

**ST2195 Programming for Data Science**

**Coursework**

**Student Registration Number: 220458971**

**Name: SEOW XIN YONG**

**Date: 3 April 2025**

**TABLE OF CONTENT**

1. Part 1 3
2. Part 2
   1. Introduction 4
   2. Data Preprocessing 4-5
   3. Modelling process and Interpretation
3. Question a 6-7
4. Question b 7-8
5. Question c 8-9

**Part 1**

For **part 1(a)**, by applying the random walk Metropolis-Hastings algorithm, a histogram and a kernel density plot, along with the f(x) graph were conducted in the same figure, to have a clearer vision on the accuracy of the estimated values of f(x). The value of the parameters, N=10000 and s=1 was set to calculate the estimations for the mean and standard deviation.

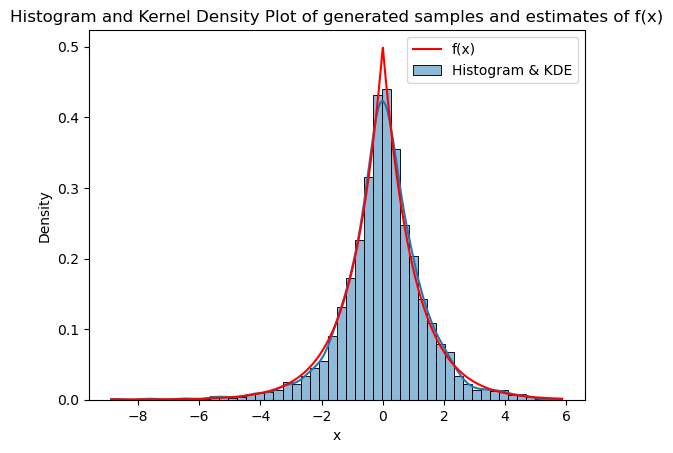
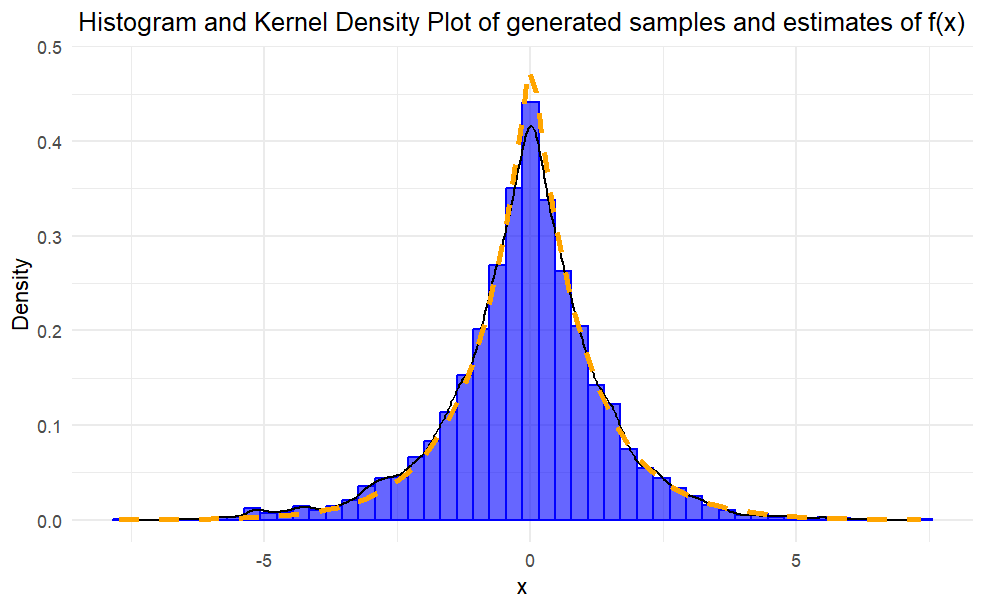
****

Figure 1a: The plot on the left is from R coding and the plot on the right is from Python, both plot are the Histogram and KDE plot.



For **part 1(b)**, under the assumption of the convergence of algorithm, the convergence diagnostic, called value, was used. To determine the strength of the convergence of the algorithm, a plot of values was plotted from across the difference values of standard deviation, s. The initial value for the parameters N=2000 and J=4 was kept constant.

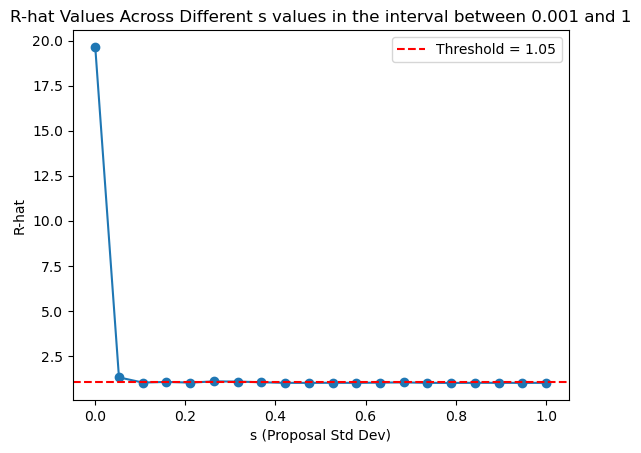
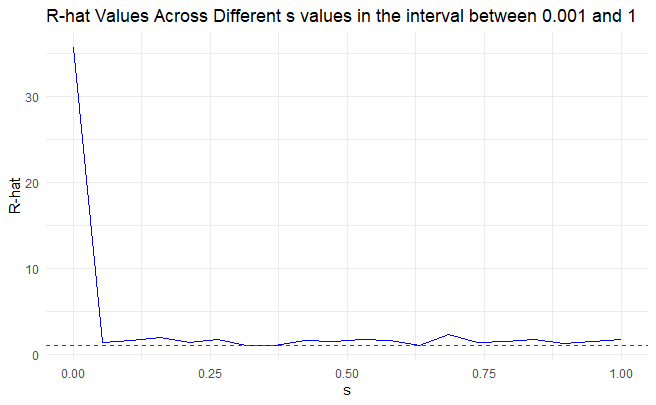
****

Figure 1b: The plot on the left is from R coding and the plot on the right is from Python, both plot are R-hat values across different s values.

****

**Part 2**

**2.1 Introduction**

In this part, the questions were focused on the flights delay in each year, such as determining the best times and days to minimise delays each year, evaluating the effect of plane usage on the delays and the features that affected the delays across the year. The questions were listed as below:

1. What are the best times and days of the week to minimise delays each year?
2. Will older planes suffer more delays?
3. Fit a logistic regression model for the probability of the diverted US flights by using as much features as possible for each year. Visualize the coefficients across the years.

The dataset used in this part was from the Harvard DataVerse (Unav, 2008). To answer the questions above, the consecutive five years chosen were from year 1996 to year 2000. To have a simpler and smoother coding process, the datasets of each year were combined together into one big dataset, named “flights\_raw” by binding the rows.

In both R and Python, the necessary packages were loaded at the beginning, before any coding was done, to ensure the smoothness of the coding process. For R coding, all questions required the packages such as tidyverse, lubridate, and ggplot2. While for Python, the packages such as pandas, numpy, matplotlib.pyplot, and seaborn.

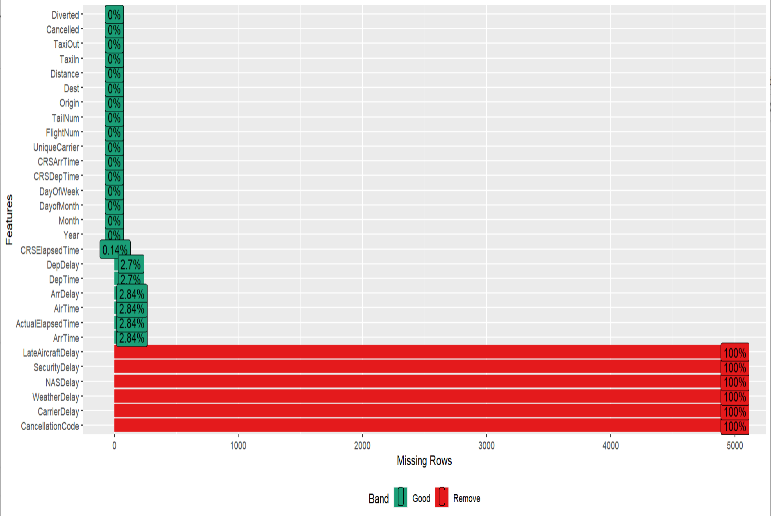
**2.2 Data Preprocessing**

First and foremost, in both R and Python, necessary libraries were loaded to carry out the exploratory data analysis on the big dataset. In R, the extra libraries such as DataExplorer and corrplot were used to visualize the missing values and the correlation matrix. While in Python, extra library like missingno was required to visualize missing values.

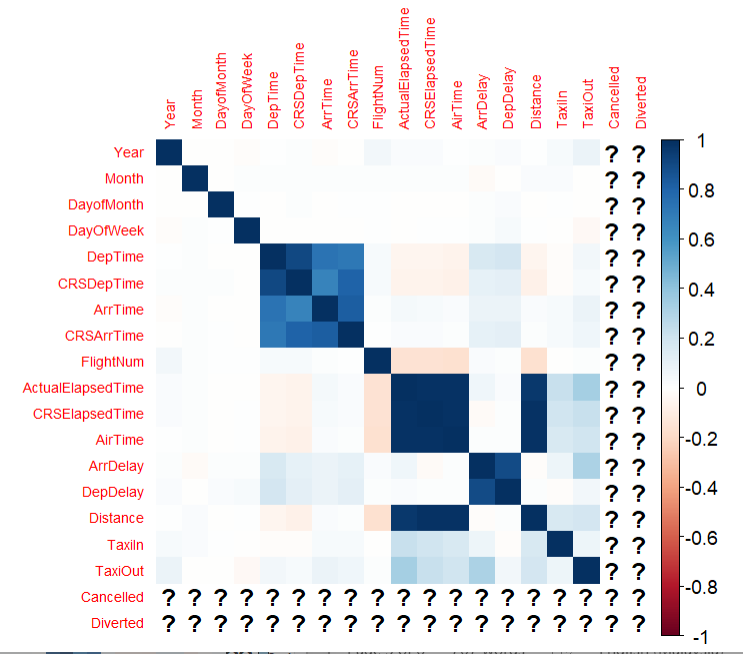
In the following, an exploratory data analysis was made on the big set of data by sampling out the first 5000 rows in the dataset. Since the dataset consists of nearly 120 million records, hence it was easier to conduct the following analysis by reducing the size to 5000 records, which was an acceptable amount. The sample formed was named “flights\_eda”, to have a clearer label of the sample formed.

The structure and the summary of the sample were being studied. From the summary, it showed that there was a sign of delayed departures and arrival in average, where the average departure delay was at 8.8 minutes and arrival delay was at 7.9 minutes.

Remarkably, there existed missing values in some of the columns in the dataset, hence the percentage of the missing values in each column was calculated and plotted in a graph to give a better vision on the missing percentage.

****The figure on the left showed the missing percentage for the last six features were 100%, hence the features were ignored in the analysis. Meanwhile, there existed less than 3% of missing percentage for all the other variables.

Since the missing percentage in the sample was negligible, thus the rows that contained missing values in the whole dataset were omitted and a new dataset was formed, named “cleaned\_flights”.

Moreover, the relationship between the features was investigated by forming the correlation matrix as shown in the figure on the right. In general, the features were having a moderately week relationship with each other, however there was an observably strong relationship between the features that were relevant to the arrival time and departure time. It was suggested to keep an eye on the moderately strong correlation to identify whether there was an effect on the answer to the questions.

Conclusively, the dataset was good enough to be used to answer the question although there existed some missing values and columns. However, it had to be further investigated the effect of the missing values on the answer to the question.

**2.3 Modelling process and Interpretation**

**Question (a)**

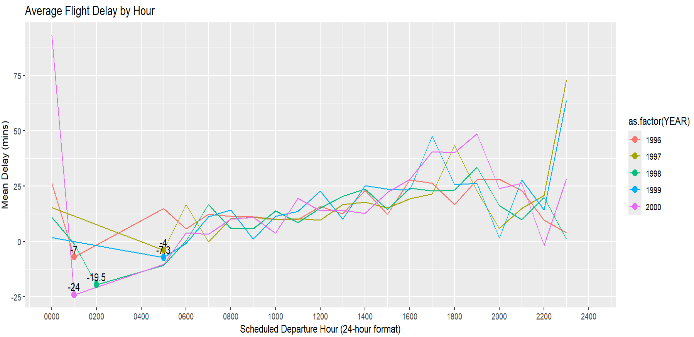
To answer question (a), the necessary libraries were loaded at the very start of the coding process. In both R and Python, there was no extra package needed for this question.

The dataset was mutated by combining the features such as “Year”, “Month”, and “DayofMonth” into one feature, named “DATE”. The feature “DayOfWeek” was also categorized according to different levels and labelled into a new feature “DAY\_OF\_WEEK”, while the feature “Year” remained as “YEAR”.

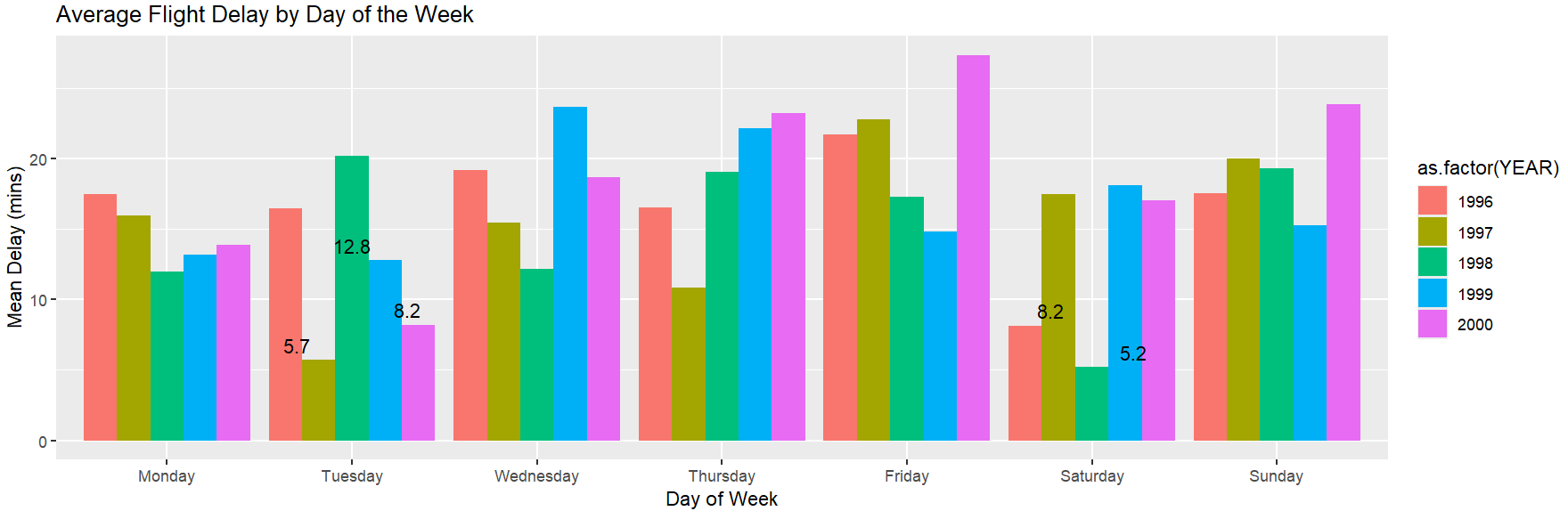
To answer the question, one more variable “TOTAL\_DELAY” was created by summing the values in the columns of “ArrDelay” and “DepDelay”.

The average delayed duration (in minutes) was defined according to hours and days for each year; the definition was used to find the minimum values of the delayed duration by hours and days in each year.

The graph of the mean delayed duration against the hours and days for each year, along with the label of the minimum value of mean delayed duration by hours and days in each year.

The line graph showed the mean delayed duration by hours, the minimum values were plotted for each year respectively.

Hence, it was clearly that the range from 0100 to 0500 hour had the lowest delayed duration in general.



From the bar chart, the best days to minimise flight delays were Tuesday and Saturday. Overall, it was a balanced separation of each year in the best days.

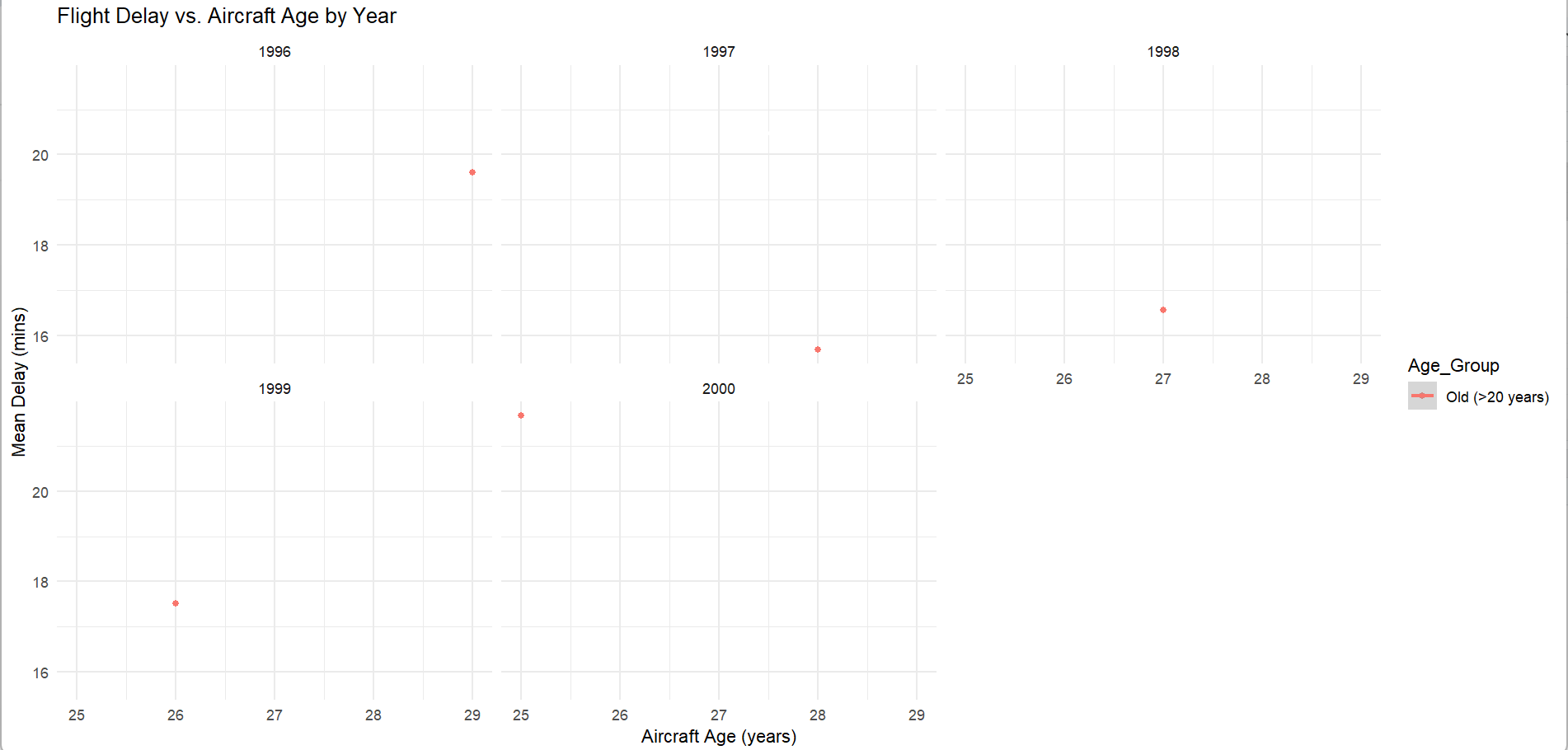
**Question (b)**

For this question, the question required to identify the relationship between the airplane age and the flights delay. Hence, it was essential to have a complete understanding on the condition of the airplanes.

To achieve this purpose, the “plane-data” file was read to get the information about the airplanes and used to determine the mean delayed time of the aged airplane in each year.

Firstly, there was no extra package needed for both R and Python coding. A new dataset was created, called “flights\_2” by joining the variable “talinum” in the “plane-data” file and the variable “TailNum” in the dataset. This was to extract more relevant information into only one single dataset for an easier control on the dataset.

Furthermore, the age of the airplanes was separated into three groups, which were “Young (0-10 years)”, “Medium (11-20 years)”, and “Old (>20 years)”. Thus, there was a better visualization on the effect of the airplane age on the flights delay. Then, the graph of the mean delayed duration for the “Old” age group was plotted for each year.



The figure above showed all the graphs which plotted the mean delayed time of “Old” age group in each year. There was a decreasing trend of the mean flight delayed time along with the decrease in the age of the airplane for each year.

However, it was noticed that there was a remarkably high mean flight delayed time of the “youngest” age for the “Old” age group in year 2000.

Thus, it was concluded that older planes suffered more flight delays, but it should be further studied on a longer timeline to identify the underlying trend.

**Question (c)**

In this question, an extra package like broom was needed to extract the coefficients of variables in a tidy format in R coding, while there was no extra package needed for Python.

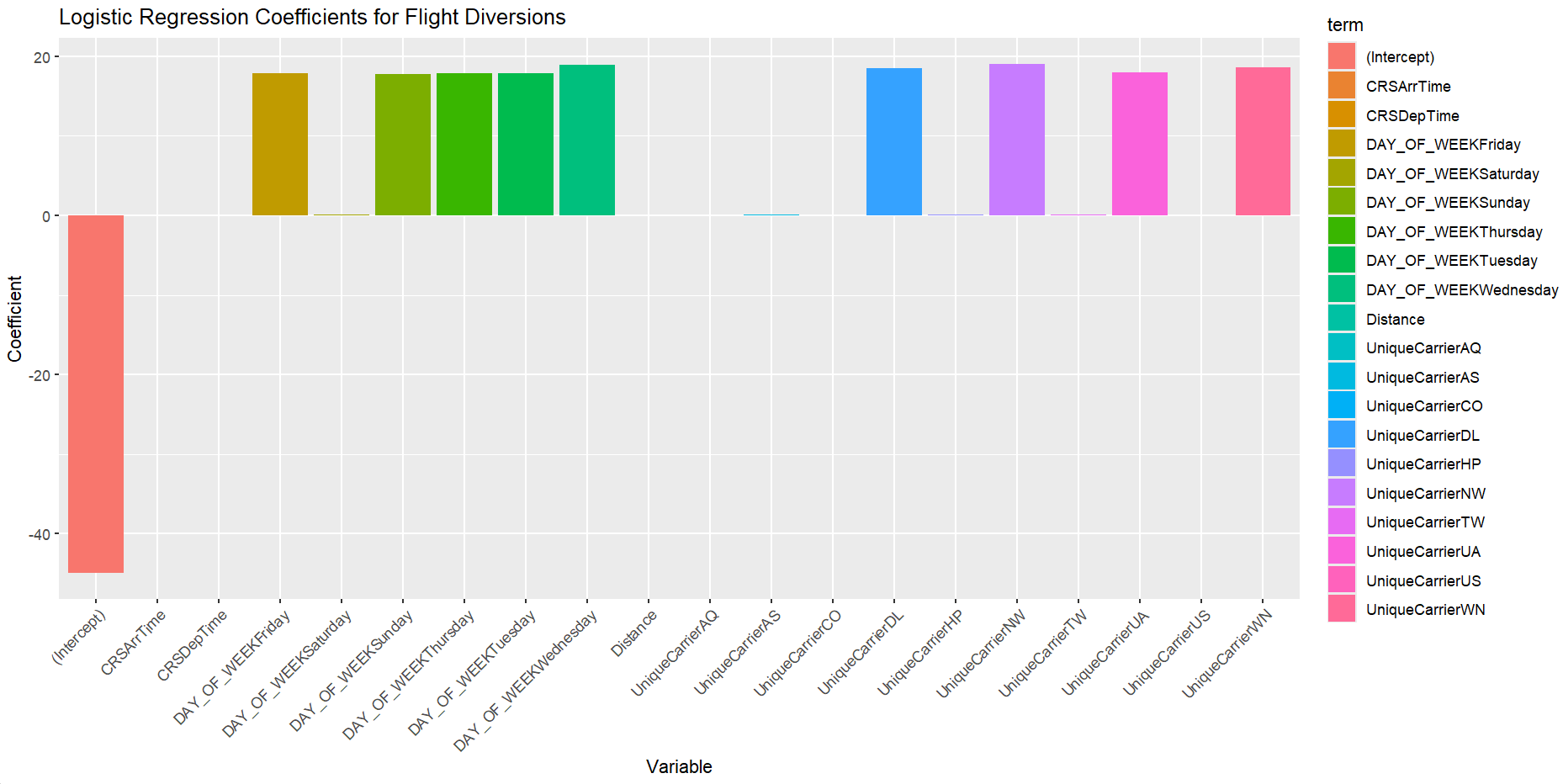
In R coding, to form a logistic regression model for the probability of diverted flights in each year, the variable “Diverted” was converted into a factor variable to group and order the values of data in factors, as this would help in better visualize the coefficients. The R coding was able to handle the factor variables automatically without manually convert it into numeric values.

However, in Python coding, the “Diverted” variable was ensured that it had already converted to a numerical variable manually. This was because logistic regression model in Python had to have a numerical dependant variable when it was dealing with a binary classification. Hence, there was a difference in the coding process for them.

There were a few variables chosen for the interest, such as “CRSDepTime”, “CRSArrTime”, “Distance”, “UniqueCarrier”, and “DAY\_OF\_WEEK”. These variables were chosen due to the previous experience, and the summary of the logistic regression model was formed.

From the summary of the model, there was a significantly low intercept (-44.99) which indicated a large standard error. While “CRSDepTime”, “CRSArrTime” and “Distance” were showing a non-significant effect from the large p-value (p-value >0.001), the carrier effects and the “DAY\_OF\_WEEK” effects were statistically significant, suggesting that both the variables had a stronger relationship with the diverted flights.

The coefficients of each variable were plotted to clearly show the relationship between each of the variable and the diverted flights.



Although some of the variables showed a significant effect on the diverted flights, yet more knowledges should be gained and more investigation was needed to be done for the factors of the diverted flights.

.