UOL Student ID: 230696581

Programming for Data Science [ST2195]

Submission Date: 3 April 2025

Table of Contents

Part 1(a)

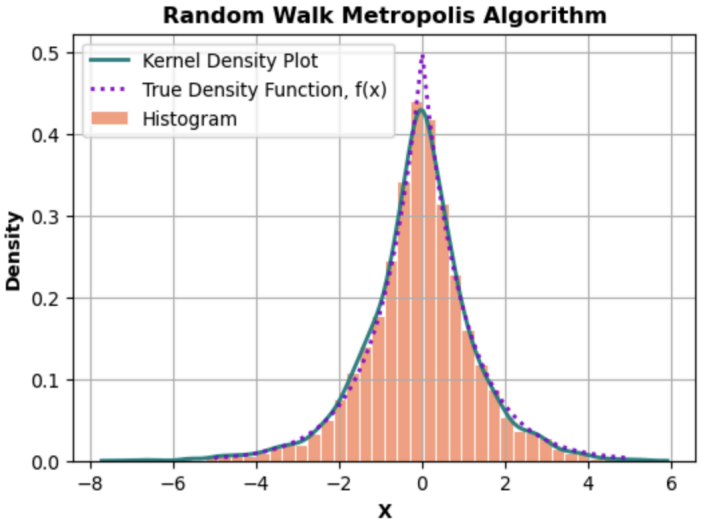
The Random Walk Metropolis (RWM) algorithm is used to generate samples from a complex probability distribution. In this case, we are sampling from a Laplace distribution, defined by:

$$f(x) = \frac{1}{2}exp(-|x|)$$

This is the target distribution, and the goal is to generate samples that match it as closely as possible.

Random Metropolis Algorithm

The algorithm generates a random proposal x^* from a normal distribution centered on the current value x. Then, the proposal is accepted or rejected based on the formula given in the question. If a random number u is less than a certain threshold r, the proposal is accepted; otherwise, the current value remains. This process is repeated for a specified number of iterations.




```
Sample Mean: -0.07881479066943044
Sample Standard Deviation: 1.3291219578604818
```

Figure 1(a): Random Walk Metropolis Algorithm

After running the algorithm, the generated samples are visualized. The histogram shows the the distribution of the samples. The Kernel Density Estimate (KDE) (blue line) provides a smoothed estimate of the sample distribution. The True Density function(dotted purple line) represents the Laplace distribution. From the plot, it is evident that the KDE closely matches the true distribution, confirming that the algorithm effectively generates samples from the Laplace

distribution. The histogram also aligns with the true distribution, further validating the method. The generated samples are also centered around zero, and their spread is close to the expected value, which is consistent with the properties of the Laplace distribution.

Part 1(b)

To evaluate how well the Random Walk Metropolis algorithm converges to the target distribution, we use the R-hat statistic. This metric evaluates the convergence of multiple chains in a Monte Carlo simulation. If multiple independent chains have converged to the same distribution, the R-hat statistic will approach 1. Values significantly higher than 1 indicate that the chains have not converged fully.

R-Hat Calculation

The R hat statistic is computed by comparing the means and variances of several independent chains. If the chains have properly converged, the R-hat statistic should be close to 1.05 or below.

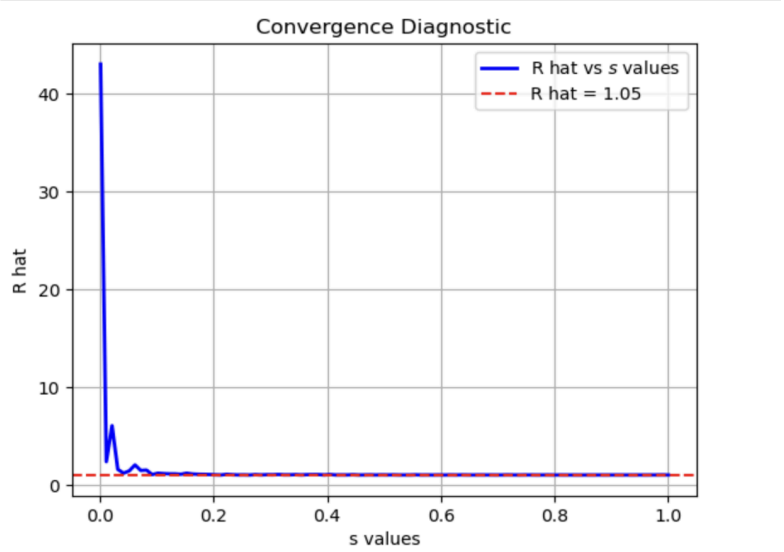


Figure 1(b): Convergence Diagnostic Plot

This plot illustrates the relationship between the R-hat statistic and the standard deviation s of the proposal distribution. For a very small s, the proposal distribution is too narrow, leading to slow exploration and high R hat values, indicating poor convergence. As the s values increase,

the chains explore the target distribution more effectively, and R-hat approaches the convergence threshold of 1.05, indicating improved convergence. The red dashed line at R-hat = 1.05 serves as a benchmark for acceptable convergence.

Part 2(a)

The focus of this section is to identify the best times of day and days of the week to fly to minimise delays. We examine the flight data from 2004 to 2008 to analyse how delays vary depending on the time of day and day of the week.

Data Preparation

We start by importing the packages needed for part 2. Then the dataset is loaded from multiple CSV files for the 5 different years that we have chosen from Harvard Dataverse, and the data for these years is concatenated into a Pandas Dataframe. This dataset contains flight-related information such as scheduled departure time, arrival delays, and the day of the week. Two additional features were created: **TimeofDay**(by categorising the scheduled departure times into morning, afternoon, evening, and night) and **Day**(representing the day of the week).

I filtered the **ArrDelay** column to remove negative delays, as negative values would imply a flight arrived earlier than scheduled, which is irrelevant for this analysis.

Grouping and Aggregation

I grouped the data by **Year**, **TimeofDay** and **Day**, and calculated the average delay for each group to provide insights into the time-of-day and day-of-week delay patterns.

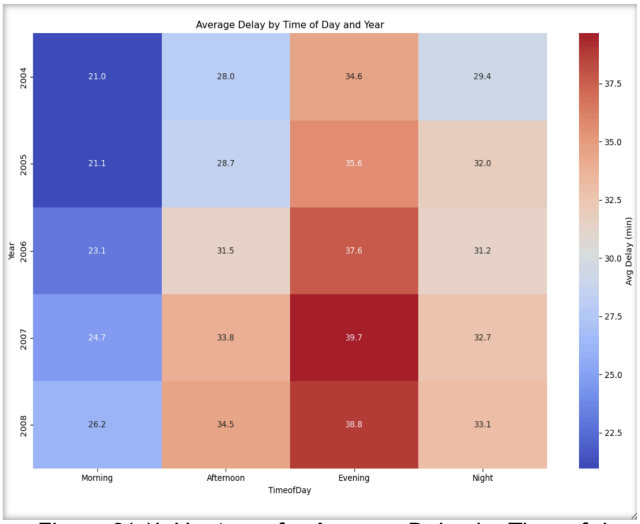


Figure 2(a)i: Heatmap for Average Delay by Time of day

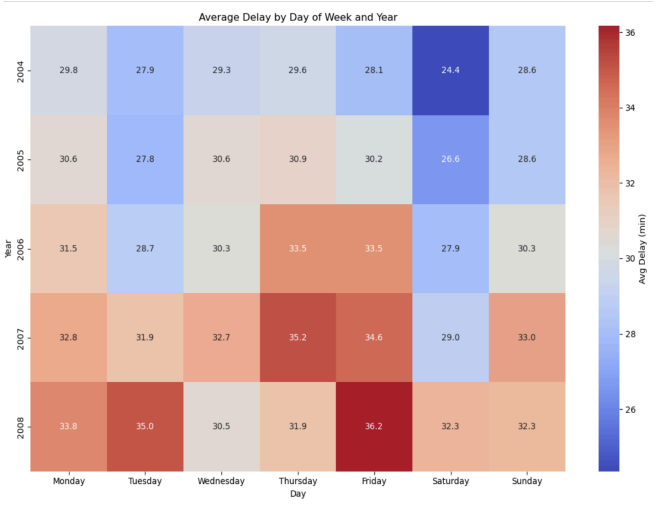


Figure 2(a)ii: Heatmap for Average Delay by Day

I used heatmaps to visualize the average delay by TimeofDay and Day for each year. The heatmaps provide a visual representation of how delays vary by time of day and day of the week. From the heatmap visuals, we can observe that certain times of day and specific days tend to have lower average delays. The first heatmap shows that the delays generally tend to be the lowest in the morning. This can be seen from the lighter blue colour in most years, indicating a more efficient period for flights. On the other hand, the delays in the Evening tend to be the highest with darker colours seen especially for 2007 and 2008. The second heatmap shows that Fridays have a noticeable spike in delays, particularly in the later years(2007 & 2008). Interestingly, weekends tend to have the fewest delays, making weekends an optimal time to fly.

Part 2(b)

This section examines the relationship between the age of aircraft and the delays they experience, using regression analysis to understand if older planes are more likely to face delays.

Data Preparation

I merged the flight data with the planes dataset to get the age of each aircraft. This is calculated by subtracting the aircraft's issue year from the flight year. After this, I filtered out the negative plane ages(planes manufactured after the flight year) as well as negative arrival delays, ensuring the data is clean and accurate.

Regression Analysis

Linear regression was performed for each year with aircraft age as the independent variable and average delay as the dependent variable. The correlation coefficient was computed for each year to assess the strength of the relationship between age and delay. The correlation results are compiled via a scatterplot, and for each year, a statement is generated based on the correlation value, explaining whether there's a strong, moderate, or weak relationship between plane age and delays. These results help to identify whether older planes are indeed causing delays or if other factors might be at play.

- **Strong positive correlation (0.87 to 1)**: Older planes tend to experience significantly more delays.
- **Moderate positive correlation (0.5 to 0.87)**: Older planes tend to have more delays, but the relationship isn't as strong.

- **Weak or no correlation**: There's little to no clear relationship between plane age and delays.

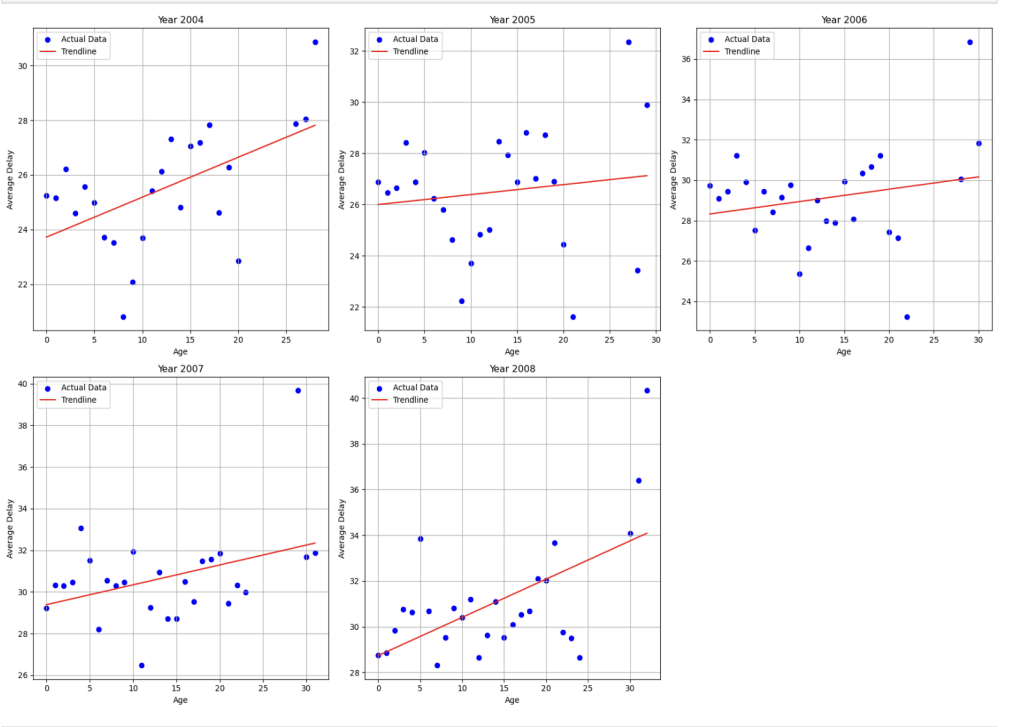


Figure 2(b)i: Scatterplot

```
A correlation coefficient of 0.5455586359652859 indicates a moderate positive correlation between age of the aircraft and average arriv
al delays. Hence, this shows that there are more delays as the plane gets older in year 2004.
A correlation coefficient of 0.1347208776849535 indicates a weak positive correlation between age of the aircraft and average arrival d
elays. Hence, this shows that delays remain relatively consistent even as the plane gets older in year 2005.
A correlation coefficient of 0.21540113155457888 indicates a weak positive correlation between age of the aircraft and average arrival
delays. Hence, this shows that delays remain relatively consistent even as the plane gets older in year 2006.
A correlation coefficient of 0.37478344367879235 indicates a weak positive correlation between age of the aircraft and average arrival
delays. Hence, this shows that delays remain relatively consistent even as the plane gets older in year 2007.
A correlation coefficient of 0.5867910044549984 indicates a moderate positive correlation between age of the aircraft and average arriv
al delays. Hence, this shows that there are more delays as the plane gets older in year 2008.
```

Figure 2(b)ii: Correlation Coefficient Results

The scatterplots of plane age versus average delay for each year include a regression line to visualize trends. The scatterplots indicate a positive correlation between aircraft age and delay in most years, with older planes generally facing higher delays. The correlation is particularly strong in 2004 and 2008, suggesting that as aircraft age, they may become less reliable, leading to more delays. However, the correlation is weaker in the intermediate years (2005–2007), implying that factors other than age may be contributing to delays.

Part 2(c)

For this final section, we build a logistic regression model to predict the likelihood of a flight being diverted based on several factors such as departure and arrival times, carrier, distance, and the location of the airports (latitude and longitude).

Data Preparation

We extracted features such as departure hour, arrival hour, carrier, and airport locations (latitude and longitude). For each year, the departure and arrival times are extracted and converted into departure hour(DepHour) and arrival hour(ArrHour). The airports dataset is merged with the flight data to add latitude and longitude for both the origin and destination airports to fill in the missing values. Categorical variables like origin airport, destination airport and carrier were encoded using LabelEncoder to convert them into numerical representations.

Logistic Regression Model

A logistic regression model was constructed to predict the probability of a flight being diverted. The features included the month, day, time of day, carrier, airport locations, and distance. I also used accuracy, confusion matrix, and classification metrics to evaluate the model. The data is split into training and testing sets, and a logistic regression model is trained using the training set. Predictions are made on the testing set, and the model's performance is evaluated using accuracy and a confusion matrix. Additionally, I visualized the logistic regression coefficients to interpret how each feature influences the likelihood of diversion.

```
Year 2004:
Accuracy: 0.6403748209844766
Confusion Matrix:
[[911646 511420]
 [  1353   1435]]
Year 2005:
Accuracy: 0.6206894378623645
Confusion Matrix:
[[884863 540476]
 [  1225   1556]]
Year 2006:
Accuracy: 0.619914798881254
Confusion Matrix:
[[883581 541583]
 [  1325   1896]]
Year 2007:
Accuracy: 0.6237938929710198
Confusion Matrix:
[[927912 559332]
 [  1457   1942]]
Year 2008:
Accuracy: 0.6069888917722102
Confusion Matrix:
[[289393 187342]
 [   456    653]]
```

Figure 2(c)i: Confusion Matrix

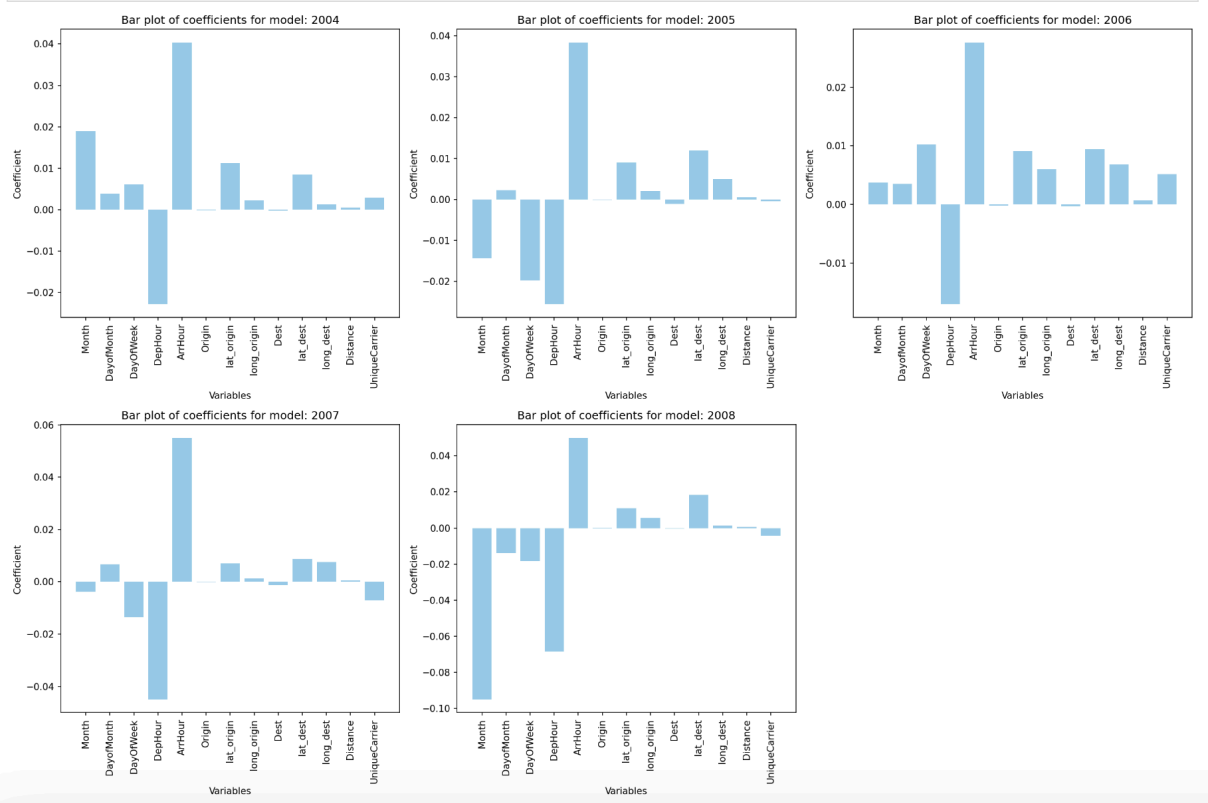


Figure 2(c)ii: Barplot of Coefficients

The accuracy value in the confusion matrix represents the proportion of correct predictions made by the model. It is calculated as the ratio of correct predictions (both true positives and true negatives) to the total predictions. Higher accuracy values indicate that the model is performing better in correctly classifying flights as either diverted or not diverted. The matrix also helps assess performance by identifying the true positives, false positives, true negatives, and false negatives. It is a detailed breakdown of the model's predictions versus the actual outcomes in a 2x2 matrix:

- **True Positives (TP)**: Flights that were diverted and predicted as diverted.
- **False Positives (FP)**: Flights that were not diverted but were predicted to be diverted.
- **True Negatives (TN)**: Flights that were not diverted and predicted as not diverted.
- **False Negatives (FN)**: Flights that were diverted but predicted as not diverted.

This matrix helps assess the performance of the model more granularly, showing how many flights were correctly and incorrectly classified. Accuracy varies slightly across years, with the highest accuracy of 64.04% in 2004 and the lowest of 60.69% in 2008. While these values are

not extremely high, they suggest that the model has some predictive power but may still benefit from further improvements.

The barplot reveals the influence of different features on the probability of diversion; some variables had stronger coefficients, indicating a larger impact on the probability of a diversion. Departure Hour(Dephour) consistently displays a positive coefficient across all years, indicating that flights departing later are more likely to be diverted. Carrier(Arthur) also implies that certain carriers or airlines have a stronger tendency to divert flights, as indicated by positive coefficients in many years. Hence, from these plots, we can conclude that the time of departure(Dephour) is the most influential factor in determining whether a flight will be diverted. Later departure hours significantly increase the likelihood of diversion across all years. Additionally, certain carriers also seem to have a higher tendency to divert flights, which might reflect operational or organizational practices. Day of the week and month have relatively smaller effects, indicating that external factors like airport congestion, weather conditions may play a larger role in determining diversion outcomes.

Conclusion

Each of these sections contributes to a deeper understanding of flight data, helping airlines and passengers make more informed decisions based on the patterns observed in delays and flight diversions. Key findings state that morning flights, particularly on the weekends, have the lowest delays. Additionally, the analysis showed a varied relationship between plane age and delay. Older aircraft tend to experience more delays, but this relationship is not always consistent across all years, suggesting that other factors suggesting that other factors, such as carrier and route conditions, play a role. Lastly, the logistic regression models revealed that factors such as departure time and carrier are strong predictors of whether a flight will be diverted. Later departure times significantly increase the likelihood of diversion.