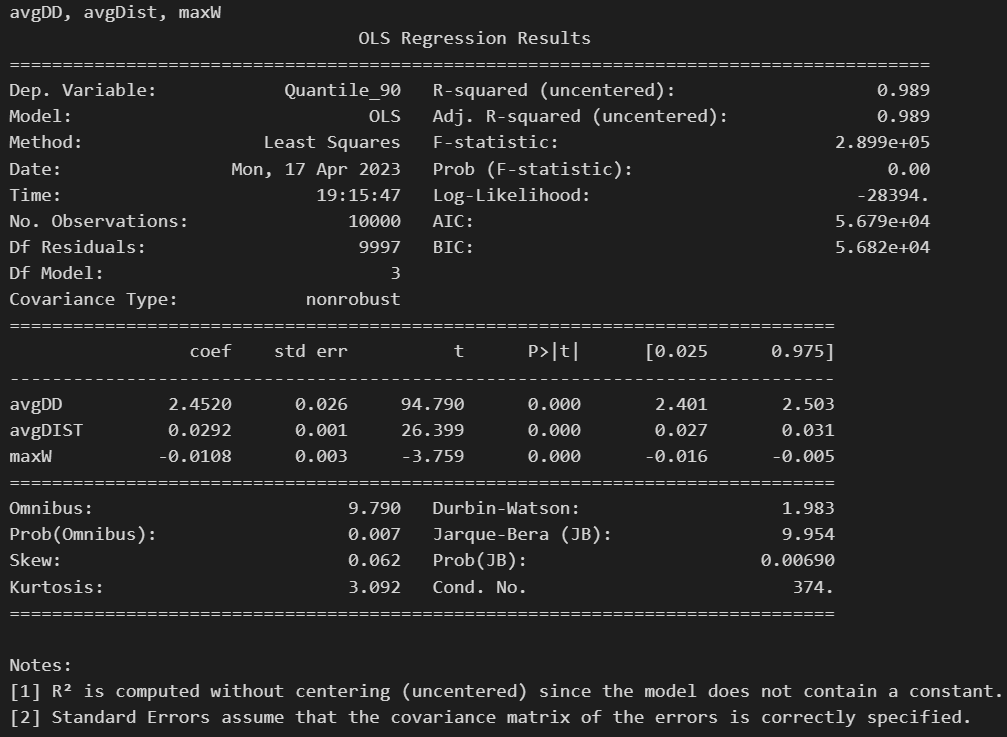
1

Median Absolute Error : 2.7713105432705305

Mean Squared Error : 16.77987577812255

Root Mean Squared Error: 4.096324667079326

2

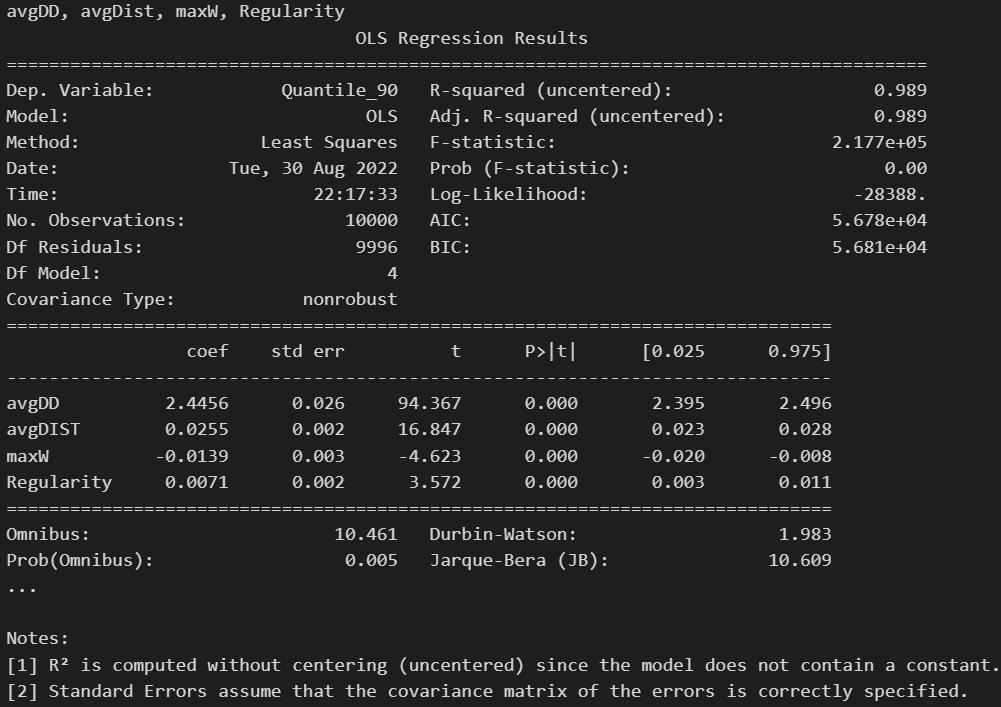


Median Absolute Error : 2.7578668368738875

Mean Squared Error : 16.736584731321653

Root Mean Squared Error: 4.091037121723739

3



Median Absolute Error : 2.756571870554321

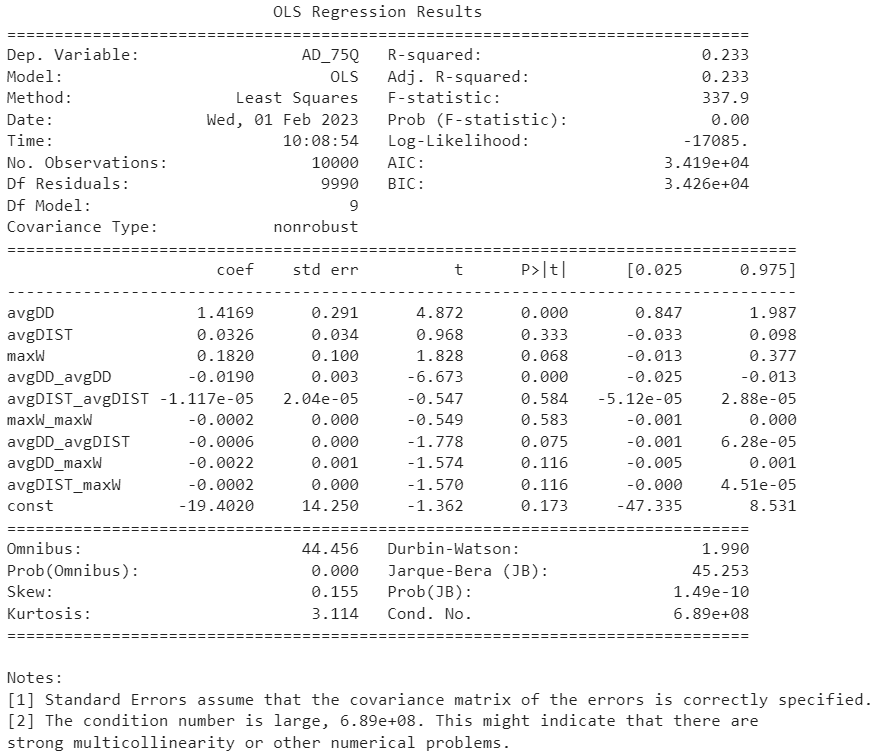
Mean Squared Error : 16.7303987006062

Root Mean Squared Error: 4.090281005090751

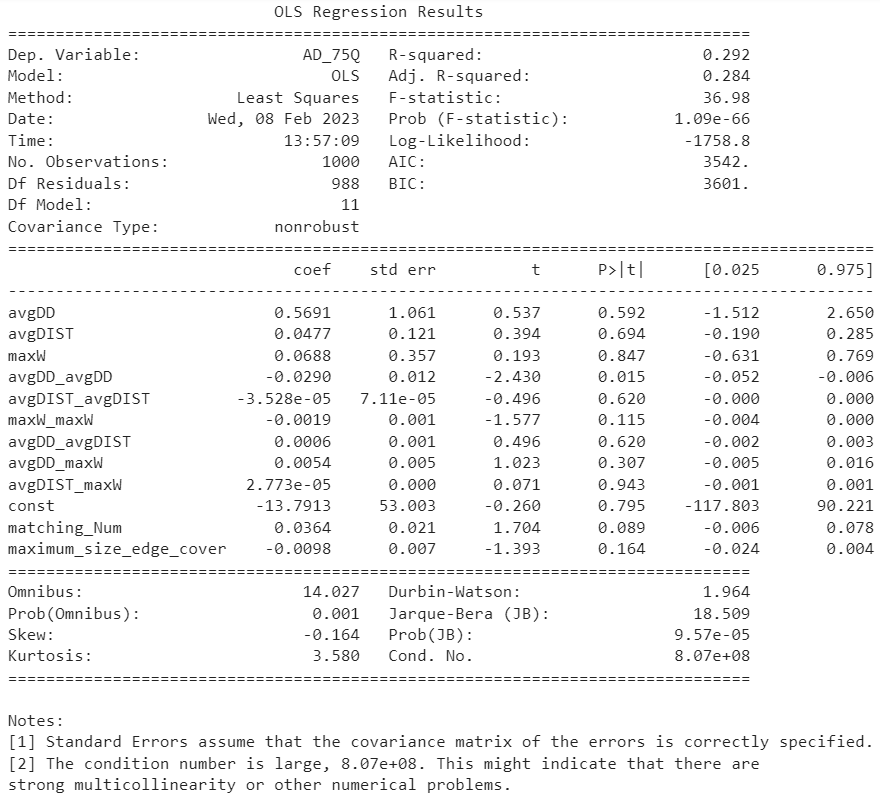
Summary 10V:

['avgDD', 'avgDIST', 'maxW', 'avgDD\_avgDD', 'avgDIST\_avgDIST', 'maxW\_maxW', 'avgDD\_avgDIST', 'avgDD\_maxW', 'avgDIST\_maxW', 'const']

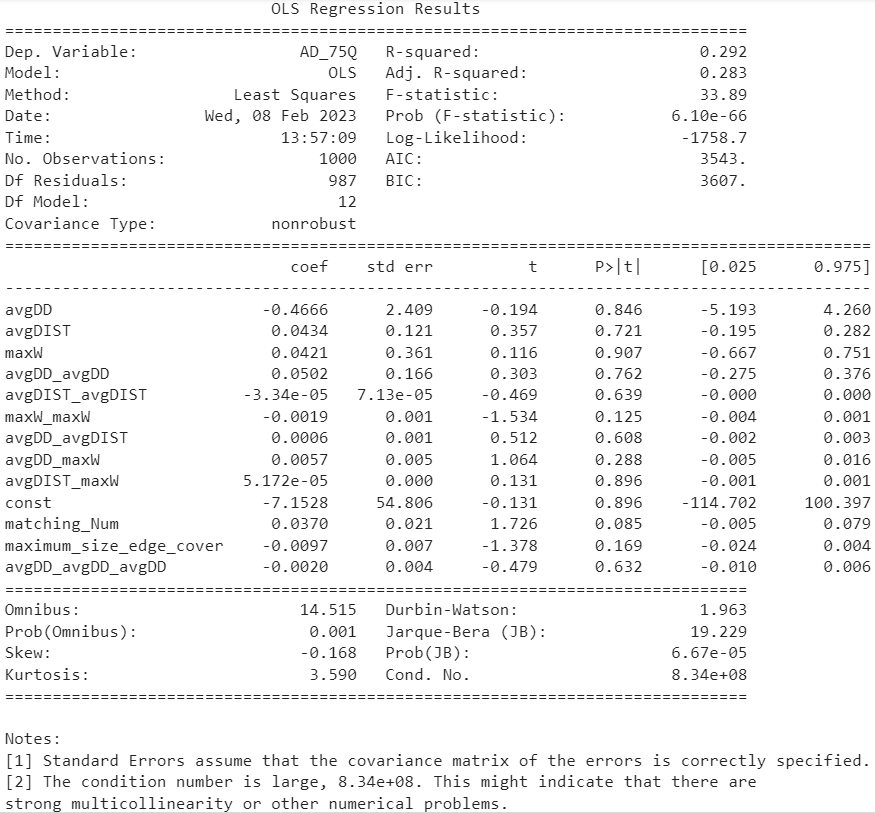
4



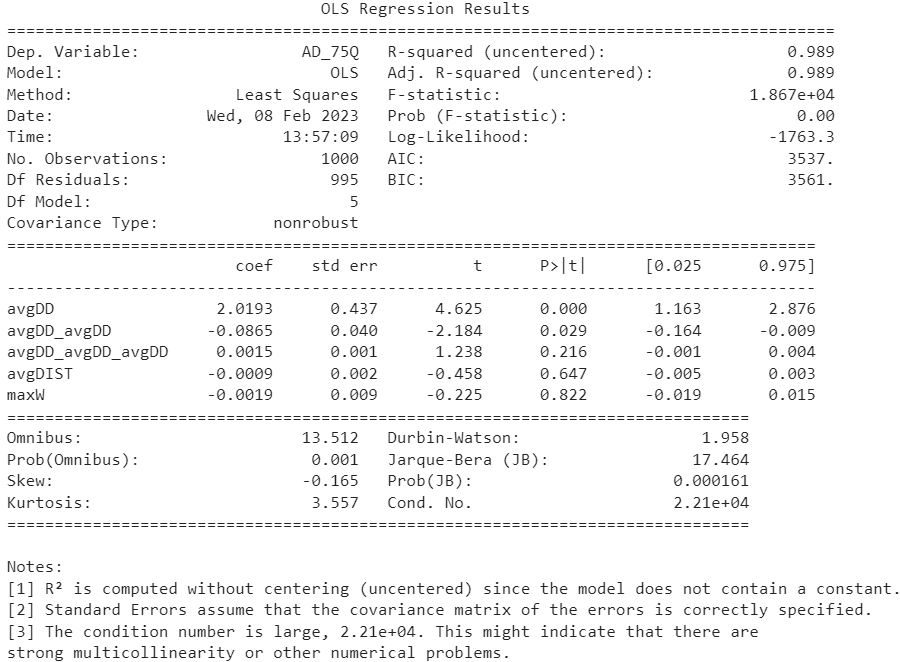
['avgDD', 'avgDIST', 'maxW', 'avgDD\_avgDD', 'avgDIST\_avgDIST', 'maxW\_maxW', 'avgDD\_avgDIST', 'avgDD\_maxW', 'avgDIST\_maxW', 'const', 'matching\_Num', 'maximum\_size\_edge\_cover']

5

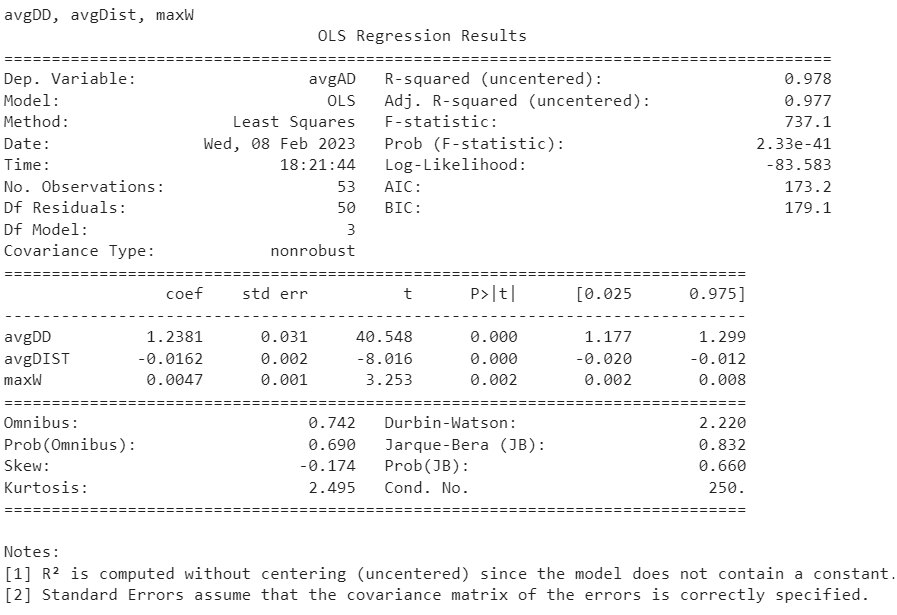
['avgDD', 'avgDIST', 'maxW', 'avgDD\_avgDD', 'avgDIST\_avgDIST', 'maxW\_maxW', 'avgDD\_avgDIST', 'avgDD\_maxW', 'avgDIST\_maxW', 'const', 'matching\_Num', 'maximum\_size\_edge\_cover', 'avgDD\_avgDD\_avgDD']

6

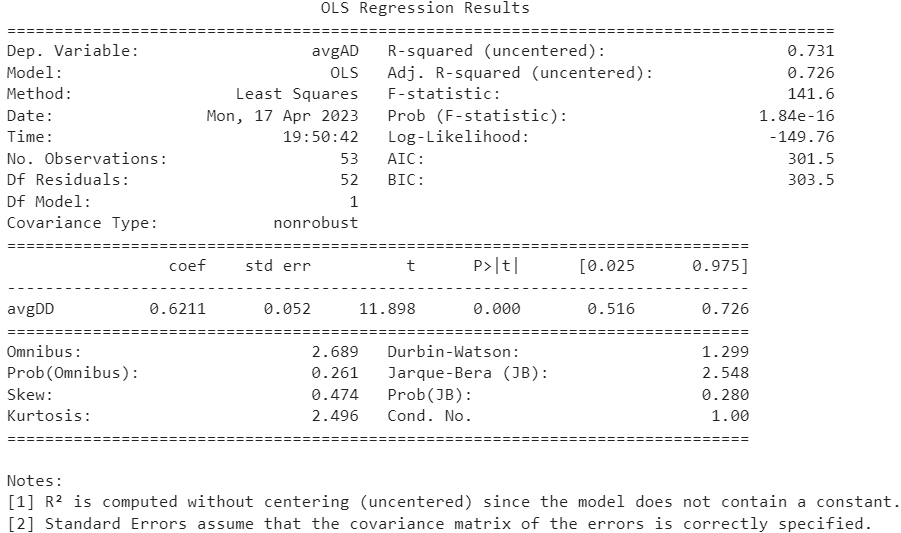
['avgDD', 'avgDD\_avgDD', 'avgDD\_avgDD\_avgDD', 'avgDIST', 'maxW']

7

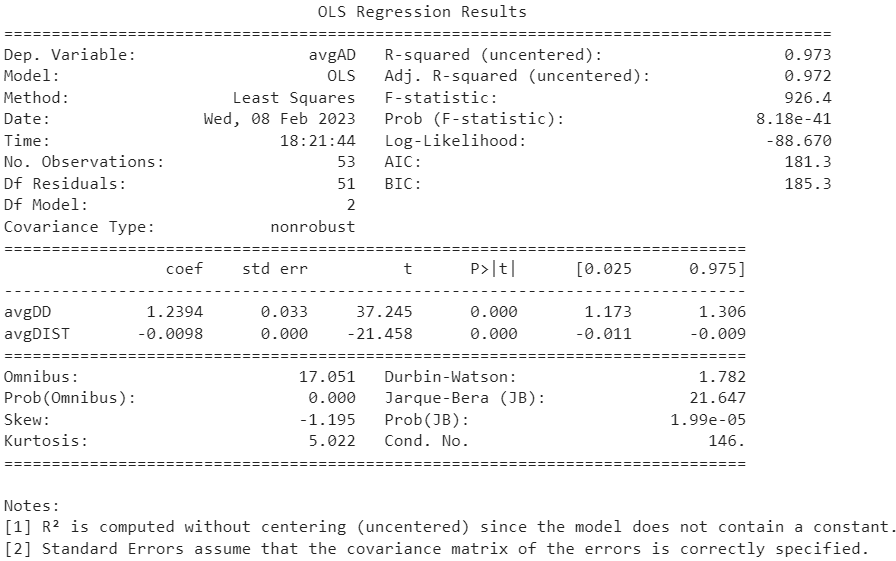
['avgDD', 'avgDIST', 'maxW'] summary\_3VOK

8

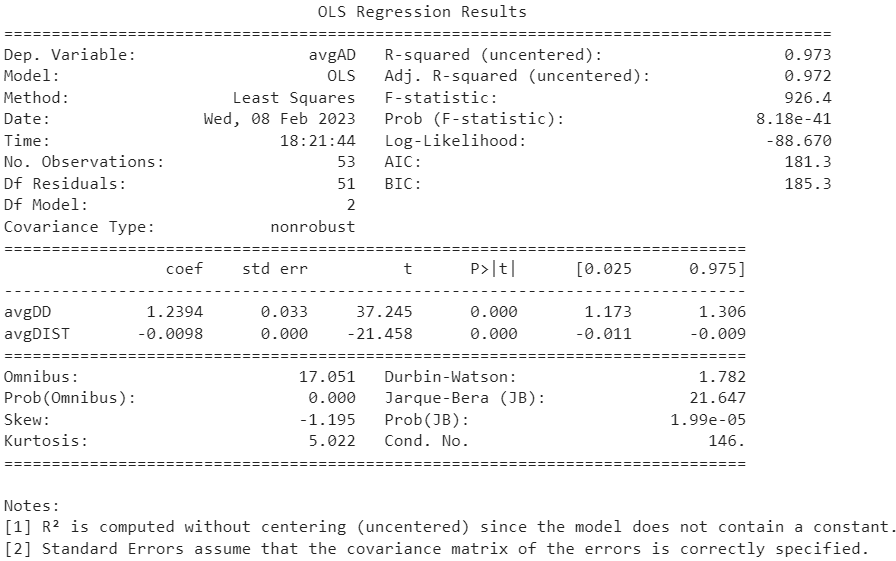
[‘avgDD’]

9

[‘avgDD’, ‘maxW’]

10

[‘avgDD’,’avgDIST’]

11

Ridge and Lasso reglarization:

X11 = ['avgDD', 'avgDIST']

X12 = ['avgDD', 'avgDIST', 'maxW']

X2 = ['avgDD', 'avgDIST', 'maxW', 'avgDD\_avgDD', 'avgDIST\_avgDIST', 'maxW\_maxW', 'avgDD\_avgDIST', 'avgDD\_maxW', 'avgDIST\_maxW', 'const']

|  |  |  |  |
| --- | --- | --- | --- |
| No Reglarization | X11 | X12 | X2 |
| Median Absolute Error | 0.8859 | 0.8823 | 0.878 |
| Mean Absolute Percentage error | 0.1516 | 0.1515 | 0.151 |
| Mean Squared Error | 1.7863 | 1.7856 | 1.7737 |
| Root Mean Squared Error | 1.3365 | 1.3363 | 1.3318 |

|  |  |  |  |
| --- | --- | --- | --- |
| Ridge Reglarization | X11 | X12 | X2 |
| Median Absolute Error | 0.8859 | 0.8823 | 0.8846 |
| Mean Absolute Percentage error | 0.1516 | 0.1515 | 0.1512 |
| Mean Squared Error | 1.7863 | 1.7856 | 1.7787 |
| Root Mean Squared Error | 1.3365 | 1.3363 | 1.3337 |

|  |  |  |  |
| --- | --- | --- | --- |
| Lasso Reglarization | X11 | X12 | X2 |
| Median Absolute Error | 0.8859 | 0.8823 | 0.9269 |
| Mean Absolute Percentage error | 0.1516 | 0.1515 | 0.1754 |
| Mean Squared Error | 1.7863 | 1.7856 | 2.2985 |
| Root Mean Squared Error | 1.3365 | 1.3363 | 1.516 |

There are several ways we can try to improve the performance of our machine learning models in Scikit-learn. Here are a few approaches we can try for each of our four models:

1. Multiple Linear Regression:

* Feature engineering: We can try adding new features or transforming existing features to better capture the relationship between the predictor variables and the response variable.
* Regularization: We can use Lasso or Ridge regression to add regularization to our model and prevent overfitting.
* Cross-validation: We can use cross-validation to estimate the performance of our model on unseen data and tune our hyperparameters accordingly.

1. Decision Tree:

* Pruning: We can use pruning to reduce the size of our tree and prevent overfitting.
* Ensemble methods: We can use ensemble methods such as random forests or boosting to improve the predictive performance of our model.
* Hyperparameter tuning: We can tune the hyperparameters of our decision tree, such as the maximum depth or minimum samples per leaf, to find the optimal settings for our data.

1. Support Vector Regressor:

* Kernel selection: We can try different kernel functions, such as linear, polynomial, or radial basis function, to find the best one for our data.
* C parameter tuning: We can tune the C parameter, which controls the trade-off between fitting the training data and allowing for more generalization, to optimize our model performance.
* Feature scaling: We can try scaling our features, such as using StandardScaler or MinMaxScaler, to improve the performance of our model.

1. Random Forest:

* Number of trees: We can experiment with different numbers of trees in our forest to find the optimal trade-off between bias and variance.
* Maximum features: We can try different settings for the maximum number of features to consider when splitting each node in the tree to find the best balance between randomness and overfitting.
* Out-of-bag error estimation: We can use the out-of-bag error estimate to evaluate the performance of our model during training and tune our hyperparameters accordingly.

To implement these approaches using Scikit-learn, we can use techniques such as GridSearchCV, RandomizedSearchCV, or Pipeline to systematically search through different hyperparameter settings and find the best combination for our data. We can also use metrics such as cross-validation scores or learning curves to evaluate the performance of our models and identify areas for improvement.

Before Optimization:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance metrics | LR | DT | SVR | RF |
| Median Absolute Error | 1.0061 | 1.7525 | 5.7978 | 1.196 |
| Root Mean Squared Error | 1.1749 | 2.5066 | 6.813 | 1.518 |
| R2 Score | 0.9692 | 0.8598 | -0.036 | 0.94856 |

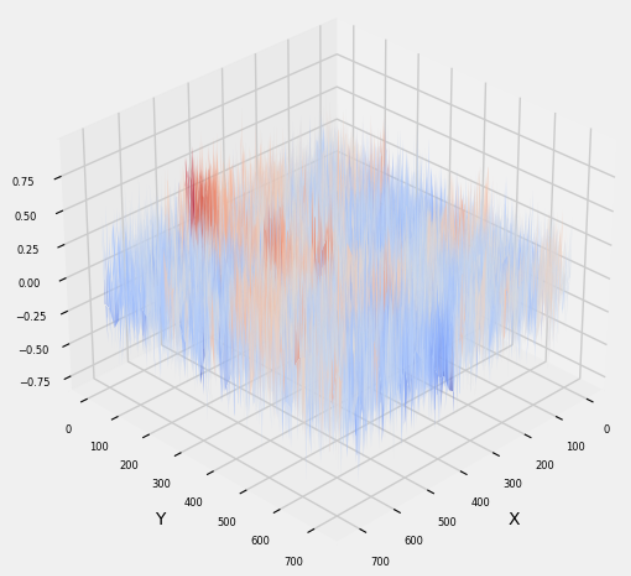
After optimization:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance metrics | LR | DT | SVR | RF |
| Median Absolute Error | 1.0062 | 1.7525 | 1.1529 | 1.3983 |
| Root Mean Squared Error | 1.1749 | 2.5066 | 1.3215 | 1.6991 |
| R2 Score | 0.9692 | 0.8598 | 0.961 | 0.9356 |

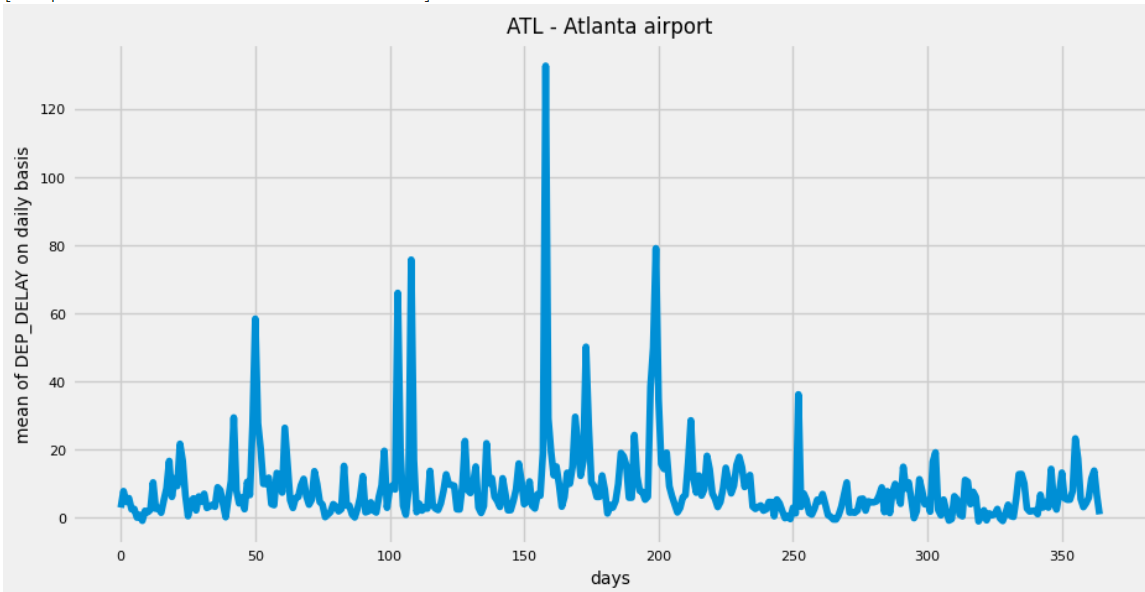
We demonstrated that the multilinear regression model fitted to the simulated data can accurately predict flight delays and capture the relationships between the variables and flight delays. This project can be extended to include more variables and a larger dataset to improve the accuracy of the model and its ability to capture the dynamics of the aviation industry.

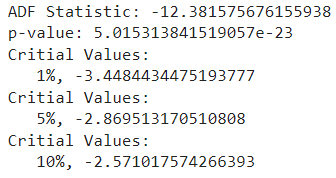
Now we are applying different machine learning model for finding avgAD using avgDD, avgDIST, maxW.

**Chapter 4  
Exploring the relation between departure delays and graph  
invariants:**



**Chapter 5  
Univariate time series analysis:**





ADF Statistic: -12.381575676155938

p-value: 5.015313841519057e-23

Critial Values:

1%, -3.4484434475193777

Critial Values:

5%, -2.869513170510808

Critial Values:

10%, -2.571017574266393

Auto arima:

RMSE: 6.919497510028782

Best parameters: {'maxiter': 50, 'method': 'lbfgs', 'order': (1, 1, 1), 'out\_of\_sample\_size': 0, 'scoring': 'mse', 'scoring\_args': {}, 'seasonal\_order': (0, 0, 0, 0), 'start\_params': None, 'suppress\_warnings': True, 'trend': None, 'with\_intercept': False}

To begin with, we will import the necessary libraries. We will be using the pandas library to handle time series data, numpy for numerical operations, and the pmdarima library for the Auto ARIMA model.

Next, we will load the data into a pandas DataFrame. Let's assume the data is stored in a CSV file called "data.csv". We will use the read\_csv() function to read the file and convert the "Date" column to a datetime object.

Since we want to use the first 350 values for training purposes, we will create a new DataFrame containing only the first 350 rows.

Now, we will use the auto\_arima() function from the pmdarima library to fit the Auto ARIMA model to the training data. We will specify the seasonal parameter to be False since we are working with univariate data.

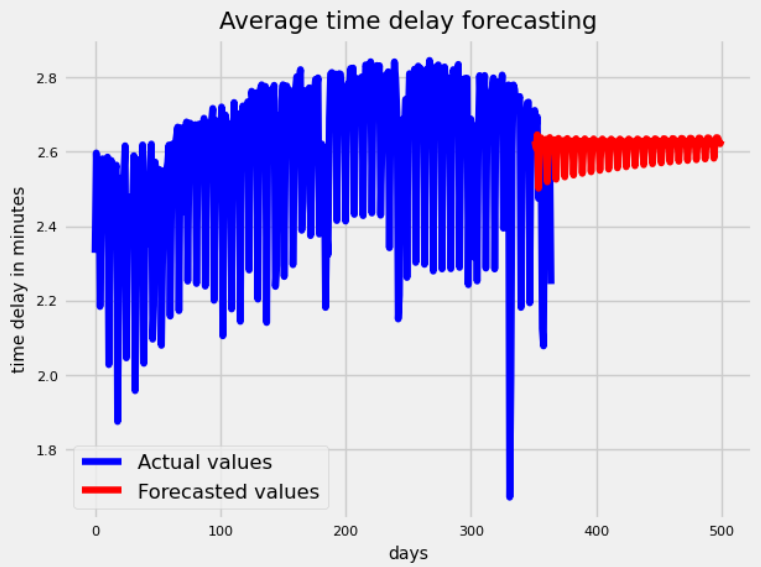
Next, we will use the predict() method to forecast the next 15 values.

Now, we will calculate the RMSE between the forecasted values and the last 15 values of X. Finally, we will print out the RMSE value and the best parameters for the Auto ARIMA model. And that's it! We have trained an Auto ARIMA model using the first 350 values of the X time series and used it to forecast the next 15 values. We then calculated the RMSE between the forecasted values and the last 15 values of X, and printed out the RMSE value and the best parameters for the Auto ARIMA model.

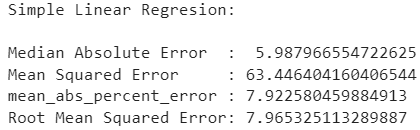


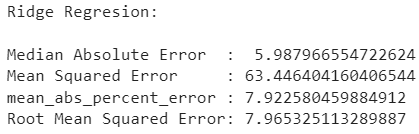
Leave ACF and PACF –

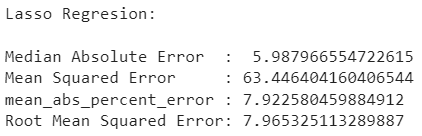
**Chapter 6  
Multivariate time series analysis:**



**Chapter 7  
Hybrid machine learning model:**





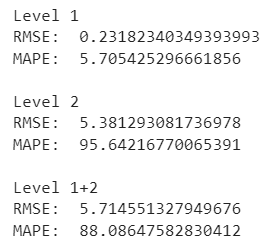


Y = a\*x+b

Where,

a = 3.3253

b = -3.3190



Observations

Results

Conclusion and Future work

References

% @Article{Einstein,

% author = "Albert Einstein",

% title = "{Zur Elektrodynamik bewegter K{\"o}rper}. ({German})

% [{On} the electrodynamics of moving bodies]",

% journal = "Annalen der Physik",

% volume = "322",

% number = "10",

% pages = "891--921",

% year = "1905",

% DOI = "http://dx.doi.org/10.1002/andp.19053221004"

% }

% @Article{Einstein,

% author = "Albert Einstein",

% title = "{Zur Elektrodynamik bewegter K{\"o}rper}. ({German})

% [{On} the electrodynamics of moving bodies]",

% journal = "Annalen der Physik",

% volume = "322",

% number = "10",

% pages = "891--921",

% year = "1905",

% DOI = "http://dx.doi.org/10.1002/andp.19053221004"

% }

% @{{\huge \bf Reference} \pra}

% \begin{enumerate}

% \item[1.] https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018?select=2018.csv\\

% \item[2.] Flight delay prediction based on deep learning and the Levenberg Marquart algorithm.

% (Maryam Farshchian Yazdi1 , Seyed Reza Kamel2\* , Seyyed Javad Mahdavi Chabok2 and Maryam Kheirabadi)-Journal of Big Data.\\

% \item[3.] Introduction to Linear Regression Analysis.

% (Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining)\\

% \item[4.] Analyzing the Regularity of Complete k-Partite Graph using Super Strongly Perfect Graphs.

% (R. Mary Jeya Jothi, Ebin Ephrem Elavathingaln)- 2015 Online International Conference on Green Engineering and Technologies (IC-GET 2015)

% \item[5.] Part (Semi Partial) and Partial Regression Coefficients

% Abdi, H. (2007). In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. pp. 736-740.

% \end{enumerate}

Abstract:

Machine learning models with better statistical analysis provide confidence in our result. Here is the real-world data of flight delay that has been used to create stochastic modeled data to get our estimates. The feature engineering involved a stochastic model giving us the average arrival delay of flights for a particular day. The Regularity of a graph is one of the derived parameters as an input/independent variable. It is an algebraic term that encapsulates graph structural property to make our results more intuitive. Various machine learning regression models are fitted on this data. The Statistical tests provide significance of this variable/parameters on a response variable (with a confidence interval of 95% taken as usual). Further studies involve the use of different graph invariants for analysis to get better results.

These days deep learning models have been used to implement flight delay prediction to achieve better performance. But abstract information of what is learned by these model, make it highly noninterpretable. A fitted model with regression analysis gives us a highly interpretable model. Multivariate time series analysis is performed to get forecasted values. For this purpose AR, MA, ARMA, and ARIMA models have been used. The seasonality component is also taken into consideration.

Flight delay is inevitable and it plays an important role in both profits and losses of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and the incomes of airline agencies. A lot of work has been done on modeling and predicting flight delays, but most of them have tried to predict delays by extracting the main features and the most relevant features. However, most of the proposed methods are not accurate enough because of the massive volume of data, dependencies, and extreme number of parameters.

Keywords: Regularity, Confidence interval, linear regression.

Revised 01:

Machine learning models that leverage statistical analysis offer a high degree of confidence in the accuracy of their results. In this thesis project, we use real-world data on flight delays to generate stochastic models that estimate average arrival delays for flights on specific days. One of the key inputs to our models is the regularity of a graph, which is an algebraic term that encapsulates the structural properties of the graph and makes our results more interpretable.

We fit various machine learning regression models to this data and use statistical tests to evaluate the significance of the regularity parameter on the response variable. Our analysis also includes the use of other graph invariants to improve the accuracy of our predictions.

While deep learning models are increasingly used in flight delay prediction, the abstract nature of these models makes them less interpretable. By contrast, our fitted regression models offer a highly interpretable approach to prediction. We use multivariate time series analysis techniques, including AR, MA, ARMA, and ARIMA models, to generate forecasted values, taking seasonality into account.

Predicting flight delays accurately is essential for airlines, as it can have a significant impact on customer satisfaction and revenue. While many methods for predicting delays have been proposed, most have focused on extracting the most relevant features, which often leads to inaccuracies due to the large volume of data, dependencies, and the sheer number of parameters involved.

1. Introduction

1. Motivation:

The forecasting of flight delays is a significant problem that has an impact on millions of passengers every year. It causes frustration, inconvenience, and potential financial losses leading to missed connections, lost time, and additional expenses to the passengers. However, understanding flight delays and their causes can help businesses improve customer service and minimize the negative impact of delays. Overall, the prediction/forecasting of flight delays is a critical problem that has far-reaching impacts on the economy, passengers, safety, and operational efficiency. Improving our ability to predict and forecast flight delays can help address these challenges and improve the overall performance of the aviation industry. These delays can lead to increased operation costs, lost revenue, and negative impacts on business travel and tourism. According to a report by the US Department of Transportation, flight delays cost the US economy around $25 billion annually. Flight delays can also impact flight safety. For example, when flights are delayed, there may be increased pressure on pilots and ground crews to complete tasks quickly, which can lead to errors or oversights. Improved prediction and forecasting of flight delays can help airlines better manage their resources and reduce the risk of accidents or incidents. Accurate prediction and forecasting of flight delays can help airlines and airport authorities optimize their operations and improve their efficiency. This can include better allocation of resources, such as gates, runways, and personnel, as well as improved planning for maintenance, fueling, and other tasks.

Flight delays can have significant impacts on businesses that rely heavily on air travel for the transportation of goods and services, or for business travel purposes. Understanding the causes of flight delays can provide businesses with valuable insights that can inform decision-making and improve operations. There are various ways flight delay data can be useful for making business decisions. Firstly, businesses can use flight delay data to plan and schedule more effectively by analyzing historical flight data to identify times and routes that are more likely to experience delays and adjust their schedules accordingly. Secondly, businesses can monitor flight delays and their causes to identify potential bottlenecks in the supply chain and take proactive measures to mitigate them. Thirdly, businesses can use flight delay data to improve customer satisfaction by providing customers with real-time updates on flight status and alternative travel options. Fourthly, analyzing flight delay data can help businesses identify the costs associated with delays and take steps to reduce them. Finally, businesses can monitor flight delay data to identify potential risks associated with delays and take proactive measures to minimize their impact. In conclusion, flight delay data can provide businesses with valuable insights that can inform decision-making and improve operations.

Businesses can use flight delay information to improve customer service and minimize the negative impact of delays. Real-time updates about flight delays can reduce anxiety and provide customers with a sense of control. Alternative travel options, such as rebooking on a different flight or arranging ground transportation, can help customers reach their destinations quickly and efficiently. Providing information about compensation and refunds can help customers feel valued and heard, even in the face of delays. Proactive communication about delays and potential solutions can demonstrate a business's commitment to customer service and build customer loyalty. Finally, customer feedback about flight delays can help businesses identify areas for improvement and prevent future disruptions. Overall, flight delay information can be useful for addressing customer concerns and improving customer service.

2. Problem description and objective:

Flight delay forecasting uses past data and other factors to predict the likelihood of a flight being delayed. The problem is to introduce the notion of graph invariants in flight delay forecasting.

Graph invariants are used to study the structural properties of graphs and to distinguish between different graphs that may look different but share some underlying structure. These invariants are important because they provide a framework for studying the structural properties of graphs and for classifying and comparing large sets of graphs. Graph invariants can be used to identify important features of a graph, measure its complexity and connectivity, and predict its behavior and functionality.

This problem statement has the following objectives:

1. Providing prior knowledge of flight delays on a particular day to ensure customer satisfaction. The service provider will inform customers if there is a likelihood of delays in their flight schedules so that they can plan accordingly.
2. To make business intelligent decisions by the service provider. The service provider will use the flight delay data to make better business decisions. For example, they may use the data to adjust their flight schedules, allocate resources, optimize their staffing levels, or improve their overall operations.
3. To compare the performances of the individual airline industry. The performance of individual airlines means their ability to manage flight delays. This could be achieved by analyzing various metrics such as the frequency and duration of delays, the percentage of flights that are delayed, and the impact of delays on customer satisfaction. By comparing the performance of different airlines, the service provider can identify areas for improvement and take appropriate measures to enhance their performance.
4. …

These could be achieved by collecting and analyzing historical flight data and other relevant data sources to predict potential delays. Overall, the objectives listed aim to improve the customer experience, increase operational efficiency, and drive business success by leveraging flight delay data.

3. Summary of contribution:

Revision 01:

In my thesis project, I conducted research on the relationship between flight delays and graph invariants. To incorporate graphical properties into our analysis, I used data simulation techniques and employed various machine-learning models and algorithms to demonstrate our findings.

One approach I used was creating a stochastic model to formulate average flight delays daily, and I performed statistical analysis on the results. I also introduced the concept of regularity and incorporated it into our statistical analysis.

To improve the performance of our models, I employed various machine-learning techniques. I also performed univariate and multivariate time series analyses using AR, MA, ARMA, and ARIMA time series models.

In our stochastic model, we considered different combinations of attributes, and we discussed our observations on the statistical significance of each attribute. We found that some attributes were statistically significant while others were not, and we considered only the significant attributes for further analysis.

To better understand the relationships between attributes, we computed a correlation matrix and discussed the dependencies between the different attributes. We also computed a partial correlation matrix to account for these dependencies and plotted a pair plot to express the distribution of each attribute.

Our univariate analysis considered each airport delay separately and discussed the output. In contrast, our multivariate analysis considered the interdependencies between each airport delay.

Overall, my contribution to this project involved applying various statistical and machine-learning techniques to study the relationship between flight delays and graph invariants, and to improve our understanding of the factors that contribute to flight delays.

4. Organization of report:

…

2. Description of data:

The USA flight dataset we have contains information on various attributes related to flights, including the origin and destination airports, scheduled and actual departure and arrival times, taxi in and out times, arrival and departure delays, and distance between origin and destination.

The purpose of this dataset is to provide a comprehensive record of flight information in the USA, which can be used for a variety of purposes. For example, airlines and airports can use this dataset to analyze their performance and identify areas for improvement. They can also use it to track trends and patterns in flight delays and cancellations and to make adjustments to their schedules and operations accordingly. Additionally, regulatory agencies can use this dataset to monitor compliance with safety and security regulations and to identify potential safety risks.

One of the primary problems that this dataset aims to solve is improving the efficiency and reliability of air travel in the USA. By analyzing the data, airlines, and airports can identify areas where they are experiencing delays or other issues, and take steps to address these issues. This can lead to better on-time performance, which can improve the overall passenger experience and reduce costs for airlines.

The insights that can be derived from this dataset are numerous. For example, airlines and airports can use the data to identify which routes are the most profitable, which airports have the highest delays, and which aircraft are the most efficient. They can also use the data to identify trends and patterns in passenger behavior, such as when people are most likely to travel and what types of amenities they prefer. Additionally, the data can be used to identify potential safety risks, such as airports with high rates of bird strikes or aircraft with recurring maintenance issues.

Moreover, this dataset can also be used by researchers and analysts to study and understand the broader trends and patterns in the US airline industry. For example, researchers may use this dataset to examine the impact of various factors, such as weather or airline mergers, on flight delays and cancellations. This information can be used to develop strategies to improve the efficiency and reliability of air travel in the USA.

Overall, the flight dataset serves as a valuable tool for airlines, regulatory agencies, and other stakeholders in the aviation industry. By analyzing the data and gaining insights into flight operations, airlines can improve their performance, reduce costs, and provide a better experience for their passengers.

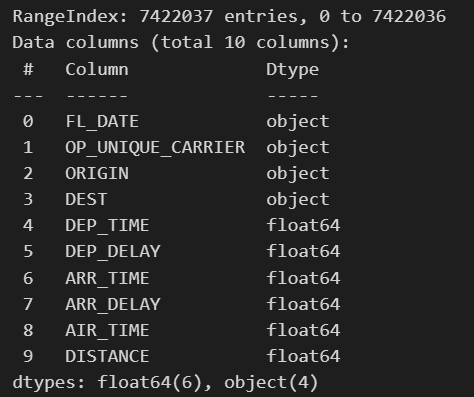
In the dataset, the abbreviation used corresponds to the origin airport and destination airport. The attributes available in the dataset are explained below,

1. OP UNIQUE CARRIER: It is a unique scheduled operating carrier code(airline industry code) that identifies a carrier.
2. OP CARRIER: The flight number assigned by the carrier.
3. Origin: It is the starting flight airport code.
4. Destination: It is the ending flight airport code.
5. Departure time: It is the planned departure time.
6. Departure delay: The total delay on departure in minutes.
7. Arrival time: The actual arrival time.
8. Arrival delay: The total delay on arrival in minutes.
9. Air time: The duration between wheels off and wheels on time.
10. Distance: The distance between airports (miles).
11. Taxi\_IN: The time duration elapsed between wheels-on and gate arrival at the destination airport.
12. Taxi\_OUT: The time duration elapsed between departure from the origin airport gate and the wheels off.
13. Wheels ON and Wheels OFF are respectively the times when an aircraft’s wheels touch down on or leave the runway.
14. Carrier delays, NAS delays, security delays, and late aircraft delays are all related to flight delays caused by different factors such as weather, air traffic control, etc.

For flight delay prediction, prior information is very much necessary. Here parameter like OP UNIQUE CARRIER, OP CARRIER, Air time, Taxi\_IN, Taxi\_OUT, Wheels\_ON, Wheels\_OFF, Carrier\_delay, NAS\_delay, security\_delay, late\_aircraft\_delay is all information dependent on that day itself for which we are going to make a prediction. This information is not possible to get in advance. Hence we are just omitting this info as it is not useful for our problem statement. Just dropping these columns using the following command,

data.drop(['OP\_CARRIER\_FL\_NUM','TAXI\_OUT', 'WHEELS\_OFF', 'WHEELS\_ON','TAXI\_IN','CARRIER\_DELAY','WEATHER\_DELAY','NAS\_DELAY', 'SECURITY\_DELAY',  'LATE\_AIRCRAFT\_DELAY'], axis=1, inplace=True)

data.info()



The OP\_UNIQUE\_CARRIER is the code for airlines and 'OP\_CARRIER\_FL\_NUM' is a corresponding number, available to study flight data wrt individual airlines. Here are the corresponding airline names for the given flight carrier codes:

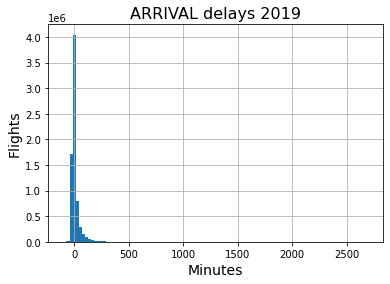
|  |
| --- |
| 9E: Endeavor Air Inc. |
| AA: American Airlines Inc. |
| MQ: Envoy Air Inc. |
| G4: Allegiant Air |
| OH: PSA Airlines Inc. |
| B6: JetBlue Airways Corp. |
| YV: Mesa Airlines Inc. |
| EV: ExpressJet Airlines Inc. |
| F9: Frontier Airlines Inc. |
| YX: Republic Airways Inc. |
| HA: Hawaiian Airlines Inc. |
| NK: Spirit Airlines Inc. |
| UA: United Airlines Inc. |
| OO: SkyWest Airlines Inc. |
| WN: Southwest Airlines Co. |
| AS: Alaska Airlines Inc. |
| DL: Delta Air Lines Inc. |

A dataset is a collection of data that is organized in a specific way and used for a particular purpose. Here are some key points about the dataset mentioned:

1. The dataset contains information on a total of 360 airports.
2. The data pertains to the year 2019 and covers one year.
3. The dataset contains several variables, including 'Air\_time', 'Arr\_delay', 'Dep\_delay', 'TAXI\_IN', 'TAXI\_OUT', 'Wheels\_ON', 'Wheels\_OFF', 'Carrier\_delay', 'NAS\_delay', 'security\_delay', and 'late\_aircraft\_delay'. All of these variables are measured in terms of minutes.
4. The variables 'Arrival\_time' and 'Departure\_time' are time instances that are recorded in 24-hour format with the HH: MM time stamp. The variable 'Distance' is measured in kilometers and 'FL\_Date' is the date of the flight in YY-MM-DD format.
5. The variables 'ORIGIN' and 'DEST' are codes that represent the names of the airports from which the flight originated and to which it was destined.

Overall, this dataset contains information about flights from 360 airports in the year 2019. It provides data on a variety of variables related to the flights, including the duration of the flight, delays, and the time of departure and arrival. The dataset can be used for various purposes, such as predicting flight delays, analyzing airport performance, or optimizing airline schedules.

The arrival delay and departure delay distribution for the 2019 dataset is shown below,

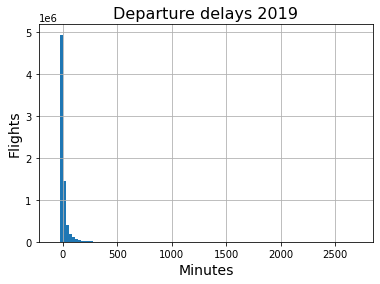


Mean value: 5.414849168270909

Median value: -6.0

Skewness: 8.724431987317192

Kurtosis: 143.7963561380509



Mean value: 10.923267333861132

Median value: -2.0

Skewness: 9.754031796638873

Kurtosis: 168.74017019405102

**Summary of contribution:**

*This thesis presents the findings of our research on the dependence of graph invariants on flight delays. The notion of graphical properties has been incorporated using some data simulation techniques. The machine learning models/algorithms used to demonstrate our output. There are different manifestations have been performed and discussed. These include different approaches as mentioned here,*

1. *Stochastic model created with the idea to formulate average flight delays on daily basis. Statistical analysis was performed and discussed.*
2. *Concept of regularity is introduced and incorporated into statistical analysis.*
3. *The machine learning techniques employed to improve model performance.*
4. *Univariate and multivariate time series analysis is performed using AR, MA, ARMA, and ARIMA time series models.*
5. *…*

*We formalize a stochastic model that has information on flight delays on daily basis. Here we have indulged different combinations of attributes in the model. All the corresponding observations have been discussed there. We show that some attributes are statistically significant and some are not. Those attributes which are significant are considered for further analysis.*

*The correlation matrix that gives a correlation between all combinations of attributes is computed and discussed. As the correlation matrix has some dependencies over other attributes, the partial correlation matrix is also computed and discussed with the observations. Pair-plot is plotted to express the distribution of each attribute. The Machine learning techniques were applied to improve the performance of the model.*

*The univariate and multivariate time series analysis is performed. In Univariate analysis, each airport delay is considered separately and its output is discussed. While multivariate analysis considers interdependencies between each airport.*

Data is information about a person, facts, and statistics that may be qualitative or quantitative. This information of a person is useful or more specifically collectively for a large group of people from the dataset. Let the example of Employees in a company has their information related to their jobs such as employment Id, Age, Skills, joining date, and work profile. These data for faster processing are being converted into different file formats like CSV, Excel, Tabular, Text format, etc.

\vspace{\baselineskip}

This data representation can be done in terms of any data structure, a pictorial representation that is in terms of graphs. Many machine learning algorithms employ various statistical, probabilistic, and optimization methods to learn from past known experiences. This learning finds useful patterns from large, unstructured, and complex datasets. These algorithms have a wide range of applications, in marketing just as customer purchase behavior detection, in the medical field for disease modeling, information theory for junk e-mail detection, and spam message detection. These applications are supervised learning variants rather than unsupervised learning. In the supervised learning method of Machine learning, decisions are made by the learning dataset. At the learning stage labeling of data is known and this information is used for making decisions on unlabeled examples. The following is the NOIR classification of data.

% (Fig.~\ref{fig:1a}).

\begin{figure}[ht]

\centering

\includegraphics[width=1.0\textwidth]{Images/a.png}

\caption{Data Classification}

\label{fig:1a}

\end{figure}

% Reference \cite{Einstein}

\chapter{Machine Learning}

This is a topic under the recent application useful in many applications. As in recent years or decades, two major changes have boosted machine learning use in a practical scenario. These are,

\begin{enumerate}

\item Advent of high-speed processors.

\item Huge amount of data being generated

\end{enumerate}

\noindent

Machine learning is the study of computer algorithms in a way that it improves/teaches itself through experience and by the use of data. It has included some parts of computational statistics and is part of Artificial intelligence. The algorithms used in machine learning are applicable in various works such as email filtering, sentiment analysis, customer feedback support (Chat Box), speech and text recognition, and even in object detection techniques. In earlier days this application cannot fulfill the system requirement and was not feasible to develop such an algorithm to perform on computers. In recent processors, these tasks become easier, even though in recent days some machine learning algorithms are taking lots of time for their learning.

\vspace{\baselineskip}

In nutshell, Machine learning more focuses on making a decision based on predictions made by computers. Not always machine learning statistics, but it also focuses on Data Mining. Data Mining is a field that has roots in exploratory data analysis. One part is machine learning i.e Deep learning which mimics the biological brain. In the field of business analytics, machine learning is often referred to as predictive analysis

\vspace{\baselineskip}

Machine learning uses programs in terms of algorithms that modify certain variables’ present data structure depending on the data provided to it. Thereafter it performs certain useful tasks. For advanced tasks, it could be impossible to create an algorithm manually. From the practical point of view, it will no longer be effective for a person to develop such an algorithm.

\vspace{\baselineskip}

\noindent

\chapter{Linear regression}

Linear regression is a machine learning algorithm based on supervised learning. It runs a regression task. Regression model targets a dependent variable based on independent variables. It is mainly used to find relationships between variables. Different regression models differ based on the type of relationship between dependent and independent variables under consideration and the number of independent variables used.

\vspace{\baselineskip}

% {{\large \bf Simple Linear Regression: } \par}

\section{Simple Linear Regression:}

Linear Regression performs the task of predicting the value of the dependent variable (y) based on a specified independent variable (x). Therefore, this regression technique finds a linear relationship between x (input) and y (output). hence the name linear regression.

% \begin{flushleft}

% \begin{minipage}[c]{0.45\textwidth}

% \centering

% \hrule

% \vspace{0.5\baselineskip}

% {\Large \bf Supervisor} \\

% \flushleft{\large \textbf{Department of Mathematics} \\

% Indian Institute of Technology, Kharagpur}

% \end{minipage}

% \end{flushleft}

{\large \bf \[ y\_i = \beta x\_i + \alpha + \epsilon\_i \]}

% \begin{center}

% \centering

% [y\_i] = response variable

% [x\_i] = input variable/ regressor

% \end{center}

\vspace{\baselineskip}

% {{\large \bf Gradient Descent: } \par}

\subsection{Gradient Descent:}

To update coefficient values to reduce the Cost function (minimizing RMSE value) and achieve the best fit line, the model uses Gradient Descent. The best fit line for a given training dataset can be find using Gradient descent in a smaller number of iterations. The idea is to start with random b0 and b1 values and then iteratively update the values, reaching the minimum cost.

\vspace{\baselineskip}

% {{\large \bf Multiple Linear Regression: } \par}

\section{Multiple Linear Regression:}

The following is a set of methods intended for regression in which the target value is expected to be a linear combination of the features.

When there are multiple predictors, the equation of linear regression is simply extended to carry more variables:

{\large \bf \[ y\_i = {\beta\_0} + {\beta\_1}x\_{1i} + {\beta\_2}x\_{2i} + ... + {\epsilon\_i} \]}

% \begin{figure}[ht]

% \centering

% \includegraphics[width=0.5\textwidth]{Images/f.png}

% \end{figure}

Instead of a line, we now have a linear model, the relationship between each coefficient and its variable (feature) is linear.

\vspace{\baselineskip}

% The Multiple linear regression model (with all variable as discuss in chapter 5) has residual term following the distribution as shown in Fig.~\ref{fig:4a}. Thus, the assumption that followed by hypothesis testing is employed.

% \begin{figure}[ht]

% \centering

% \includegraphics[width=0.7\textwidth]{Images/g.png}

% \caption{Normal distribution of residual term}

% \label{fig:4a}

% \end{figure}

\chapter{Experimental Details}

% {{\large \bf Steps involved in the experiment:} \par}

\section{Steps involved in the experiment:}

\hrule

\begin{enumerate}

\item Simulate Data using Poisson distribution:\\

• Generate Poisson distribution with lambda = 750 (average number of lights per day)\\

• Perform an “iter” number of iterations, such that in each iteration, select the nth day, giving n[i] several flights.\\

• Randomly select n[i] a number of flights (data points) from the original data.\\

\item Compute the following variables from the above simulated data:\\

• Average departure delay\\

• Average distance\\

• Maximum Indegree of graph\\

• Regularity of graph\\

\item Fit the Multilinear regression model

\item Analyse performance

\end{enumerate}

\hrule

\vspace{\baselineskip}

Variables used for training are average departure delay, average distance, maximum Indegree of flight, and regularity of graph.

\vspace{2\baselineskip}

% {{\large \bf Variables Used:} \par}

\section{Variables under consideration:}

\begin{enumerate}

\item Average departure delay (avgDD):\\

Average departure delays of all flights scheduled for a given 24 hours

\item Average distance (avgDIST):\\

Average arrival delays of all flights scheduled for a given 24 hours

\item Maximum Indegree of flight (maxIndegree):\\

A maximum number of flight arrivals on that day at a particular airport.

\item Regularity of graph (Regularity):\\

Regularity of graph is a graph theoretic invariant used to capture all geometric structural information. Macaulay2 is software used for regularity computation.

\item Average arrival delay(response variable-avgAD):\\

Average arrival delays of all flights scheduled for a given 24 hours

\end{enumerate}

\begin{figure}[ht]

\centering

\includegraphics[width=0.7\textwidth]{Images/b.png}

\caption{Computation time for Regularity}

\label{fig:5a}

\end{figure}

Regularity calculation is computationally very expensive as shown in Fig.~\ref{fig:5a}. Computed for each day with graph having edge weight as the maximum number of flight arrival at that airports in a given day. As number of nodes increases computational time increases exponentially.

\vspace{\baselineskip}

% \section{Distribution of residual term}

The Multiple linear regression model (with all variable as discuss in chapter 5) has residual term following the distribution as shown in Fig.~\ref{fig:4a}. Thus, the assumption that require by hypothesis testing is followed.

\begin{figure}[ht]

\centering

\includegraphics[width=0.7\textwidth]{Images/g.png}

\caption{Normal distribution of residual term}

\label{fig:4a}

\end{figure}

% {{\large \bf Coefficient of Correlation and Partial correlation coefficient:} \par}

\section{Coefficient of Correlation and Partial correlation coefficient:}

The R2 score was used earlier but it is coming out to be large. The additional used parameter does not lead to any significant improvement. Hence, this work is done by considering some other performance measures, like the coefficient of correlation and partial correlation coefficient. Finding relationships between variables in a dataset helps us to eliminate redundant variables for prediction, and how rich the dataset is explained by hypothesis testing.

\vspace{\baselineskip}

One of the main issues rising concerns the dependency on different random variables. We can account for this issue using a correlation matrix, linear fit, and other different methods to find the relationship and use it in our model. But what if we have two variables that have a strong correlation between them but the truth is they are both tightly related to a third variable that increases the strength of this relationship?

\vspace{\baselineskip}

This issue can undermine our assumptions and lead to incorrect results. One possible technique for explaining this problem is to use a measure called partial correlation. In contrast to the Pearson correlation, the partial correlation takes into account the presence and "control" over third variable.