Airlines On-time Performance Analysis

Yash Kamoji	34032599	ykamoji@umass.edu
Karthik Ravichandran	34049791	kravichandra@umass.edu
Deepika Singari Velu	34045711	dsingarivelu@umass.edu
Deepa Rukmini Mahalingappa	34040788	drukminimaha@umass.edu

INTRODUCTION

In the realm of aviation operations, the experience of observing the passage of time while stationed at the departure gate due to the delay in aircrafts is a familiar scenario. Airline delays happen a lot and affect millions of travelers every year. But what causes these frustrating delays? That's what we want to find out by looking beneath the surface. "Airline delays represent the failure of flights to depart or arrive on time, stemming from various reasons such as technical issues or adverse weather conditions. These disruptions, impacting millions of travelers yearly, prompt a closer examination beyond the surface to uncover the underlying causes affecting passenger experiences." Ultimately, examining Airlines Delay On-time Performance allows for a comprehensive analysis aimed at improving reliability and the airline's efficiency, mitigating disruptions, and enhancing the overall travel experience for passengers. "Various methodologies such as statistical models including regression analysis, machine learning techniques like decision trees, binary classification prediction using supervised Learning, and ensemble methods such as Random Forests have been explored by researchers to comprehend and predict Airlines Delay On-time Performance trends." This statistical project report aims to investigate and analyze the comprehensive performance data available for flights in the US aviation industry. By utilizing statistical methods, the report aims to quantify some of the conclusions and inferences about the delays, shedding light on the validity of certain preconceived claims.

SOURCE

`https://www.transtats.bts.gov/Homepage.asp`

DATA EXPLORATION

Since June 2003, the airlines that report on-time data also report the causes of delays to the Bureau of Transportation Statistics.For the purpose of this analysis, we have aggregated the on-time flight performance raw data of the past year. Since there are plethora of flights scheduled daily, we will focus on our analysis to just the top four highest Airline's and extrapolate our conclusions for the rest: American Airlines Inc. (AA), Delta Air Lines Inc. (UA), Delta Air Lines Inc. (DL), Skywest Airlines Inc. (OO).

To ensure the integrity of our findings, we performed a rudimentary level of data cleaning and preproccessing, addressing issues such as missing data, format discrepancies, and removing possible outliers from sample pool.

```
 DAY_OF_WEEK      FL_DATE          OP_UNIQUE_CARRIER  ORIGIN_AIRPORT_ID    ORIGIN          DEST_AIRPORT_ID     DEST
Min.   :1.000   Length:3231213     Length:3231213     Min.   :10135     Length:3231213     Min.   :10135   Length:3231213
1st Qu.:2.000   Class :character   Class :character   1st Qu.:11292     Class :character   1st Qu.:11292   Class :character
Median :4.000   Mode  :character   Mode  :character   Median :12478     Mode  :character   Median :12478   Mode  :character
Mean   :3.983                                        Mean   :12632                        Mean   :12633
3rd Qu.:6.000                                        3rd Qu.:14027                        3rd Qu.:14027
Max.   :7.000                                        Max.   :16869                        Max.   :16869

  DEP_DELAY         DEP_DEL15        ARR_DELAY          ARR_DEL15      ACTUAL_ELAPSED_TIME    AIR_TIME        DISTANCE
Min.   : -79.00   Min.   :0.0000   Min.   : -96.000   Min.   :0.0000   Min.   : 16.0       Min.   :  9.0   Min.   :  45.0
1st Qu.:  -5.00   1st Qu.:0.0000   1st Qu.: -15.000   1st Qu.:0.0000   1st Qu.: 92.0       1st Qu.: 66.0   1st Qu.: 422.0
Median :  -2.00   Median :0.0000   Median :  -6.000   Median :0.0000   Median :130.0       Median :103.0   Median : 733.0
Mean   :  13.79   Mean   :0.1983   Mean   :   8.232   Mean   :0.2052   Mean   :147.6       Mean   :120.8   Mean   : 886.8
3rd Qu.:   8.00   3rd Qu.:0.0000   3rd Qu.:   9.000   3rd Qu.:0.0000   3rd Qu.:183.0       3rd Qu.:154.0   3rd Qu.:1156.0
Max.   :4413.00   Max.   :1.0000   Max.   :4405.000   Max.   :1.0000   Max.   :749.0       Max.   :697.0   Max.   :4983.0

CARRIER_DELAY     WEATHER_DELAY       NAS_DELAY       SECURITY_DELAY    LATE_AIRCRAFT_DELAY
Min.   :   0.0    Min.   :   0      Min.   :   0.0    Min.   :  0.0     Min.   :   0.0
1st Qu.:   0.0    1st Qu.:   0      1st Qu.:   0.0    1st Qu.:  0.0     1st Qu.:   0.0
Median :   6.0    Median :   0      Median :   0.0    Median :  0.0     Median :   0.0
Mean   :  32.4    Mean   :   5      Mean   :  11.3    Mean   :  0.1     Mean   :  29.1
3rd Qu.:  29.0    3rd Qu.:   0      3rd Qu.:  15.0    3rd Qu.:  0.0     3rd Qu.:  31.0
Max.   :3957.0    Max.   :1653      Max.   :1708.0    Max.   :885.0     Max.   :3581.0
NA's   :2568313   NA's   :2568313   NA's   :2568313   NA's   :2568313   NA's   :2568313
```

Figure 1: Data summary

DATA DICTIONARY

DAY_OF_WEEK : 1 (Monday) - 7 (Sunday)

FL_DATE : Scheduled date

OP_UNIQUE_CARRIER : Unique carrier code

ORIGIN_AIRPORT_ID : Origin IATA code

ORIGIN : IATA(International Air Transport Association) airport code

DEST_AIRPORT_ID : Destination IATA code

DEST : Destination IATA code

DEP_DELAY : Difference in minutes between scheduled and actual departure time (in minutes)

DEP_DEL15 : Indicates delay in departure (0 = No, 1 = Yes)

ARR_DELAY : Difference in minutes between scheduled and actual arrival time

ARR_DEL15 : Indicates delay in arrival (0 = No, 1 = Yes)

ACTUAL_ELAPSED_TIME : Actual time an airplane spends in the air(in minutes) with TaxiIn/Out

AIR_TIME : Flight Time (in minutes)

DISTANCE : Distance between airports (miles)

CARRIER_DELAY : Flight delay due to carrier(e.g. maintenance or crew problems, aircraft cleaning, fueling, etc), 0 = No, yes = (in minutes)

WEATHER_DELAY : Flight delay due to weather, 0 = No, yes = (in minutes)

NAS_DELAY : Flight delay by NSA(National Aviation System), 0 = No, yes = (in minutes)

SECURITY_DELAY : Flight delay by this reason, 0 = No, yes = (in minutes)

LATE_AIRCRAFT_DELAY : Flight delay by this reason, 0 = No, yes = (in minutes)

1 Delay Proportion Two Sample Test

The purpose of this analysis is to see if we have enough evidence to support this claim *weekends have the same amount of delays as weekdays.*

We introduce a new column *Day* to indicate if the day is weekend or weekday. Also, note for the purpose of simplicity, we consider weekends to be only Saturday and Sunday.

1.1 Preliminary Analysis

Before the search for the evidence for the claim, we need first perform some preliminary analysis about the number delays per day of the arlines. We take a *random* sample of the dataset by picking the flights of the same origin and destination airports to make our conclusions more accurate.

Assumptions:

1. The sample is random.

   We picked a random set of airports from the total list for this analysis. A total of approx 365 data points of delays are collected.

2. The delays are normally distributed.

   Generated different random sample groups and picked the sample which is the closest match to a normal distribution.

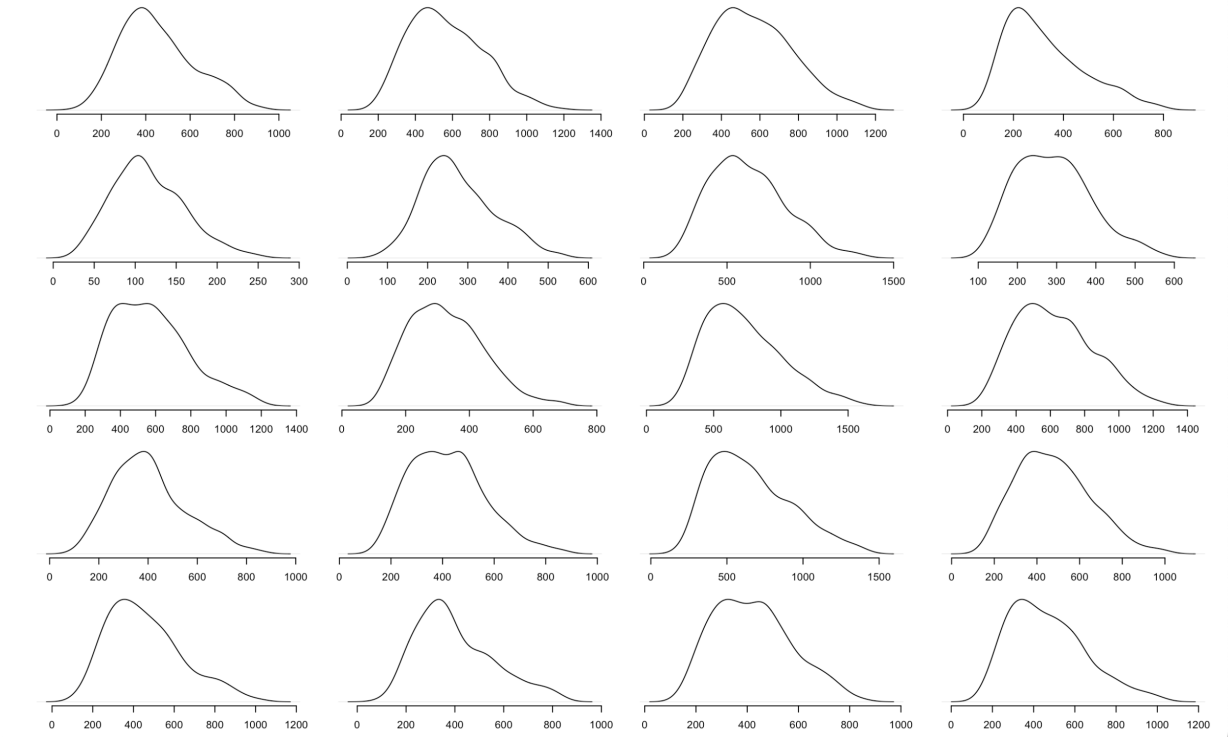
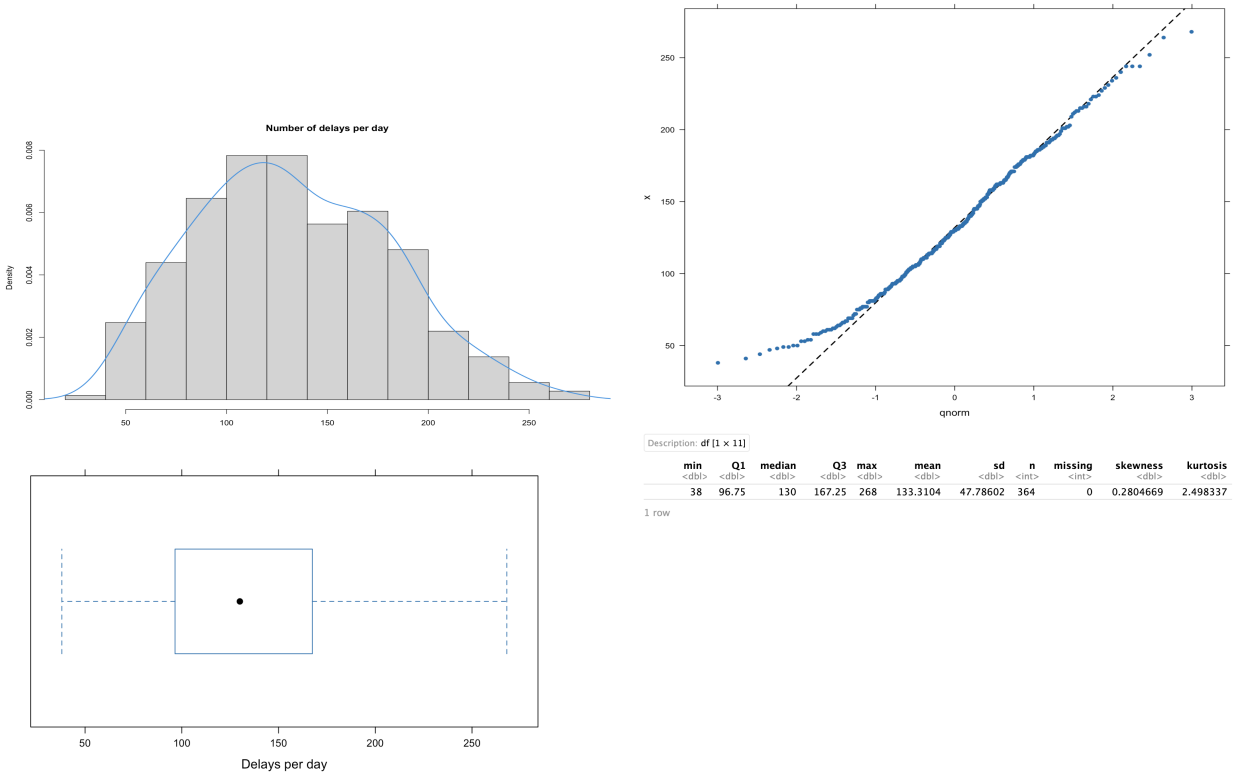


Figure 2: Random airlines delays density matrix plots

1.2 Sample Dataset

Based on the these random density plots, we select the best sample close to the normal population for our analysis.



Description: df [1 × 11]

min	Q1	median	Q3	max	mean	sd	n	missing	skewness	kurtosis
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>
38	96.75	130	167.25	268	133.3104	47.78602	364	0	0.2804669	2.498337

1 row

Figure 3: Sample dataset analysis

Analysis:

1. The data shown above is after we remove the $|Z_{score}| > 3$ outliers.
2. Histogram analysis.
   - 2.1 The distribution has a skewness of **0.28**, it is slightly skewed to the right, however, since it lies in this $[-0.5, 0.5]$ bound, we can treat this distribution to be nearly symmetrical.
   - 2.2 The distribution has a kurtosis of **2.49**, its a platykurtic thin-tailed, meaning that outliers are infrequent. Since its normal distribution has a kurtosis of 3.0, its is approximately a normal distribution.
3. The qqplot shows very few departures near both the ends, however, majority of the data points follow closely to the normal line, hence we can conclude the distribution is a normal distribution.
4. Box plot analysis
   - 4.1 The box plot shows mean (**133.3104** min) and median (**130** min). Since mean > median, the distribution is slightly skewed to right, however its very little.
   - 4.2 The inter-qartile range is $Q3 - Q1 = 167.25 - 66.75 = 100.5$ min. We can infer that there is good enough spread of the distribution.

We conclude that our sample is very close to a normal distribution and can proceed with the data preparation for the delays over weekends and weekdays.

1.3 Hypothesis Test Preparation

Now we separate our data into two independent sets, the number of delays on weekdays (MON - FRI) and weekends (SAT-SUN). Using the new column *Day*, we tabulate the total number of delays over weekdays and the

total number of delays for weekends. After separating the data points, we get approximately 364 (days) of data points. Once again we perform few preliminary analysis to check if the assumptions for the Hypothesis test are met.

1.4 Independence between Weekdays & Weekends Delays

Its would be prudent to check if there is any relationship between the weekdays/weekends & delays/no delays of the airlines. If there is no relationship between the delays on weekdays & weekends, then our hypothesis testing is more accurate as the proportions of the delays will be also independent. To check for independence, we perform a chi-squared test and analyze the results.

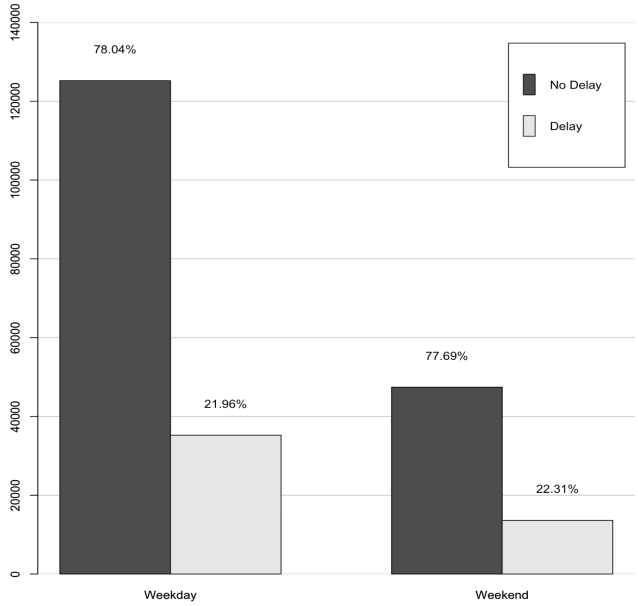
H_0 : Days and delays have no relationship.

H_A : Days and delays have relationship.

```
       day_of_week
delay    Weekday Weekend
  0       125228   47431
  1        35240   13624
  Total   160468   61055

        Pearson's Chi-squared test

data:  tally(delay ~ day_of_week)
X-squared = 3.2154, df = 1, p-value = 0.07295
```



Assumptions Checklist:

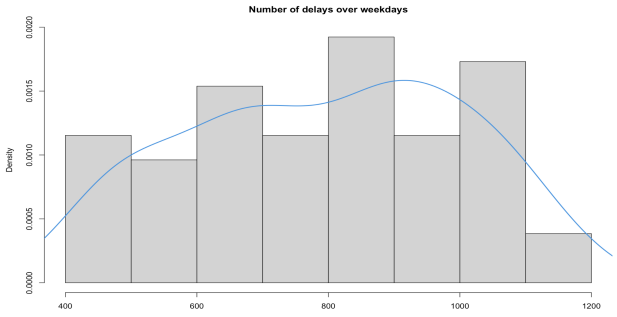
1. Data are counts for the categories of a categorical variable. The first data is in a Weekday/Weekend category, while the second data is in Delay/No delay category. This checks out.
2. The counts in the cells should be independent of each other. The data is essentially a Yes/No answer, so the data in the cells are independent of each other.
3. We should have a random sample. This checks out since we collected a random set of data points of the airlines.
4. We should expect to see at least 5 individuals in each cell. Every cell have enough data to perform the test.

Conclusion:

Since the P-value $0.0729 > 0.05$, we will do reject H_0 and conclude that days and delays are independent for this sample. This implies that the reasons of delays over weekdays and weekends are also independent, which will give more credibility to our hypothesis testing of delay proportions.

1.5 Delay Analysis Of The Two Groups

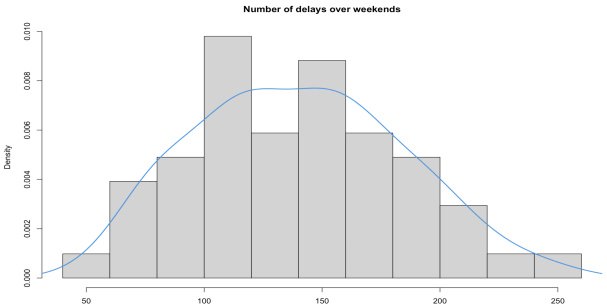
This is our final analysis before the hypothesis testing. As we have concluded that weekday delays are independent of weekends delays, we also need to check for the two sample distributions for normality. For this part of analysis, we collect the number delays/no delays from MON-FRI and total delays/no delays from SAT-SUN per week. We get a time series delay data for both the groups over approx 52 weeks. We also check for any outliers in each of the groups.



Description: df [1 × 11]

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>	skewness <dbl>	kurtosis <dbl>
440	649.5	814	955.25	1166	794.8269	206.6585	52	0	−0.1135632	1.899385

1 row



Description: df [1 × 11]

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>	skewness <dbl>	kurtosis <dbl>
59	113.75	141	171	268	142.8462	46.4232	52	0	0.4262299	2.799522

1 row

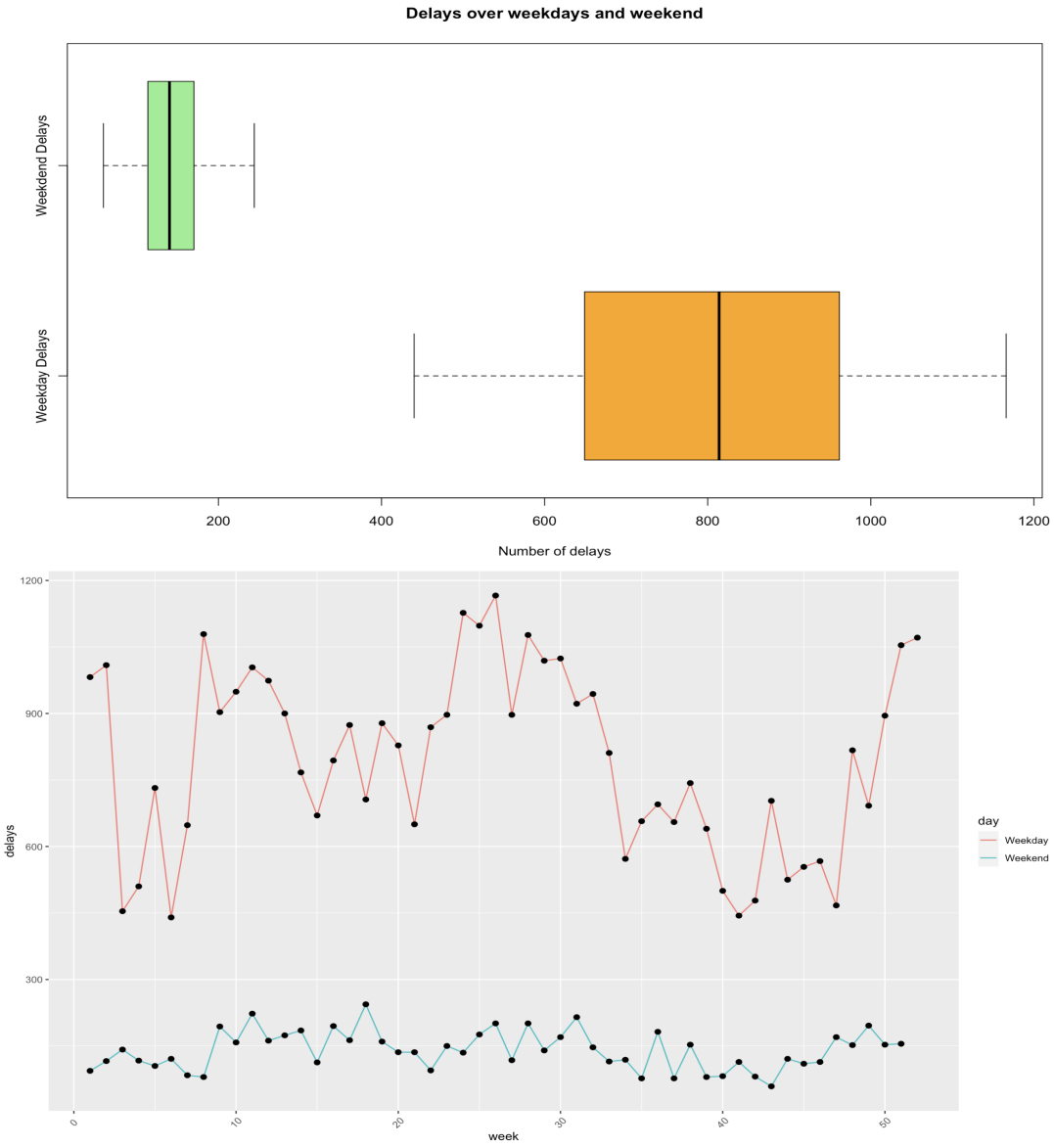


Figure 4: Two groups comparison

Analysis:

1. Histograms plot:
   1.1 Weekday delays has a skewness of **-0.113** (slightly left skewed) and kurtosis of **1.9** (medium tailed). We can consider this distribution to be nearly symmetrical and approximately normal.
   1.2 Weekday delays has a skewness of **0.426** (slightly right skewed) and kurtosis of **2.8** (thin tailed). We can again consider this distribution to be nearly symmetrical and approximately normal.
2. Box plot:
   2.1 We see that weekend delays have very small spread compared to weekday delay.
   2.2 Weekday delays are bounded between $[450, 1200]$ and weekend delays are bounded between $[50, 300]$.
   2.3 At a glance we see that weekends have less delays on average than weekdays. However, we will not conclude this yet since there is one more factor we have yet to consider.
3. Delay time series shows that on average every week has more weekday delays than weekend delays.

Conclusion:

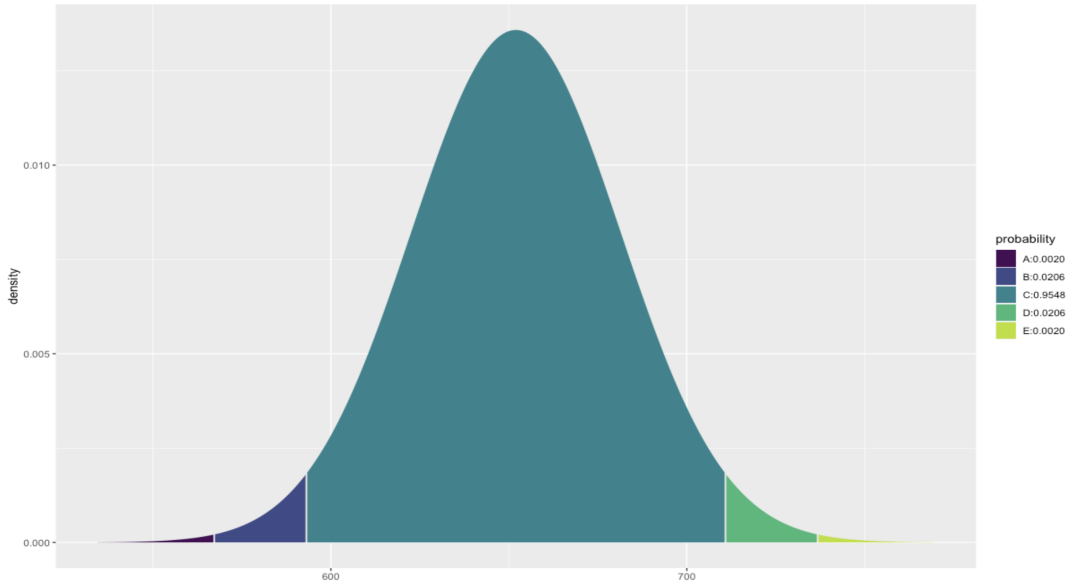
We can conclude that the two groups are independent as well as normally distributed. We perform the hypothesis testing on the two groups.

1.6 Two Sample T-test for means

$$H_0 : \mu_{weekdaydelay} - \mu_{weekenddelay} = 0$$
$$H_A : \mu_{weekdaydelay} - \mu_{weekenddelay} \neq 0$$

```
        Welch Two Sample t-test

data:  weekday_delays and weekend_delays
t = 22.197, df = 56.134, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 593.1436 710.8180
sample estimates:
mean of x mean of y
 794.8269  142.8462
```



As the P-value $2.2e-16 \sim 0$, we reject H_0 and conclude that we have enough evidence to show that the means of weekday delay is not the same as the mean of weekend delays. Since difference of means lies in $[593.14, 719.8]$ which is always positive, we have strong reason to believe that the average delays over weekdays is significantly more than the average delays over weekends,

However, this test does not back up the claim of our hypothesis. This is because we compared the number of delays without considering the total number of flights over weekdays and weekends. Its plausible that weekends have less flights scheduled by the airlines. And due to less number of flights, there is less number of delays as well. This implies that while he have higher average number of delays on weekdays, it might be not true for the proportion of delays.

1.7 Two Sample Proportion Test

Instead of testing means, we compare the proportion of delays over weekdays and weekends.

$$\text{Proportion of weekday delays} = \frac{\text{Number of weekday delays}}{\text{Total flights scheduled on weekdays}}$$

$$\text{Proportion of weekend delays} = \frac{\text{Number of weekend delays}}{\text{Total flights scheduled on weekends}}$$

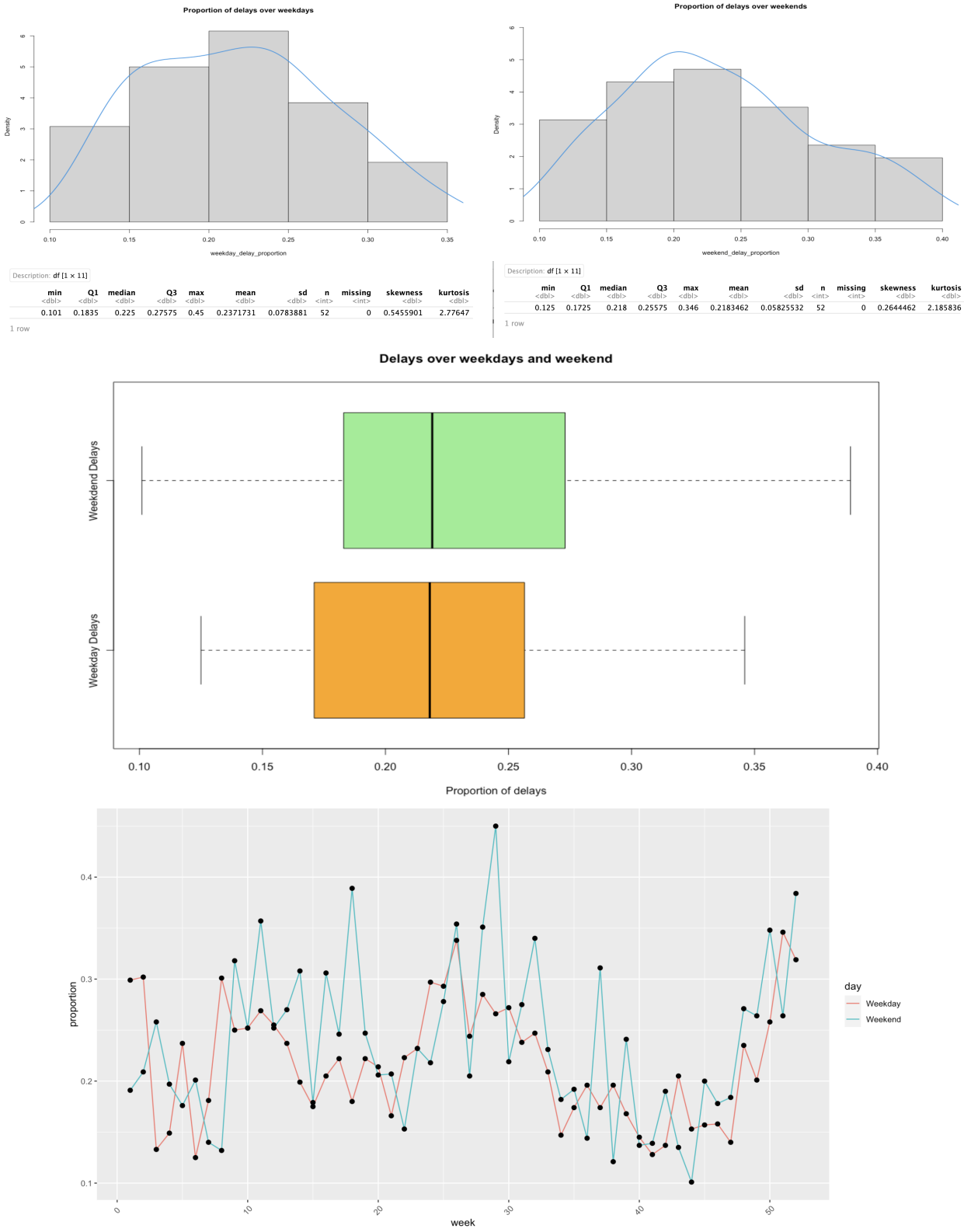


Figure 5: Two groups proportion comparison

Analysis:

1. Histograms plot:
   1.1 Weekday proportion delays has a skewness of **0.54** (slightly right skewed) and kurtosis of **2.77** (medium tailed). We can consider this distribution to be nearly symmetrical and approximately normal.
   1.2 Weekday proportion delays has a skewness of **0.26** (slightly right skewed) and kurtosis of **2.18** (thin tailed). We can again consider this distribution to be nearly symmetrical and approximately normal.
2. Box plot:
   2.1 We see that weekend proportion delays median is almost same compared to weekday delay.
   2.2 weekend proportion also has similar IQR, spread.
3. When comparing the proportions of delays per week, we can see now that the gap between delays of weekdays and weekends has decreased. Hence, this test is more appropriate the check if there are more delays between the two groups or not.

Conclusion:

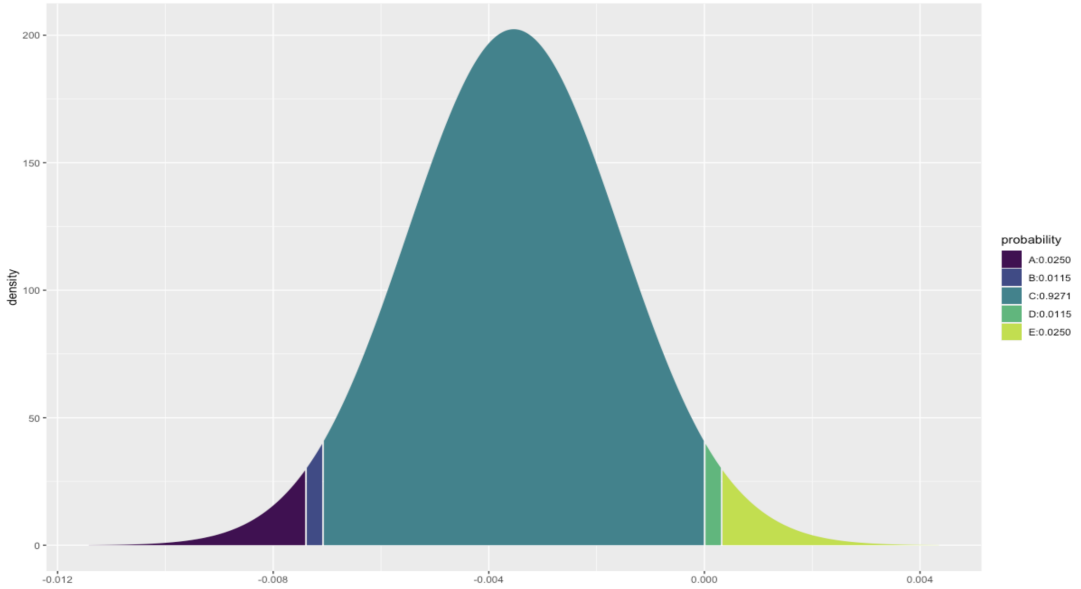
We can conclude that the two groups are independent as well as normally distributed. We perform the hypothesis testing on the two groups.

$$H_0 : P_{weekdaydelay} - P_{weekenddelay} = 0$$
$$H_A : P_{weekdaydelay} - P_{weekenddelay} \neq 0$$

```
        2-sample test for equality of proportions without continuity correction

data:  c out of cp1 out of n1p2 out of n2
X-squared = 3.2154, df = 1, p-value = 0.07295
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0074096344  0.0003387942
sample estimates:
   prop 1    prop 2
0.2196076 0.2231431
```

Figure 6: Two groups proportion comparison



As the P-value 0.07295 > 0.05, we do not reject H_0. We have don't have enough evidence to disprove that *weekends have the same amount of delays as weekday.*

1.8 Conclusion

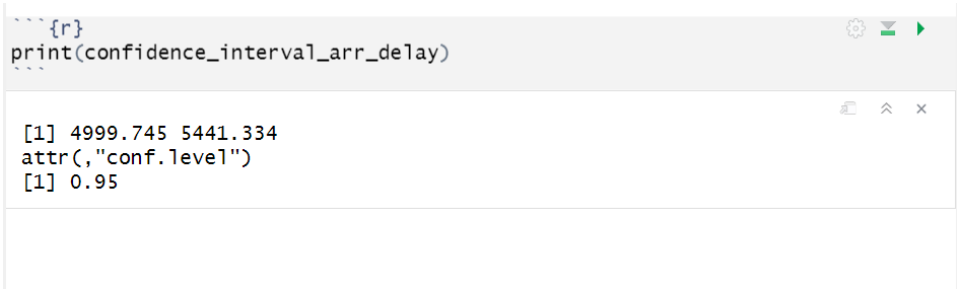
To summarise, while the average number of delays over weekdays are more, the proportion of delays is nearly same for weekdays & weekends delays.

2 ESTIMATE THE MEAN WITH A CONFIDENCE INTERVAL

2.1 ANALYSIS

We're calculating the mean and its confidence interval specifically for the 'arr_delay' column in the given dataset. We extract the column and compute the mean and confidence interval using the t.test() function in R.This code uses a 95

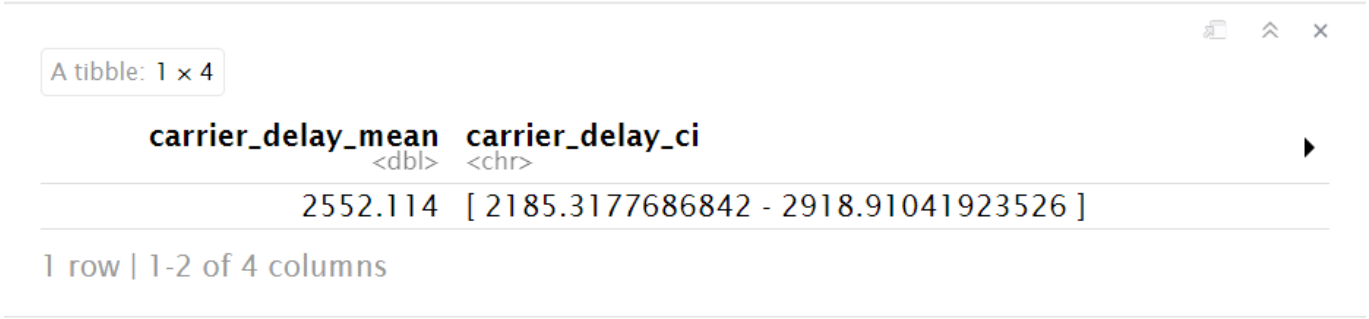
This will display the confidence interval for the 'arr_delay' variable in the R console.



The confidence interval represents a range of values within which the true delay parameter (in our case, the true mean of 'arr_delay') is likely to fall with a certain level of confidence (e.g., 95 percent confidence level). The values 4999.745 and 5441.334 represent the confidence interval for arr_delay(likely the mean of a dataset). This interval ranges from approximately 4999.745 to 5441.334.

The attr(,conf.level") line shows that the confidence level associated with this interval is 0.95, which means it's a 95 percent confidence interval. We can see that the lower bound is 4999.745 and the 5441.334 . It implies that we are 95 percent confident that the true delay means lies within the interval [4999.745, 5441.334].

Now, we are carrying out an analysis of filtering the airports to be "AUS", "MSP", "BNA", "PIT" and we mutate 2 delays(carrier_delay and the weather_delay). Now we try to calculate the mean and confidence interval for 'carrier_delay' and 'weather_delay' for the filtered airports. We can see the carrier_delay_mean and the carrier_delay_ci .



Likewise, we calculate for the weather_delay_mean and the carrier_delay_ci

1. For 'weather_delay':
   - Mean: The mean value for 'weather_delay' is approximately 404.2671 minutes. Confidence Interval: The confidence interval for
   - 'weather_delay' is given as [311.83528465823 - 496.698943529689].

- This means that we are 95 percent confident that the true proportion mean of 'weather_delay' lies within this range. The lower bound of this interval is around 311.83, and the upper bound is around 496.70.

2. For 'carrier_delay':
   - Mean: The mean value for 'carrier_delay' is approximately 2552.114 minutes.
   - Confidence Interval: The confidence interval for 'carrier_delay' is given as [2185.3177686842 - 2918.91041923526].
   - This interval indicates a 95 percent confidence level that the true population mean of 'carrier_delay' is within this range. The lower bound of this interval is around 2185.32, and the upper bound is around 2918.91.

These values provide estimates of the means for 'weather_delay' and 'carrier_delay', along with ranges (confidence intervals) that are likely to contain the true proportion means with a 95 percent confidence level based on the observed sample data for the specifically selected airports. The intervals give us a measure of the precision and uncertainty associated with our estimates of the population means for these delays.

2.2 Interpretation

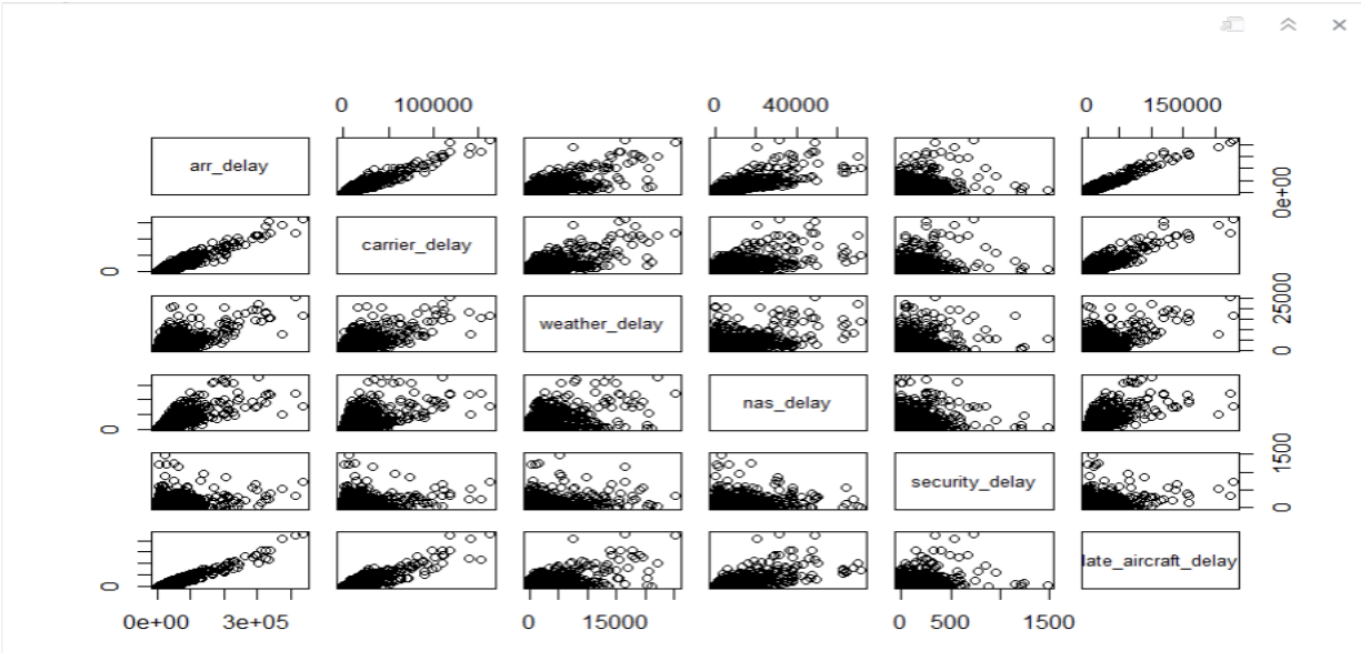
The wider the confidence interval, the more variability and uncertainty in our estimate of the population mean. The narrower the interval, the more precise and confident we are in our estimate of the true population mean.

2.3 Conclusion

By comparing the mean delays and their confidence intervals, we can assess the relative impact of weather-related delays versus carrier-specific delays for the selected airports. Any overlapping or non-overlapping portions between the confidence intervals indicate significant differences or similarities between these types of delays. Overall, these calculations provide valuable information about average delays and their variability, helping in understanding and potentially addressing the factors contributing to delays at these specific airports.

3 SLR Analysis on Numerical Data

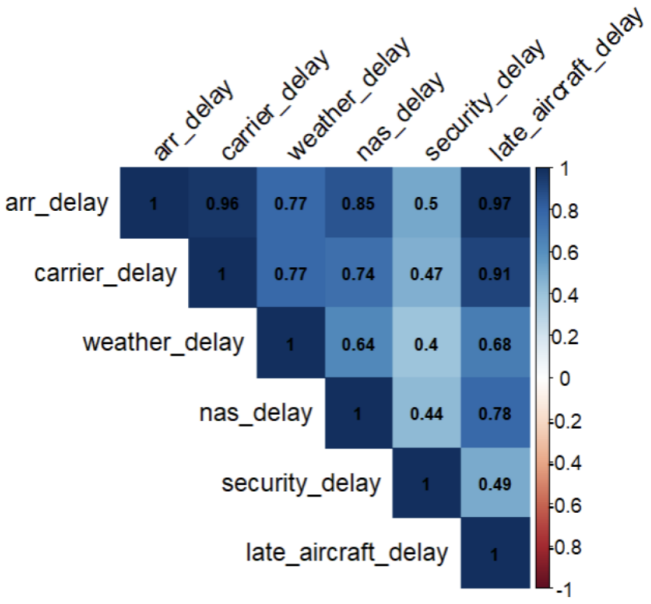
Here we trained a linear regression model considering the arrival_delay as dependent variable and the other delays as the explanatory variables. The results are as follows: The linear regression graph is found by taking into account the arr_delay vs all the other delays(carrier_delays, weather_delay, nas_delay, security_delay, and the late_airfact_delay.



Now, we find the correlation_matrix of the arr_delay.

```
                    arr_delay carrier_delay weather_delay nas_delay
arr_delay           1.0000000     0.9646240     0.7726996 0.8547342
carrier_delay       0.9646240     1.0000000     0.7711933 0.7435485
weather_delay       0.7726996     0.7711933     1.0000000 0.6367367
nas_delay           0.8547342     0.7435485     0.6367367 1.0000000
security_delay      0.5035365     0.4719425     0.3971712 0.4371275
late_aircraft_delay 0.9728146     0.9141268     0.6841897 0.7789705
                    security_delay late_aircraft_delay
arr_delay                0.5035365           0.9728146
carrier_delay            0.4719425           0.9141268
weather_delay            0.3971712           0.6841897
nas_delay                0.4371275           0.7789705
security_delay           1.0000000           0.4922447
late_aircraft_delay      0.4922447           1.0000000
```

Correlation plot



3.1 Model Diagnostics

The summary of the linear regression model is inferred taking all the delays and this is the summary .

```
Call:
lm(formula = arr_delay ~ carrier_delay, data = Airline_Delay_Cause)

Residuals:
   Min     1Q Median     3Q    Max
-75658   -552   -150    248 103704

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.504e+02  3.122e+01   4.818 1.46e-06 ***
carrier_delay 2.645e+00  5.029e-03 526.046  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4270 on 20669 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.9305,    Adjusted R-squared:  0.9305
F-statistic: 2.767e+05 on 1 and 20669 DF,  p-value: < 2.2e-16
```

We can see that the Intercept (Estimate): The intercept estimate is 150.4(1.504e+02 in scientific notation). In the context of this model, it suggests that when the carrier_delay (in minutes) is zero, the predicted arr_delay is 150.4

minutes. However, it's important to note that the value of zero for carrier_delay might not be practically possible. However, we find the slope estimate of the carrier delay.

Slope (carrier_delay): The slope estimate is 2.645(2.645e+00 in scientific notation). For each additional minute increase in carrier_delay, the predicted arr_delay is estimated to increase by approximately 2.645 minutes, assuming all other variables remain constant.

The R square value is 0.9305 or 93.05 %. This means that approximately 93.05 % of the variance in the 'arr_delay' can be explained by the linear relationship with 'carrier_delay'. In other terms, we can say the 'carrier_delay' variable in this model explains about 93.05% of the variability observed in the 'arr_delay'. The higher the R^2 value(closer to 1) the better the model fits the data in explaining the variability of the dependent variable. This quantifies the strength of the relationship between variables in the model.

3.2 Assumptions used for linear regression

1. Linearity: TRUE. The scatter plot of carrier_delay versus arrival_delay in portion demonstrates this. They are linked in a straight line.

2. Residuals are distributed normally: FALSE. The residuals distribution is biased i.e skewed to the left (the left side of the figure contains more data points). From the QQ plot the observed data points clearly deviate from the straight line representing the theoretical quantiles of a normal distribution. This confirms that the distribution of the residuals in your data is not normal.

3. Relative independence: FALSE. The scatter plot clearly shows that there is no association between the various residuals.

4. The residual variance is constant: FALSE. The spread of the residuals appears to be somewhat wider at higher fitted values. This could be an indication of non-constant variance, but also be due to the outliers or a non-linear relationship between the independent and dependent variables. There seems to be a slight tilting of the residuals upwards as the fitted values increase. Hence, the scatter plot of carrier_delay vs arrival_delay (fitted model) shows that the variance of residuals is not constant.

5. The expected value of the residuals must be zero: TRUE. According to the histogram, the distribution of residuals appears to be centered around zero, implying that the mean residual is near zero.

3.3 Conclusion

In conclusion, the linear model's assumptions are false. As a result, the linear model is applicable. Overall, the linear regression model fails to meet several assumptions, including normality of residuals, independence among residuals, and constant variance. As a result, the model might not be suitable or reliable for making inferences about the relationship between 'carrier_delay' and 'arrival_delay'.

4 ANOVA ANALYSIS

Here we're try to find out is there a difference in arr_delay among the different airports

4.1 OBJECTIVE

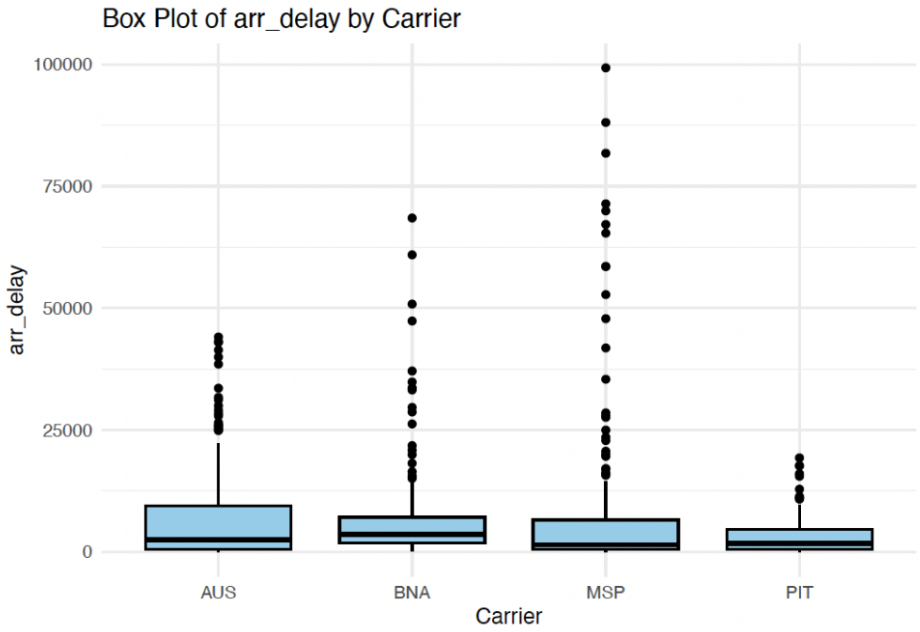
Null Hypothesis (H_0): There is no significant difference in the mean "arr_delay" among the different airports.

Alternative Hypothesis (H_a): There is a significant difference in the mean "arr_delay" among at least some of the airports.

Explanation: H0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ (where μ represents the mean "arr_delay" for each airport) H1: At least one pair of means μ_i and μ_j is different.
Here, $\mu_1, \mu_2, \mu_3 ... \mu_k$ represent the mean "arr_delay" for each airport category. If the p-value is less than the significance level (commonly used in our Stat501 class: 0.05), we reject the null hypothesis and conclude that there is evidence to suggest that at least some of the means are different.

4.2 ANALYSIS W/ OUTLIERS

First, we have to check the box plot to understand the distribution of each airport and check for the outliers.



Through this box plot, we can see that we have some outliers. So, we're making ANOVA analysis with and without Outlier to make the correct conclusion

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## airport       3 2.789e+09 929607159   7.144 9.84e-05 ***
## Residuals   741 9.642e+10 130126503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2.1 INTERPRETATION:

- Df (Degrees of Freedom): There are three degrees of freedom for the factor "airport" and 741 degrees of 2 freedom for residuals.
- Sum Sq (Sum of Squares): This represents the sum of squared differences between the observed values and the mean. For "airport," it is 2.789e+09, and for residuals, it is 9.642e+10. Mean Sq

- (Mean Square): Mean Squares are calculated by dividing the Sum of Squares by the corresponding degrees of freedom.
- F value: The F statistic is a ratio of the variance between groups to the variance within groups. Here, it is 7.144.
- Pr(>F): This is the p-value associated with the F statistic. It is extremely small (9.84e-05), indicating that there is a significant difference in mean "arr_delay" among at least two airports.

Since we are rejecting the Null Hypothesis, we want to investigate further. The method used for investigating is Tukey's post-hoc test

4.3 Tukey's post-hoc test

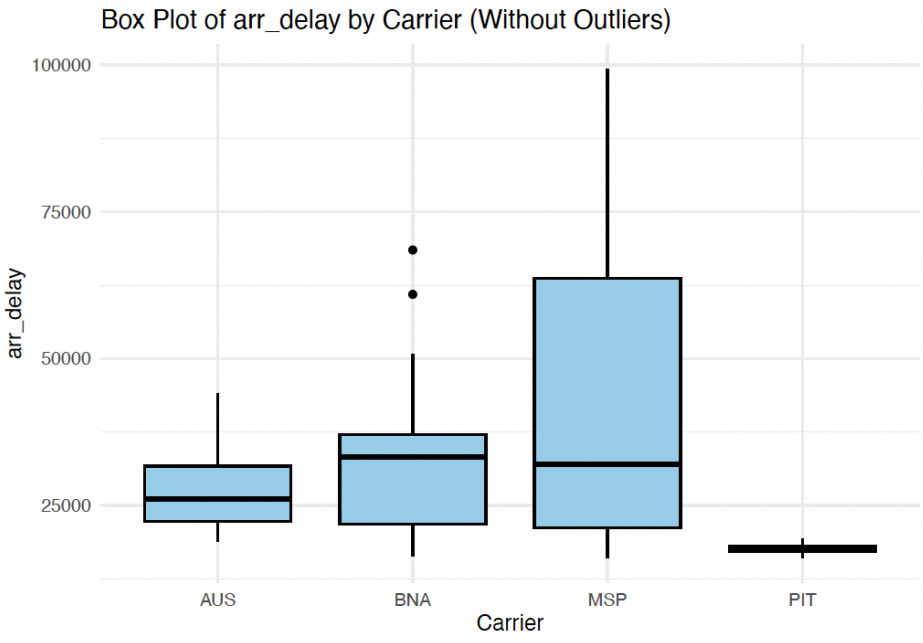
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = arr_delay ~ airport, data = raw_top4_airport)
##
## $airport
##                diff       lwr        upr     p adj
## BNA-AUS  -338.3429 -3380.036  2703.3500 0.9918084
## MSP-AUS   989.2646 -2052.428  4030.9575 0.8366112
## PIT-AUS -4109.3429 -7151.036 -1067.6500 0.0029908
## MSP-BNA  1327.6075 -1718.160  4373.3750 0.6758373
## PIT-BNA -3771.0000 -6816.767  -725.2325 0.0081176
## PIT-MSP -5098.6075 -8144.375 -2052.8401 0.0001087
```

4.3.1 Interpretation:

- PIT-AUS and PIT-MSP have differences in mean of "arr_delay" that are statistically significant. P values are 0.00299 and 0.0001087 (<0.005) respectively.

4.4 Analysis w/o Outliers

Box plot to analyze:



The Anova analysis report:

```
##              Df   Sum Sq   Mean Sq F value  Pr(>F)
## airport       3 4.03e+09 1.343e+09   4.519 0.00584 **
## Residuals    72 2.14e+10 2.972e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4.1 Interpretation:

The p-value (0.00584) is still less than 0.05, indicating that there is a significant difference in mean "arr_delay" among at least two airports

4.5 Tukey multiple comparisons of Means

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = arr_delay ~ airport, data = flight_data_no_outliers)
##
## $airport
##                diff        lwr       upr     p adj
## BNA-AUS   5540.235  -8310.797 19391.268 0.7195213
## MSP-AUS  14405.346   2158.554 26652.139 0.0146183
## PIT-AUS -11017.500 -35202.977 13167.977 0.6300320
## MSP-BNA   8865.111  -5278.157 23008.379 0.3584351
## PIT-BNA -16557.735 -41756.652  8641.182 0.3168312
## PIT-MSP -25422.846 -49776.865 -1068.827 0.0374062
```

4.5.1 Interpretation:

- The p-value for the pair MSP-AUS is 0.0146183 and PIT-MSP 0.0374062, which are less than 0.05. This suggests a significant difference in mean "arr_delay" between Minneapolis (MSP) and Austin (AUS), PIT and MSP respectively .

4.6 Anova Conclusion:

- Thus we conclude that there is a significant difference in the mean "arr_delay" among at least some of the airports and we reject the Null hypothesis.

5 Conclusion

With the ensemble of analysis conducted here, we have covered the below statistical analysis.

- Two sample mean testing between weekend & weekday delays.
- Two sample proportion testing between weekend & weekday delays.
- Chi square test of weekend / weekday & delays / No delays.
- Confidence Interval of delays.
- Simple Linear Regression model to predict delays.
- Anova analysis for comparing delays between different carriers.