**UNIVERSITY OF LONDON**



**COURSEWORK REPORT**

**PROGRAMMING FOR**

**DATA SCIENCE**

**Student:** Nguyen Dinh Khoi

**ID:** 200665977

**Pages:** 10/10

**HO CHI MINH CITY, March, 2022**

Table of Contents

[Introduction 3](#_Toc99018987)

[Data Wrangling 3](#_Toc99018988)

[Question 1: When is the best time of day, day of week, and time of year to fly to minimize delays? 4](#_Toc99018989)

[Question 2: Do older planes suffer more delays? 6](#_Toc99018990)

[Question 3: How does the number of people flying between different locations change over time? 7](#_Toc99018991)

[Question 4: Can you detect cascading failures as delays in one airport create delays in others? 9](#_Toc99018992)

[Question 5: Use the available variables to construct a model that predicts delays? 11](#_Toc99018993)

[Conclusions 11](#_Toc99018994)

[References 13](#_Toc99018995)

Coursework Report

# Introduction

The report uses the flights’ data of 1999 and 2000, with more than 11 million records in total. To answer the 5 questions, initially, the data will be checked and applied filter to remove Na values and select the necessary columns. Subsequently, depending on each question, the modified data will be further transformed to analyze and draw conclusions.

# Data Wrangling

The data used consist of information about American flights in 1999 and 2000. The first step was to download the data from the original website to the storage folder. After that, the data was imported to IDE (Rstudio or JupyterNotebook) to analyze. From the raw data of two consecutive years, we merged them into one united group. After checking, there were 11,210,931 records, in which NA values appeared in some columns as shown below.

|  |  |
| --- | --- |
| Table  Description automatically generated |  |

By observing some samples in the data, we can see that the DepTime value is NA whenever the flight is canceled (Cancelled = 1). Additionally, there is a difference between DepTime NA values and ArrTime NA values. This indicates that for some flights, the airplanes took off then diverted (Diverted = 1); therefore, the time they arrived at the scheduled destination was not recorded. After filtering out data that were canceled and diverted, and choosing columns with ArrDelay greater than 0, we examine if there are still NA values or not.

|  |  |
| --- | --- |
| A picture containing text  Description automatically generated | A black screen with white text  Description automatically generated with low confidenceA black screen with white text  Description automatically generated with low confidence |

As displayed in the table, the data were clean and ready to analyze. Moreover, when talking about the delays, what people actually think of is how the actual time they arrive at the destination differs from the scheduled time. Even if there is a delay in departure time, it is still acceptable as long as the flight arrives on time. Therefore, we may focus on the Arrival Delay only.

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

**Best time of day**

|  |  |
| --- | --- |
| Table  Description automatically generated | A black screen with white text  Description automatically generated with low confidence |

The table above demonstrates the average and total delay on various time frames in a day. Each time interval lasts for four hours, starting from 0 AM to 4 AM is the first time slot. Additionally, the delay was calculated based on the arrival time, and only those with late arrival were taken into account to reduce unnecessary information. According to the data, the difference between each time frame was not very significant, with the gap of the least and greatest delay being about 6 minutes. The average delay during 4 AM and 8 AM was lowest at approximately 18.1 minutes, while from 4 PM to 8 PM, the figure was highest at around 34.2 minutes. However, when calculating the total delay, 0 AM to 4 AM had the smallest figure with 3,476,366 minutes, about 44,823,970 minutes less than the most significant delay from 4 PM to 8 PM. Consequently, the best traveling time in a day to minimize the effect of delay is from 0 AM to 12 PM.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart  Description automatically generated |

This bar chart illustrates the average number of minutes of delay during the day. According to the visualized graph, it is clear that the best time to fly is from 4 AM to 8 AM, following is 8 AM to 12 PM and 0 AM to 4 AM. Flights from 4 PM to 8 PM experienced the highest delay; therefore, people should avoid flying in this time frame to minimize the delaying effects.

**Best day of Week**

|  |  |
| --- | --- |
| Table  Description automatically generated | Graphical user interface, text, application  Description automatically generated |

In terms of days of the week, we also compare based on the average minutes of delay. Note that in the table, Monday was assigned to number 1, and Sunday was 7. According to the data, the delay between days in the week did not differ a lot, with the gap being approximately 6 minutes. Additionally, the most favorable day to fly was Tuesday, and the least preferable day was Friday, with minutes of delay being around 24.8 and 30.9, respectively.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | A picture containing icon  Description automatically generated |

The bar graph displayed the average minute's delay in a week. According to the chart, Tuesday had the lowest delay with around 24.83 minutes. The delay increased slightly during the week and reached the highest figure on Friday with 30.88 minutes. Consequently, people should travel on Tuesday to avoid delay.

**Best time of Year**

|  |  |
| --- | --- |
| Table  Description automatically generated with medium confidence | Graphical user interface, text  Description automatically generated with medium confidence |

The gap in average delay between months also shared a similar trend to the two categories above, with the difference being approximately 6 minutes. As presented in the table, October had the lowest delay with around 23.6 minutes, while June obtained the highest minutes with almost 33 minutes. Consequently, people should avoid traveling during mid-year.

|  |  |
| --- | --- |
| Chart, bar chart, histogram  Description automatically generated | Chart, bar chart  Description automatically generated |

According to the bar chart, the delays were most significant in January, December, and especially in June. For other months, we can observe an upward trend in the delay from February to June, while there was a downward slope from July to November.

# Question 2: Do older planes suffer more delays?

|  |  |
| --- | --- |
| Table  Description automatically generated | Graphical user interface, text  Description automatically generated |

To calculate the relationship between the plane’s age and delays, we calculated the average time of delay and the number of delays. As presented in the table, older planes suffered more average time delays, with planes produced in 1956 having around 40.9 minutes of delay. The minutes of delay decreased as the plane was newer. However, the number of delays fluctuated despite the plane’s production year due to the different number of flights. As proof, planes produced in 1989 stood out with 169,409 times of delay, ranked second after 1990 with 198,899 times.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |

The scatter plot showed an inverse relationship between the plane production year and the average minutes of delay. The graph had a gentle slope, indicating the slight difference between different years. From this plot, we can conclude that older planes suffered more delays.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |

The scatter plot demonstrated the relationship between the plane production year and the number of delays. There was a positive correlation, illustrated by the upward line in the graph. Therefore, we can conclude that the delay also depends on the number of flights in a particular year.

# Question 3: How does the number of people flying between different locations change over time?

The data did not record exactly how many people are on a flight. Therefore, to analyze the number of people flying between different locations, we may focus on the number of flights instead.

Since the data is about the flights in US only. Hence, we can count the number of flights in different states of the USA only.

|  |  |
| --- | --- |
| Table  Description automatically generated | A black screen with white text  Description automatically generated with low confidence |

The table displayed data about the number of flights flying to top 6th states in 1999 and 2000. The numbers were ranked in descending order, starting from the state with the most flights visited. California was the most popular traveling place in 1999 with 640,212 flights flying in, and it still maintained its position in 2000 with 678,908 flights. Following California were Texas, Florida, Illinois, Georgia, and Missouri. Interestingly, only California and Florida had an increasing number of flights, while other states experienced a downward trend.

|  |  |
| --- | --- |
| Table  Description automatically generated | Graphical user interface, table  Description automatically generated |

The table provides data about states that had more flights flying in. After calculating, there were 25 states that had more flights, with California and Florida leading the list with 38,696 and 14,390 more flights, respectively.

|  |  |
| --- | --- |
| Table  Description automatically generated | Graphical user interface, table  Description automatically generated with medium confidence |

This table consists of information about states that had fewer flights flying in. Note that the differences between 2000 and 1999 were in absolute values. From the first table of question 3, Missouri (MO) was one of the states with the most flights flying in. However, it was also the state with the fewest flights coming in compared to the previous years. Then, we use the "usmap" package to make the visualization.

|  |  |
| --- | --- |
| A picture containing map  Description automatically generated | A diagram of a house  Description automatically generated with medium confidenceA picture containing text  Description automatically generated |

The maps gave clear illustrations of the visiting trends in 1999 and 2000. There were not any significant changes between the two consecutive years.

|  |  |
| --- | --- |
| A picture containing diagram  Description automatically generated | Map  Description automatically generatedMap  Description automatically generated with low confidence |

The map on the left demonstrates the states with more flights, while the right one shows the states with fewer flights. According to the maps, we can observe that visitors travel to the East and West of America more frequently than the central states.

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

|  |  |
| --- | --- |
| Table  Description automatically generated | Table  Description automatically generated |

According to the number of flights, the aircraft N513 was used the most. Therefore, we may take a look at its routine only.

|  |  |
| --- | --- |
| Table  Description automatically generated  Table  Description automatically generated | A picture containing calendar  Description automatically generated |

Observing from the data above, the departure delay and arrival delay had a strong relationship with each other. Whenever there was a departure delay, arrival delay also has high chance occurring.

|  |  |
| --- | --- |
| Table  Description automatically generated | A screenshot of a computer  Description automatically generated with low confidence |

This table provides data about the number of airports that were affected by the delay in the departure place. Chicago O’Hare International Airport (ORD) had the highest figure with 98 airports being affected, causing 219,545 flights to be delayed. Consequently, there were cascading failures because delays in one airport can cause delays in others.

|  |  |
| --- | --- |
| Diagram  Description automatically generated | A picture containing text, map  Description automatically generated |

The map shows the relationship of delay between each airport. Based on the visualization, there was significant cascading failure as delays in one airport had create severe delays in other airports.

# Question 5: Use the available variables to construct a model that predicts delays?

As mentioned above in the Data Wrangling part, the scope of this report is the Arrival Delay, hence, when we build a model to predict delay, the target value of the models is the Arrival Delay.

Before building the model, data pre-processing is required to clean and prepare necessary data inputs. Observing from the data wrangling process, we can remove all the NA values just by applying the filter. Therefore, in this step, we can also utilize the filter to eliminate null values and select relevant features, which are months, days of the month, days of the week, unique carrier, origin, destination as categorical features, and CSR departure time, CSR arrival time, distance as numerical features. Next, due to the insufficient home computer capacity, we have to take a random sample with the size n = 100000 to run the model. After that, we split the data into two parts, including 70% for training and 30% for testing.

|  |  |
| --- | --- |
|  | Graphical user interface, text, application  Description automatically generated |

The table shows the data about three models, namely the random forest, the linear regression, and the ridge regression. According to the table on the left (which are generated by R), the random forest (ranger) had the lowest mean squared error (mse) with 1109.125, followed by the ridge regression (glmnet) and the linear regression (lm), with 1153.472 and 1153.498, respectively. Consequently, the random forest is the most suitable model to predict delays in R. On the other hands, in Python, the ridge regression has the lowest mse, following by linear regression. Random forest, surprisingly, has the biggest mse with 1207.837. There is also a slight difference when building models in R and Python. The mse of linear and ridge regression in Python are approximately 1179 and 1177, which are nearly 20 units more than in R.

# 

# Conclusions

To reiterate, the best traveling time to minimize the delaying effects are 4 AM to 8 AM, on Tuesday, and during October. Additionally, people should avoid flying from 4 PM to 8 PM, on Friday, and in June. We had also analyzed that there was an inverse relationship between the airplane’s age and the average time of delays. In other words, the older the plane, the more delays it may cause to the flights. However, in terms of the number of flights being delayed, the plane’s age displayed a positive correlation. Aside from that, people’s traveling behaviors had also changed over the years. Although the top 5 most visited states remained the same, there were considerable differences in the number of visitors to each state. People tended to travel to the outer East and West more often than the central parts of America. Utilizing the departure and arrival time, we can also conclude that delays in one airport can cause severe cascading failures. Lastly, among the three different models, the Random Forest was the most suitable one to predict delays due to its small mean squared error in R, while Ridge Regression is the best one in Python with the smallest mean squared error.

# References

- Udacity (2021). How to Write a Professional Data Analysis Report

Available at: <https://career-resource-center.udacity.com/portfolio/data-science-reports>

- David Weedmark (2021). How to Write a Data Report

Available at: <https://smallbusiness.chron.com/write-data-report-61330.html>

- Yugesh Verma (August 29, 2021). Why Data Scaling is important in Machine Learning & How to effectively do it

Available at: <https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/#:~:text=Scaling%20the%20target%20value%20is,learn%20and%20understand%20the%20problem.&text=Scaling%20of%20the%20data%20comes,algorithms%20in%20the%20data%20set>.

- Plotly (2021), Lines on Maps in Python

Available at: <https://plotly.com/python/lines-on-maps/#us-flight-paths-map>

- Zhiyi Guo and Fan Wu (2021). Different Ways of Plotting U.S. Map in R

Available at: <https://jtr13.github.io/cc19/different-ways-of-plotting-u-s-map-in-r.html>