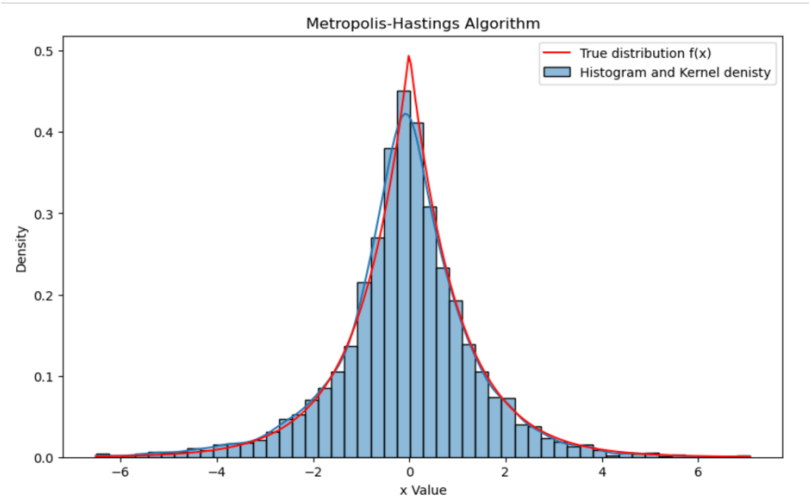
ST2195 Coursework Project

Pang Christopher - 220457882

Table of Contents

1. Metropolis- Hastings algorithm




```
Sample Mean: -0.08202418768851337
Sample Standard Deviation: 1.4247290465611733
```

Figure 1

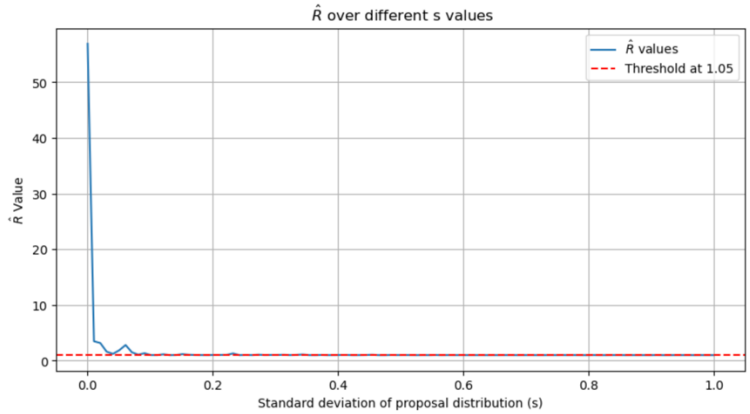


Figure 2

Figure 1 allow us to understand the effectiveness of the Metropolis-Hastings Algorithm from a specific distribution, comparing with the generated sample distribution to a theoretical one. It displays a histogram with Kernel Density Estimate of the sampled given by Metropolis-Hasting Algorithm. The graph determines how closely the sampled distribution matches the true distribution, which serves as a guide to analyse the accuracy of how well the sampled data matches the expected distribution.

Figure 2 shows the method that is used often to generate a sequence of samples that take off a pattern when direct sampling becomes challenging. The graph shows what is the best jump size to use for this method. Small jump size represent that it may take too long to cover the range of the samples while large jump size shows that the graph may move too drastically and not settle into the pattern we want. What we want is an ideal jump size that will get the test to produce the correct pattern efficiently.

2. Methodology and Analysis

2.1 Data Pre-Processing

In this analysis, we are using data from 1991 to 2000, adding it with relevant data from the airports and plane data excels. After importing the data using a loop to read the CSV file for each year from 1991 to 2000, we create SQL tables named “ontime” and “planedata”.

2.2 When is the best time to fly to minimize delay?

To find out the best day to fly, a filtered query was done where departure time was taken where both arrival and departure delays exceeded 15 minutes. The data is thenconverted to integer. This query ensures that we are taking the flight that has delays since the formal definition of delays is greater than 15 minutes. The data with ‘NA’ was then filtered out. Departure times from these delayed flights were then categorized into dinstinct daily segments such as Early Morning (0200-0600), Morning (0600 - 1000) and so forth using the time segment function in the codes. Subsequently, the frequency of delays within each segment was counted and stored into dataframes. These data are then plotted on to a bar chart showing the number of delays per time period from 1991 to 1999. This allows us to visualise the distribution of flight delays with ease. Legends on time segment was plotted on the top left of the graph to display the time period for each segment for clearer interpretation. Example of the barchart in 1991 is shown in figure 3.

The line graph shown in figure 4 displayed the least delayed time segment for each year from 1991 to 1999 which was retrieved from the datas stored in a list in Figure 3’s findings. X-axis displays the years while y-axis displays the number of flight delays. By using the line graph, this allows a year-by-year comparison of flight delays, highlighting the time segments that are consistently having low flight delays. By analysing these trends, we are able to identify the most optimal time of the day to fly.

2.2.1 Analysis of Discovery

We can see that across all the lowest delayed flights displayed, all of the least delays are in the early morning segments. The most flight delayed in the figure is 3046 (Early Morning) in 1996. This implies that flying during the early morning hours could be the most reliable option to reduce delay.

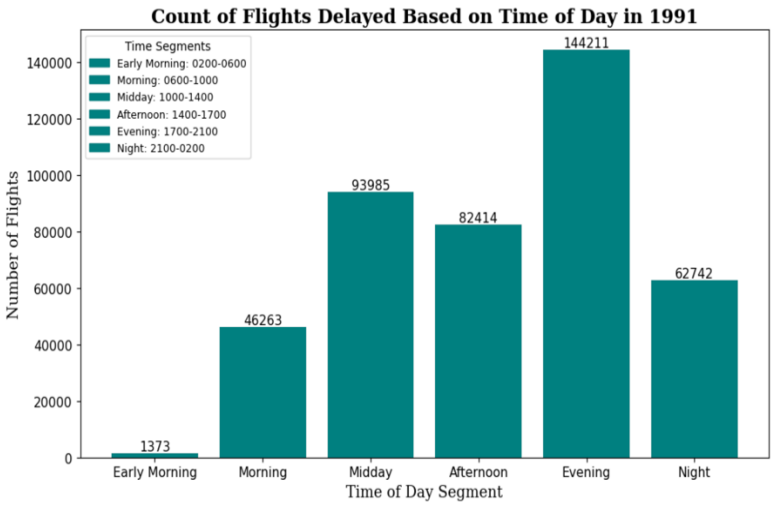


Figure 3

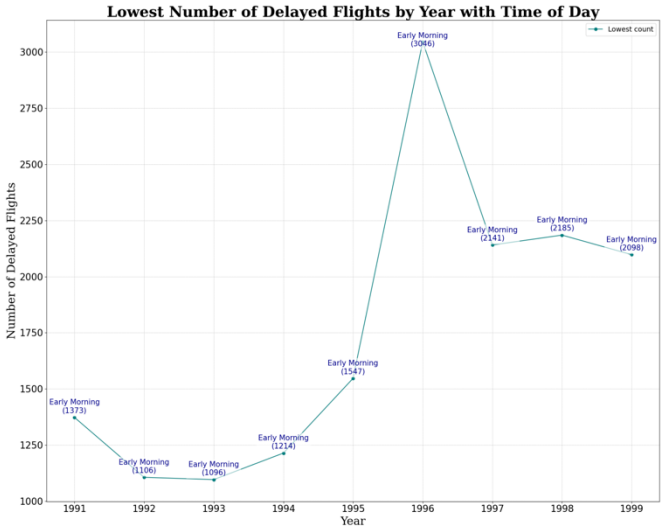


Figure 4

2.3 When is the best day of the week to minimize delay

This analysis is done by first executing a filtered query that only selects the day of the week that experiences both arrival and departure delays greater than 15 minutes. The results are then counted and grouped by the day of the week which are then extracted into a list called 'count'. This is done for every year. The minimum delay days for each year are identified and these data points were stored in a list and used to plot another graph shown in figure 6.

Figure 5 displays a bar-chart that shows the count of flights delays across all the weeks in 1991. The days of the week are plotted on the x-axis while the number of delayed flights are on the y-axis. On each bar, the number of delayed flights for that specific day is displayed. This allows a comparison of flight delays, thus offering a clear metric to determine the most opportune day to travel. As seen in the graph, the lowest delayed flight of the weeks are Fridays with 50, 675 counts.

The line graph in Figure 6 shows the minimum flight delays on the day of a week for each year from 1991 to 1999. X-axis displays the years while y-axis displays the number of flight delays. Each data point on the graph is annotated with both the quantity of delays and the

specific day of week to which it corresponds, This allows us to identify the lowest delayed flights across the 9 years that we are looking at.

2.3.1 Analysis of Discovery

The data revealed in Figure 6 highlights an upward trend where Friday is consistently seen with the fewest flight delays. Notably, only in 1997, Mondays showed the lowest number of delayed flights. Over 9 years, the lowest number of delayed flights are 45,075 flights on a Friday in the year 1992 while the highest number of delayed flights are 88, 825 on a Friday in the year 1996. This pattern indicates that choosing flights on a Friday could enhance the reliability and efficiency of travellers to avoid delays.

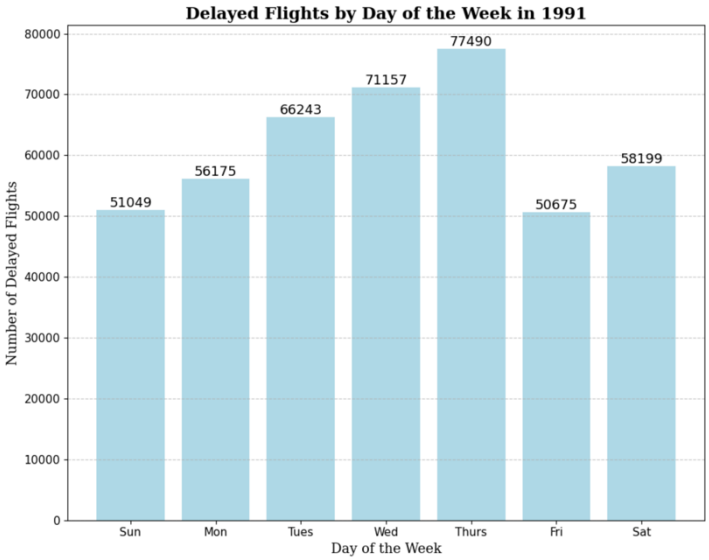


Figure 5

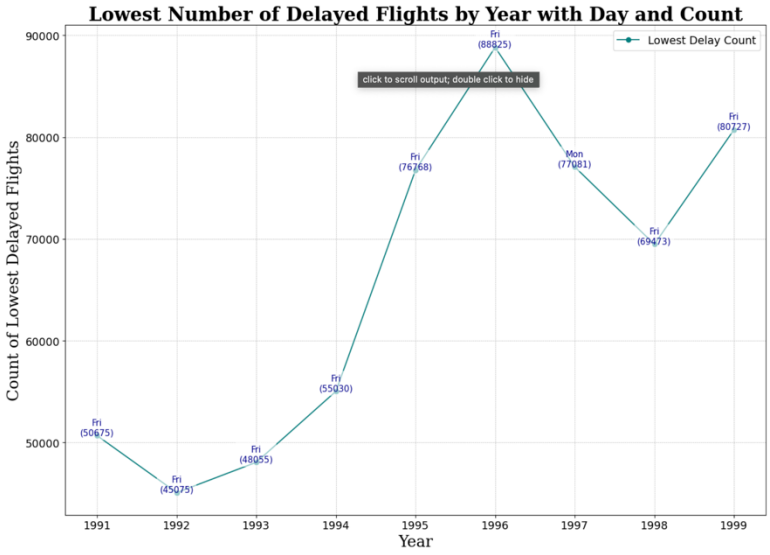


Figure 6

2.4 Do older planes suffer more delay?

At the start of the analysis, a filtered query was performed annually to extract the manufacturing year of the plane where both arrival and departure delays exceeded 15 mins. The 'PlaneData' database was joined with 'ontime' to align plane ages with delay occurrences. The data with null values were then filtered out, leaving us with sufficient data to work with. To calculate the age of each aircraft, the manufacture year of the aircraft is subtracted from the year when the aircraft is still operating.

Once the data is prepared, it is visualized using a histogram for each year, plotting the number of flights experiencing significant delays against the age of planes in service for that year. Figure 7 shows an example of the graph for 1991. From this graph, it is evident that the

age group experiencing the most delays is 1 year old aircraft, which is 135,372 delays. Since each graph displays individually displays data for a single year, the data point representing the highest number of delays along with the corresponding ages from each year will be extracted and the insights will be compiled into a comprehensive graph, figure 8.

2.4.1 Analysis of discovery

After analyzing the annual data, a pattern can be seen indicating a correlation between plane age and flight delays. However, the analysis did not follow the hypothesis that older planes suffer more delays. On the contrary, the year-to-year data showed that most of the years displayed a higher incidence of delays in younger aircraft. To understand the broader trend, it is crucial to look at a consolidated graph in Figure 8. The graph shows the most number of delayed flights throughout the 9 years plotted against the corresponding age of the plane with a correlation coefficient plotted in the graph. A correlation is calculated to justify the relationship between the plane age and number of delays. As seen in the graph, the correlation is -0.89. This shows that there is a strong negative correlation, indicating that as the age of the planes increases, the number of delays actually decreases significantly. It suggests that newer planes, despite their advanced technology and efficiency, might be more prone to delays due to factors such as intensive usage of the aircraft or more stringent testing protocols.

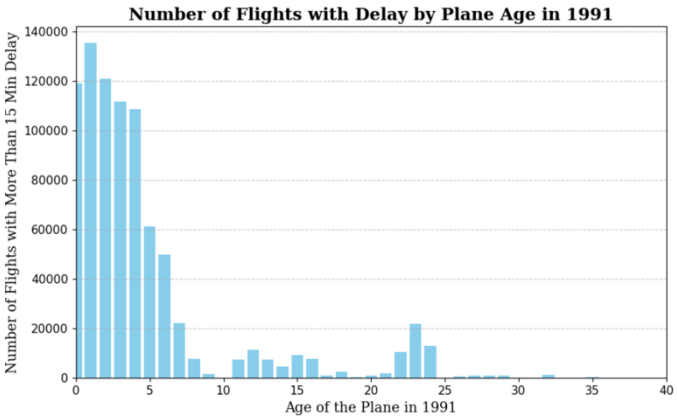


Figure 7

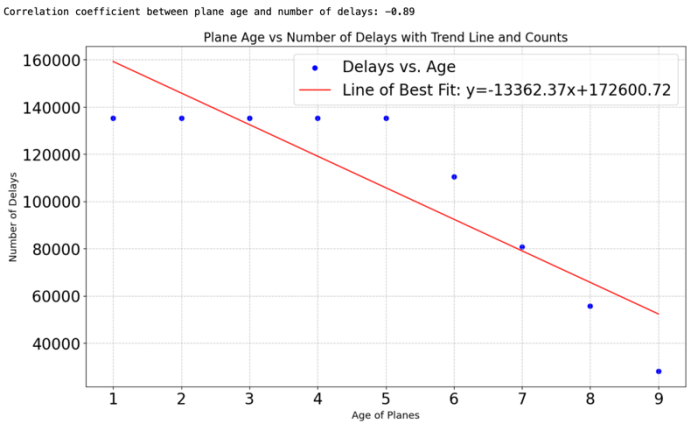


Figure 8

2.5 Logistic regression

Figure 9 shows logistic regression used to analyze trends and calculates the probability of flight diversion throughout the years based on departure time, arrival time and distance. The data preparation for the graph was done by executing a loop over the years 1991 to 1999 where data is filtered and the missing values in the feature are replaced with the mean of each column. Further more, the data is split into training and testing sets with an 80/20 ratio. Some

important findings are that the trends in coefficient over the years can indicate changing patterns in flight operations or policies. For example, if the coefficient for departure time becomes more positive over the years, it could suggest that later departures have more likely to be diverted.

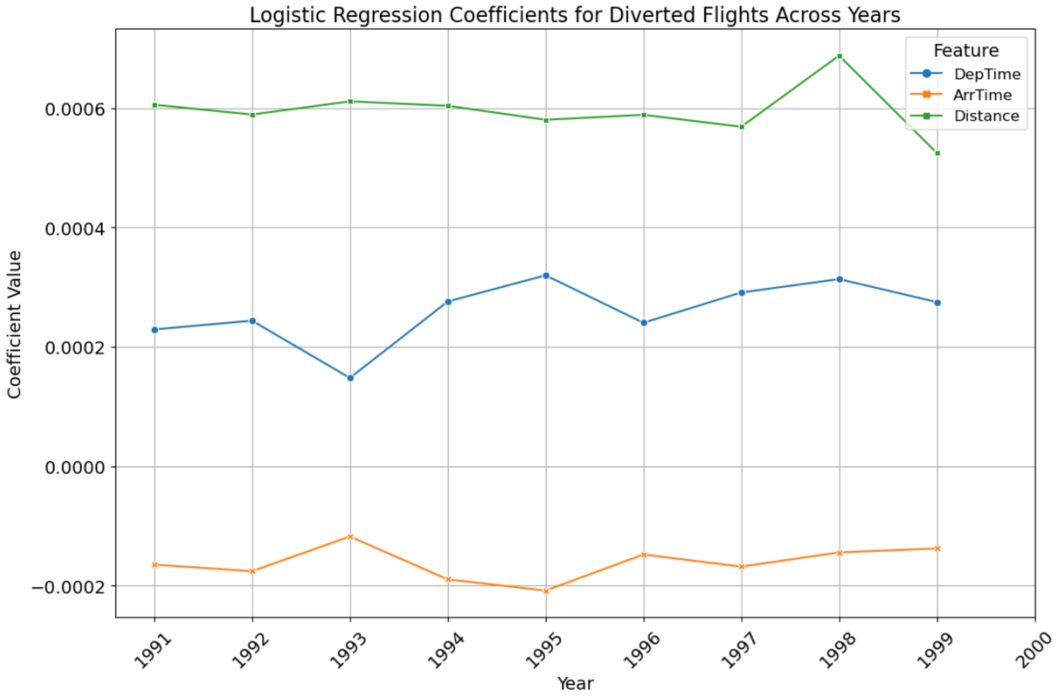


Figure 9