PROJECT REPORT

Statistical Analysis of Airline Delays

Group Members

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

INTRODUCTION

The dataset under study is the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) dataset, specifically the monthly Air Travel Consumer Report. This report tracks the on-time performance of domestic flights operated by large air carriers in the United States. The dataset provides valuable information on various aspects of flight operations, including the number of on-time, delayed, canceled, and diverted flights, as well as the factors contributing to these delays.
The research question we aim to answer is: What are the primary factors contributing to flight delays in the domestic air travel industry? We also compare two most popular domestic airlines: Delta Airlines and American Airlines.
This analysis provides insights into the relationship between the delay factors and the arrival delay, enabling stakeholders in the aviation industry to understand the factors contributing to delays, prioritize their efforts, and implement targeted strategies for reducing delays and improving overall performance.
Flight delays not only inconvenience passengers but also have significant economic implications. Delays can lead to missed connections, disruption in travel plans, increased costs for airlines, and reduced customer satisfaction.
Hence, by analyzing the dataset and examining the different delay factors such as carrier delay, weather delay, NAS (National Airspace System) delay, security delay, and late aircraft delay, we can gain insights into the causes of flight delays. Understanding these factors is crucial for airlines, airport authorities, policymakers, and passengers in order to improve overall flight operations, minimize delays, and enhance the travel experience.

DATA

The dataset contains information on flights operated by large air carriers in the United States. Each row in the dataset corresponds to a specific flight in a given month and includes the following variables:

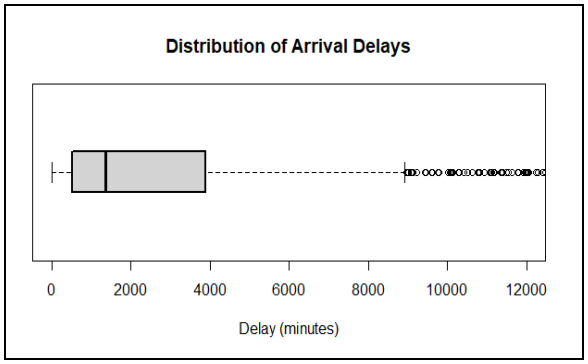
Variable	Description
year	the year in which the flight took place
month	the month in which the flight took place
carrier	the airline carrier code
carrier_name	the name of the airline carrier
airport	the airport code

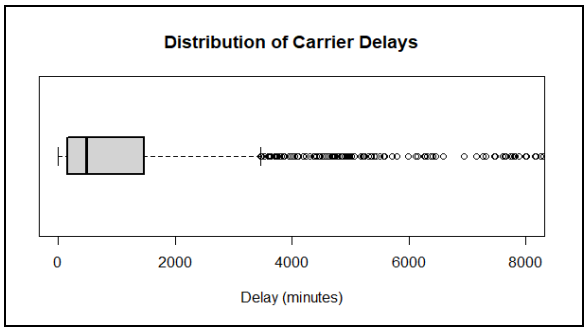
airport_name	the name of the airport
arr_flights	the total number of flights arriving at the airport
arr_del15	the number of flights arriving at the airport delayed by 15 minutes or more
carrier_ct	the average number of delayed flights due to the carrier
weather_ct	the average number of delayed flights due to weather conditions
nas_ct	the average number of delayed flights due to the National Aviation System (NAS)
security_ct	the average number of delayed flights due to security issues
late_aircraft_ct	the average number of delayed flights due to late aircrafts
arr_cancelled	the number of canceled flights
arr_diverted	the number of diverted flights
arr_delay	the total number of minutes of delay for all arriving flights
carrier_delay	the total number of minutes of delay due to carrier issues
weather_delay	the total number of minutes of delay due to weather conditions
nas_delay	the total number of minutes of delay due to the National Aviation System (NAS)
security_delay	the total number of minutes of delay due to security issues
late_aircraft_delay	the total number of minutes of delay due to late aircrafts

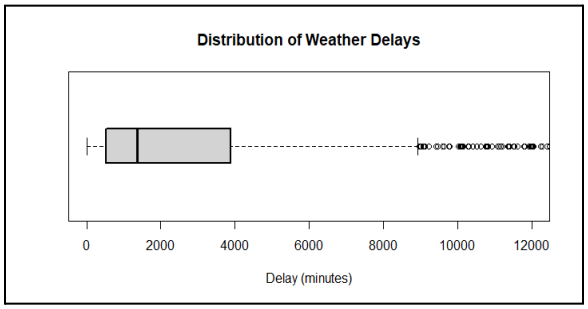
In the given airlines dataset, the response variable is "**arr_delay**" which represents the arrival delay in minutes. All the other variables are the explanatory variables, also known as independent variables or features. These explanatory variables provide contextual information and potential factors 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.

Box Plots :

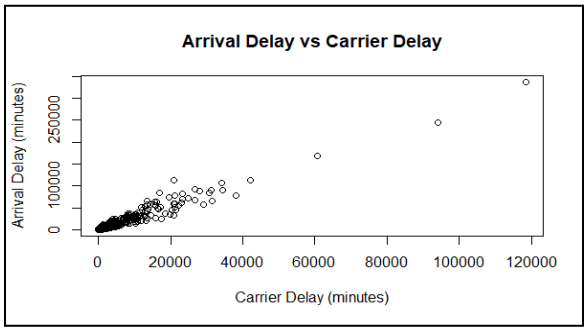
We have created box plots to visualize the distribution of delays (e.g., "arr_delay", "carrier_delay", etc.) for all flights. These plots help in understanding the spread and skewness of the delay data, along with identifying any outliers.

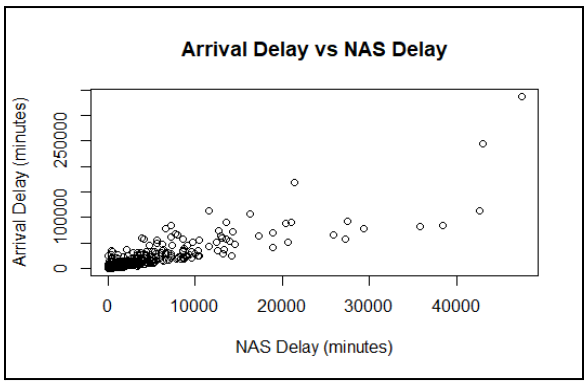






Scatter Plots :





ANALYSIS

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.

Linear Regression Model : Arrival Delay vs Carrier Delay

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

Residuals:
   Min     1Q Median     3Q    Max
-29412   -604   -149    293  54829

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   180.75503   99.83028   1.811   0.0704 .
carrier_delay   2.78345    0.01627 171.084   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3703 on 1535 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9502,    Adjusted R-squared:  0.9501
F-statistic: 2.927e+04 on 1 and 1535 DF,  p-value: < 2.2e-16
```

- The slope is 2.78345. This indicates that for every 1 minute increase in carrier delay, the arrival delay is expected to increase by approximately 2.78345 minutes.
- The R-squared value is 0.9502 suggests that approximately 95.02% of the variability in arrival delay can be explained by the variation in carrier delay and the use of a regression model.

Linear Regression Model : Arrival Delay vs Weather Delay

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

Residuals:
   Min     1Q Median     3Q    Max
-99865  -2261  -1723   -120 135325

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2285.4096   301.7539   7.574 6.23e-14 ***
weather_delay   12.7093     0.3059  41.551  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11380 on 1535 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5294,    Adjusted R-squared:  0.5291
F-statistic:  1727 on 1 and 1535 DF,  p-value: < 2.2e-16
```

- The slope is 12.7093. This indicates that for every 1 minute increase in weather delay, the arrival delay is expected to increase by approximately 12.7093 minutes.
- The R-squared value is 0.5294 suggests that approximately 52.94% of the variability in arrival delay can be explained by the variation in weather delay and the use of a regression model.

Linear Regression Model : Arrival Delay vs NAS Delay

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

Residuals:
   Min     1Q Median     3Q    Max
-70152  -1018   -650    278 145697

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 940.01907  214.94217   4.373 1.31e-05 ***
nas_delay     4.02077    0.05655  71.103  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8006 on 1535 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.7671,    Adjusted R-squared:  0.7669
F-statistic:  5056 on 1 and 1535 DF,  p-value: < 2.2e-16
```

- The slope is 4.02077. This indicates that for every 1 minute increase in NAS delay, the arrival delay is expected to increase by approximately 4.02077 minutes.
- The R-squared value is 0.7671 suggests that approximately 76.71% of the variability in arrival delay can be explained by the variation in NAS delay and the use of a regression model.

Linear Regression Model : Arrival Delay vs Security Delay

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

Residuals:
   Min     1Q Median     3Q    Max
-87625  -2920  -2219   -486 249614

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3189.12     367.06   8.688   <2e-16 ***
security_delay   209.86       8.16  25.717   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13870 on 1535 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.3011,    Adjusted R-squared:  0.3007
F-statistic: 661.4 on 1 and 1535 DF,  p-value: < 2.2e-16
```

- The slope is 209.86. This indicates that for every 1 minute increase in security delay, the arrival delay is expected to increase by approximately 209.86 minutes.
- The R-squared value is 0.3011 suggests that approximately 30.11% of the variability in arrival delay can be explained by the variation in security delay and the use of a regression model.

Linear Regression Model : Arrival Delay vs Late Aircraft Delay

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

Residuals:
   Min     1Q Median     3Q    Max
-18709   -776   -486    110  35644

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         641.18025   94.79012   6.764  1.9e-11 ***
late_aircraft_delay   2.24662    0.01255 179.021  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3547 on 1535 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9543,    Adjusted R-squared:  0.9543
F-statistic: 3.205e+04 on 1 and 1535 DF,  p-value: < 2.2e-16
```

- The slope is 2.24662. This indicates that for every 1 minute increase in late aircraft delay, the arrival delay is expected to increase by approximately 2.24662 minutes.
- The R-squared value is 0.9543 suggests that approximately 95.43% of the variability in arrival delay can be explained by the variation in late aircraft delay and the use of a regression model.

Correlation Analysis :

We examine the relationships between different delay factors by calculating correlation coefficients. This analysis helps us provide insights into the interdependencies of delay factors.

	carrier_delay	weather_delay	nas_delay	security_delay	late_aircraft_delay
carrier_delay	1.0000000	0.7780220	0.7845271	0.4968192	0.9379380
weather_delay	0.7780220	1.0000000	0.5472747	0.3780986	0.6385579
nas_delay	0.7845271	0.5472747	1.0000000	0.5261967	0.8062205
security_delay	0.4968192	0.3780986	0.5261967	1.0000000	0.5424562
late_aircraft_delay	0.9379380	0.6385579	0.8062205	0.5424562	1.0000000

Analyzing Delta vs American Airlines：Chi-Square test: weather count delays vs carrier delays:

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
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 = 13282, df = 13064, p-value = 0.08943
```

Since the p-value obtained (0.08943) is greater than 0.05, we do not have enough evidence to Reject the null hypothesis at the 0.05 significance level. This means that there is no significant association between the variables "weather_ct" (delays due to weather) and the airlines (Delta Air Lines Inc. and American Airlines Inc.) in the dataset.

Chi-Square test: National Airline Services delays vs carrier delays:

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 20294, df = 20022, p-value = 0.08747
```

Since the p-value obtained (0.08747) is greater than 0.05, we do not have enough evidence to reject the null hypothesis at the 0.05 significance level. This means that there is no significant association between the variables "nas_ct" (delays due to National Aviation Services) and the airlines (Delta Air Lines Inc. and American Airlines Inc.) in the dataset.

Chi-Square test: Security delays vs carrier delays:

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 730, df = 710, p-value = 0.2934
```

Since the p-value obtained (0.2934) is greater than 0.05, we do not have enough evidence to reject the null hypothesis at the 0.05 significance level. This means that there is no significant association between the variables "security_ct" (delays due to security) and the airlines (Delta Air Lines Inc. and American Airlines Inc.) in the dataset.

Chi-Square test: Late Arrival delays vs carrier delays:

```
        Pearson's Chi-squared test

data:  delta_table
X-squared = 20270, df = 19880, p-value = 0.02589
```

Since the p-value obtained (0.02589) is less than 0.05, we 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 airlines (Delta Air Lines Inc. and American Airlines Inc.) in the dataset, and distribution of late flight arrival delays differs significantly across the two airlines.

CONCLUSIONS

From the *Linear Regression Model* 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 *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.

CODE (APPENDIX)

Linear Regression Model -

```
# Load the necessary libraries
library(dplyr)
library(ggplot2)

airlines_data <- read.csv("Airline_Delay_Cause.csv")

# Select the relevant columns for the linear regression analysis
airlines_data_lm <- airlines_data %>%
  select(arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay)

# Fit a linear regression model
arr_carrier_model <- lm(arr_delay ~ carrier_delay , data = airlines_data_lm)
arr_weather_model <- lm(arr_delay ~ weather_delay , data = airlines_data_lm)
arr_nas_model <- lm(arr_delay ~ nas_delay , data = airlines_data_lm)
arr_security_model <- lm(arr_delay ~ security_delay , data = airlines_data_lm)
arr_aircraft_model <- lm(arr_delay ~ late_aircraft_delay , data = airlines_data_lm)

# Print the model summary
print(summary(arr_carrier_model))
print(summary(arr_weather_model))
print(summary(arr_nas_model))
print(summary(arr_security_model))
print(summary(arr_aircraft_model))
```

Chi-square test -

```r
delta <- subset(data, carrier_name == "Delta Air Lines Inc.")
american <- subset(data, carrier_name == "American Airlines Inc.")

# carrier_ct vs weather_ct
delta_table <- table(delta$carrier_ct, delta$weather_ct)
american_table <- table(american$carrier_ct, american$weather_ct)
chi_sq <- chisq.test(delta_table, american_table)
print(chi_sq)

# carrier_ct vs nas_ct
delta_table <- table(delta$carrier_ct, delta$nas_ct)
american_table <- table(american$carrier_ct, american$nas_ct)
chi_sq <- chisq.test(delta_table, american_table)
print(chi_sq)

# carrier_ct vs security_ct
delta_table <- table(delta$carrier_ct, delta$security_ct)
american_table <- table(american$carrier_ct, american$security_ct)
chi_sq <- chisq.test(delta_table, american_table)
print(chi_sq)

# carrier_ct vs late_aircraft_ct
delta_table <- table(delta$carrier_ct, delta$late_aircraft_ct)
american_table <- table(american$carrier_ct, american$late_aircraft_ct)
chi_sq <- chisq.test(delta_table, american_table)
print(chi_sq)
```

Correlation Matrix -

```r
# Load the necessary libraries
library(dplyr)

# Select the relevant columns for correlation analysis
data <- airlines_data %>%
  select(carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay)

# Calculate correlation coefficients
correlation_matrix <- cor(data , use = "complete.obs")

# Print the correlation matrix
print(correlation_matrix)
```