The London School of Economics and Political Science

ST2195 Programming for Data Science

Coursework Project

Contents

1. Introduction

The purpose of this report is to analyze the flight arrival and departure details for all commercial flights on major carriers within the USA, from October 1987 to April 2008. The data is obtained from the 2009 ASA Statistical Computing and Graphics Data Expo and consists of nearly 120 million records in total, taking up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed. The complete dataset along with supplementary information and variable descriptions can be downloaded from the Harvard Dataverse.

The dataset used for this analysis is a subset of 2004-2006 years and supplementary information provided by the Harvard Dataverse. The questions to be answered are as follows:

1.When is the best time of day, day of the week, and time of year to fly to minimize delays?

2.Do older planes suffer more delays?

3.How does the number of people flying between different locations change over time?

4.Can you detect cascading failures as delays in one airport create delays in others?

5.Use the available variables to construct a model that predicts delays.

The report details the steps taken to answer each question, including data wrangling and cleaning, data visualization, and modeling techniques. The code used in R and Python for each question is also included for replicability

Data Sources

In order to analyze flight delay data, we need to collect and preprocess the data from reliable sources. In this analysis, we will be using available data from the from the Harvard Dataverse . The data is available in CSV format and can be downloaded from their website (https://doi.org/10.7910/DVN/HG7NV7).To speed up the process of data manipulation and for faster we have set up a database using SQLite. Before you start to use the code make sure to download SQLite (https://www.sqlite.org/download.html). After downloading SQLite

Data Cleaning and Preprocessing

Before analyzing the data, it is important to preprocess and clean the data. In our analysis, we performed the following data cleaning and preprocessing steps:

Removed missing values: We removed any rows that had missing values for important columns such as "Arrival Time" or "Departure Time".

Based on our need we converted Year,Month,DayOfMonth into a single date format which later was used to plot data against time

Created new columns: We created new columns such as "Plane Life", "season", and "TimeofDay" to help with our analysis.

Removed unnecessary columns: For our analysis we used only columns that were necessary such as Year, Month, DayofMonth, CRSDepTime, ArrDelay, DepDelay We removed columns that were not needed for our exploroatory analysis up until The modeling part to reduce the memory footprint (columns: “TaxiIn”, ”Taxiout”, ”UniqueCarrier”,”Distance”).

Summary Statistics:

Here are some summary statistics of the data after preprocessing:

Number of flights: 20,984,369

Number of airports: 3245

Number of cities: 2675

Average Arrival delay time: 7.45 minutes

Average Departure delay time: 8.85 minutes

Maximum Arrival delay time: 1925 minutes

Maximum Departure delay time: 1930 minutes

We will use this cleaned and preprocessed data for further analysis in the following sections of the report.

2.When is the best time of day, day of the week, and time of year to fly to minimize delays?

2.1 Introduction

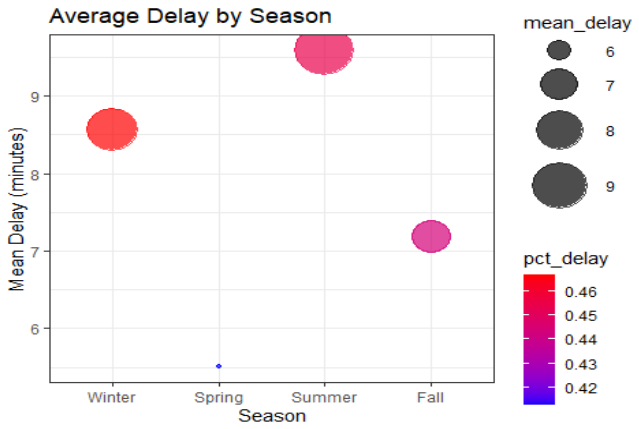
In this section, we will explore the data to determine the best time of day, day of the week, and time of year to fly to minimize delays. The purpose of this analysis is to provide insights to help travelers plan their flights better and avoid delays.

2.2 Methodology

To answer this question, we will use a combination of data exploration and visualization techniques. We will analyze the flight data to determine the average arrival delay time during different times of the day, days of the week, and months of the year. We will look for Arrival Delays grouped by time days,months,seasons,timeofday and also We will look for percentage of delayed flights based on grouped data in order to check which grouped data has the best observations.

2.3 Analysis of the result

Figure 1: Average Arrival Delay by Season



Based on Figure 1 We observe that during **spring** season we have less delays based on the percentage of flights that had delays is **41.3%** and the mean time of the delays is the lowest at **5.51** minutes compared to other seasons

Figure 1.1: Average Arrival Delay and percentage of delayed flights by Day of week

The bar plot to the right shows percentage of Delayed Flights in total flights on total flights. Based on the graph Saturday has the least percentage of delayed flights (**38.65**%) followed by Tuesday(**41.62**%).
From the line plot in figure 1.1 we notice that the Average Arrival Delay by Day of Week is the lowest on Saturday (**3.7 minutes**) followed by Tuesday (**5.7 minutes**)

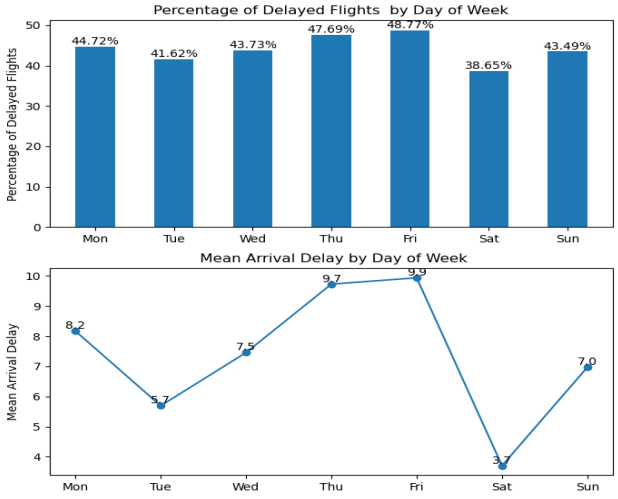
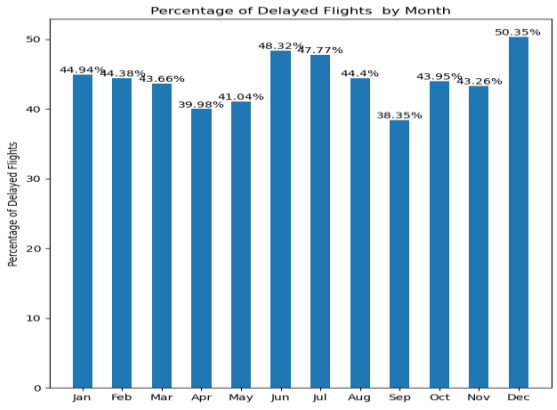
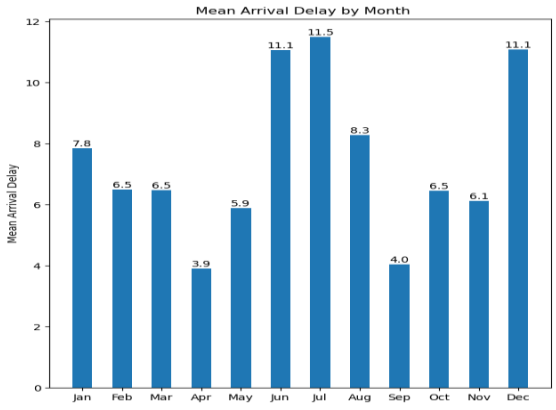


Figure 1.2 Average Arrival Delay by Month





From figure 1.2 We notice that Percentage of Delayed Flights by month is the lowest on the month of September reaching **38.35**% followed by month of April (**39.98**%). We notice

also that the Average Arrival Delay is the lowest on the month of April (**3.9 minutes**) followed by September (**4.0 minutes**)

Figure 1.3: Average Arrival Delay and percentage of Delayed Flights by Group of Day

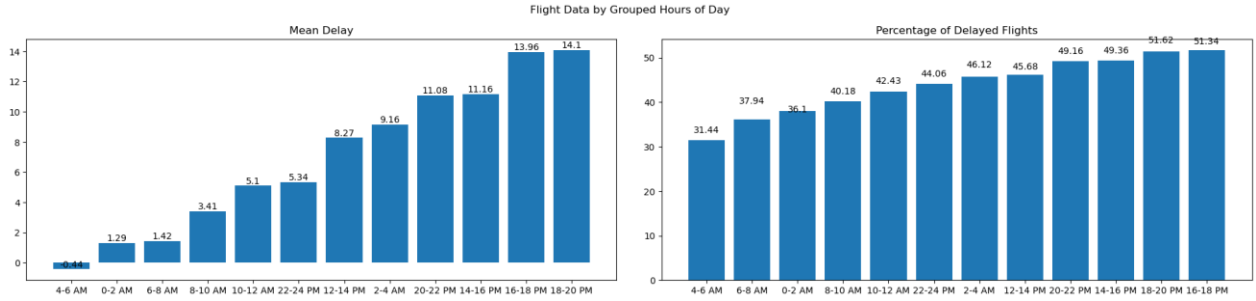
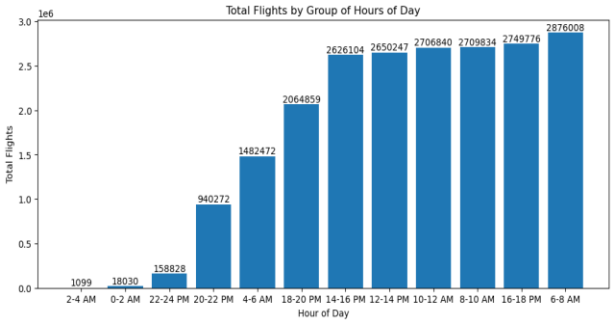


Figure 1.4: Count of Flights by Group of Hours



From figure 1.3 we notice that between **4-6 am** the average of arrival delay for flights is the lowest at **-0.44 minutes** and the percentage of flights that experience delay is **31.44%.** This is followed by flights that are scheduled between **6-8 am** which experience **1.42 minutes** delay and **37.94%** of flights experience delay during this timeframe. On the left plot the number of flights between **4-6 am** and between **6-8 am** is significant and these timeframes based on the data experience the least amount of delays.

3.Do Older Planes Suffer More Delays?

3.1 Introduction

In this section, we will explore the data to determine if older planes suffer more Delays. The purpose of this analysis is to provide insights and assess if the age of the planes correlates with plane having more delays.

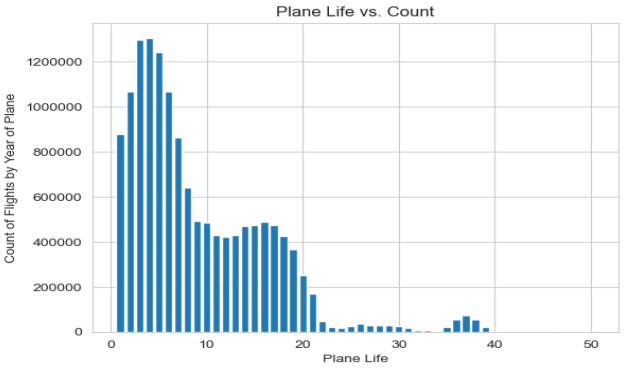
3.2 Methodology

To effectively answer this question, we will employ a rigorous methodology that combines data exploration and visualization techniques. Our first step will involve setting up the relevant data in a structured database using the SQLite platform. This will enable us to efficiently manage and analyze the large dataset of flight data.

Once the data has been properly structured and organized, we will conduct a thorough analysis of the flight data to determine the number of flights for each age of plane. We will also plot the delays versus the age of the plane to uncover any patterns or trends that may emerge.

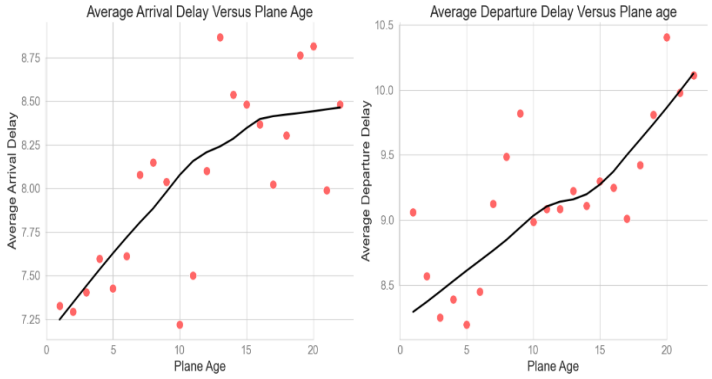
2.3 Analysis of the result

Figure 2: Count of number of flights versus age of plane



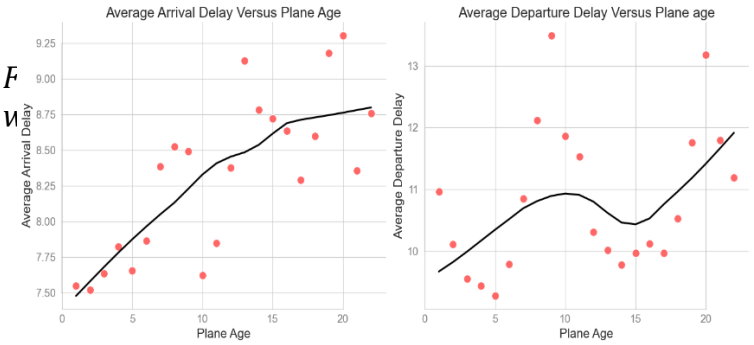
After analyzing the flight data, we observed that the count of flights decreases as the age of the plane increases. However, we also noted that the number of flights for planes with more than 22 years is not significant and can potentially skew our observations. To address this, we decided to narrow our focus to only the subset of planes that are in service between **1-22 years**, which we believe provides a more accurate representation of the flight data.

Figure 2.1: Average Arrival Delay for all planes versus age of plane



Our analysis of the relationship between plane age and delays reveals a clear positive association between these variables. Specifically, we observe that as the age of the plane increases, so does the average arrival and departure delays of the flights operated by that planes

F
w



Specifically, when analyzing only the flights with delays (left two plots), we observe that the average arrival and departure delays increase as the age of the plane increases. This finding underscores the need for proactive maintenance and replacement strategies for aging aircraft to mitigate the negative effects of aircraft age on flight delays

4.How does the number of people flying between different locations change over time?

4.1 Introduction

In this section, we will explore the data to explore how the number of people flying between different locations change over time

4.2 Methodology

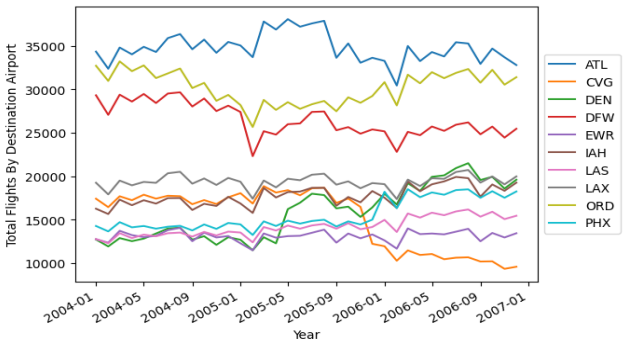
In order to answer the question at hand, we will employ a combination of data exploration and visualization techniques. Given that we do not have access to data on the number of passengers per flight, we will use the number of flights as a proxy for the number of people flying between different locations.

In order to answer the question at hand, we will employ a combination of data exploration and visualization techniques. Given that we do not have access to data on the number of passengers per flight, we will use the number of flights as a proxy for the number of people flying between different locations.

Our first step will be to analyze the flight data to determine the number of flights for each year. This will allow us to identify any trends or patterns in the overall volume of air travel over time. To achieve this, we will leverage data exploration techniques such as data filtering, aggregation, and statistical analysis. Next, we will use the same data exploration and visualization techniques to identify the top destinations for air travel based on the number of flights. By identifying the most popular destinations, we can gain insight into the underlying drivers of air travel demand, and potentially uncover any regional or global trends in travel patterns.

4.3 Analysis of the result

Figure 3 Line plot of total flights by destination through time

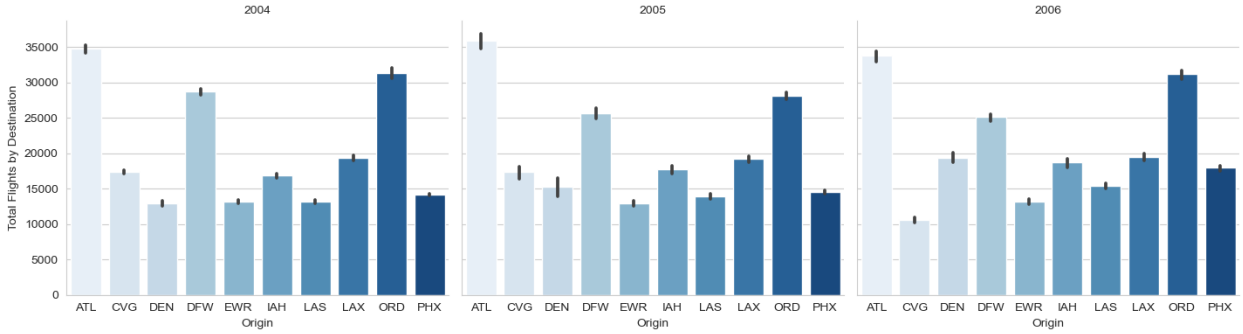


The line graphs presented above depict the number of flights to the top U.S airports from January 2004 to December 2006. A visual inspection of the plot reveals clear seasonal patterns in air travel demand. Specifically, we observe a sharp decrease in the number of flights between January and February, followed by a moderate increase in March. This trend is likely attributable to the seasonal nature of air travel demand, with many individuals returning home from holiday travels in January and February, and the spring break season commencing in March. As such, a more comprehensive analysis of

air travel demand would need to consider a broader range of factors in order to generate a more complete understanding of travel patterns and trends.

On the figure below we plotted number of flights by top 10 airports. We didn't notice any pattern or change in number of flights compared year to year.

Figure 3.1 Number of flights for top 10 airports



5.Can you detect cascading failures as delays in one airport create delays in others?

5.1 Introduction

In this section, we will explore the data to determine if delays in one airport create delays in other. The purpose of this analysis is to provide insights and assess if the age of the planes correlates with plane having more delays.

5.2 Methodology

The analysis of cascading failures due to delays in one airport creating delays in others was conducted by creating a network of airports using the flight data. The network was analyzed for its structure, and the top airports based on their betweenness centrality were identified. These top airports were then removed one by one, and the impact on the rest of the network was observed. Additionally, data on flight delays and cancellations was collected and added as vertex attributes to the network. Additionally we plotted correlation matrix of top 5 highest correlation of departure delays of top 10 airports with other departure delays of other airports.

5.3 Analysis of the results

Betweenness centrality is a measure of a node's centrality in a network. It quantifies the number of shortest paths that pass-through a given node. In the context of airports, betweenness centrality measures the importance of an airport in connecting other airports.

In this analysis, the betweenness centrality was calculated for all airports in the network, and the top 10 airports with the highest betweenness centrality were identified. These top airports are likely to have a greater impact on the overall network, as they serve as important hubs connecting multiple airports.

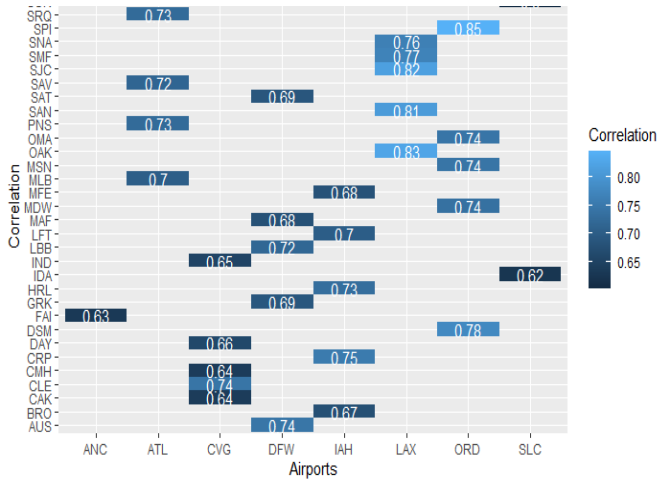
The betweenness centrality values for the top airports are presented in the centrality data frame, with the airport name and corresponding betweenness centrality score. The results show that (ATL) has the highest betweenness centrality, followed by Dallas/Fort Worth (DFW) and (SLC).

name	betweenness
ORD	8577.914
DEN	5512.786
DFW	10375.391
IAH	7842.847
ATL	20668.460
LAX	3827.182
MSP	4053.779
SLC	9226.564.
ANC	6407.876
CVG	6908.310

Table 1: Effects of removing Airports to the network

## [1] "Removed ORD	number of connected components: 2"
## [2] "Removed DEN	number of connected components: 3"
## [3] "Removed DFW	number of connected components: 4"
## [4] "Removed IAH	number of connected components: 4"
## [5] "Removed ATL	number of connected components: 15"
## [6] "Removed LAX	number of connected components: 1"
## [7] "Removed MSP	number of connected components: 3"
## [8] "Removed SLC	number of connected components: 4"
## [9] "Removed ANC	number of connected components: 7"
## [10] "Removed CVG	number of connected components: 3"

Figure 4 : Correlation of top 10 airports with other airports



Our analysis of the correlation of departure delays between the top 10 airports and the departure delays of the remaining airports reveals interesting findings. Notably, we observe a high correlation between the departure delays of LAX (Los Angeles International) and those of Oakland International, San Diego International-Lindbergh, San Jose International, and Sacramento International, all of which are located within the state of California. Similarly, we observe a high positive correlation between the departure delays of ORD (Chicago O'Hare International) and those of SPI (Capital), both located in the state of Illinois.

It is important to note that correlation does not imply causation. However, these findings can provide insights into potential common factors that may contribute to departure delays across airports. Further research is needed to identify the underlying causes of the observed correlations and their potential implications for air travel

6. Use Available data to construct a model that predicts delays?

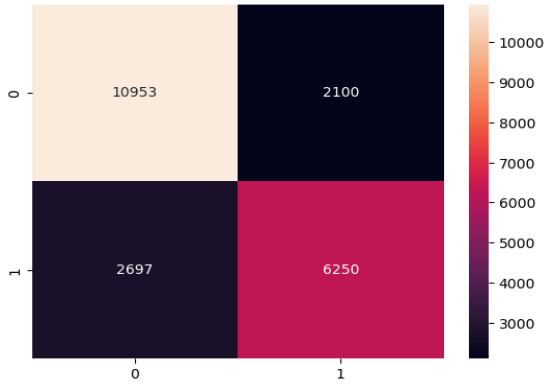
6.1 Introduction

In this section, we will describe the methodology used to construct a model that predicts delays. Specifically, our target variable is the Arrival Delay (ArrDelay), which is measured in minutes. To make this a classification report, we have encoded the variable using 1s and 0s. We have chosen the XGBoost model as it is known for its high performance in predictive modeling. Due to computational limitations, we have randomly sampled 110,000 observations from the data.

6.2 Methodology

First we have selected our target variable which is Arrival Delay (ArrDelay). Since ArrDelay is in minutes we have encode it with 1 and 0 to make this a classification report. We choose xgboost model. Because of computation reasons we choose a random sample of 110,000 . Since in the first iteration we noticed that the model predicted more negative values since it was an unbalanced dataset we decided to balance the dataset

6.3 Analysis of the results




```
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0.0	0.80	0.84	0.82	13053
1.0	0.75	0.70	0.72	8947
accuracy			0.78	22000
macro avg	0.78	0.77	0.77	22000
weighted avg	0.78	0.78	0.78	22000

XGBoost model achieved an overall accuracy of 78% on the test set. In terms of precision, the model correctly identified 80% of the negative cases (class 0) and 74% of the positive cases (class 1). This means that when the model predicted that a flight has delay , it was correct 74% of the time, and when it predicted that a flight will not have delay , it was correct 80% of the time.In terms of recall, the model correctly identified 84% of the negative cases and 70% of the positive cases. This means that of all the actual flight delays, the model correctly identified 70% of them, and of all the actual non flight delays , the model correctly identified 84% of them. The F1 score is the harmonic mean of precision and recall, and it provides an overall measure of the model's performance. The F1 score for class 0 is 0.82, and for class 1, it is 0.72. The weighted

average of the F1 score is 0.78, which is the same as the overall accuracy of the model. The macro-average F1 score is 0.77, which is the same as the macro-average precision and recall. The macro-average is the average of the F1 score for each class, and it gives equal weight to each class, regardless of their support (the number of instances in each class). The weighted average of precision, recall, and F1 score takes into account the support of each class and provides a more accurate measure of the model's performance when the classes are imbalanced, as in this case. Finally, the ROC curve area is 0.77, which indicates that the model has moderate discrimination ability to distinguish between flight delay and non-flight delay.