Data 5300 Project 2

Ngunjiri, Justus JG

November 30, 2022

Executive Summary

There are various factors that can be significant when it comes to studying flight operations. Therefore, investigating and understanding the significance based on certain attributes becomes essential. In this study, the significance of two main variables for United Airlines was investigated. These variables were gain and gain per hour. Gain defined how quicker the flight ended up being than it was planned while gain per hour was just the gain divided by the airtime in hours. The investigation for gain was conducted for various flight groups, starting with the study of the difference of gain between flights that had a departure delay that was greater than zero (late) and those that were not late, flights that were delayed more than 30 minutes (very late) versus those that were not. Afterward, the gain per hour was calculated by dividing the gain by the airtime in hours. The difference in the derived gain per hour for flights was then investigated for late versus not late group, very late versus not very groups and lastly for Long versus short flights. Long flights were defined as flights with airtime of more than 3 hours, while short was for flights with airtime of less than 3 hours. Lastly, we also summarized the top five destinations for United Airlines.

To determine whether there was a difference between groups, a hypothesis test was used to study the gain and gain per hour variables for flights with delays that were greater than zero (late) and those that were not late, flights that were delayed more than 30 minutes (very late) versus those that were not. Several options for hypothesis testing existed including the use of a standard *Ttest* for hypothesis as well as standard bootstrap test method. The choice between these two methods was determined by two main attributes, the first being the samples size and the secondly, the skewness in the distribution of the variables across groups under consideration. We determined there were 57782 observations for UA flights. Because the data size had a large sample size for every group ($n > 30$), we determined that a standard T test would yield good results. The specific splits in groups was 7550 UA flights for very late while 50232 were not very late. Likewise, 27125 UA flights were late while 30657 were not. In addition, our study was based on the difference of two means and thus from theory, the skewness would cancel out. Use of bootstrap test that caters for skewness was thus not necessary.

After implementing the hypothesis test in *Rprogramming*, the various probability of occurrences when determined and compared with a significance value of 0.05. The mean gain

for late group was 7.56 minutes and not late was 9.46 resulting in a mean gain difference of 1.89 minutes. Similarly, the mean gain for the very late group was 6.86 minutes and not very late group was 8.70 minutes resulting in a mean gain difference of 1.84 minutes. The probability values from the hypothesis using t-test values were computed as $2.2e-16$ for late vs not late group, $3.215e-10$ for very late group versus not very late groups. Since these p values were very small, the null hypothesis was rejected. As a result, we determined that there were differences in gain for the respective cases.

The mean gain per hour for the late group was calculated as 3.14, for not late flights was 3.99, resulting in a gain per hour difference of 0.89. Similarly, the mean gain per hour for the very late group was 3.69, not very late group was 3.69 resulting in a mean gain per hour difference of 0.63. The mean gain per hour for the long flights was 1.86, short flights groups was 5.86 resulting in a mean gain per hour difference of 0.63. The gain per hour probability values from the hypothesis using t test values were computed as $2.2e-16$ for late vs not late group, $1.372e-06$ for very late group versus not very late groups. Since these p values were very small, the null hypothesis was rejected. As a result, we determined that there were differences in gain per hour for the respective cases. For Long and short flights differences study for gain per hour, the confidence interval was determined to be (-4.145229,-3.848538). Because this interval did not include zero, we concluded that the mean gain per hour for long and short flights is not the same. Therefore, the general conclusion was that UA would have to investigate all the parameters further to improve efficiency and customer satisfaction.

Contents

List of Figures

List of Tables

1 Introduction

This report serves as a continuation of a previous project titled Data 5300 Project 1 that evaluated various relationships to the departure delay variable in the $nyc13flights$ data set for United Airlines(UA). In this study, we report on findings from a study conducted on two additional variables namely gain and gain per hour. These variables were not part of the original data set, however, they were derived from original variables contained in the original data set. In addition, the study techniques employed in this project were different-hypothesis testing and confidence intervals were used as opposed to permutation tests.

In this study, gain defines how quicker the flight ended up being than the was planned, which is calculated by subtracting the arrival delay from the departure delay. Gain per hour is then calculated by dividing gain by airtime in hours. Having defined our variables of interest, we then investigate 4 main relationships. Firstly, we investigate whether the average gain differs for flights that departed late versus those that did not. We also investigate the same for flights that departed more than 30 minutes late versus those that did not. Secondly, we conduct an exploratory data analysis (EDA), to determine the five most common destination airports for United Airlines flights from New YorkCity. We then describe the average gain and the corresponding distribution for the top five destinations. Thirdly, we investigate whether the average gain per hour differs for flights that departed late versus those that did not and similarly, for flights that departed more than 30 minutes late versus those ones that did not. Lastly, we investigate whether there is a difference in the gain per hour for long and short flights. A long flight is defined as a flight with a duration of more than three hours whereas a flight of three hours or lower is considered short.

Two techniques for statistical significance testing namely Confidence interval and Hypothesis testing were applied for investigation. The confidence intervals technique evaluates whether the resulting calculation contains zero in the interval or not. If the interval contains zero, we conclude that there is no evidence that there is a difference in the parameter(s) we are testing. On other hand, if the interval does not include zero, we conclude that there is a difference at a specified confidence interval for the variables being tested. For hypothesis testing, we set a significance level and then propose a hypothesis to evaluate whether it is true based on our desired probability. This significance was set to 0.05. Therefore, if we evaluate the parameters and the probability of that occurring happens to be less than 0.05, we conclude that there is evidence of a difference. Likewise, when the probability is more than 0.05, we conclude there is no evidence of a difference and subsequently, we failed to reject the null hypothesis. These tests were conducted using *Rprogramming* statistical software.

Rprogramming test statistic was used to implement our techniques in order to simplify our investigation in both techniques. However, from theory, skewness and sample size affects the choice of our test statistic. Determining skewness helped us to identify the choice of the test method. For instance, it is known from theory that the T-test would not perform well on skewed data involving one sample but would perform well in a test for difference in two means. For this to be an accurate implementation, the number of observations must be greater than 30. This requirement was met in our study because there were sufficient observations in either category.

2 Study of gain and gain per hour for UA data

Before applying a test statistic for each variable, the necessary visualizations were carried out to determine key attributes that affect the performance of the test statistic. These distributions were determined from the resulting EDA. The nature of this study involved comparing gain and gain per hour based on categories discussed in the introduction section. For instance, gain for departure delay less than zero and those greater than zero, long versus short flights, departure delays greater than 30 versus less than 30, etc. Due to the comparative nature of the study, suitable visualizations that were utilized included box plots, histograms, and qqnorm plots to visualize and understand the data as well as formulation of the hypothesis in order to form the basis of criticism. The null hypothesis between the statistics in question were then addressed. Comparative box plots helped us to explore the relationship of the categories under consideration by visualizing their median, minimum, maximum, upper and lower quartiles. In summary, the spread of our data based on category. In addition, the box plots allowed us to visualize any potential outliers.

The use of histogram and qqnorm for visualization allowed us to investigate the distribution of our data and in turn, identify potential skeweness in data. Skewness in this case refers to whether we have extended observations on either of the tails of the histogram. Determining skewness helped us to identify the choice of test method.

2.1 Gain for UA flights with departure delay

As discussed earlier gain was calculated by subtracting the arrival delay from the departure delay. Figure 1 shows the resulting distribution of gain. We observe that the distribution is left-skewed and uni-modal.

2.1.1 Average gain for flights that departed late versus not late

Having noted that the overall distribution of gain is left-skewed, we reviewed the distribution of gain based on the categories, departed late vs not late. A quick glimpse at the data indicated that 27125 UA flights were late while 30657 were not. In addition, we were also interested in understanding whether any of the observation followed a normal distribution. To achieve this, we plotted a qqnorm plot as shown on Figure 2.

From both the histogram and QQ plots, it is evident that the nature of gain being left-skewed extends to both categories. In a similar manner to how we interpret the tails of a histogram, the qqline helps us to visualize the observations that fall below the fitted line that assumes a normal distribution. We observe from the qqplot, that points near the origin of our graphs fall below the fitted normal line which means that our data is left-skewed. A further EDA was carried out to visualize the spread of gain depending on the group as shown on Figure 3. We observe that flights that departed early or on time have a slightly higher gain. The mean gain for $Departed-on-time$ group was 9.46 minutes while the gain for the

.

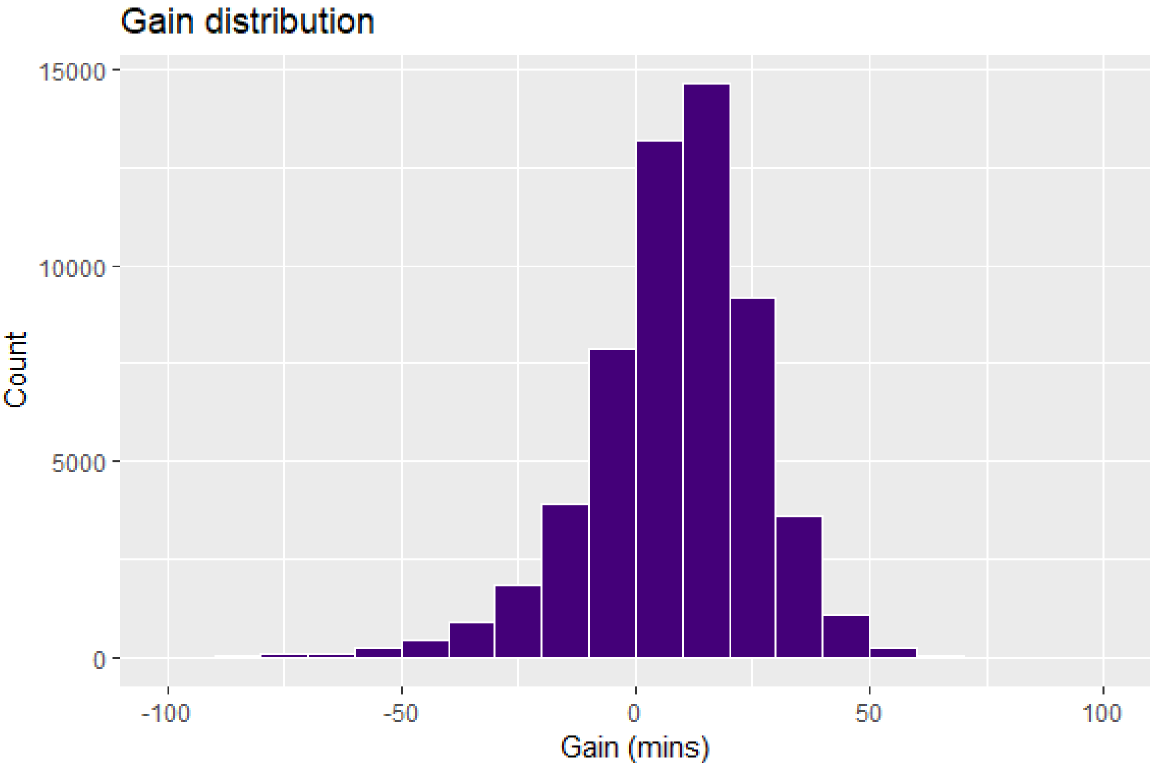

Figure 1: Histogram depicting the distribution of gain.

$Departed - late$ group was 7.56 minutes. The resulting difference in average gain was 1.89 minutes.

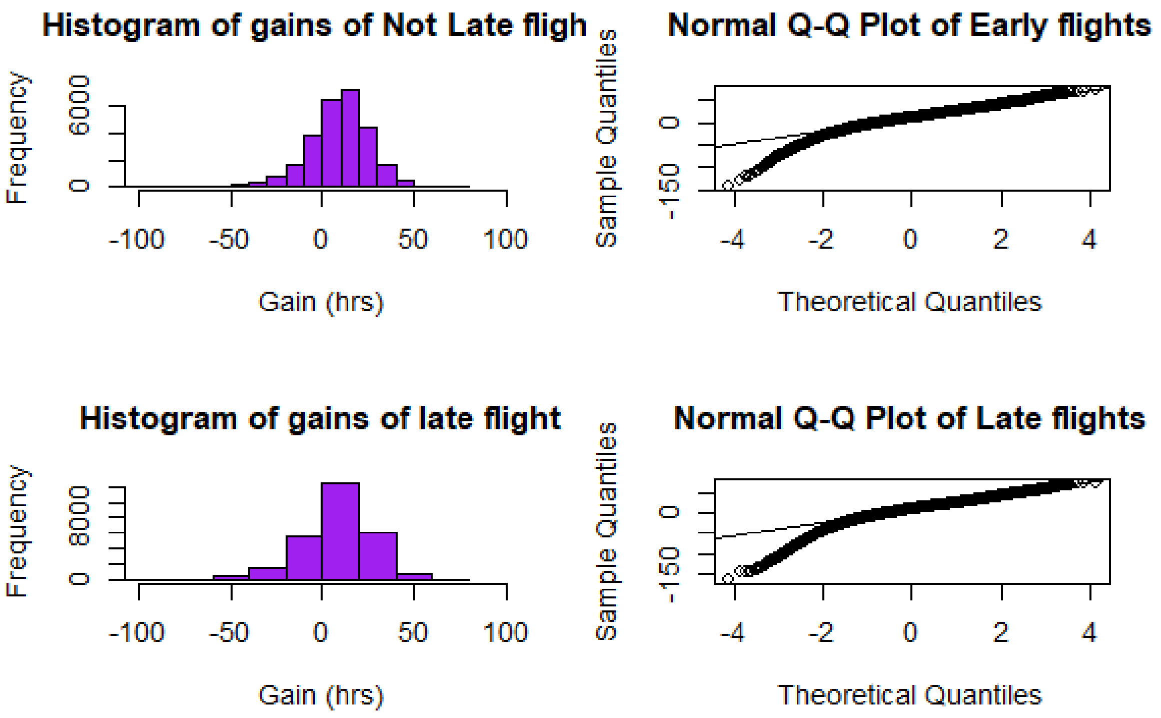

Figure 2: Histogram and qqnorm plots depicting the distribution of gain based on late vs not late categories.

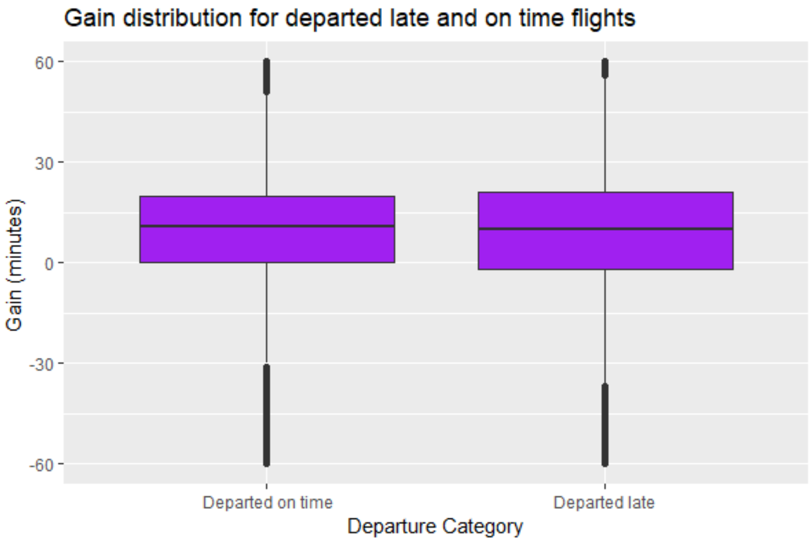

Figure 3: Box plot depicting the spread of gain in the departure delay late classification.

After determining that the observed difference was around 2 minutes, we decided to test the probability of that happening. This was achieved by formulating two conjectures as follows.

- Null hypothesis ,H_o, mean gain difference($\mu_{ontime} - \mu_{late}$) is zero. That is, the mean gain of the departed on time and departed late groups are the same.

- Alternative hypothesis ,H_a, mean gain difference($\mu_{ontime} - \mu_{late}$) is not zero. That is, the mean gain of the departed on time and departed late groups are not the same.

This test was then implemented using the *t.test* function in *Rprogramming*. The resulting probability of this test was $2.2e - 16$. Since we had assumed a significance of 0.05, we rejected the null hypothesis. Therefore, there is evidence that the mean gain between flights that departed on time is different from the gain that departed late.

2.1.2 Average gain for flights that departed more than 30 minutes late versus those that did not

Similar to section 2.1.1, we review the distribution of gain for flights that departed more than 30 minutes late vs those that did not. In this study, flights that departed more than 30 minutes late are labeled $very - late$ while those that did not are identified as $very - late$. Examining the data showed that 7550 UA flights were very late while 50232 were in not very late. We then reviewed the distribution of gain based on the categories, departed very late vs not very late. In addition, we were also interested in understanding whether any of the observations followed a normal distribution. To achieve this, we plotted a qqnorm plot as shown on Figure 4.

As was the case with $late$ vs $on - time$ groups from section 2.1.1, the histogram and QQ plots suggest that the nature of gain is left-skewed for both groups. A further EDA was carried out to visualize the spread of gain depending on the group as shown on Figure 5. We observe that flights that did not depart very late have a slightly higher gain. The mean gain for $not - very - late$ group was 8.70 minutes while the gain for the $Departed - late$ group was 6.86 minutes. The resulting difference in average gain was 1.84 minutes. We also observe that the very late group has a higher spread.

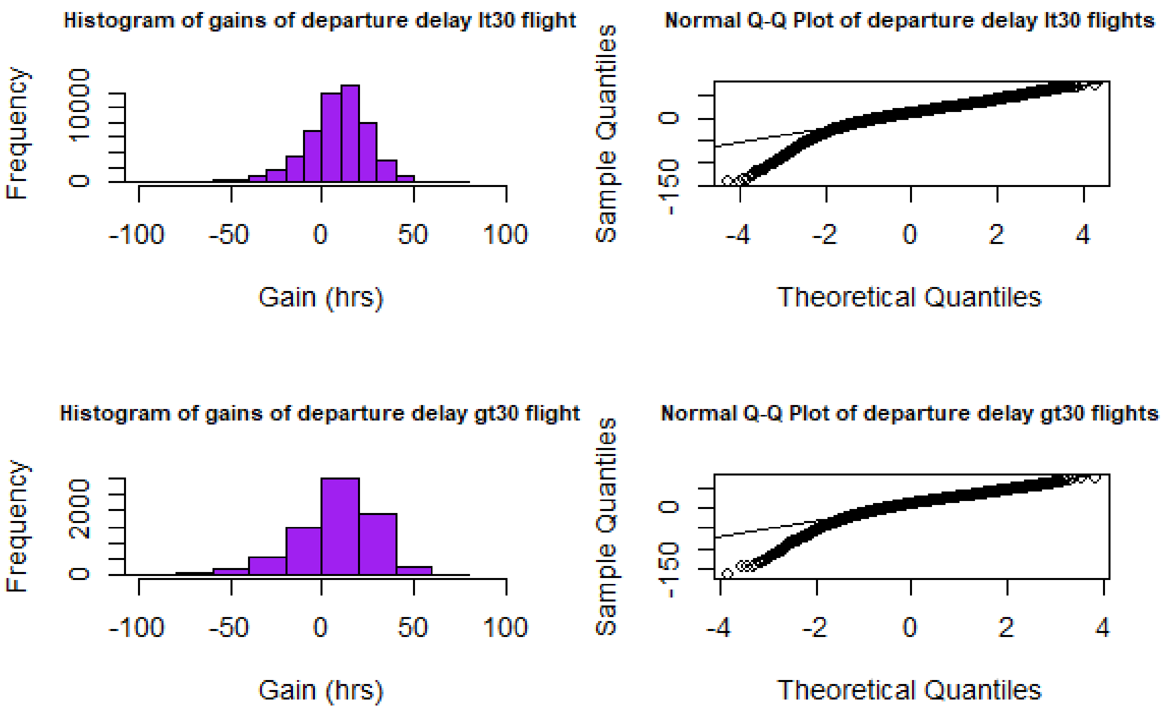

Figure 4: Histogram and qqnorm plots depicting the distribution of gain based on late vs not late categories.

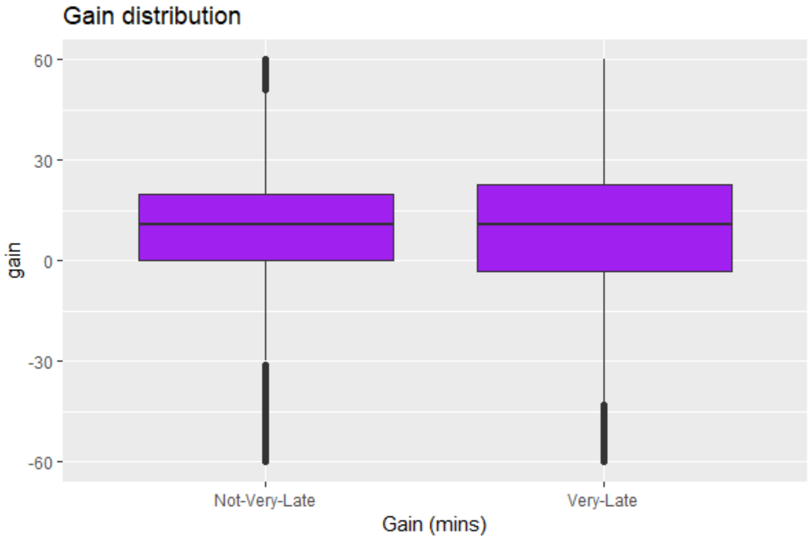

Figure 5: Box plot depicting the spread of gain in the departure delay late classification.

After determining that the observed difference was around 2 minutes, we decided to test the probability of that happening. This was achieved by formulating two conjectures as follows.

- Null hypothesis ,H_o, mean gain difference($\mu_{notverylate} - \mu_{verylate}$) is zero. That is, the mean gain of flights that did not depart very late and the very late groups are the same.

- Alternative hypothesis ,H_a, mean gain difference($\mu_{notverylate} - \mu_{verylate}$) is not zero. that is, the mean gain of flights that did not depart very late and the very late groups are not the same.

After testing the hypothesis using *t.test* in *Rprogramming*, the resulting probability was $3.215e - 10$. Since we had assumed a significance of 0.05, we rejected the null hypothesis. Therefore, there is evidence that the mean gain between flights that departed very late (more than 30 minutes late) is different from the mean gain for the ones that did not.

2.2 Top Five Destination Of UA Flights from the City of New York

To determine the top 5 destinations for UA flights from New York, we grouped the flights by the destination variable,*dest*, and then counted them while sorting in descending order. We then limited our view to the first five destinations. Table 1 summarizes these results while Figure 6 visualizes the top five destination of UA flights. The average gain for these destinations are also shown on Table 1

Table 1: Top five destination of UA flights summary

Destination	Frequency	Average gain
IAH	6814	6.86
ORD	6744	7.78
SFO	6728	8.70
LAX	5770	7.82
DEN	3737	7.30

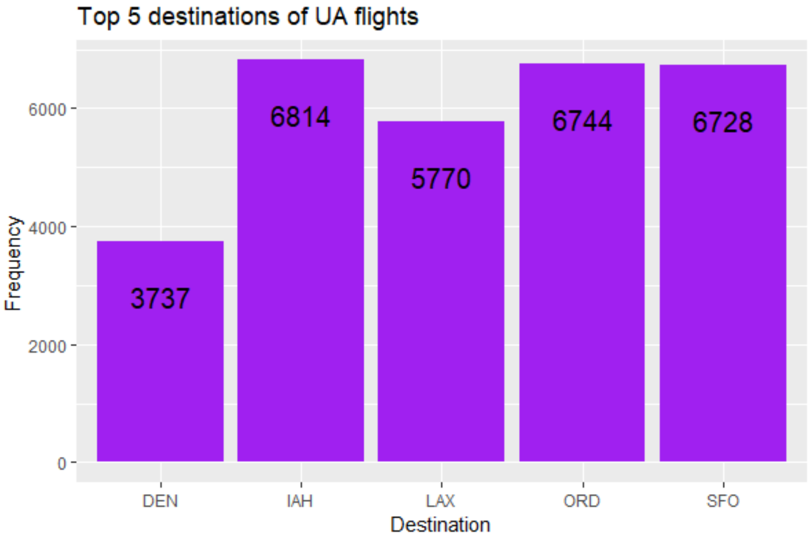

Figure 6: Bar graph visualizing the top five destinations of UA flights.

From Table 1, IAH which is the most frequent destination has the lowest average gain followed by DEN which is the least frequent destination. There does not seem to be any trend between the frequency and the resulting average gains.

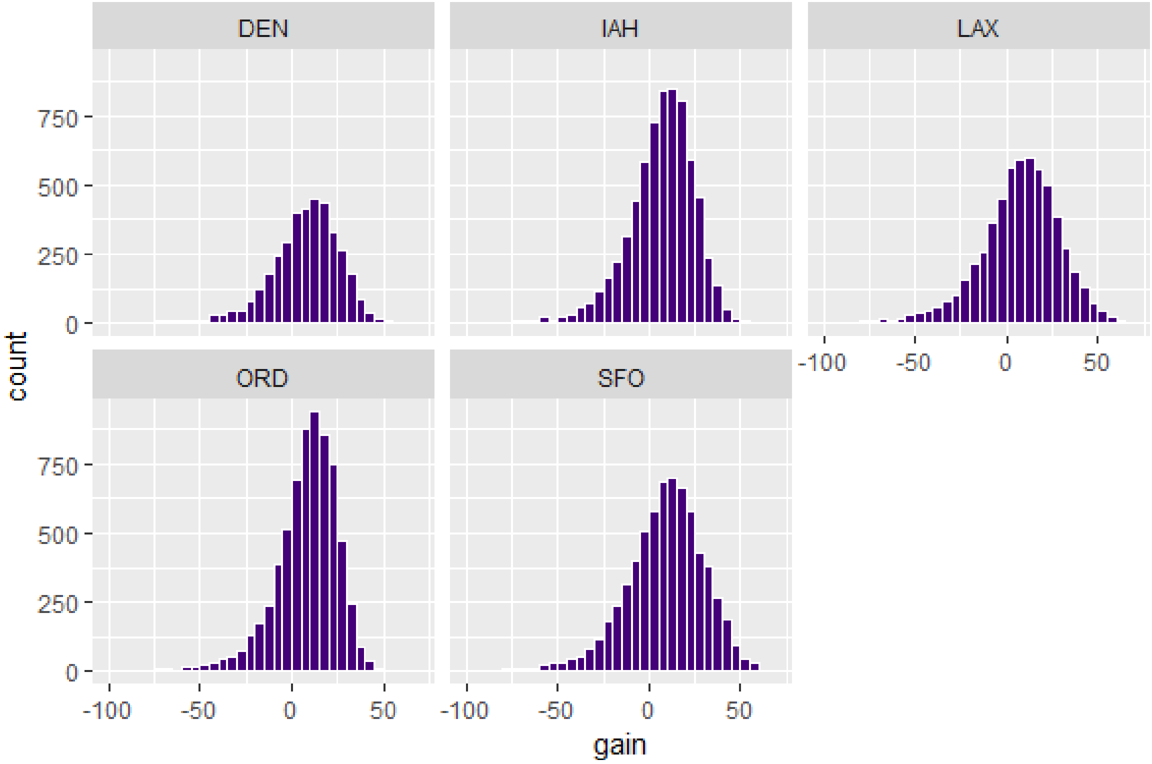

Figure 7: A Histogram visualization of the distribution for the top five destinations of UA flights.

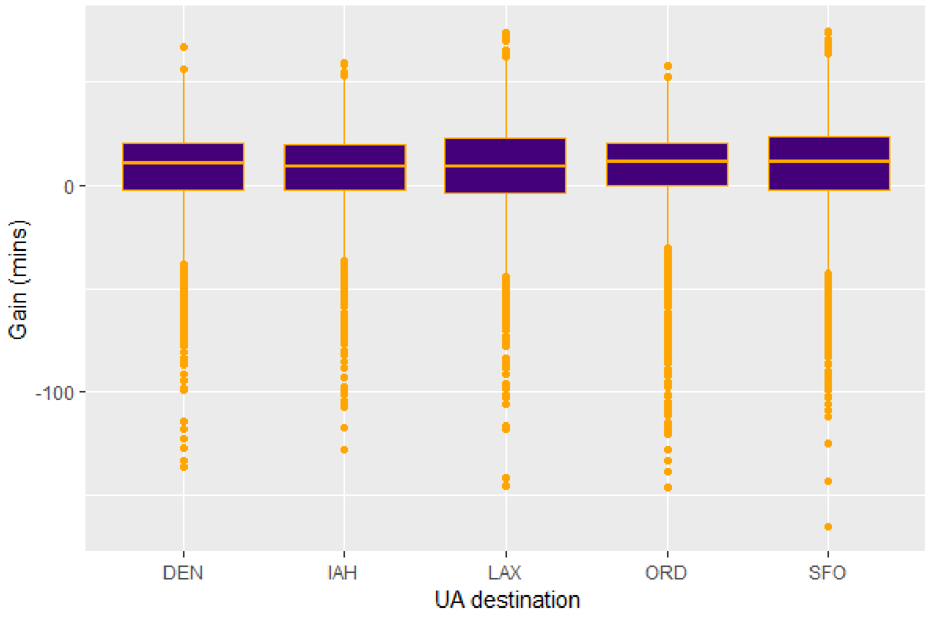

Figure 8: Boxplot visualization of averages for the top five destinations of UA flights.

Figures 7 and 8 visualize the resultant distributions and average gains for the top five destinations of UA airlines respectively. As shown on Figure 7, all of the gain distributions have some skewness to the left. On the other hand, the box plot on Figure 8 shows that the mean gain for the various destinations is very close. However, LAX and SFO appear to have a larger spread in gain compared to DEN,IAH and ORD.

2.3 Gain Per Hour For UA flights with departure delay

Gain per hour was calculated by dividing the gain variable by the airtime in hours.

2.3.1 Average gain per hour for flights that departed late versus not late

The mean gain per hour for on-time flights was 3.99 for on-time flights while the mean for late flights was 3.14. The resulting mean difference was 0.89 hours. However, is the difference in mean gain per hour between the on-time flights and late flights significant? To answer this question, we conducted a hypothesis test. The hypothesis were as follows.

- Null hypothesis ,H_o, mean gain per hour difference($\mu_{ontime} - \mu_{late}$) is zero. That is, the mean gain per hour of the departed on time and departed late groups are the same.
- Alternative hypothesis ,H_a, mean gain per hour difference($\mu_{ontime} - \mu_{late}$) is not zero. That is, the mean gain per hour of the departed on time and departed late groups are not the same.

Use of the $t.test$ function in $Rprogramming$ resulted in a probability of $2.2e-16$. Since we had assumed a significance of 0.05, we rejected the null hypothesis. Therefore, there is evidence that the mean gain per hour between flights that departed on time is different from the gain per hour of those that departed late. Notably, the probability obtained in this conjecture is the same that was obtained from the gain hypothesis of these groups. The significance of the gain difference does not seem to be affected strongly by the airtime because we ended up dividing the gain per hour by airtime in both groups.

2.3.2 Average gain per hour for flights that departed more than 30 minutes late versus those that did not

In this analysis, flights that departed more than 30 minutes late were classified as very late, while the ones that departed less than or equal to 30 minutes were classified as not very late. The calculated gain per hour was 3.06 min/hr and 3.69 min/hr for the very late and not very late groups respectively. The resulting difference was therefore 0.63. To understand whether this difference was significant, we conducted a hypothesis test. The two hypotheses were proposed as follows.

- Null hypothesis ,H_o, mean gain per hour difference($\mu_{notverylate} - \mu_{verylate}$) is zero. That is, the mean gain per hour for flights that did not depart very late and the very late groups are the same.

- Alternative hypothesis ,H_a, mean gain per hour difference($\mu_{notverylate} - \mu_{verylate}$) is not zero. That is, the mean gain per hour of flights that did not depart very late and the very late groups are not the same.

The calculated probability was $1.372e - 06$. Since we had assumed a significance of 0.05, we rejected the null hypothesis because our probability is very small. Therefore, there is evidence that the mean gain per hour between flights that departed very late (more than 30 minutes late) is different from the mean gain per hour of the ones that did not. However, we note that the probability for gain per hour is smaller compared to that of gain in the two groups. Nevertheless, the probabilities are very low in both cases leading to similar conclusions. Having reviewed the mean differences between various categories for gain and gain per hour, we study one more potential difference in the following section. The difference of gain per hour for long and short flights.

2.3.3 Gain Per Hour For Long UA Flights Versus Short Flights

In this study, in order to test whether the gain per hour for long and short flights was the same or not, the flights were categorized based on air time. Flights with air time exceeding three hours were classified as long while flights that were less or equal to three hours were classified as short. In the data, there were 32,512 in the long flight category and 25,270 in the short flight category. Figure 9 shows a histogram depicting the distribution of gain per hour in the two categories. From the figure, we observe that the distribution of short flights has longer trains with a stronger skew to the left. This is particularly clear from the Q-Q plot. Similarly, the histogram for the Long flights shows a skewed distribution. However in this case, the gain per hour has a longer tail to the left as supported by the qqplot.

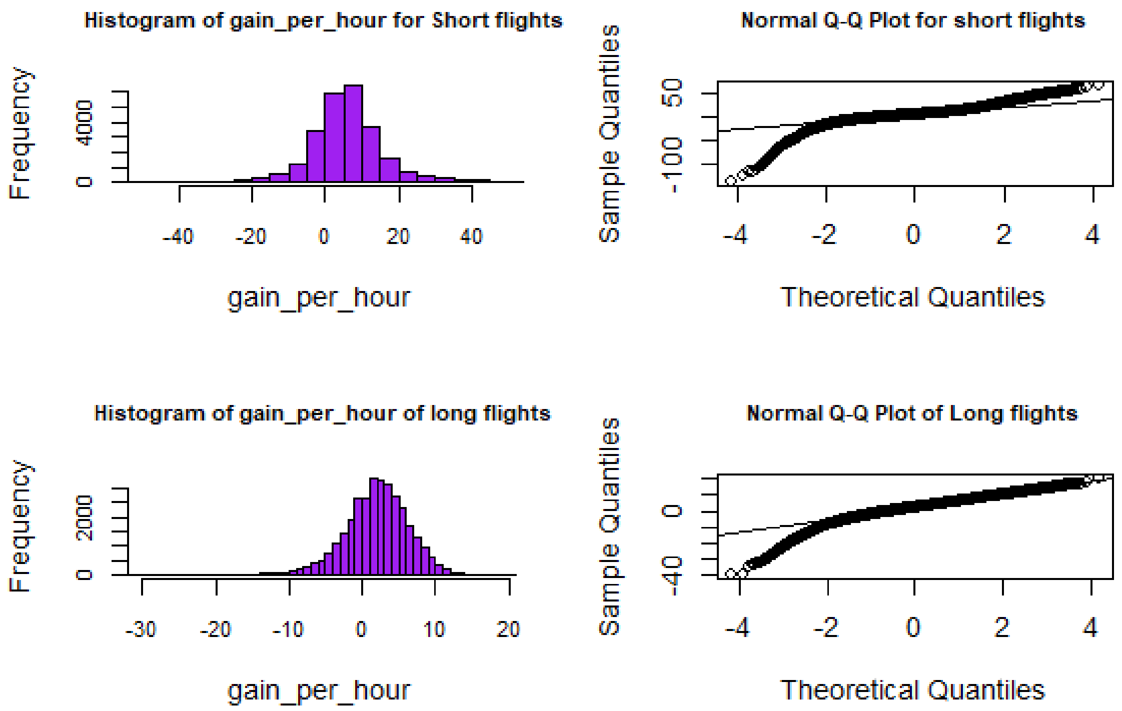

Figure 9: Histogram of flight of long and short flights.

Next to understand the spread and average gain of the average gain per hour, a box plot was plotted as shown on Figure 10. From the figure, it is obvious that gain per hour for the short flights group is higher compared to the long flights group. The exact difference in the mean gain per hour was calculated as 3.996 min/hr. The means for the short flights was 5.86 while that for long flights groups was only 1.86. This difference is quite huge but is it significant? To test this, we conducted a confidence interval test. Again because we are working with a difference of two means, with a large sample size, we conducted the confidence interval test in R. The resulting interval was (-4.145229 ,-3.848538). Because this interval does not include zero, we concluded that the mean gain per hour for long and short flights are not the same. We are therefore 95% confident that the mean gain per hour for long flights is between (4.145229, 3.848538) lower than the short flights.

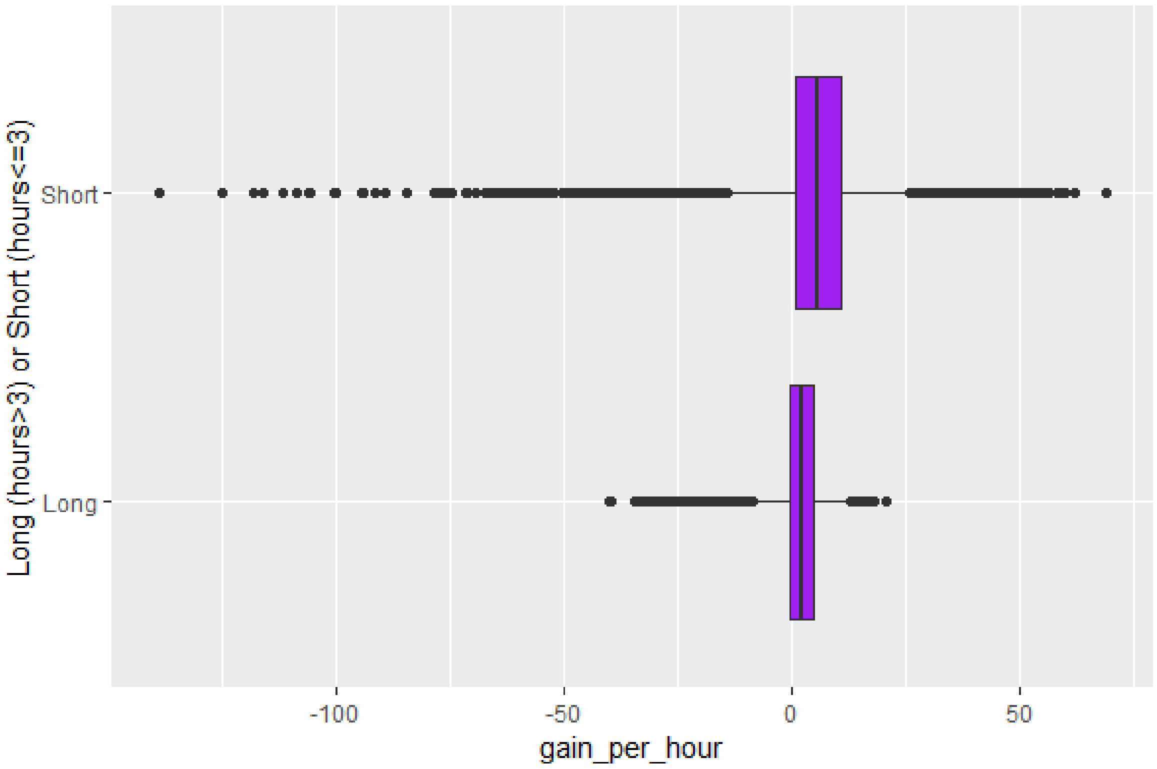

Figure 10: Box plot for long and short flights groups.

3 Conclusion

This study was interested in two derived variables, gain and gain per hour for United Airlines from the $nycflights13s$ data set. We were interested in determining whether the average of these variables differ across several groups. These groups were late vs not late, very late and not very late, short vs long flights. We were also interested in determining the top 5 destinations and the resulting distribution for UA flights from the city of New York. To address these we used both hypothesis testing and confidence intervals testing. In addition, various EDAs were conducted to visualize the data. These visualizations included histograms qqnorm plots and box plots. We conducted hypothesis testing for late vs not late, very late and not very late groups and confidence intervals for Long vs Short flights analysis on our difference of interest. For hypothesis testing, we stated a significance level and evaluated the resulting probability to reject or not reject the null hypothesis. For confidence intervals, we used the inclusion of zero in the interval to determine whether the difference was significant or not.

For all the groups we studied,the results were significant and UA would need to investigate these parameters further if efficiency and customer service were to be improved. The top five destinations for UA flights were determined to be AH,ORD , SFO , LAX , DE. If inefficiency in these five destinations were to get improved by UA, perhaps there would be significant consequently improving their customer service. The results from the confidence interval showed that the average difference in gain per hour between long and short flights was also different because the interval did not include zero. In conclusion, all the groups we compared had significant differences. Significant differences meant that the differences were unusual and thus potential subjects for further investigation by United Airlines.