Statistical Analysis of Airline Delays

501 Final Project Report

Group Members

Aman Singh Thakur - amansinghtha@umass.edu
Astha Baranwal - abaranwal@umass.edu
Samriddhi Raj - samriddhiraj@umass.edu

Table of Contents

1. Introduction

We will be examining the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) dataset, specifically the monthly Air Travel Consumer Report. This report focuses on the punctuality of domestic flights operated by major airlines within the United States. It contains valuable information regarding the number of flights that were on time, delayed, canceled, or diverted, as well as the contributing factors to these delays.
Our research aims to address the following question:

What are the main reasons behind flight delays in the domestic air travel industry?

For our hypothesis testing, we'll be using two popular airlines: Delta Airlines and American Airlines. Further, This analysis will provide us with insights into the correlation between delay factors and the duration of arrival delays. Such understanding will enable stakeholders within the aviation industry to identify the causes of delays, prioritize their efforts, and implement targeted strategies to minimize delays and enhance overall performance. Flight delays not only inconvenience passengers but also have significant economic implications. They can result in missed connections, disruptions in travel plans, increased costs for airlines, and decreased customer satisfaction.

Therefore, by analyzing the dataset and examining various delay factors like carrier delay, weather delay, NAS (National Airspace System) delay, security delay, and late aircraft delay, we can gain valuable insights into the factors contributing to flight delays. Understanding these factors is crucial for airlines, airport authorities, policymakers, and passengers alike, as it allows for the improvement of flight operations, reduction of delays, and enhancement of the overall travel experience.

2. Data & Visualization

The dataset contains recent information on flights operated by large air carriers in the United States from Jan 2023 to March 2023. There are in total **4612 total row** in the dataset corresponds to a specific flight in a given month and includes the following **21 variables**:

Column	Description	Categories	Min	Median	Mean	Max
year	the year in which the flight took place	-	2,023.00	2,023.00	2,023.00	2,023.00
month	the month in which the flight took place	-	1.00	2.00	2.00	3.00
carrier	the airline carrier code (Character Type)	15.00	-	-	-	-
carrier_name	the name of the airline carrier (Character Type)	15.00	-	-	-	-
airport	the airport code (Character Type)	341.00	-	-	-	-
airport_name	the name of the airport (Character Type)	341.00	-	-	-	-
arr_flights	the total number of flights arriving at the airport	-	1.00	93.00	352.10	18,504.00

arr_del15	the number of flights arriving at the airport delayed by 15 minutes or more	-	0.00	20.00	75.04	3,529.00
carrier_ct	the average number of delayed flights due to the carrier	-	0.00	7.67	25.14	904.87
weather_ct	the average number of delayed flights due to weather conditions	-	0.00	0.64	2.47	163.00
nas_ct	the average number of delayed flights due to the National Aviation System (NAS)	-	0.00	4.01	21.44	999.84
security_ct	the average number of delayed flights due to security issues	-	0.00	0.00	0.22	14.55
late_aircraft_ct	the average number of delayed flights due to late aircrafts	-	0.00	5.45	25.77	1,550.65
arr_cancelled	the number of canceled flights	-	0.00	1.00	5.80	580.00
arr_diverted	the number of diverted flights	-	0.00	0.00	0.81	84.00
arr_delay	the total number of minutes of delay for all arriving flights	-	0.00	1,206.00	4,997.00	337,375.00
carrier_delay	the total number of minutes of delay due to carrier issues	-	0.00	437.00	1,802.00	118,554.00
weather_delay	the total number of minutes of delay due to weather conditions	-	0.00	30.00	280.60	20,423.00
nas_delay	the total number of minutes of delay due to the National Aviation System (NAS)	-	0.00	159.00	1,002.00	50,792.00
security_delay	the total number of minutes of delay due to security issues	-	0.00	0.00	10.78	1,477.00
late_aircraft_delay	the total number of minutes of delay due to late aircrafts	-	0.00	369.00	1,901.00	155,262.00

Table 1 - Dataset Statistics

In the given airlines dataset, the response variable is "**arr_delay**" which represents the arrival delay in minutes. All the other variables are the explanatory variables, also known as independent variables or features. These explanatory variables provide contextual information and potential factors that could contribute to the arrival delay of flights. We are mainly considering these explanatory variables for our analysis : **carrier_delay , weather_delay , nas_delay , security_delay , late_aircraft_delay.** These are the potential factors that could contribute to the arrival delay of flights.

We have created box plots to visualize the distribution of delays (e.g., "arr_delay", "carrier_delay", etc.) for all flights. These plots help in understanding the spread and skewness of the delay data, along with identifying any outliers.

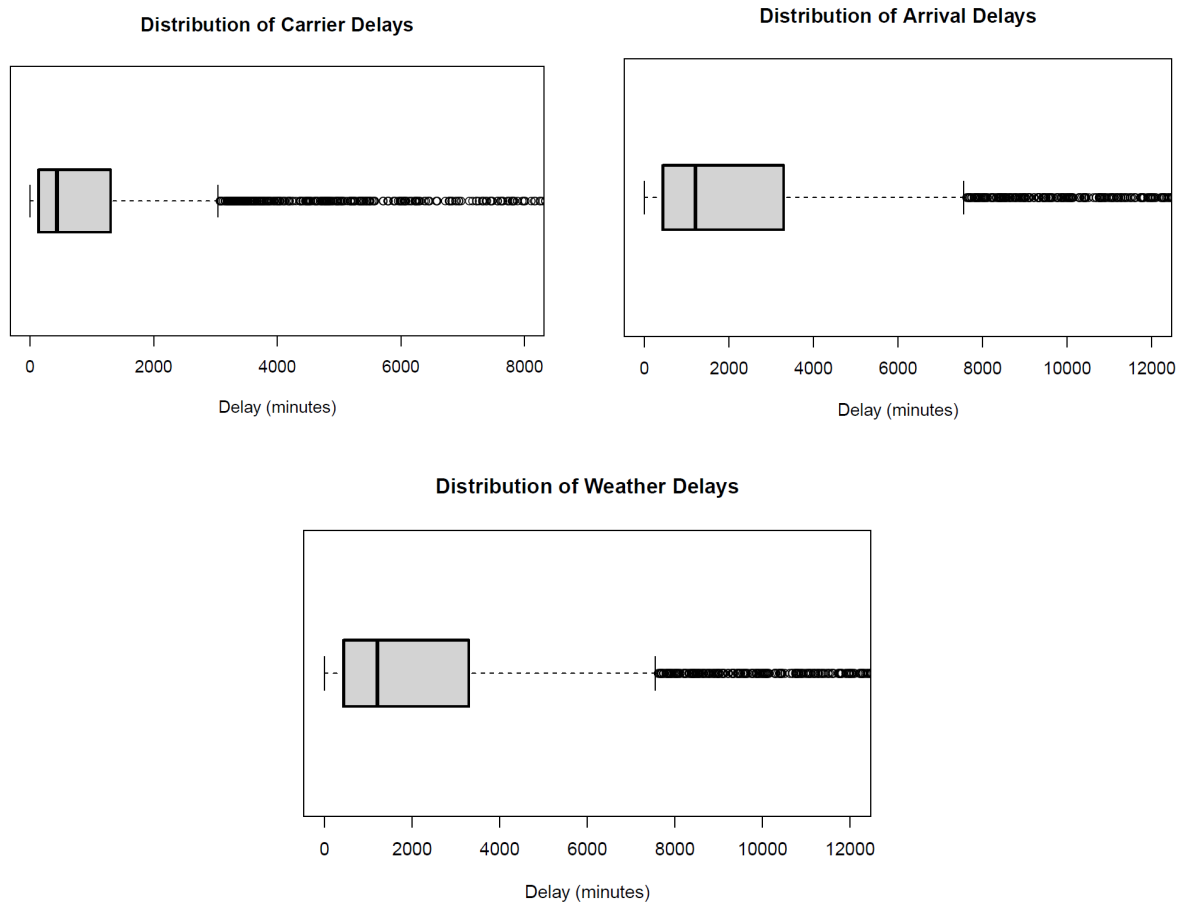


Figure 1 - Box Plots for Visualization of Critical Delay Causing Variables

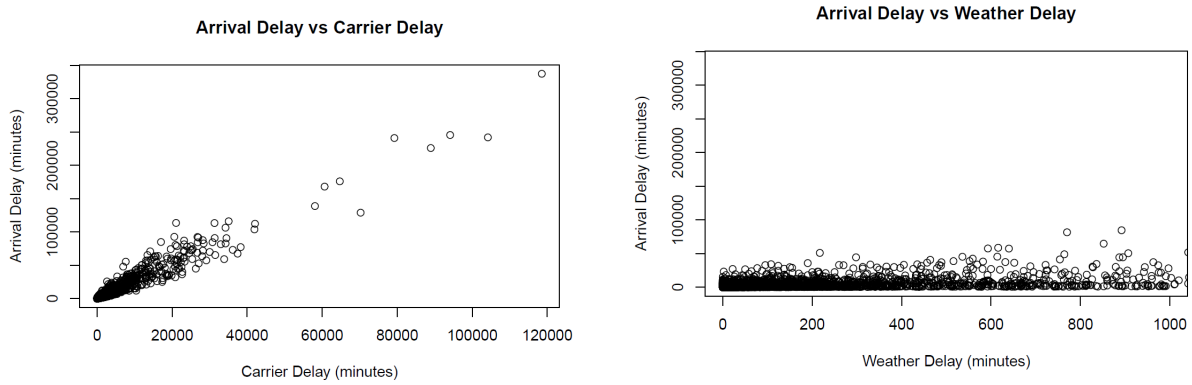
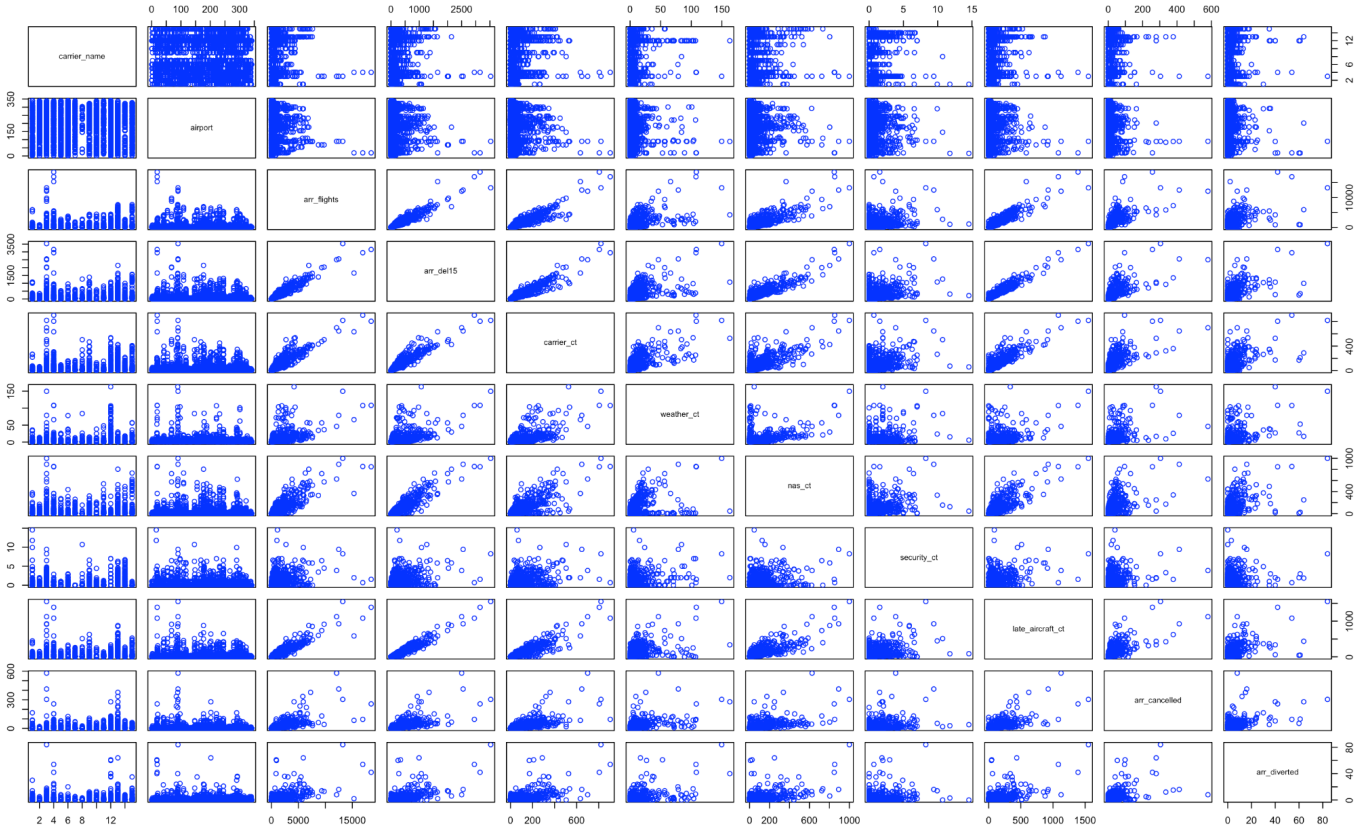


Figure 2 - Scatter plot for Visualizing patterns in Arrival Delays

3. Analysis

Correlation Analysis



	carrier_delay	weather_delay	nas_delay	security_delay	late_aircraft_delay
carrier_delay	1.0000000	0.7406985	0.7743065	0.4501237	0.9234710
weather_delay	0.7406985	1.0000000	0.5202790	0.4013624	0.5888050
nas_delay	0.7743065	0.5202790	1.0000000	0.3883256	0.8175952
security_delay	0.4501237	0.4013624	0.3883256	1.0000000	0.4364896
late_aircraft_delay	0.9234710	0.5888050	0.8175952	0.4364896	1.0000000

Figure 3 - Correlation Matrices for all Critica Variables Causing Delays

These two correlation matrices reveal variations in the frequency of flight delays among different airports and carriers. It suggests that certain airports and carriers tend to experience a higher number of delays compared to others. Additionally, the scatter plot matrix indicates a positive relationship between the number of airplane arrivals and delays across multiple factors. Similarly, the table matrix demonstrates a similar pattern when comparing the total duration of delays across different factors.

Linear Regression Analysis

For our analysis, we perform a **linear regression analysis** to predict the arr_delay (arrival delay) based on the delay factors carrier_delay, weather_delay, nas_delay, security_delay, and late_aircraft_delay.

Through our R code, we fit 5 linear regression models using the lm function, specifying the response variable (arr_delay) and the predictor variables (carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay) from the dataset.

This analysis helps assess the relationship between the delay factors and the arrival delay, and it provides information about the significance and magnitude of the effects of each delay factor on the arrival delay. From the below figure, we can use the Slope and intercept to build a linear regression model using equation -

$$\hat{Y}_{arr\_delay} = Slope \times \hat{X}_{variable} + Intercept$$

For example, for model arr_delay ~ security_delay, slope is 141.454 indicating that for every 1 minute increase in security delay, the arrival delay is expected to increase by approximately 141.454 minutes.

Linear Regression Model	Slope	Intercept	R-Square Value (%)	Correlation Coefficient
arr_delay ~ carrier_delay	2.65278	215.55778	94.15	0.9703092291
arr_delay ~ weather_delay	10.3403	2095.2375	48.74	0.698140387
arr_delay ~ weather_delay	4.02231	967.63751	77.07	0.8778952101
arr_delay ~ security_delay	141.454	3471.691	21.26	0.4610856753
arr_delay ~ late_aircraft_delay	2.34028	547.24086	94.73	0.9732933782

Table 2 - Linear Regression Model Summaries

Further, we can use the below R compiled R output to find the R^2 and the correlation coefficient, summarized above. **We can see that carrier_delay, weather_delay and late_aircraft_delay would be good fit for linear regression values as they have high correlation coefficients and comparatively fewer outliers.** For example, late_aircraft_delay has an R2 value of 0.9473, indicating that approximately 94.73% of the variability in the arrival delay can be explained by the variation in late_aircraft_delay and the use of regression model.

```
Call:
lm(formula = arr_delay ~ carrier_delay, data = airlines_data_1m)

Residuals:
   Min     1Q Median     3Q    Max
-57616   -548   -191    223  57544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   215.55778   54.86562   3.929 8.66e-05 ***
carrier_delay   2.65278    0.00974 272.365  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3528 on 4605 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9416,    Adjusted R-squared:  0.9415
F-statistic: 7.418e+04 on 1 and 4605 DF,  p-value: < 2.2e-16
```

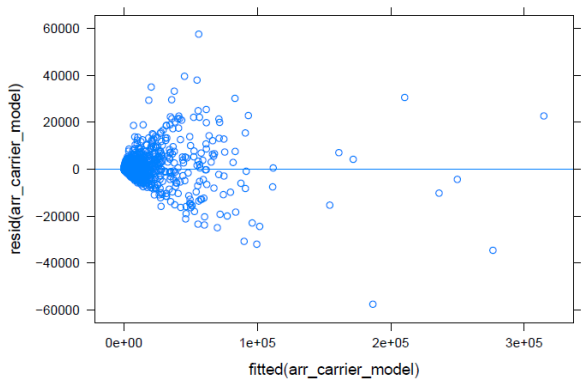
```
Call:
lm(formula = arr_delay ~ weather_delay, data = airlines_data_1m)

Residuals:
    Min      1Q  Median      3Q     Max
-110151   -2062   -1577    -206  172750

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2095.2375   160.0484   13.09   <2e-16 ***
weather_delay   10.3403     0.1562   66.18   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 10450 on 4605 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.4875, Adjusted R-squared:  0.4874
F-statistic:  4380 on 1 and 4605 DF,  p-value: < 2.2e-16
```

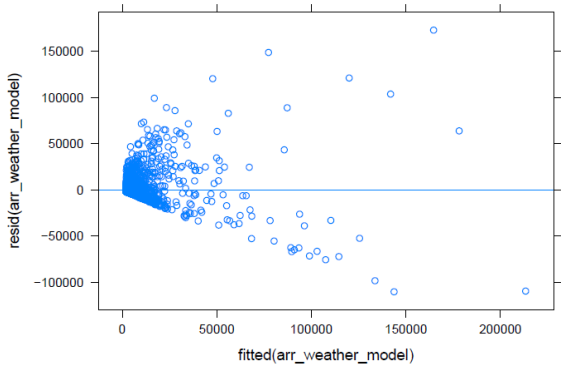
```
Call:
lm(formula = arr_delay ~ nas_delay, data = airlines_data_1m)

Residuals:
   Min     1Q Median     3Q    Max
-70239  -1025   -718    176 145597

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 967.63751  107.92052   8.966   <2e-16 ***
nas_delay     4.02231    0.03233 124.428   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6988 on 4605 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.7707,    Adjusted R-squared:  0.7707
F-statistic: 1.548e+04 on 1 and 4605 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = arr_delay ~ security_delay, data = airlines_data_1m)

Residuals:
    Min      1Q  Median      3Q     Max
-188983   -3170   -2502    -806  276897

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3471.691    195.605   17.75   <2e-16 ***
security_delay  141.454      4.009   35.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12950 on 4605 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.2128,    Adjusted R-squared:  0.2126
F-statistic:  1245 on 1 and 4605 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = arr_delay ~ late_aircraft_delay, data = airlines_data_1m)

Residuals:
   Min     1Q Median     3Q    Max
-26529   -672   -416    113  50317

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         547.24086   51.70081   10.59   <2e-16 ***
late_aircraft_delay   2.34028    0.00813  287.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3349 on 4605 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9473,    Adjusted R-squared:  0.9473
F-statistic: 8.286e+04 on 1 and 4605 DF,  p-value: < 2.2e-16
```

Figure 4 - Linear Regression Model R Output for Analyzing Airline Delays





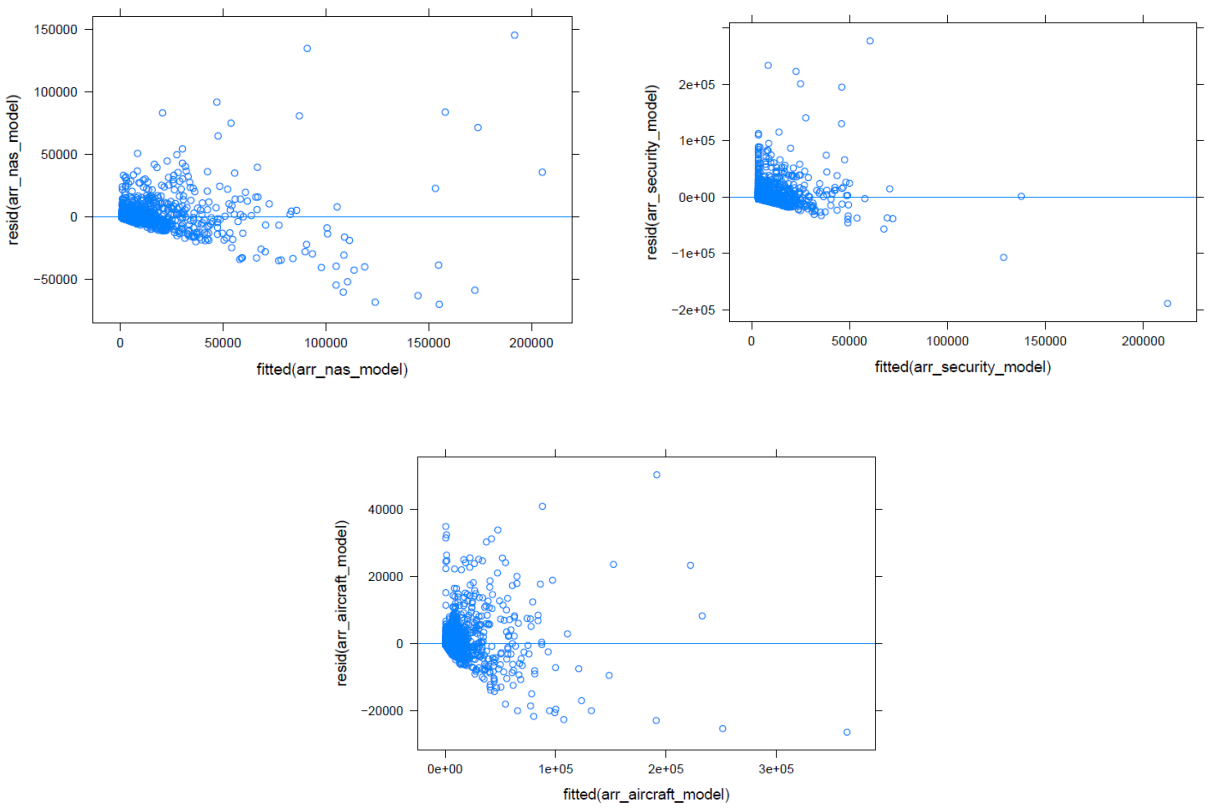
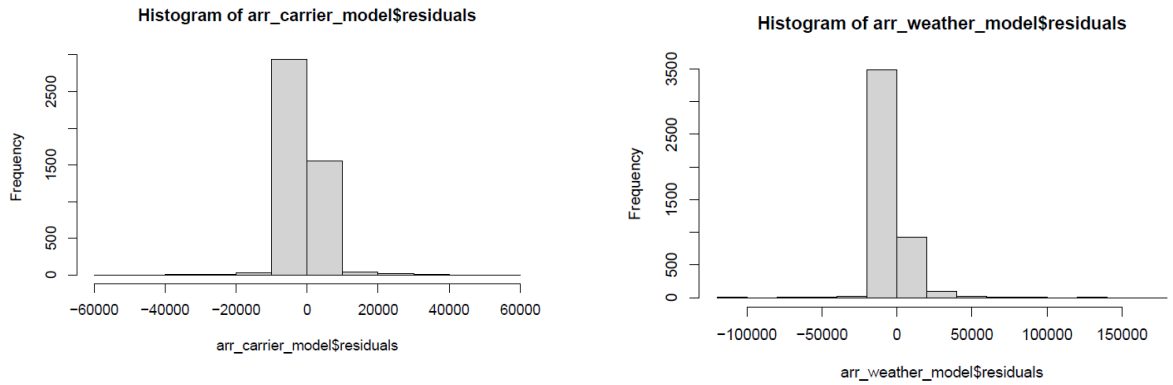


Figure 5 - Scatter plots between residual and Fitted Models



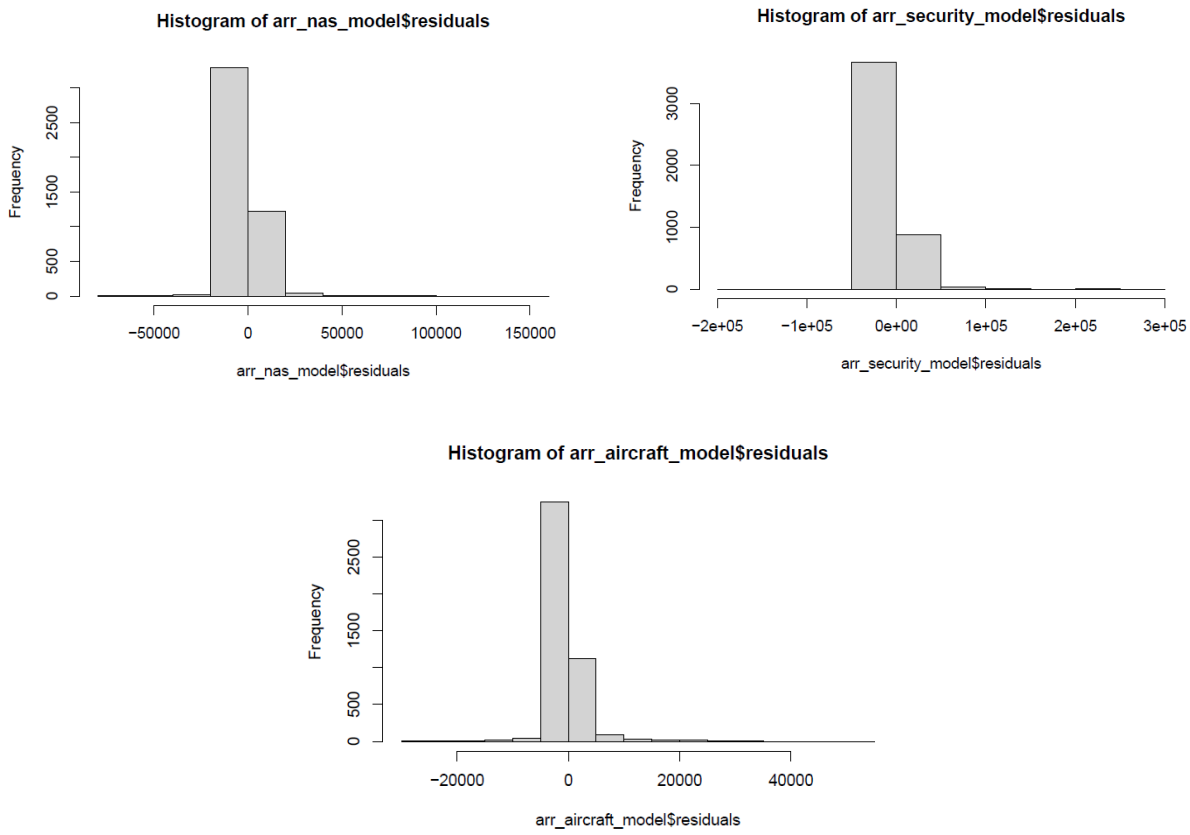


Figure 5 - Histogram Plots of Residuals for Linear Regression Models

4. Hypothesis Testing

Sampling

In order to avoid the challenges of testing with the entire dataset, we will opt for a random sample. This **random sample** will specifically include flights operated by Delta and American airlines (~800 data points). By selecting a representative subset, we can still obtain meaningful insights while reducing computational burdens. It is important to note that the random sample will adhere to the below assumption required for performing hypothesis testing.

a. Random Sample - Since we don't expect airlines delays to be carrier dependent, we extract this sample randomly.

b. Independence - Since the sample is large and diverge and from the above scatter plots of LR models, we can see that the data points follow no particular pattern and are independent.

c. Normality - The population drawn must be normally distributed as the sample size is large and it follows Central Limit Theorem.

d. Homogeneity of Variance - From the histogram plots of LR models, we can see that key delay causing variables are evenly distributed around the mean and thus does have homogeneity of variance.

A random sample following all the assumptions ensures that the results from hypothesis testing obtained from the sample can be generalized to the larger population of flights operated by these airlines. Using random sample, we generate the below hypothesis -

Null Hypothesis: There is no significant relationship across the carriers and the delays due to weather conditions
Alternate Hypothesis: There is significant relationship across the carriers and the delays due to weather conditions

Chi Square Tests

a) Weather count delays vs Carrier delays:

A significant relationship between the variables would suggest that the distribution of weather-related delays differs significantly across the two airlines. This implies that one airline is more affected by weather conditions than the other, or that their operations or strategies in response to weather events vary.

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 90954, df = 85360, p-value < 2
```

Since the p-value obtained is close to zero, we have enough evidence to reject the null hypothesis at the 0.05 significance level. This means that there is significant association between the variables "weather_ct" (delays due to weather) and the carriers (Delta Air Lines Inc. and American Airlines Inc.) in the dataset, and distribution of weather delays differs significantly across the two airlines.

b) National Airline Services delays vs carrier delays:

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 146650, df = 138904, p-value < 2
```

Since the p-value obtained is close to zero, we have enough evidence to reject the null hypothesis at the 0.05 significance level. This means that there is significant association between the variables "nas_ct" (delays due to National Aviation Services) and the carriers (Delta Air Lines Inc. and American Airlines Inc.) in the dataset, and distribution of NAS delays differs significantly across the two airlines.

c) Security delays vs carrier delays:

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 7303.2, df = 6596, p-value = 1.
```

Since the p-value obtained is close to zero, we have enough evidence to reject the null hypothesis at the 0.05 significance level. This means that there is significant association between the variables "security_ct" (delays due to security) and the carriers (Delta Air Lines Inc. and American Airlines Inc.) in the dataset, and distribution of security delays differs significantly across the two airlines.

d) Late Arrival delays vs carrier delays:

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 147352, df = 137740, p-value <
```

Since the p-value obtained is close to zero, we have enough evidence to reject the null hypothesis at the 0.05 significance level. This means that there is a significant association between the variables "late_aircraft_ct" (delays due to late arrival of aircrafts) and the carriers (Delta Air Lines Inc. and American Airlines Inc.) in the dataset, and distribution of late flight arrival delays differs significantly across the two airlines.

4. Conclusions

From the ***Linear Regression Model & Correlation*** **Analysis** , we can conclude the following :

- Carrier delay, NAS delay, and late aircraft delay have relatively high adjusted R-squared values, indicating that they explain a significant portion of the variability in arrival delay.
- Weather delay and security delay have lower adjusted R-squared values, suggesting that they might have a relatively smaller impact on the overall arrival delay compared to other factors.

From the ***Chi Square test***, we can draw the following conclusions :

- The chi-squared test result suggests a significant relationship between the variables “weather_ct”, “nas_ct”, “security_ct”, “late_aircraft_ct” with the “carrier_ct” variable (number of delays due to carriers) of Delta Air Lines Inc. and American Airlines Inc. This implies that there is a notable difference in the distribution of these delay factors between these two airlines.
- Based on this finding, one could infer that there might be a higher likelihood of experiencing delays due to weather, NAS, security and late aircraft arrivals with one airline compared to the other.

From the ***Correlation Matrix***, we can draw the following conclusions :

- Carrier delay, weather delay, NAS delay, and late aircraft delay are positively correlated with each other to varying degrees. This suggests that an increase in one delay factor is associated with an increase in the others, indicating interdependencies between these factors.

- Security delay shows weaker correlations with other delay factors compared to carrier delay, weather delay, NAS delay, and late aircraft delay. This may suggest that security-related issues have less influence on the overall delay times compared to other factors.

- The highest correlation exists between carrier delay and late aircraft delay (0.938), indicating a strong positive relationship. This suggests that carrier-related issues and late aircraft arrivals have a significant impact on the overall delay times.

- Carrier delay, weather delay, and NAS delay show relatively strong positive correlations with each other, indicating a shared influence on flight delays. This suggests that addressing these factors could potentially help mitigate overall delay times.

5. Appendix

Github Repository
https://github.com/singh96aman/Airline-Delays-Investigation

R Code
https://github.com/singh96aman/Airline-Delays-Investigation/blob/main/AnalysisAirlines.Rmd

R Output
https://github.com/singh96aman/Airline-Delays-Investigation/blob/main/AnalysisAirlines.pdf

Dataset
https://github.com/singh96aman/Airline-Delays-Investigation/blob/main/Airline_Delay_Cause.csv