Data 5300 Project 1

Ngunjiri, Justus JG

October 28, 2022

1 Introduction

This report addressed several relationships for Data 5300 course project 1. In this project it was required that data from *nycflights13* package be explored from a United Airlines(carrier code UA) analyst perspective in order to improve customer satisfaction and efficiency. This was to be achieved by addressing the relationship between departure delays and the following variables:

1. Time of the day
2. Time of the year
3. Temperature
4. Wind speed
5. Precipitation
6. Visibility

The above named factors were addressed using the original *dep-delay* variable. However, there were additional options of using *late* and $very_late$ variables that had been created in previous assignments. To successfully complete the project, the necessary analysis was carried out in *Rprogramming* and the results of the analysis are hereby discussed.

2 Relationships

The necessary analysis included visualization to understand the data as well as formulation of hypothesis in order to form the basis of criticism. The null hypothesis between the statistics in question were then addressed using permutation tests. The main aim of the permutation test is to estimate the distribution and hence the probability of observing the same result that was observed from the population after carrying several randomised simulations. For this report, we set our signifance level at 5%, that is, if we conduct our hypothesis and the *pvalue* happens to be less than 0.05, we reject the null hypothesis of the proposition, otherwise we do not. We start off with the first relationship, departure delay and time of the day in the following section. To address various relationships, firstly we filtered the data as required to only include UA flights in our analysis. In this filter and to avoid misleading statistics, we also filtered out negative departure delays as they are not quite important to our analysis of improving efficiency and customer efficiency. Negative delays indicate that the flight left early which we can argue that it does not affect the customer(unless it is too early and the customers happen to miss their flights etc). To avoid contradiction, we assume that flights that left early had all customers on board an thus they would not be good indicators of the

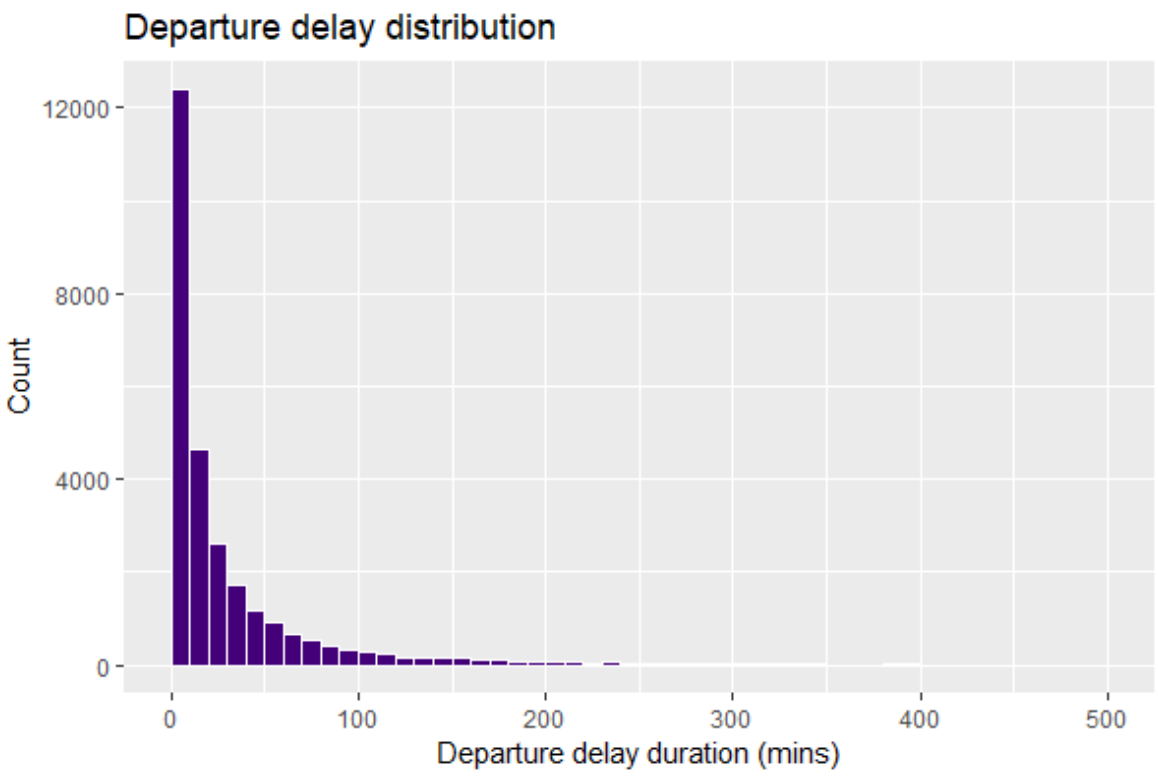

Figure 1: Histogram depicting the distribution of positive departure delays.

required improvement. In addition, for this report, we analyse all the data at the granular level, for instance at a given time, origin etc. This is because variables at this level show the best raw form of representation. By doing so, we avoid issues that would arise while using over summarised statistics. We examine the distribution of the departure delays under consideration in figure 1. Clearly, the departure delays are right skewed and suggest to follow an exponential distribution.

2.1 Time of day and departure delay

To understand the relationship, we ask the following question for this variable..how does length of departure delay compare with the time of the day? We visualised the data to check this depending on the time of the day that the flight was scheduled. We are also interested in knowing where these delays are from and hence we plotted colored them by origin.

After examining the dataset, it was noted that flights are scheduled between 0500h to 2345h. To investigate the time of the day, a scatter plot was used while limiting the time of the day to between 0500h and 2345h. Other alternatives existed of addressing the time parameter such as the actual departure time but since we are seeking to improve the operations, we decided to base everything on the scheduled departure time. The boxplot is shown in Figure 2. Since it is likely that departure delays are affected by other factors such as conditions at the origin, the scatter plot was colored by origin. The origins for UA flights are EWR, JFK and LGA. From the plot, it is clear that most of the delays as well as all the delays after 2100h are from EWR. This can be as a result of the EWR being the hub or other factors.
From the same figure, it can be seen that flight delay times are both shorter and less variable in the morning hours, and increase in the afternoon hours. There are also shorter flight delays

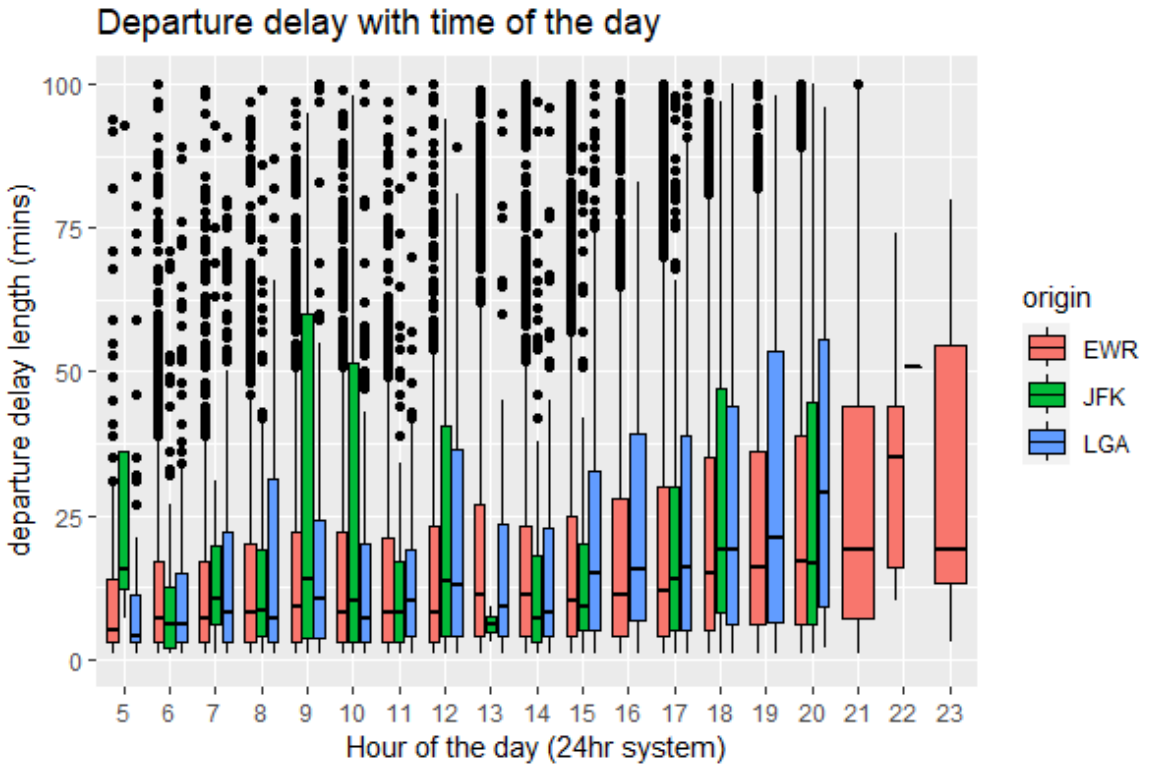

Figure 2: Time of the day boxplot.

as we approach the last hours of the scheduled departure times.

From the box plot, it appears like the flight delay lengths in the late afternoon and evening hours are different from the one for early morning hours and early afternoon hours. However, is this actually true? To answer this question a permutation test was carried out to test this hypothesis. The null hypothesis, (H_o) was that the departure delay lengths are the same for both halves of the day, while the alternative hypothesis (H_a) was that they are not the same. The first half was grouped as times between 0500h and 1420h with the second half being between 1420h and 2345h. The observed mean difference of mean departure delays in these two halves was 11.23 minutes. The permutation test was taken with N=10000. A p_value of 0.002 was obtained which is way too low when assuming a significance of 0.05. The null hypothesis that the mean delays for early morning early afternoon (EMEA) and Late afternoon Evening flights (LTEF) departure delay lengths are the same was thus rejected. The difference between mean delay of the two halves is actually different and significant. Subsequently, if the airline is struggling with late LFEV fligths sales, they can possibly attribute it to customers knowing that LFEV flights are usually delayed in the hours that fall under this category. They could also investigate other factors in order to eliminate delay hours in the second half of the day. For instance, why are departures after 2100hr only for flights originating from EWR?This can be addressed in order to improve customer satisfaction and efficiency.

2.2 Time of year and departure delay

Just like with the time of the day analysis, initially, a scatter plot was used to visualize the relationship between departure delay and time of the year (in months) as shown in Figure 3. From the plot it seems like the departures have varying delay lengths over the months of the

year. It is also apparent that the departure delays are consistently high between the month of June and September, and low for February and October. However, there is no clear trend on the relationship between delays and time of the year.

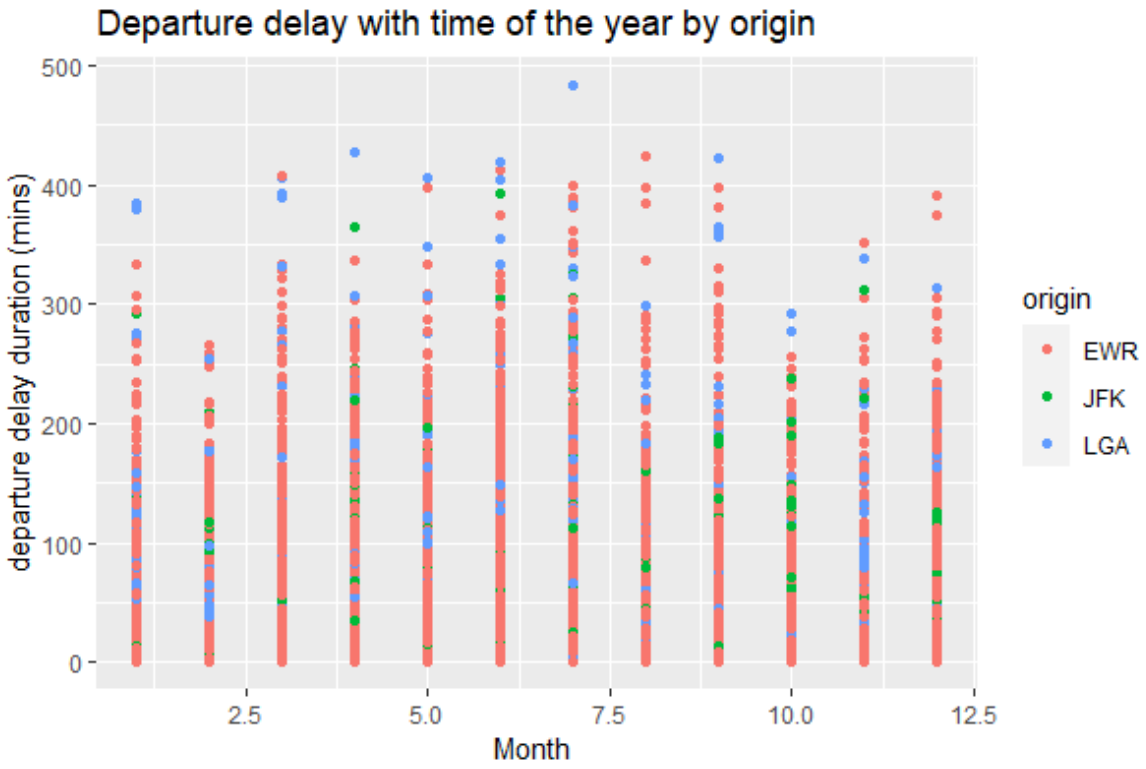

Figure 3: Time of the year scatter plot.

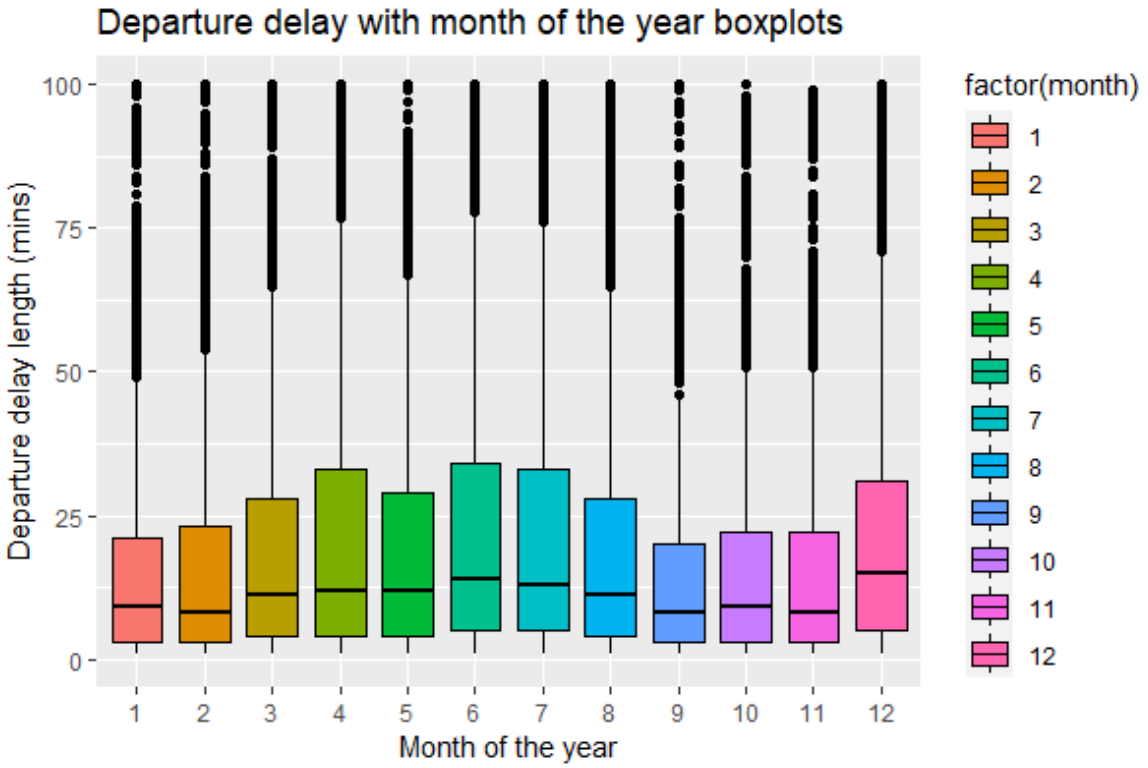

Figure 4: Time of the year box plot.

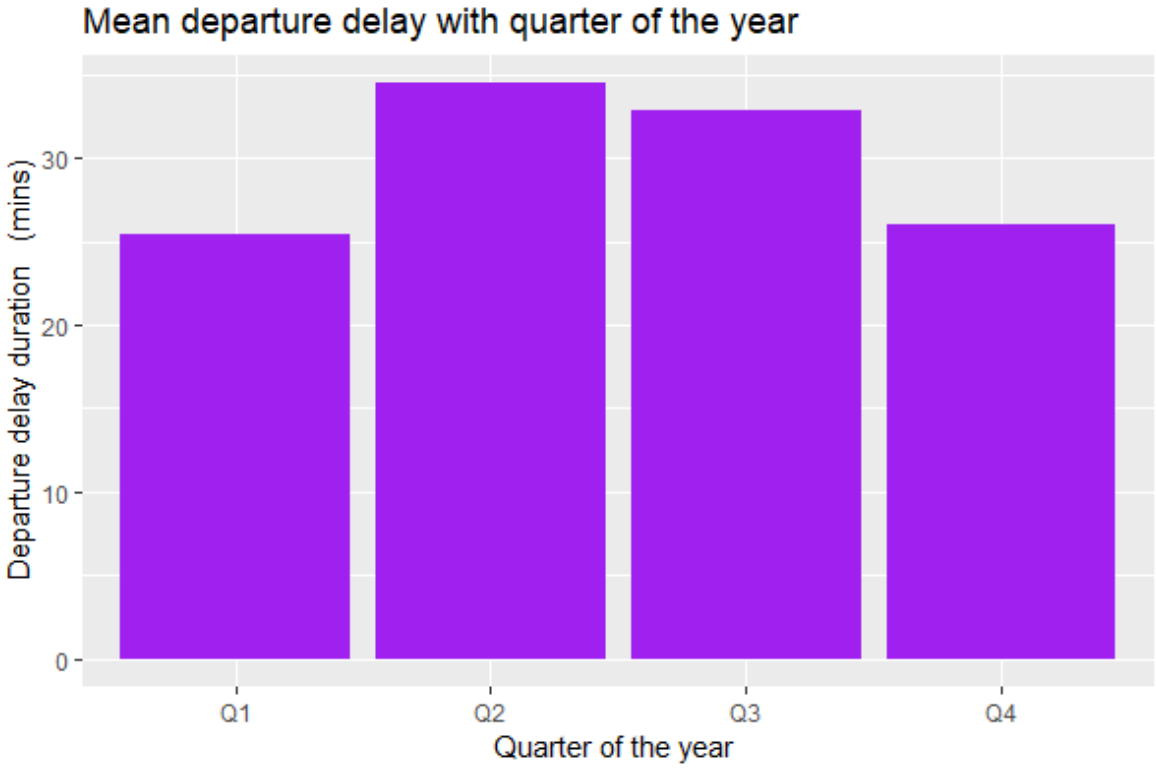

Figure 5: Time of the quarter of the bar plot.

To seek more insights, a boxplot was used to visualize the delay over a longer aggregating period over the year (quarters). This is helpful to visualise the five number summary statistically. Since some means were really close, a bar graph was added to compliment the box plot over the four quarters as shown in Figure 5 and 6.

When using these two supplementary visualisations, it is evident that the mean departure delays are higher for quarter(Q1) and quarter4 (Q4) compared to quarter 3 (Q3) and quarter 2 (Q2). However, the mean departure delays in the paired quarters are very close.

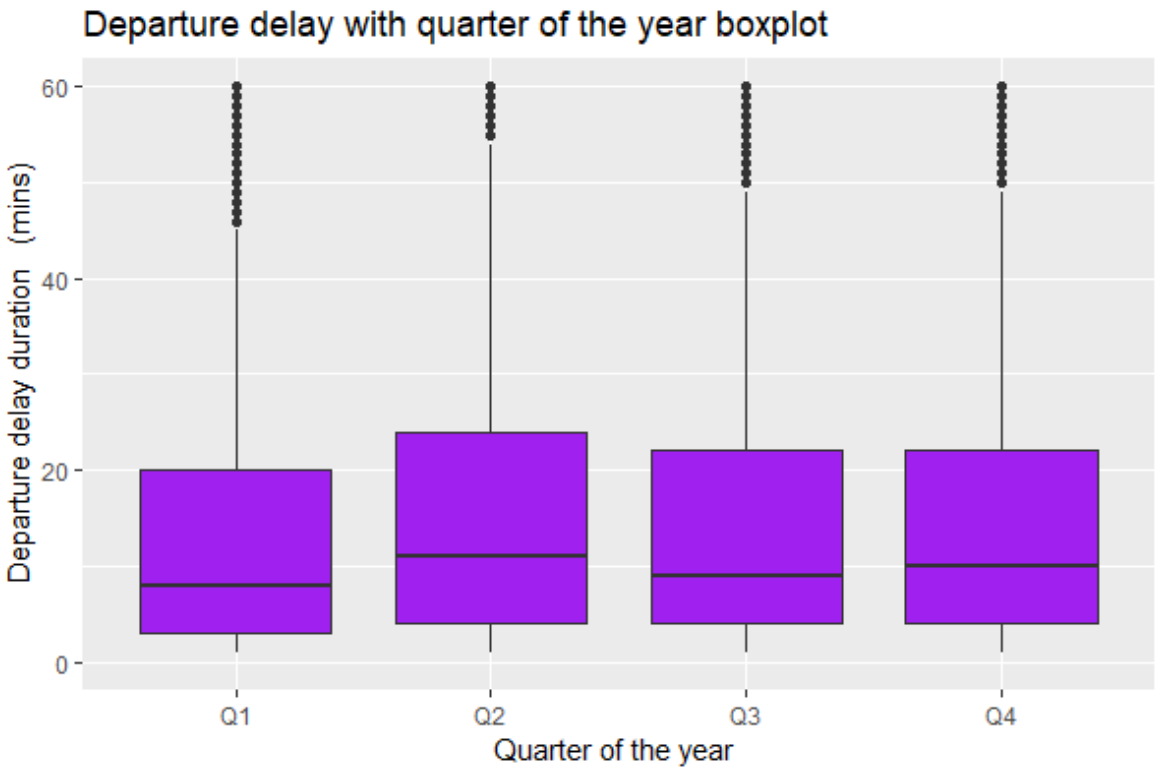

Figure 6: Quarter of the year box plot.

A further step was thus carried out to determine if the aggregate mean departure delay difference between the two halves of the year split as (Q1+Q4) and (Q2+Q3) are the same. The null hypothesis (H_o) was that the departure delay lengths are the same for both halves of the year, while the alternative hypothesis (H_a) was that they are not the same. A permutation test was carried out with N=10000. The permutation test yielded p-value of 0.0002. Since the pvalue is less than 0.05, we rejected the null hypothesis that the mean differences in the combined quarters are the same. As a result UA should consider investigating their flight delays over the months in these quarters. A similar process can be followed to compute the the difference over the various months of the year in order to yield more granular results.

2.3 Temperature and departure delay

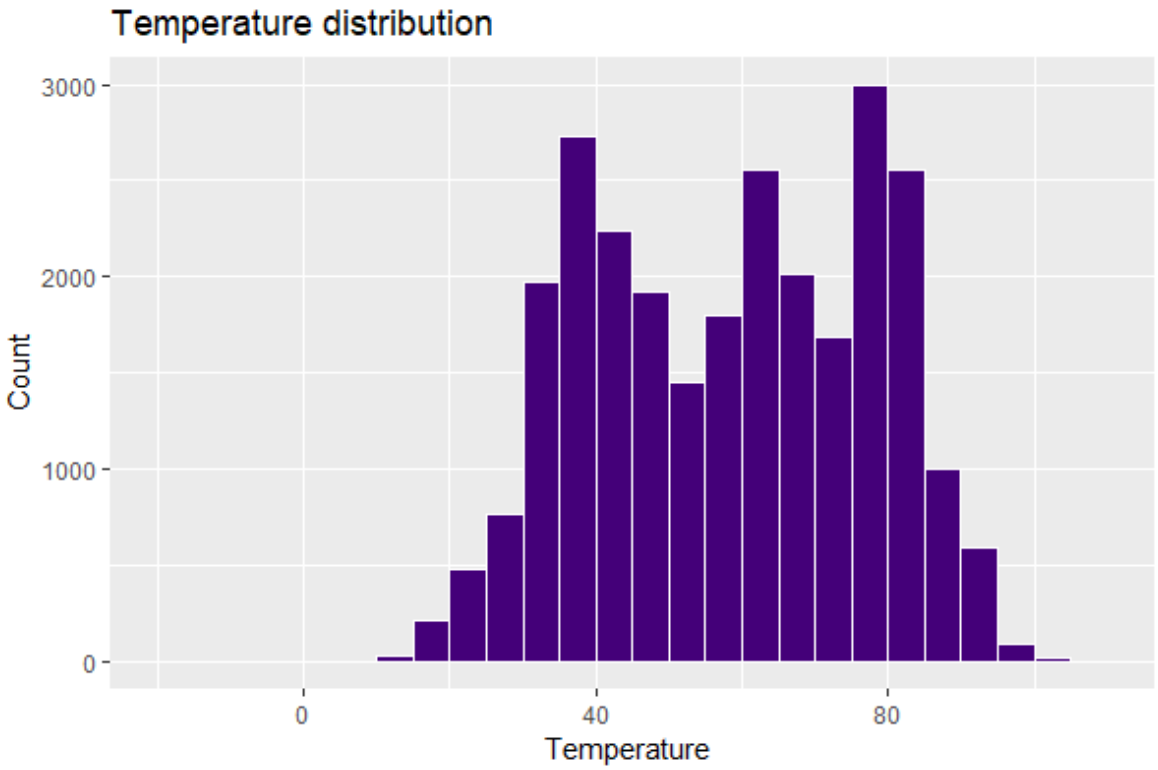

Figure 7: A histogram depicting the distribution of temperature.

Firstly, the temperature data was visualized independently to understand its distribution as shown in Figure 7. From the histogram, temperature distribution is multimodal (trimodal) with a mean of 58.41 and a median of 59. Having understood the distribution of temperature independently, we then visualise its relationship with departure delay. To investigate how the combined relationship of temperature and departure delays compare over the year, the relationship was visualised over the quarters as used in the time of the year investigation. The bar graph showing the relationship between mean temperature,mean departure delay and the year quarter is shown in Figure 9. To achieve the values used in the bar graphs, granule level data was used where temperatures were firstly grouped by origin, and then by year quarter to ensure it was not over summarised. From the figure, it can be seen that higher temperatures in Q2 and Q3 tend to result in longer flight delays.

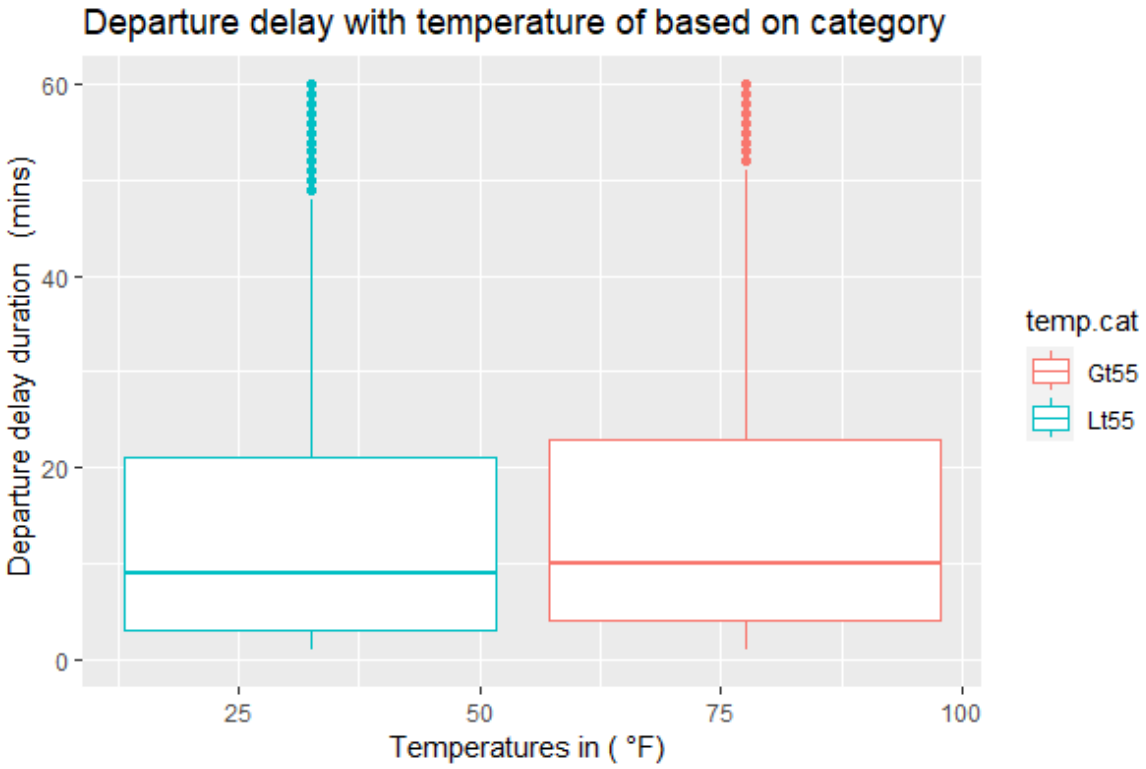

Figure 8: A histogram depicting the distribution of temperature.

We observe that this occurs in the second and third quarter of the year (which have warmer temperatures as well). To avoid the bias of using quarter of the year factor for our relationship analysis it was further decided to investigate the relationship between the two variables directly. Based on a scatter plot not included in this report(because it does not add value), it appeared as if temperatures above 55 degrees had longer departure delays. But is there a real difference? We grouped the two zones as $gt55$ and $Lt55$, where the less than investigates temperatures that are equal or less than 55 degrees. A box plot visualising these two categories is shown in Figure 8. The plots were limited to the first one hour in order to get a clearer view of the differences. Close examination of the box plots, indicates that the minimum flight delay for the $Gt55$ category has higher mean, minimum and maximum values for the departure delay minutes compared to the $lt55$ category. The spread is also higher, which indicates higher variability in the departure delays times for the $Gt55$ category.

To investigate the observed difference further, a permutation test was conducted to determine whether there is indeed enough evidence that higher temperatures (above 55) tend to result in longer flight delays or not. This was done by using the null hypothesis where H_o was that the mean difference between departure delay of the grouped higher temperature quarters are the same. The alternative hypothesis H_a was that the mean difference between departure delay of the combined higher temperature quarters are not the same. This was done by running the simulation for 10,000 times. A $p-value$ of 0.002 was obtained, and since it is lower than 0.05, the null hypothesis that the mean difference in diparture delay are the same. United Airlines should thus investigate further on why that is the case if improvements are to be realised. They should focus their improvements on the flights scheduled in these months with higher mean temperatures depending on the origin. Next, we study the relationship between wind and departure delay to gather more insights on the factors that affect flight

delays.

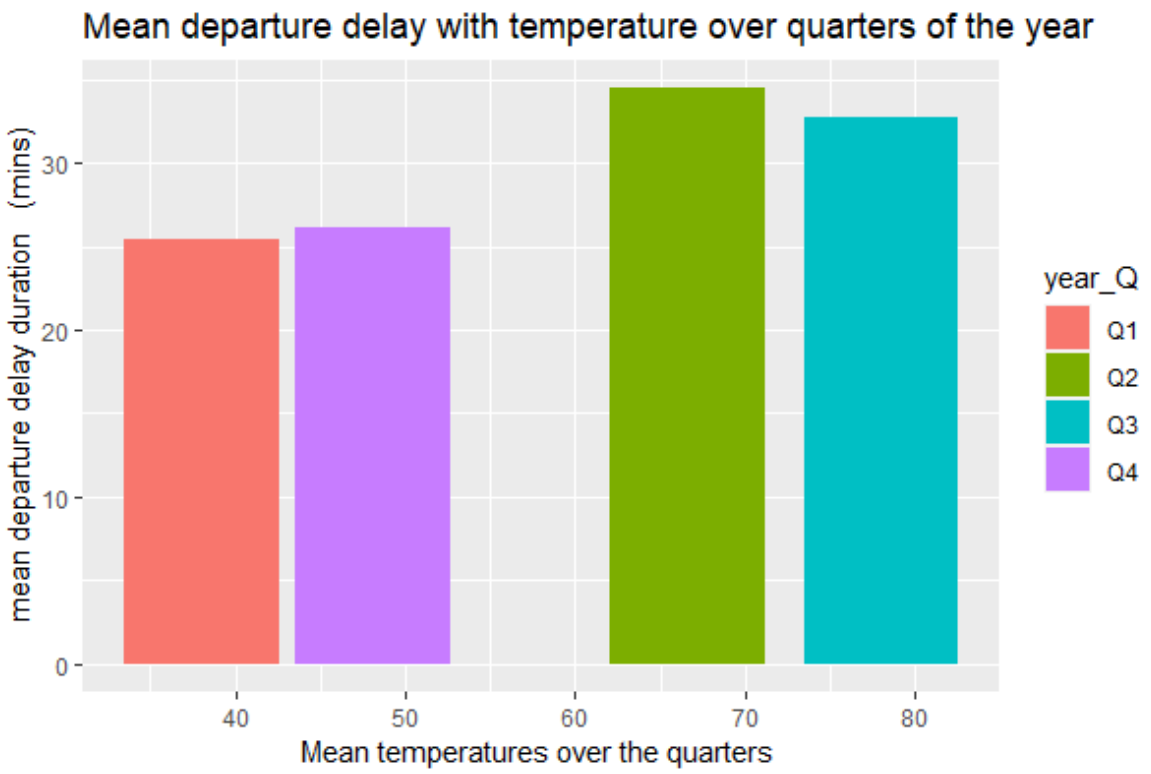

Figure 9: A bar graph depicting the relationship between mean delay and mean temperature over quarters of the year.

2.4 Wind and Departure delay

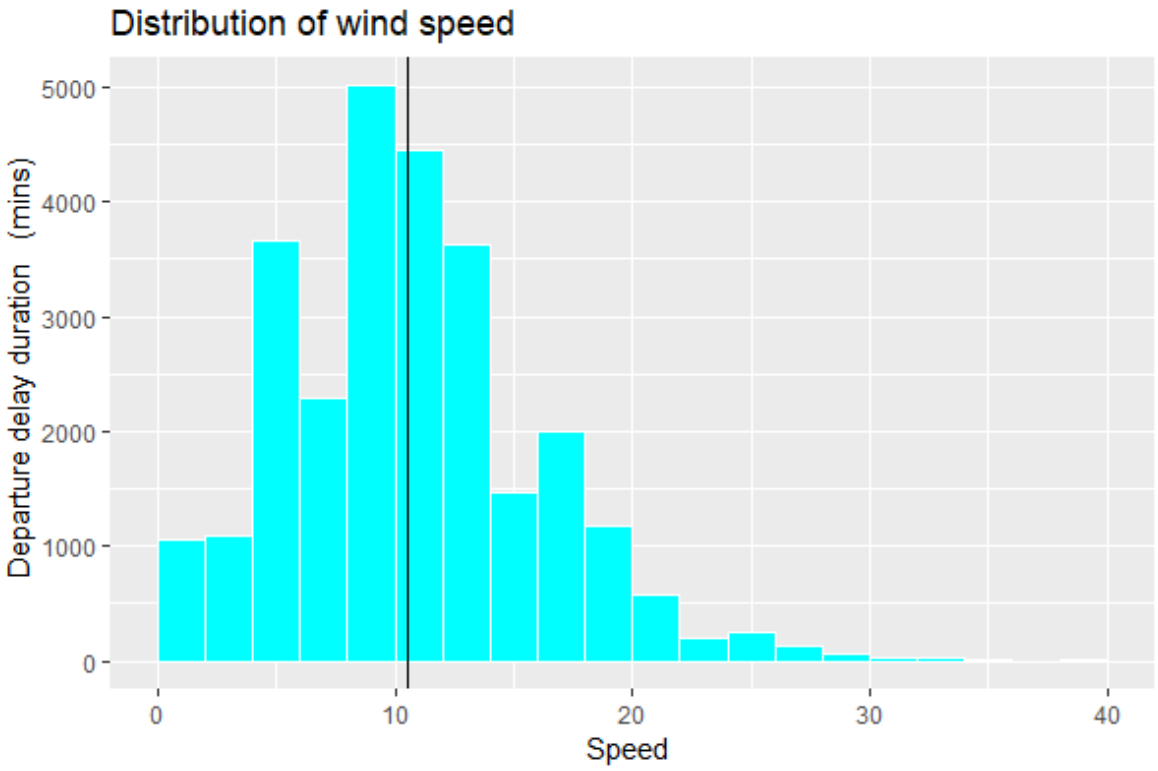

Figure 10: A histogram depicting the distribution of wind speed.

Firstly, we review the distribution of wind speed in the flights dataset by plotting a histogram shown on Figure 10. As shown, wind speed suggests to be multimodal. Secondly, we plot a preliminary scatter plot with wind speed on the x axis and departure delay on the y axis as shown in Figure 11. The plot suggests that most of the delays occur when the wind speed is low but it is not adequate in deriving insights. From the figure, there is no clear relationship or insights that can be derived from this plot. Since wind speed is also dependent on the location, a box plot for departure delay and wind speed grouped by origin for visualization was also plotted as shown in Figure 12. From this box plot, it is clear that the mean flight delays for flights originating from EWR is higher. In addition, EWR has a higher mean wind speed compared to LGA and JFK.

To de-noise the data, wind speed was categorized into two groups near the possible point of investigation(it seems like there are both shorter and fewer flights delayed when the wind speed exceeds 20). But is this really the case? To obtain more insights on the direct relationships, wind speed was summerised into two categories. Wind speed greater than 20 $Gt20$ and $Lt20$, for wind speed less than 20. The less than category also included wind speed that equals 20. The resultant box plot is shown in Figure 13. From the figure, we observe that mean departure delay for wind speed greater than 20 has a slightly higher and the associated flight delays have a higher variability. But how significant is this difference?

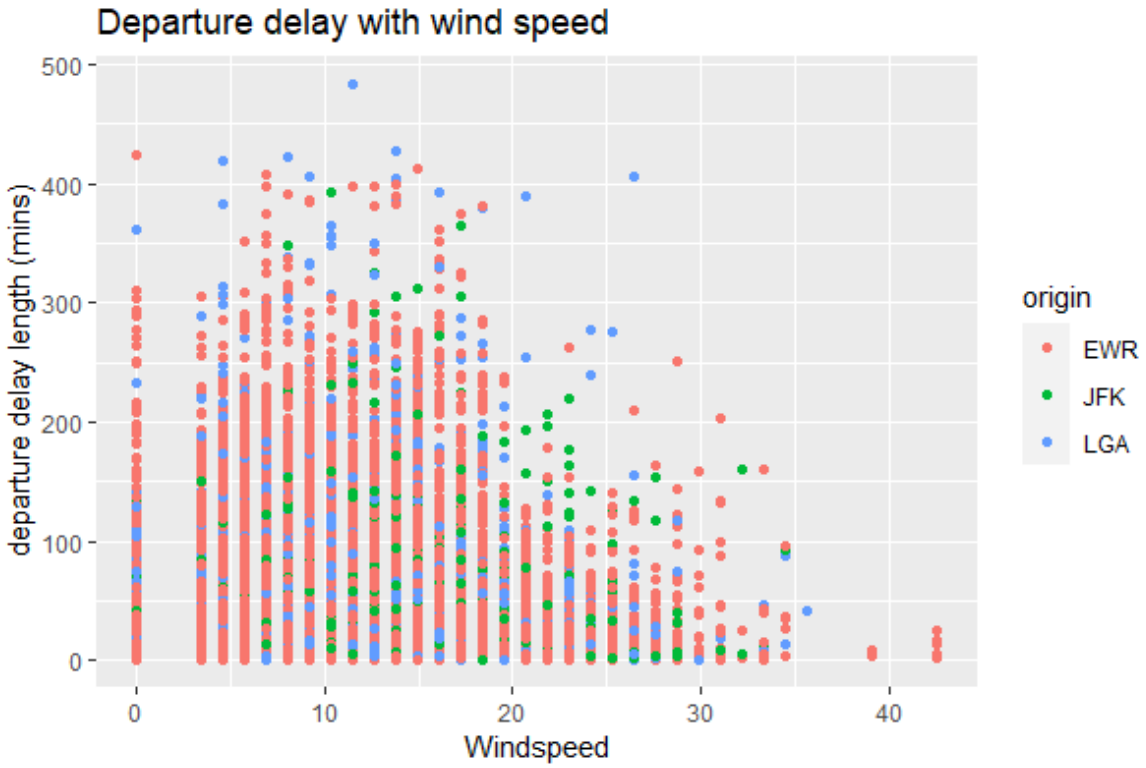

Figure 11: Scatter plot of the wind speed.

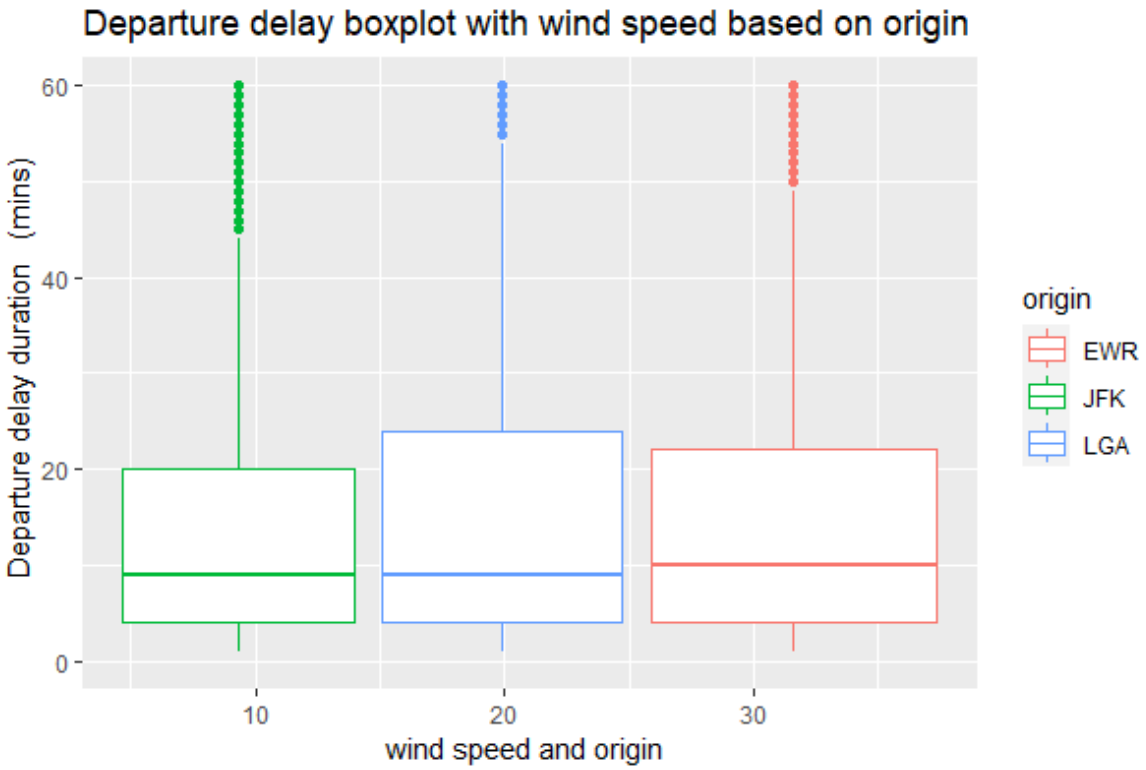

Figure 12: Box plot of the wind speed by origin.

To investigate this, we formulated a null hypothesis H_o that the mean differences of the flight delays between the two groups are the same. We then carried out a permutation test with N=10000 and calculated the proportion where the results would be unusual from the observed difference. The resulting p value wa 0.474. Because the p value is greater than 0.05, we do not reject the null hypothesis. It is normal to have the observed difference. The actual cause of the difference could thus be attributed to other factors such as wind direction etc. The airline should investigate this further to improve its operations.

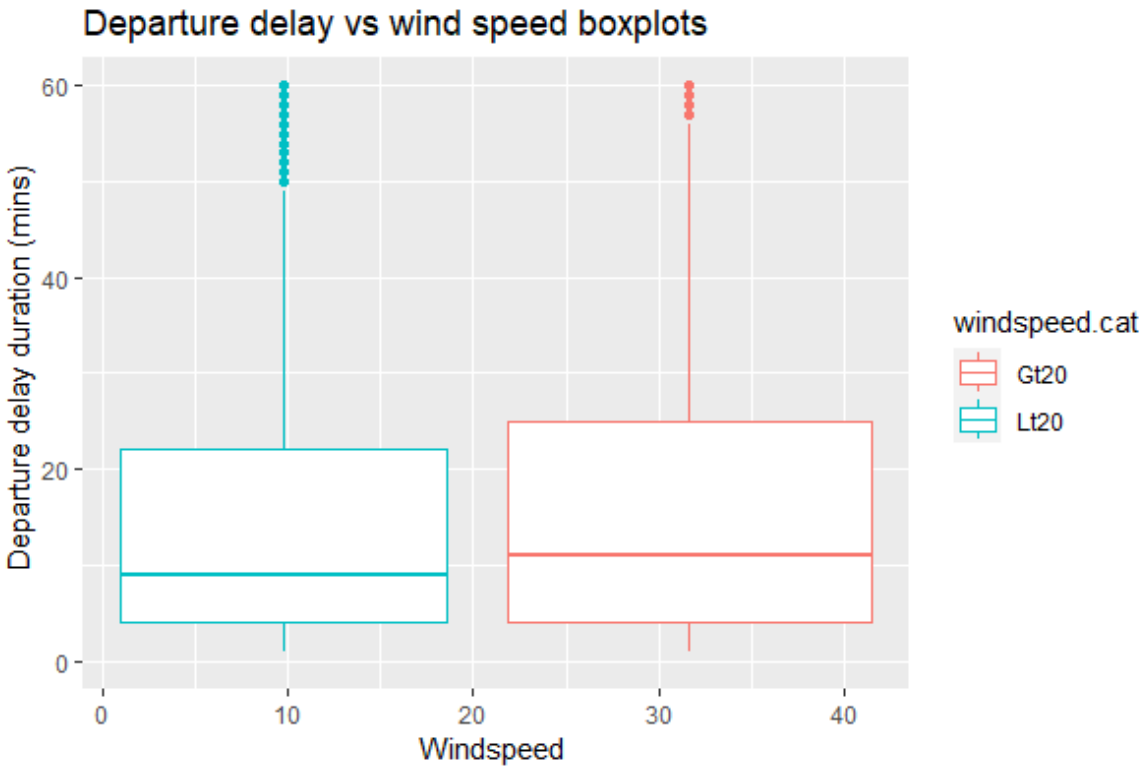

Figure 13: Box plot of the wind speed based on category.

2.5 Visibility and Departure delay

A visualization of the visibility data contained in the dataset is shown on Figure 14.

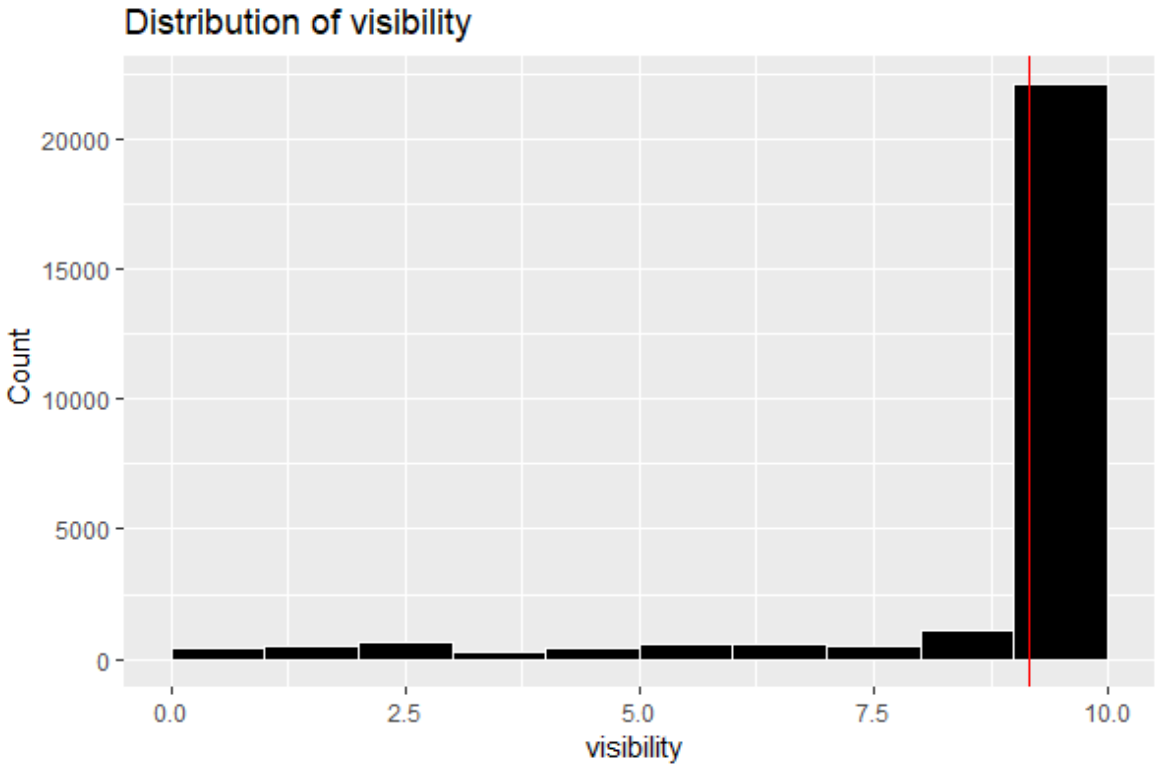

Figure 14: A histogram depicting the distribution of visibility for the flights dataset.

We observe that the mean visibility is close to 10 (the actual mean value is 9.15). It is also evident that the distribution of visibility is left skewed, with most of the observations concentrated around the mean. The mean is also very close to the maximum value of visibility contained in the data set. We then conduct a preliminary visualization for departure delay and visibility using a scatter plot as shown in Figure 14. Clearly, there is no apparent trend and it is just a little of too much noise. However, it seems like most of the flights are delayed when the visibility is high.

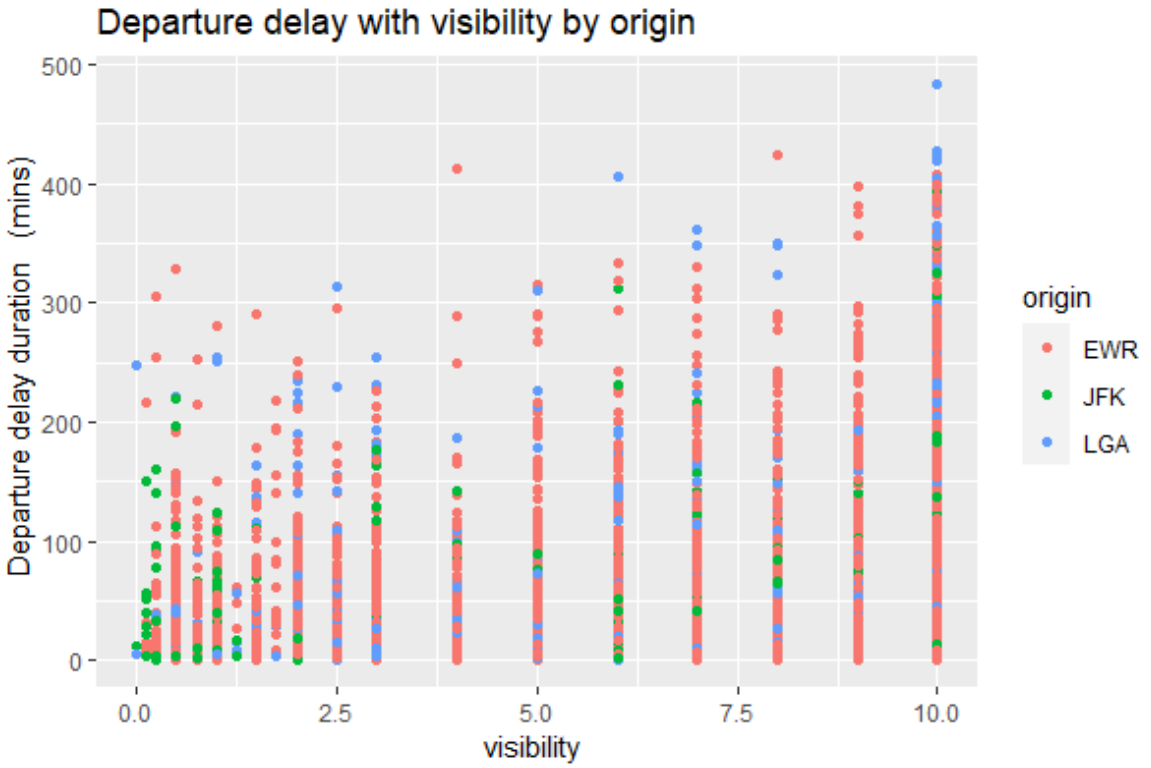

Figure 15: A scatter plot for flight delay and visibility of the flight data set.

To investigate this further, the data was grouped between the potential point of difference. This was where the visibility variable was greater than or equal to five while the other group was where it was less than or equal to 5. A box plot was then plotted as shown in Figure 16. It is worth noting that the delay was limited to the first 60 minutes in order to zoom in for a better look at the differences. From the boxplot, it appears that the category where the visibility is greater than five have shorter flight delays. Also, the departure delay statistics in this group are less variable (indicated by the span of the boxplot, inter quantile range and minimum and maximums). But is there an actual association ? To investigate this , a null hypothesis H_a that the mean difference between departure delays of visibility greater than five and that of less than five are the same. The alternative H_a is that they are not the same. A permutation test using N=10000 was then carried to see how likely that this would be the case (by calculating the proportion). The permutation test yielded a *pvalue* of 0.002. Because the p value is less than 0.05, we do not reject the null hypothesis. Therefore, this factor should be subjected to further investigation by UA. In the next section, the last relationship of departure delay and precipitation was examined.

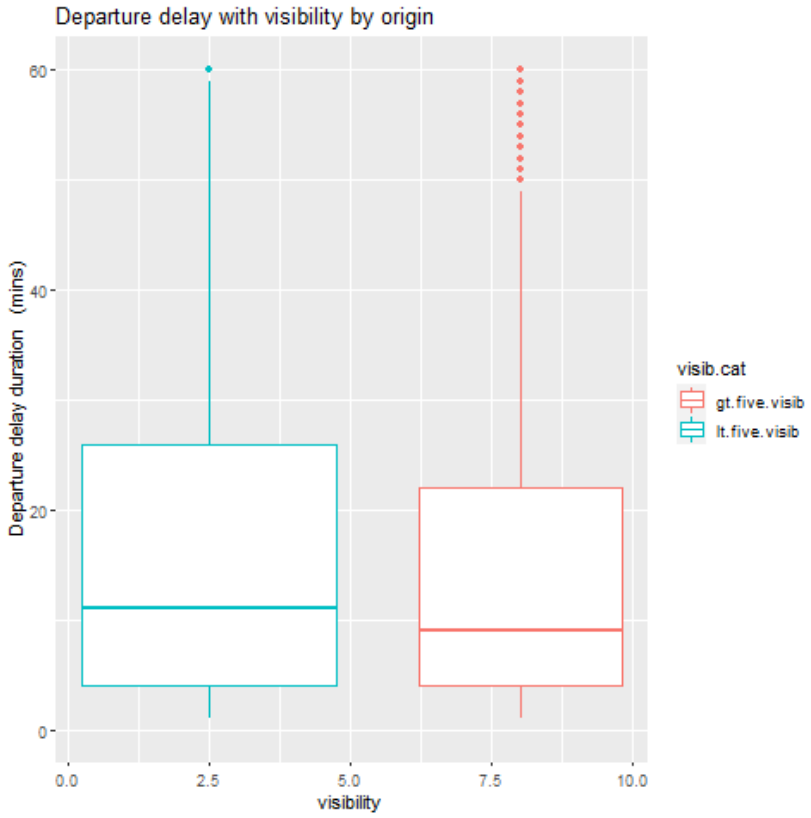

Figure 16: Box plot of visibility based on category.

v

2.6 Departure delay and precipitation

Firstly we examine the distribution of the the precipitation variable as shown in Figure 17. We observe that the the distribution is right skewed with most of the observations being very close to zero. To examine the relationship between precipitation and departure delay, a preliminary visualization was carried out by having a scatter plot of precipitation against the departure delays. The plot did not not provide meaningful insights, therefore the data was divided into two categories.

To realise these categories, the mean of precipitation was computed which was 0.605. Precipitation was then categorized into observations of values groups of greater than the mean and those ones that were less than the mean. This was followed by a boxplot as shown on Figure 18. For a closer look, we limited the departure delay observations to 100 mins. The boxplot suggests that flight delays for the group with high precipitation experience longer delays. The difference of the departure delay between means is also quite high between the two groups. We also observe that the mean for the category greater than the mean is very close to the maximum observation as shown on the box plot.

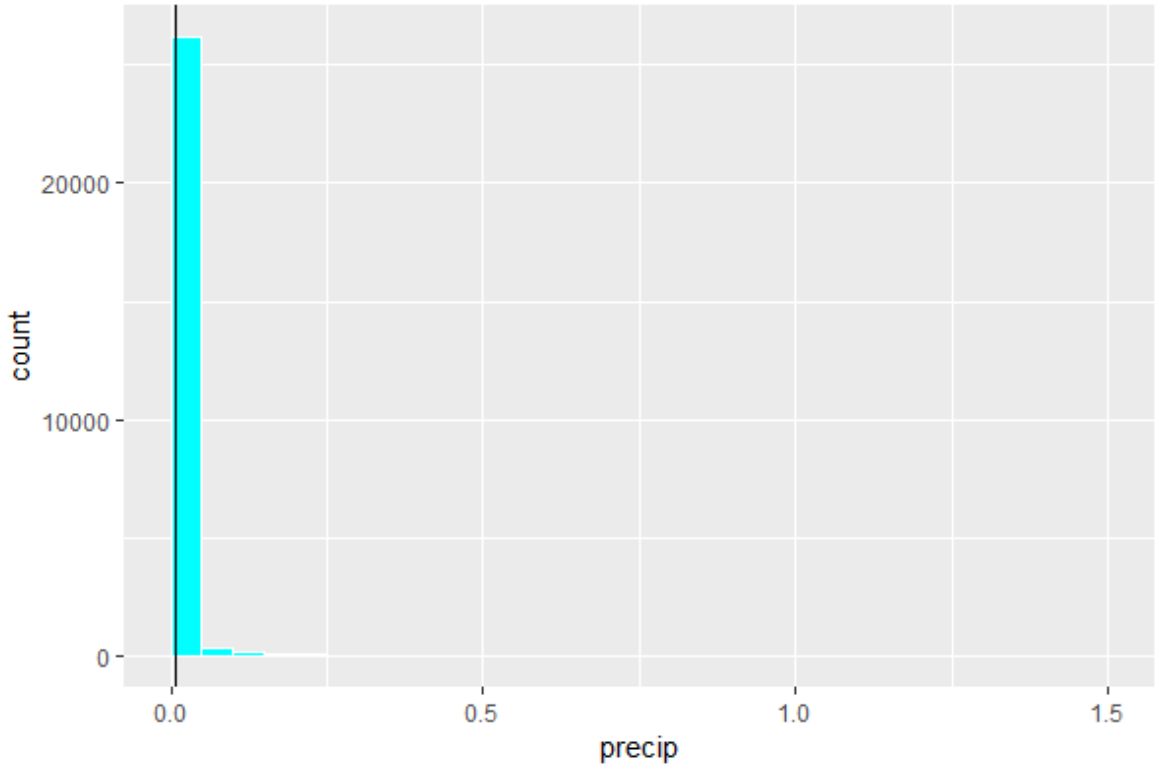

Figure 17: Distribution of precipitation data.

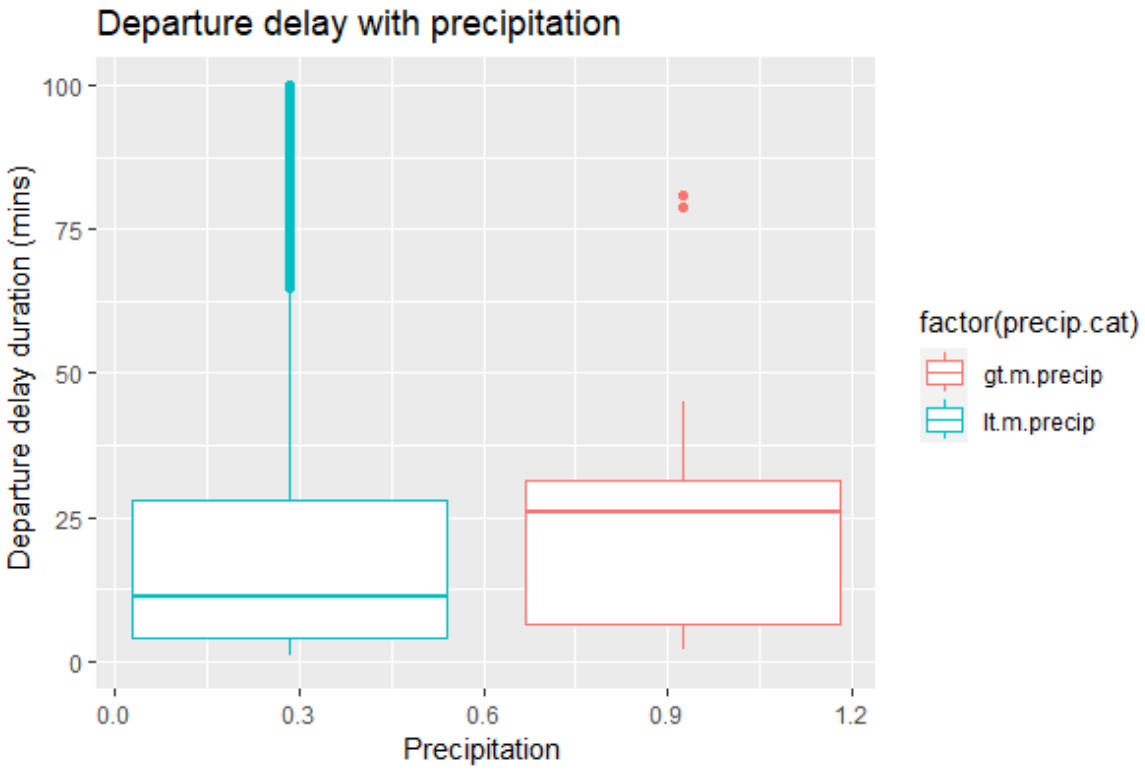

Figure 18: Box plot of precipitation based on category.

The readings from the box plot suggest that there is a relationship between the observations in the two categories based on the difference. However, is this the case? How often can we expect to have the observations greater than the mean precipitation result in longer departure delays? A permutation test with N=10,000 was carried out to answer this question. To complete the null hypothesis, the null hypothesis H_a was that the mean difference of departure delays in the two groups was the same. The alternative hypothesis was that the mean difference in departure delays are not the same. The resulting proportion was 0.008. Assuming a significance level of 0.05, we reject the H_a. The difference in flight delays arising from the two groups are not the same. We can therefore say that the difference is unusual. The airline should look for better ways of dealing with precipitation related delays. If improvements are to be realised, they could consider using predictive precipitation data and reschedule the flights accordingly in order to avoid delays when the probability of precipitation is high. They can also inform customers in advance to avoid inconveniences and high level of customer dissatisfaction.

3 Conclusion

This report intended to investigate six relationships between departure delays and time of the day, time of the year as well as the relationship with weather variables namely,temperature, wind speed, visibility and precipitation. To examine the relationships as required, the variables were visualised on a preliminary basis by use of scatter plots. However, because the scatter plots did not show the relationship clearly, the data was categorized into groups and visualised again using a box plot and bar graphs where necessary. In addition, the data was grouped and visualised at the granular level before categorizing it for box plot plotting. For instance, because weather conditions such as temperature, wind , precipitation and visibility are very dependent on the location, we visualised them and colored them based on origin.

Grouping the data was mainly guided by the point of potential difference as observed from the initial scatter plot containing all the observations. Visualisations of each variable were also carried out in order to determine their corresponding distributions. However, the weather data was contained in a different database. We thus combined them based on shared attributes such as time, hour etc. A permutation test was then carried out on to determine how unusual the observations between the two groups would be by simulating several instances. Out of the six variables that were used to examine the departure delay, we rejected or failed to reject the null hypothesis based on the grouping. We rejected null hypothesis for the time of the day, time of the year, temperatures, visibility and precipitation. However, we did not reject the null hypothesis for the relationship of departure delays with wind. United airlines can therefore subject all the variables where the null hypothesis was rejected while incorporating the recommendations discussed under each variable. Identifying why the results are unusual was a good starting point to enhance UA with statistics required for the next steps towards improving customer satisfaction and the overall efficiency.