**AIRLINE TRAFFIC ANALYSIS**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Jaisan Sam**

**Gopika Jayaraj**

**Karthik P Hareendran**

**Sarika Sasi A**

**Sivakrishna B Nambiar**

****

**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

**List of Figures**

| **Figure Number** | **Description** |
| --- | --- |
| Figure 1 | Arrival Delay Chart |
| Figure 2 | Departure Delay Chart |
| Figure 3 | Distribution |
| Figure 4 | Overview of delay by distance |
| Figure 5 | Overview of delay in minutes |
| Figure 6 | Density arrangement of delays |
| Figure 7 | Overview of delay impact |
| Figure 8 | Pair plots showing various delays |
| Figure 9 | Layout of data points |
| Figure 10 | Average delay based on carrier |
| Figure 11 | Delay Distribution by carrier |
| Figure 12 | Delay Distribution by origin |
| Figure 13 | Correlation |

**List of Abbreviations**

| **Abbreviation** | **Full Name** | **Description** |
| --- | --- | --- |
| FL\_DATE | Flight Date | Date of operation of aircraft |
| OP\_UNIQUE\_CARRIER | Operator Unique | Unique Aircraft Registration Code |
| OP\_CARRIER\_FL\_NUM | Operator Carrier Flight Number | Full Unique Aircraft Registration Code |
| NAS | National Aviation Services | the type of weather delays that could be reduced with corrective action by the airports or the Federal Aviation Administration. |

*Table 1.*

**Table of Contents**

| **Particulars** | **Pg.No** |
| --- | --- |
| List of Figures | 2 |
| List of Abbreviations | 3 |
| Abstract | 5 |
| 1.Problem Definition | 6 |
| 1.1.Overview | 6 |
| 1.2.Problem Statement | 6 |
| 2.Introduction | 7 |
| 2.1 Dataset Details | 7 |
| Literature Survey | 8 |
|  |  |

*Table 2.*

**Abstract**

Airline travel is starting to notice a steep increase after the pandemic. The increased influx and efflux of aircraft consequently means a proportional increase in delays. The occurrence of such delays causes inconvenience to passengers as well as enormous losses to the companies as well. In this project, we attempt to explore the selected raw data and convert the same to meaningful patterns and help derive insights into the reasons for these circumstances. Additionally, we will also use these insights to build a machine learning model which can forecast the probability of such delays along with possible remedies to resolve them. This project will contain exploratory data analysis combined with hyperparameter tuning of machine learning models to tweak the results to the accuracy and outcome we need.

**1. Problem Definition**

**1.1 Overview**

Air Transport has become one of the major methods of transportation in the present world and unlike its counterparts in land and sea air transport enjoys the benefit of relatively unobstructed paths and has proven to be safer. In this project we aim to analyze the reason for delays in air travel and ways to reduce it.

**1.2 Problem Statement**

The increase in airline traffic has increased delays for airplanes to reach their destination on time. Main reasons for such delays are :

1)Delay due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).

2)Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight .

3)Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.

In this project our aim is to analyze the reasons for airline delays, and to derive insights that may help us to solve these issues. Additionally, we will also use these insights to build a machine learning model which can forecast the probability of such delays along with possible remedies to resolve them.

**2. Introduction**

Flight delay has become a major challenge to global aviation industries. For example, data from the US Bureau of Transportation (BTS) indicates that, in 2012, 14.69% of the flights in the US domestic market were delayed. In 2019, this ratio increased to 20.8% (BTS, 2020). According to Ball et al. (2010), flight delay led to a total cost of as much as US $32.9 billion for the US economy in 2007. Delay is one of the most performance indicators of any transportation system. Notably, commercial aviation players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by the diﬀerence between scheduled and real times of departure or arrival of a plane. Country regulatory authorities have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential subject in the context of air transportation systems.

Flight delays have negative impacts, mainly economic, for passengers, airlines, and airports. Given the uncertainty of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip costs, to ensure their arrival on time. On the other hand, airlines suﬀer penalties, ﬁnes and additional operation costs, such as crew and aircrafts retention in airports. To better understand the entire ﬂight ecosystems, vast volumes of data from commercial aviation are collected every moment and stored in databases. Submerged in this massive amount of data, analysts and data scientists are intensifying their computational and data management skills to extract useful information from each datum. The flight delay and delay propagation problems always are paid attention by both domestic and international aviation industry insiders. The overseas research on these aspects is prior to us. This is related to the start of the civil aviation industry. Most of the research is about flight delays in domestic and their flight schedules.

**2.1 Dataset Details**

The dataset procured for this project is obtained from U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics via the website Kaggle. <https://www.kaggle.com/sherrytp/airline-delay-analysis>.  
We will be analyzing the 2019 version of the dataset whose details are given below:

Size: 809.33 Mb  
Rows: 16497871  
Columns: 20

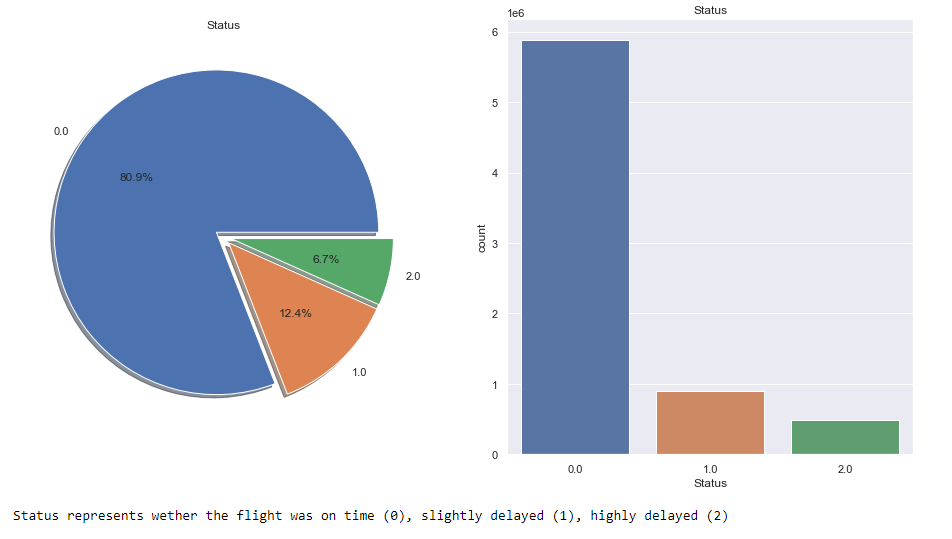
The dataset contains details on the total aircraft arrival and departure information combined with the delays occurred and the timings included. The arrival and departure delays are given in positive as well as negative values indicating delays for the former and early timings for the latter respectively. In our Project departure delay is considered as a target.

**3. Literature Survey**

Tan, Xinlong, et al [1] serves an explorative examination on the delay propagation in the Chinese airline market,which is the world's second largest airline market.China has experienced serious flight delay problems. Firstly, they studied explorative analysis, mainly to quantify and depict an overall delay propagation pattern in the Chinese airline market. Airlines strategic behaviors in ground and flight buffer choices have yet been modeled and jointly estimated. Brueckner et al. [2] have proposed a comprehensive theoretical and empirical research framework to integrate the airlines’ buffer choices in economic analysis on delay propagation, which was also verified with the US market data.

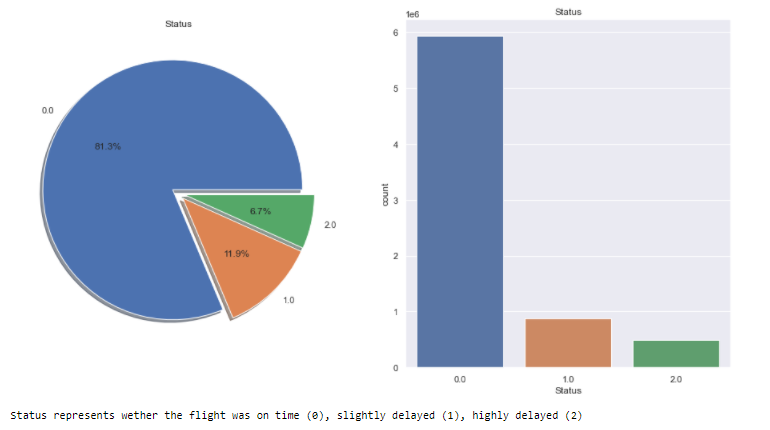
Taxonomy and summarization of the initiatives used to address the ﬂight delay prediction problem is proposed by Sternberg, Alice, et al. [3] according to scope, data, and computational methods, giving particular attention to an increased usage of machine learning methods. Besides, we also present a timeline of signiﬁcant works that depicts relationships between ﬂight delay prediction problems and research trends to address them.

**4. Data Analysis**

****

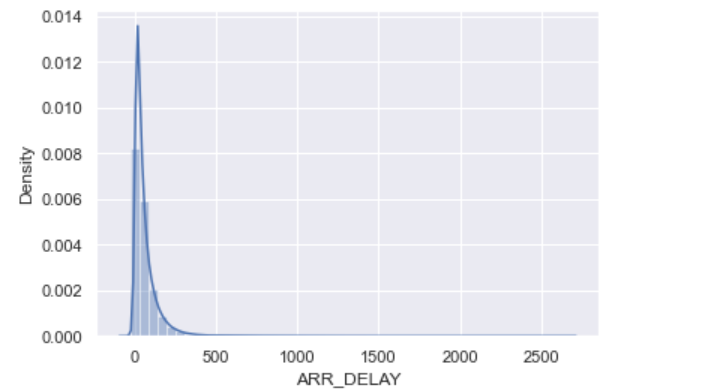
*Figure 1. - Arrival Delay Chart*

*Major portion of the flights have arrived on time.   
Although the percentage of delays are less, we need to find out if there is a pattern to them*

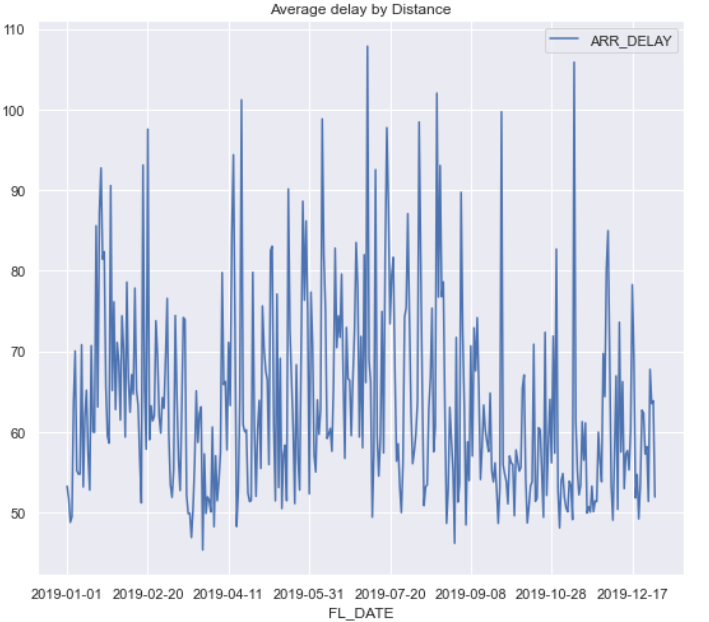
****

*Figure 2. - Departure Delay Chart*

*Major portion of the flights have arrived on time.   
Although the percentage of delays are less, we need to find out if there is a pattern to them*

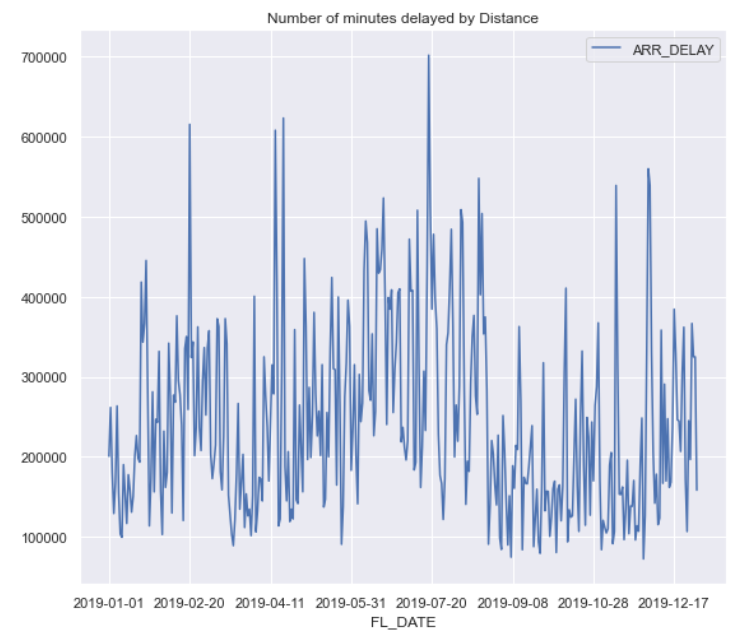
**

*Figure 3. - Distribution  
We can see that the distribution is skewed. More values seem to be concentrated in the earlier portion of the dataset. This will help us to handle any missing values as required.*

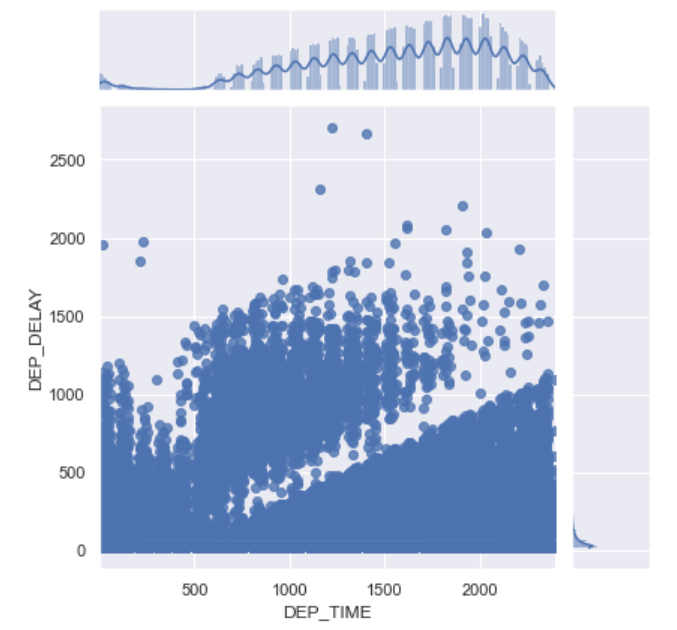
****

*Figure 4. - Overview of delay by distance*

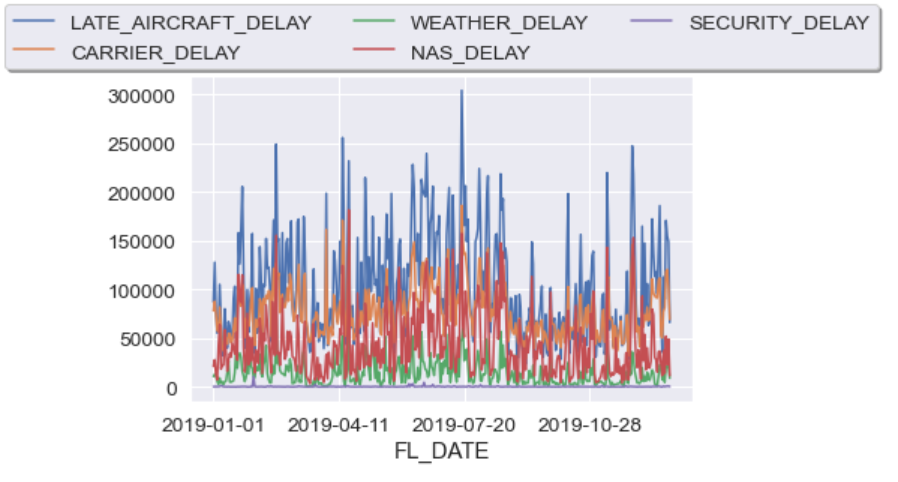
*Delays may occur due to the distance the aircraft is hoping to traverse. More distance means more resources and more projections into weather conditions, slight miscalculation of any will lead to delays.*

****

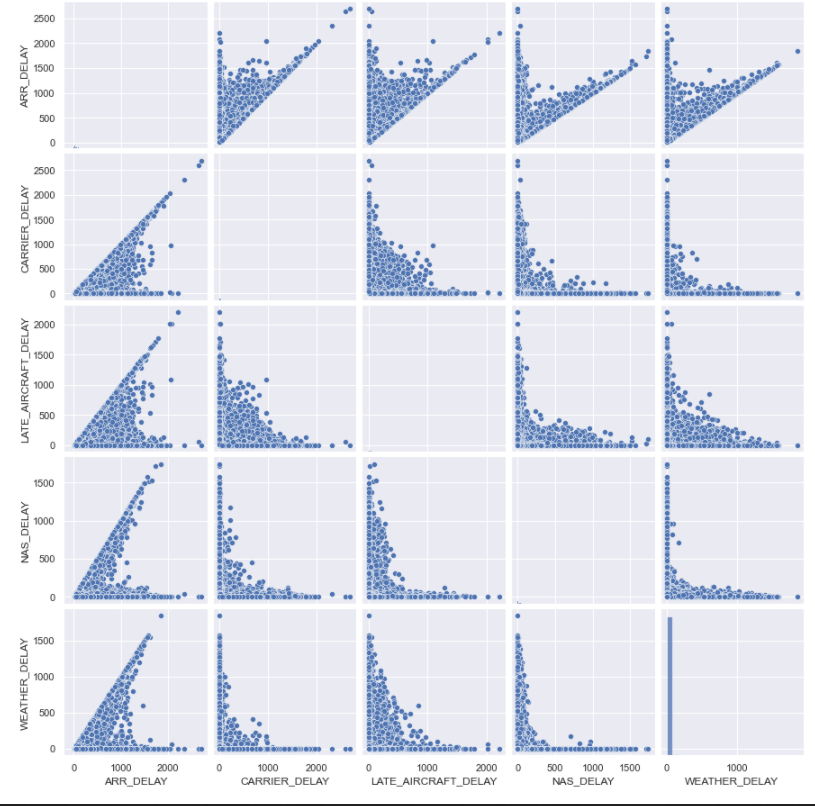
*Figure 5. - Overview of delay in minutes*

****

*Figure 6. - Density arrangement of delays*

****

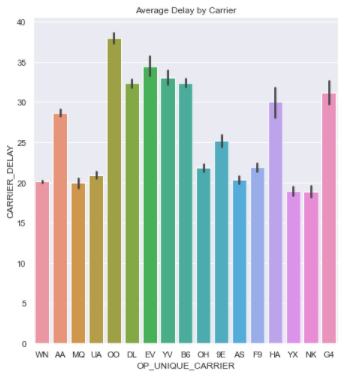
*Figure 7. - Overview of delay impact  
As we can see, the blue graph has the maximum impact - LATE AIRCRAFT DELAY.  
During the course of our project we will be taking into account this finding to provide more accuracy to the model.*

****

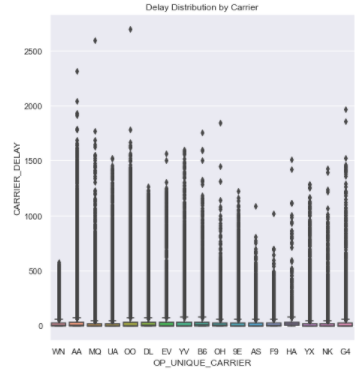
*Figure 8. - Pair plots showing various delays*

****

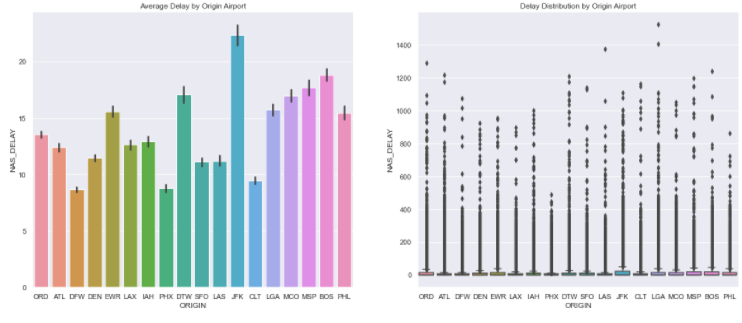
*Figure 9. - Layout of data points  
The plots which have data points more than the others have skewness towards the left.*

****

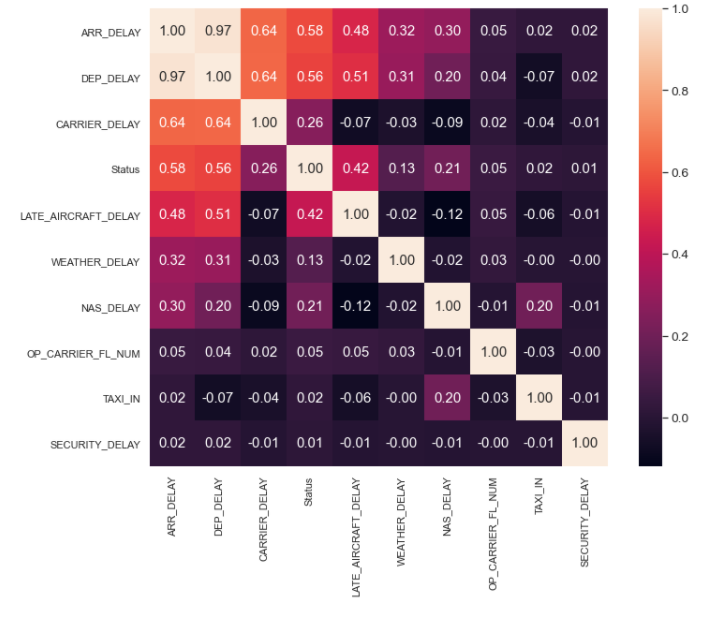
*Figure 10. - Average delay based on carrier  
Carriers OO and EV have the maximum delay reports on average.  
YX and NK have less delays*

****

*Figure 11. - Delay Distribution by carrier  
We can see outliers in each carrier, however these data points are unique and need to be considered to enhance prediction.*

****

*Figure 12. - Delay Distribution by Origin  
Based on the first plot, we can easily say that the origin airport of the carrier plays a significant role in the cause of delays - JFK being the highest.*

****

*Figure 13. Correlation  
High correlation can be seen for the features - Arrival delay and departure delay.   
Upon analysis, we will be removing one of them as a part of feature reduction.*

**5. Result**

* The data set is Skewed - Some of the features to the left and some to the right.
* Each of the five delays have different contributing factors
* Due to the high correlation between Arrival and Departure delays, we have the option to drop one of them to avoid redundancy.
* Although it appears to be a classification problem, due to the feature set, we can determine that this is a regression problem.

**6. Conclusion**

Exploratory data analysis has been completed. The dataset is now ready for the next step - Preprocessing. Univariate and bivariate analysis has been performed and the results have been included on the report. The target variable will be Departure Delay.

**References**

[1] Tan, Xinlong, et al. "An Exploratory analysis of flight delay propagation in China." *Journal of Air Transport Management* 92 (2021): 102025.

[2] Brueckner, Jan K., Achim I. Czerny, and Alberto A. Gaggero. "Airline mitigation of propagated delays: theory and empirics on the choice of schedule buffers." (2019).

[3]Sternberg, Alice, et al. "A review on flight delay prediction." *arXiv preprint arXiv:1703.06118* (2017).