Flight Delays Analysis Report using Python and R

UNIVERSITY OF LONDON

ST 2195 Programming for Data Science 2021-22

Student Number: 200615086

Table of Contents

1 Introduction

1.1 Background

In the 21st century, vast amounts of data are constantly pouring into the world. Using this data, humans have been able to do endless things that they could not do before the Internet age. It seems to be an accurate expression that human do not do something on their own but can be applied to our lives with the help of computers. Computer programming has made it easier for us to find the desired values and results, and even allows us to predict future and make the models which be learnt on own through machine learning and deep learning.

1.2 Current issue

Among the various fields today that analyze various things through data-based programming, the field to be covered in this report is the Aviation and Flight field. This field has long been a field that utilizes computing programs, and flight is a part that requires accurate statistical analysis because it costs a lot of money and is directly related to passengers' safety. Among the various statistical analyses, the analysis to be covered in this report is about delay. Aircraft delays are also a big issue in this field. Delay causes complaints from passengers to airports and airlines that provide aviation services, and the variable of delay may be dangerous for safety. Aircraft standing on the airport lane without take off on time can cause confusion to the control tower, which is likely to lead to accidents. Therefore, it would be nice if there were no delay, but this is impossible. Thus, we should deal with delays by anticipating and predicting them.

1.3 Purpose of the report

The purpose of this report is to predict and minimize delays by analyzing the pattern of delays based on past flight data. This report will present the analysis of delays using data on all flights in the United States provided by the 2009 ASA Statistical Computing and Graphics Data Expo.

1.4 Scope of study

This analysis will find out how departure and arrival delays have a relationship with other flight information, and how the delay pattern of past flights has been. The body of report will proceed as a process of exploring and solving five questions. The following is the five questions,

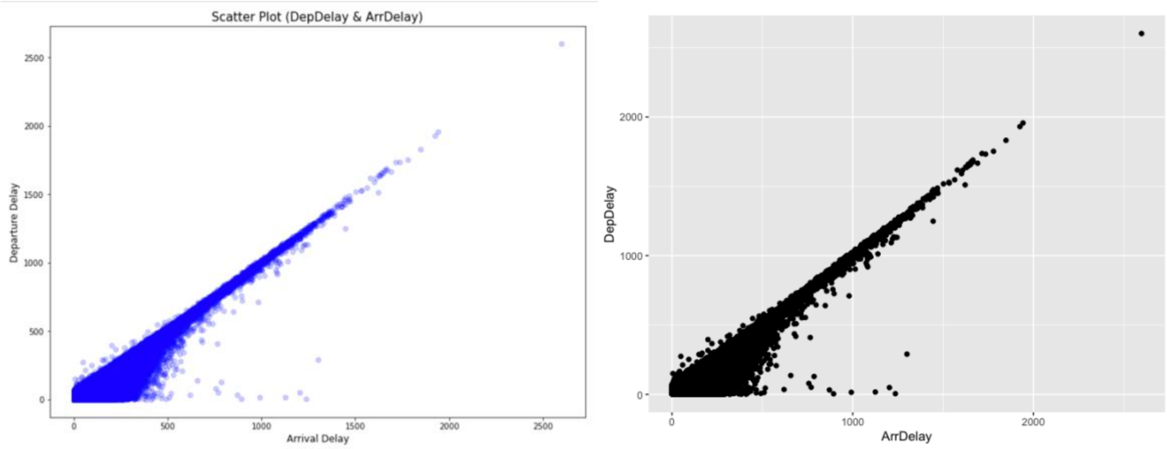
1. When is the best time of day, day of the week, and time of year to fly to min imize delays?
2. Do older planes suffer more delays?
3. How does the number of people flying between different locations change over time?
4. Can you detect cascading failures as delays in one airport create delays in others?
5. Use the available variables to construct a model that predicts delays.

At the end of this analysis report, it will be possible to identify the cause of delay, a variable different from the schedule, and to predict how much delay will occur in future flights.

2 Data Analysis

2.1 Data Understanding

Before the analyzing, the data will be briefly introduced for understand. Data files consist of flight data, carrier data, airport data, and plane data. Flight data contains information about all flights in the United States for the year. Carrier data shows which aircraft and airline each code means. Airports data contains location information for all airports in the United States. Plane data contains information of aircraft. In this analysis, the flight data for a total of three years from 2005 to 2007 were used to identify what factors can affect flight delay and introduce a machine learning model that can predict delay. Cancelled flights were not required for analysis, so they were excluded. In addition to this, prior to the questions, the correlation between departure delay and arrival delay was analyzed. In general, if there is a delay in departure, the arrival time is also delayed, so it was expected to be a strong positive relationship.



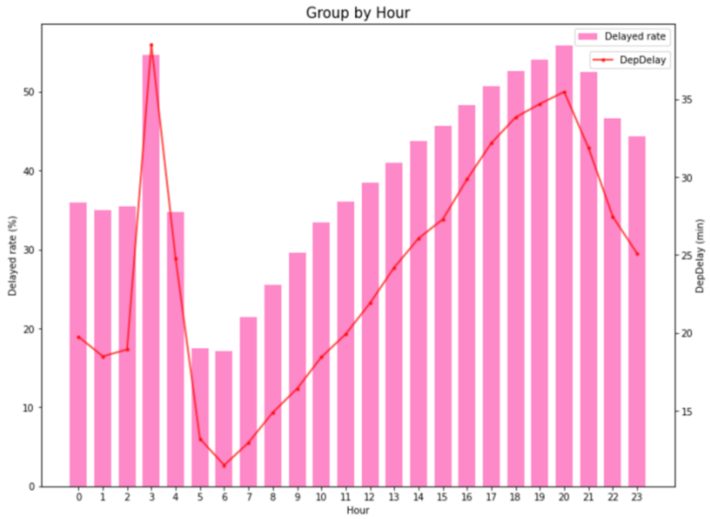
<Figure 1, 2>

Looking at the scatterplot graph, the results were as expected. The accurate value which identified through programming is 0.92. This result means that if there is a delay in departure, arrival delays occur in almost all flights. In the author's opinion, only one of the two was judged to be considered because the correlation coefficient is high, and among the two delays, the arrival delay had a greater influence and was more important than the departure delay, so it was decided to analyze only the data with the arrival delay.

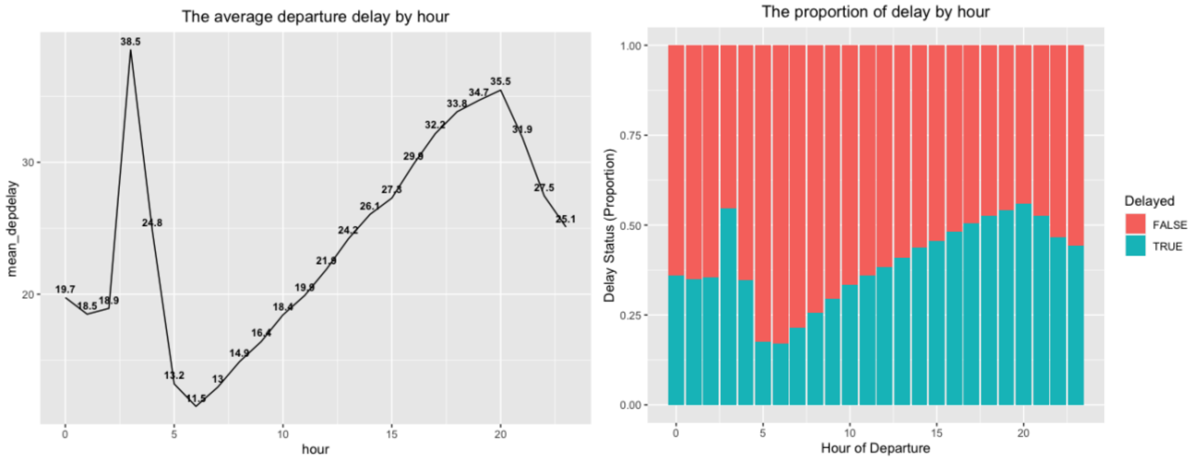
2.2 Question 1: When is the best time of day, day of the week, and time of year to fly to minimize delays?

The first question is that to find the best time to minimize the delay. The purpose of this question is to find the best start time. Thus, the data with departure delay were dealt with. Two ways were approached to find the best time. The first is to find the time when delayed flights account for the lowest proportion. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time *(Sarojanie, 1970)*. Therefore, the author was also considered to be delayed if the departure delay was more than 15 minutes, and not delayed if it was less than or equal to. The second way is to find the time with the lowest average departure delay. As given in the question, the criteria for classifying the time were divided into three categories: hour (0-24), day of week (Monday-Sunday), and month (January-December).

2.2.1 Group by Hour



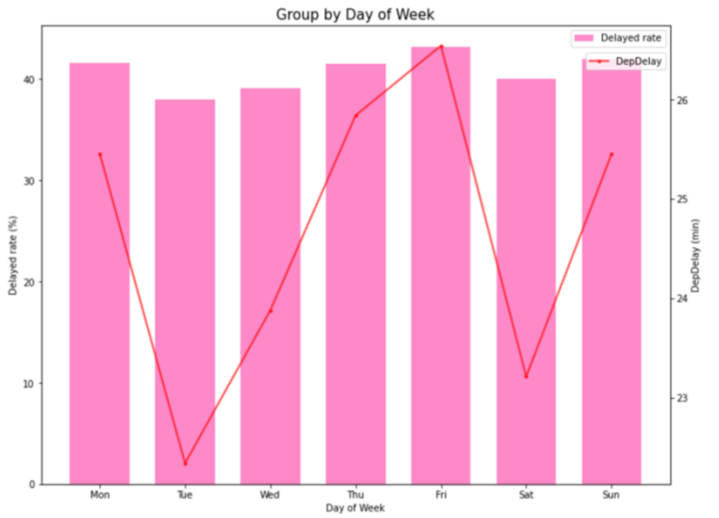
<Figure 3>



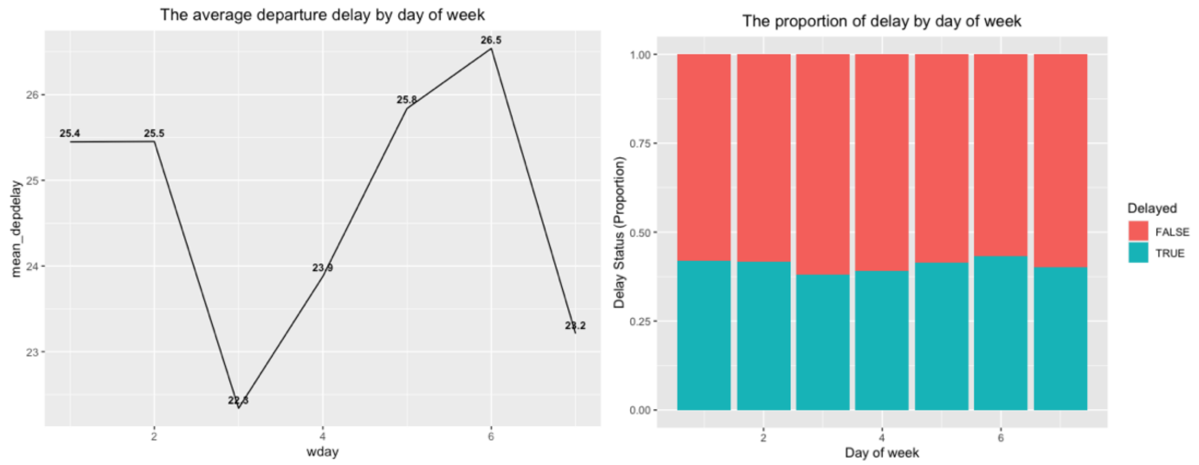
<Figure 4,5>

Looking at the graph 'Group by hour', it can be said that 6 am, which has the lowest result value in both the analyzing method, is the best time to minimize the delay. After that time, the delay proportion and the delay average time continue to rise. The worst time was 3 am.

2.2.2 Group by Day of Week



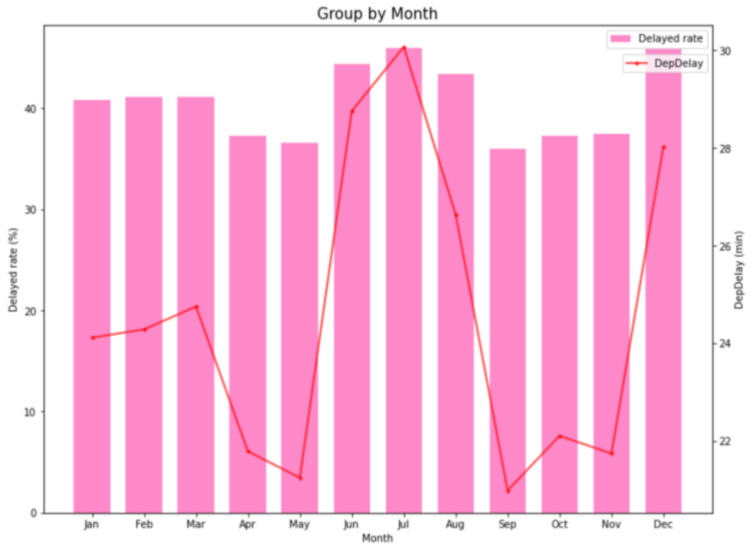
<Figure 6>



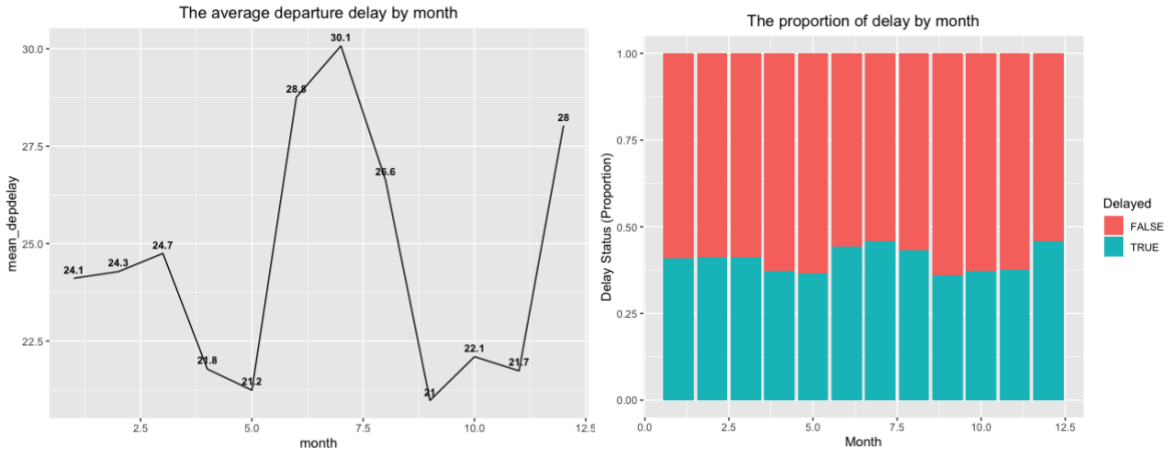
<Figure 7, 8>

Next, the results of ‘Group by day of week’ showed that Tuesday had a minimum in both ways. The proportion of delay remained in the 40% range for most months, but Tuesday had the lowest value at about 38%. The average delay time was also the lowest at 22.3 on Tuesday. Therefore, the best day of week to minimize delay is Tuesday.

2.2.3 Group by Month



<Figure 9>

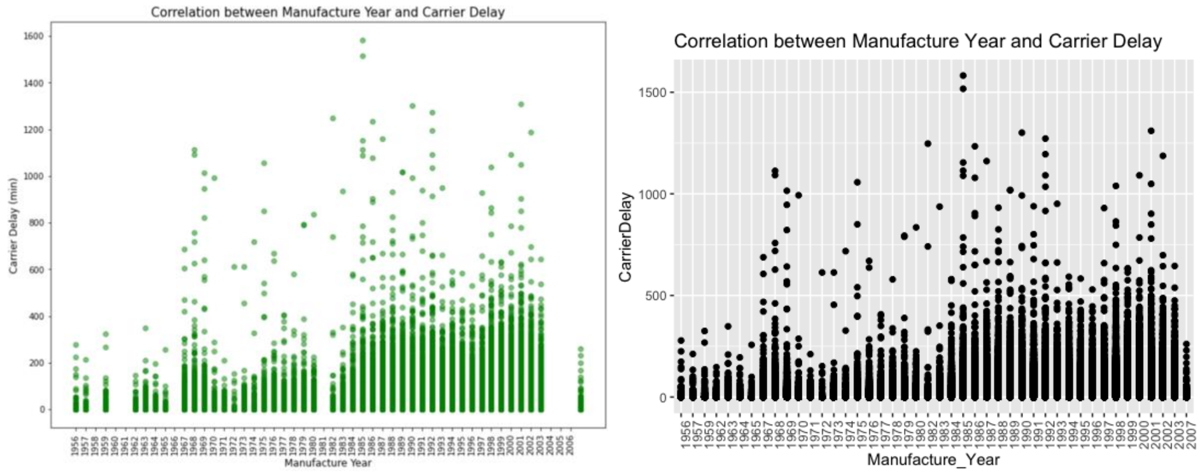


<Figure 10, 11>

As the graph 'Group by Month', the values of May and September are similar in both ways, making it difficult to judge only by looking at the graph. Comparing from the average departure delay first, May was 21.2 and September was 21, which was slightly lower in September. The proportion of delay was not labeled on the graph, so the result of programming was used. It was confirmed that May was delayed to 36.5% and September was delayed to 36%. In conclusion, September was low in both ways, so the best month to minimize the delay in September.

2.3 Question 2: Do older planes suffer more delays?

The second question is to explore whether there is a relationship between the manufacture year of the plane and delay. To distinguish whether the plane is an old plane, the 'year' column in the plane data was added to the flight data. Delays could be used for departure or arrival delays, but the question is to find out the delay due to the condition of the aircraft, so carrier delay in flight data was used. Carrier delay was delayed due to aircraft defects or problems, and it was expected to occur mainly in old aircraft, so it was judged to be more suitable than other variables for this question.



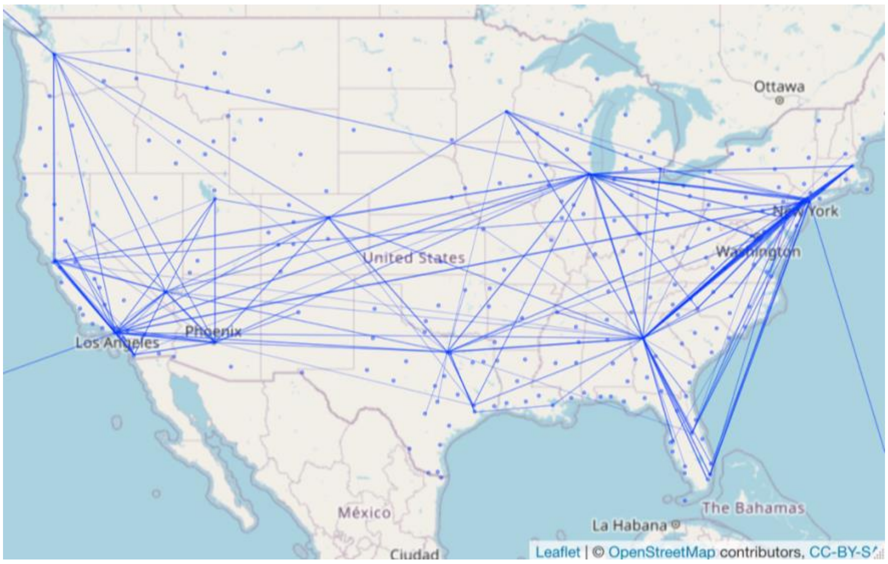
<Figure 12, 13>

The graph above is a scatterplot with the x-axis as the manufacture year and the y-axis as the carrier delay. Recently produced airplanes seem to have more carrier delay, but there is a difference in the number of flights between new and old planes. Thus, it is difficult to say that the x-axis and y-axis are related. In fact, the correlation coefficient between manufacture year and carrier delay is 0.01, so two variables are irrelevant.

2.4 Question 3: How does the number of people flying between different locations change o ver time?

The third question is to analyze population movement patterns over time. First of all, in this analysis, the flow of time is divided into years, flights in 2005, flights in 2006, and flights in 2007. People's movement was expressed as a route by combining the origin and destination. In the visualization, it was difficult to see if all routes are displayed, so only more than 2,000 paths were expressed. And the larger the number of flights, the thicker the line to express the difference in the amount of movement.

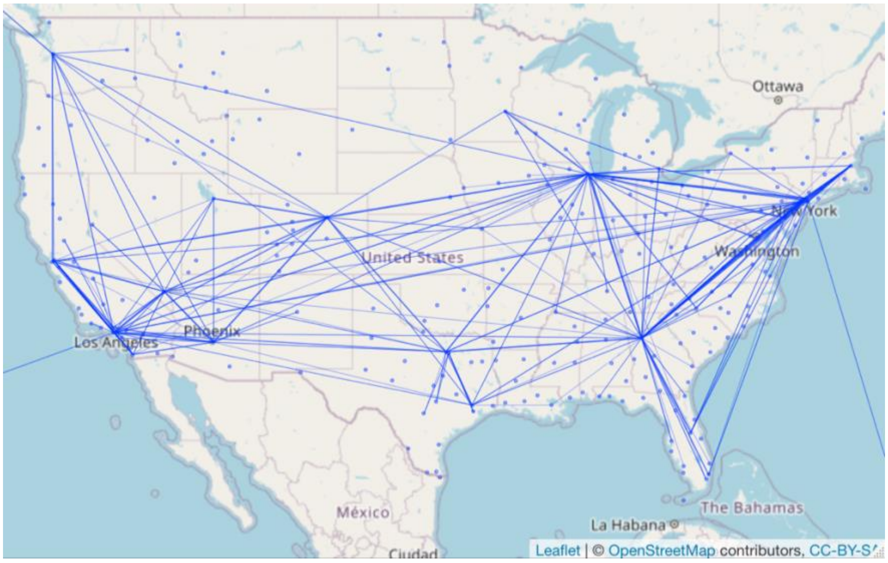
2.4.1 Inflow and Outflow in 2005



<Figure 14>

In 2005, there were flights around certain large cities. Large cities, where lines gather, include LA, New York, Atlanta, and Chicago.

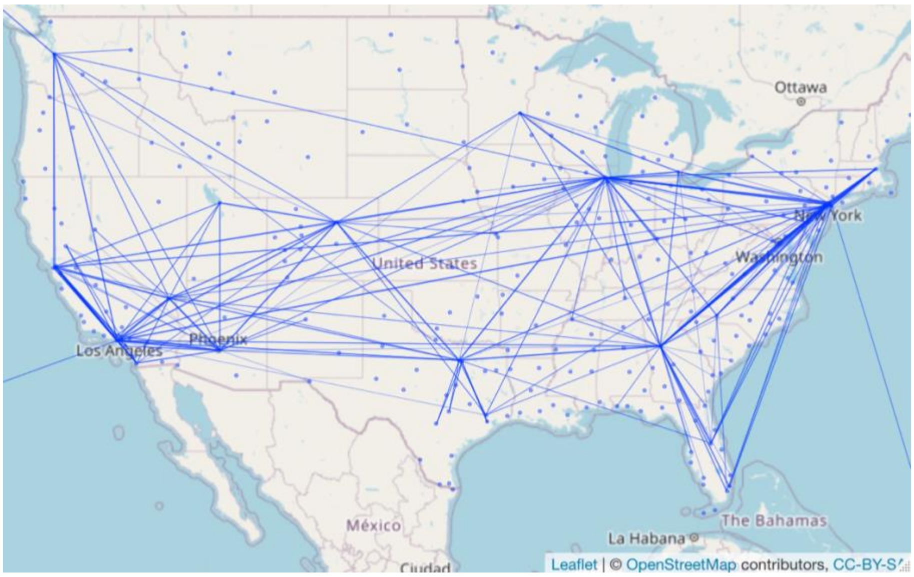
2.4.2 Inflow and Outflow in 2006



<Figure 15>

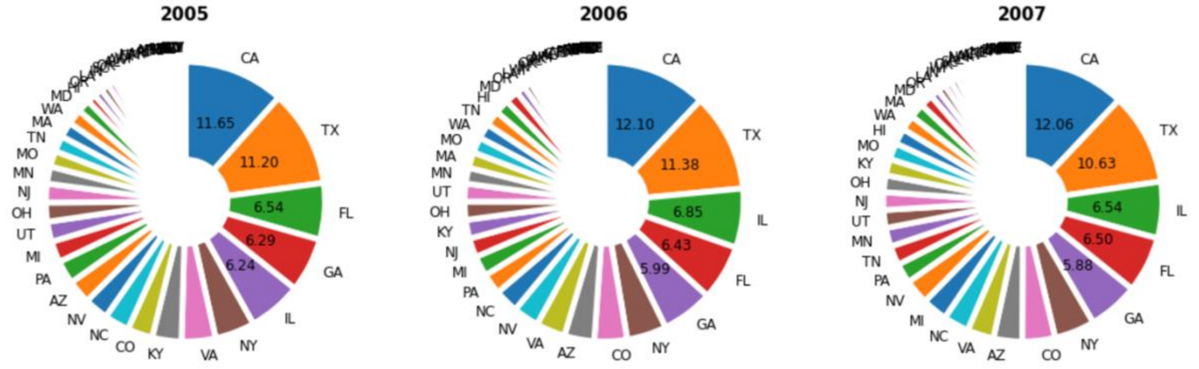
In 2006, the overall thickness of the line became thicker than the previous year, indicating that the number of flights increased. In addition to large cities, which originally had a lot of inflow and outflow, flight volume also increased in some cities, such as Dallas and Denver.

2.4.3 Inflow and Outflow in 2007



<Figure 16>

Flights in 2007 show a similar pattern of flights those in 2006. Some new lines have been added and slightly thickened lines are also visible. It seems to have increased slightly compared to the previous year, but the difference is large compared to 2005.



<Figure 17>

The above three pie charts are charts to check the share of total of inflow and outflow by state by year. First, the total inflow and outflow is 6123653 in 2005, 6390421 in 2006, and 6901955 i n 2007. As compared on the map, the amount of flight increased over time. Looking closely at the proportion by state, California and Texas ranked first and second, respectively, with more than 10% over three years. Illinois, Florida, and Georgia were followed by more than 5%. Looking at the pie charts, which are the results of accurate figures, there was a slight difference from the results analyzed through the map. As the maps, it seemed that New York had much more population movement than Texas, but the actual data results were the opposite. the value of Texas was more than twice that of New York.

2.5 Question 4: Can you detect cascading failures as delays in one airport create delays in others?

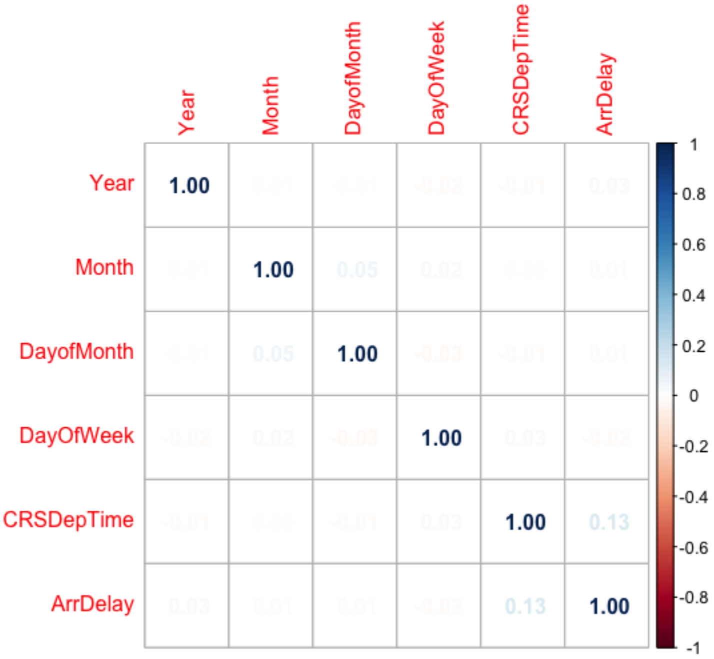
In this analysis, the cascading failures of the fourth question was regarded as delay propagation. Delay propagation means that the arrival delay of one flight affect to the departure of the next flight in same airport, and the arrival is also delayed because of the departure delay. The variable associated with this question in the flight data is Late Aircraft Delay. This means the similar as delay propagation, indicating the arrival delay caused by the late arrival of previous flight.

The process of solving this question is as follows. Firstly, among the flight with late aircraft delay, count if there is a late aircraft delay on the next flight at same airport. Next, divide the count over the total number of flights with late aircraft delay to calculate the probability of the cascading failures.

First, the total number of flights was 9,714,043, of which 2,141,332 flights had a late aircraft delay. And out of 2,141,332 flights, the author ran up to the 400,324th index due to the lack of memory, a total of 125,197 were counted, indicating that there were cascading failures. In conclusion, when late aircraft delay occurs, the probability of cascading failures is 125,197/400,324*100% = 31.27%.

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

Since there is a predictable variable which is delays in question 5, the supervised learning techniques among the three types of machine learning will be used. To predict delays, regression models belonging to supervised learning techniques will be built.



	Year	Month	DayofMonth	DayOfWeek	CRSDepTime	ArrDelay
Year	1.00	0.01	-0.01	-0.02	-0.01	0.03
Month	0.01	1.00	0.05	0.02	0.00	0.01
DayofMonth	-0.01	0.05	1.00	-0.03	-0.01	0.01
DayOfWeek	-0.02	0.02	-0.03	1.00	0.03	-0.02
CRSDepTime	-0.01	0.00	-0.01	0.03	1.00	0.13
ArrDelay	0.03	0.01	0.01	-0.02	0.13	1.00

<Figure 18>

First, an analysis of the correlation between each variable is presented. The correlation c oefficient between all variable and arrival delay is close to zero.

In question 5, three models will be introduced as the regression models for predicting d elays.

First, linear regression is a model that predicts y by creating a line representing the rela tionship between y (dependent variable) and one or more x (independent variable). As a result of learning by dividing data into train sample and test sample, a model's performance value, R squared, was usually around 0.015.

The Random Forest model was used as the second regression model. It uses the base in the Ensemble model as the Decision Tree model. The R squared of the Random Forest was usually 0.03, showing slightly better performance than the Linear Regression model. The last model is Support Vector Regression (SVR), which applies Support Vector Machines (SVMs) to regression. As a result of the test, R squared usually had a value between linear regression and the R squared value of the random forest.

In conclusion, although the results may be different by randomly dividing samples, R squared was mostly higher in the order of linear regression, SVR, and random forest. Accordingly, in part 2, it may be said that the random forest model shows the best performance. However, it cannot be said that the performance is good because all R squared values less than 0.05. There is a need to build more sophisticated models.

3 Conclusion

Figures and results that could not be found with existing raw data were easily found and visualized by programming using Python and R and the desired results were obtained. It was also interesting that even with the same data, different results and visualizations can be shown if approached in different ways. In question 3, the amount of movement was applied to the map using R, and Python briefly showed the share of inflow and outflow by state. Just as maps are easy to understand the overall data at once, and pie charts are easy to figure out the exact values of each data, there are advantages and disadvantages for each visualization. Therefore, before approaching the problem, it is important to understand and interpret what the desired result in question and how the data is organized. Currently, there is a lot of data in our society, and it is still being generated. We can transform the uncertain future into a predictable future if we treat these data meaningfully and build machine learning models applied in real life.

4 List of Figures

5 Reference

1. SaroJanie, S. A. N. (1970, January 1). *Identification of possible reasons that affect departure flight punctuality*. UoM IR. Retrieved from http://dl.lib.uom.lk/handle/123/12183