R Notebook

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

Null Hypothesis (H0): There is no significant difference in the mean "arr_delay" among the different airports.

Alternative Hypothesis (H1): 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 mu represents the mean "arr_delay" for each airport) H1: At least one pair of means mu_i and mu_j is different. Here, mu_1, mu_2, ..., 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.

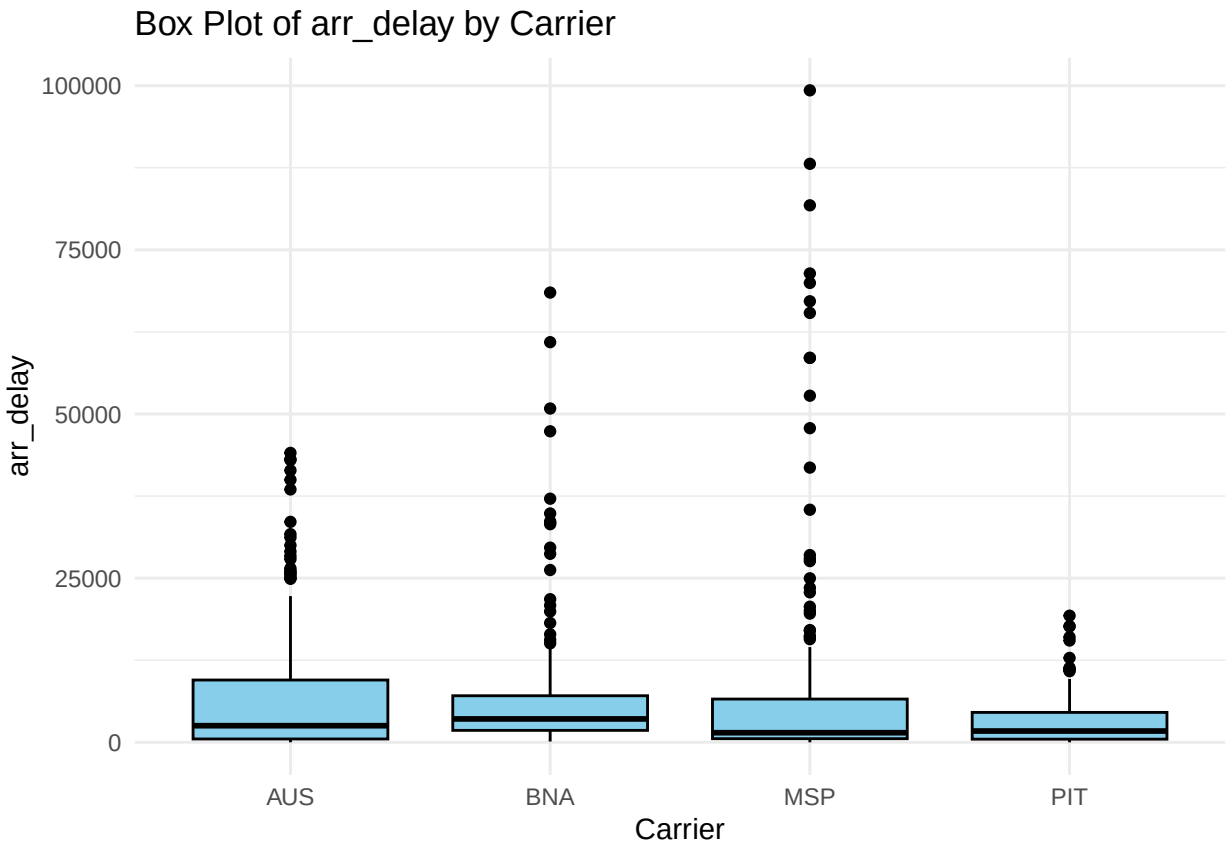
Loading the data set only for 4 Top Airports: AUS, MSP, BNA, PIT

```
library(readr)
raw_top4_airport <- read_csv("~/Documents/Stat501/project/raw_top4_airport.csv")
```

```
## New names:
## Rows: 745 Columns: 22
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (4): carrier, carrier_name, airport, airport_name dbl (18): ...1, year, month,
## arr_flights, arr_del15, carrier_ct, weather_ct,...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
library(ggplot2)

# Create a box plot
ggplot(raw_top4_airport, aes(x = airport, y = arr_delay)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Box Plot of arr_delay by Carrier",
       x = "Carrier",
       y = "arr_delay") +
  theme_minimal()
```



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

a) With Outliers, the Anova results are:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v stringr   1.5.0
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
anova_result_w_outliers <- aov(arr_delay ~ airport, data = raw_top4_airport)

print(summary(anova_result_w_outliers))
```

```
##              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
```

Interpretation:

Df (Degrees of Freedom): There are three degrees of freedom for the factor "airport" and 741 degrees of

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 investigae fruther. The method used for investigating the pairs is Tukey's post-hoc test:

```
# Perform Tukey's post-hoc test
tukey_result_w_outliers <- TukeyHSD(anova_result_w_outliers)

# Display Tukey's post-hoc results
print(tukey_result_w_outliers)
```

```
##   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
```

diff (Difference): The estimated difference in means between the pairs of airports for "arr_delay." lwr and upr (Lower and Upper Confidence Intervals): The 95% confidence interval for the difference in means. p adj (Adjusted p-value): The p-value adjusted for multiple comparisons (Tukey's correction). Interpretation: BNA-AUS: The difference in mean "arr_delay" between Nashville (BNA) and Austin (AUS) is not statistically significant (p = 0.9918).

MSP-AUS: The difference in mean "arr_delay" between Minneapolis (MSP) and Austin (AUS) is not statistically significant (p = 0.8366).

PIT-AUS: The difference in mean "arr_delay" between Pittsburgh (PIT) and Austin (AUS) is statistically significant (p = 0.00299). The negative difference suggests that Austin has a higher mean delay than Pittsburgh.

MSP-BNA: The difference in mean "arr_delay" between Minneapolis (MSP) and Nashville (BNA) is not statistically significant (p = 0.6758).

PIT-BNA: The difference in mean "arr_delay" between Pittsburgh (PIT) and Nashville (BNA) is statistically significant (p = 0.00812). The negative difference suggests that Nashville has a higher mean delay than Pittsburgh.

PIT-MSP: The difference in mean "arr_delay" between Pittsburgh (PIT) and Minneapolis (MSP) is statistically significant (p = 0.0001087). The negative difference suggests that Minneapolis has a higher mean delay than Pittsburgh.

In summary, the adjusted p-values indicate whether the differences in mean "arr_delay" are statistically significant after correcting for multiple comparisons.

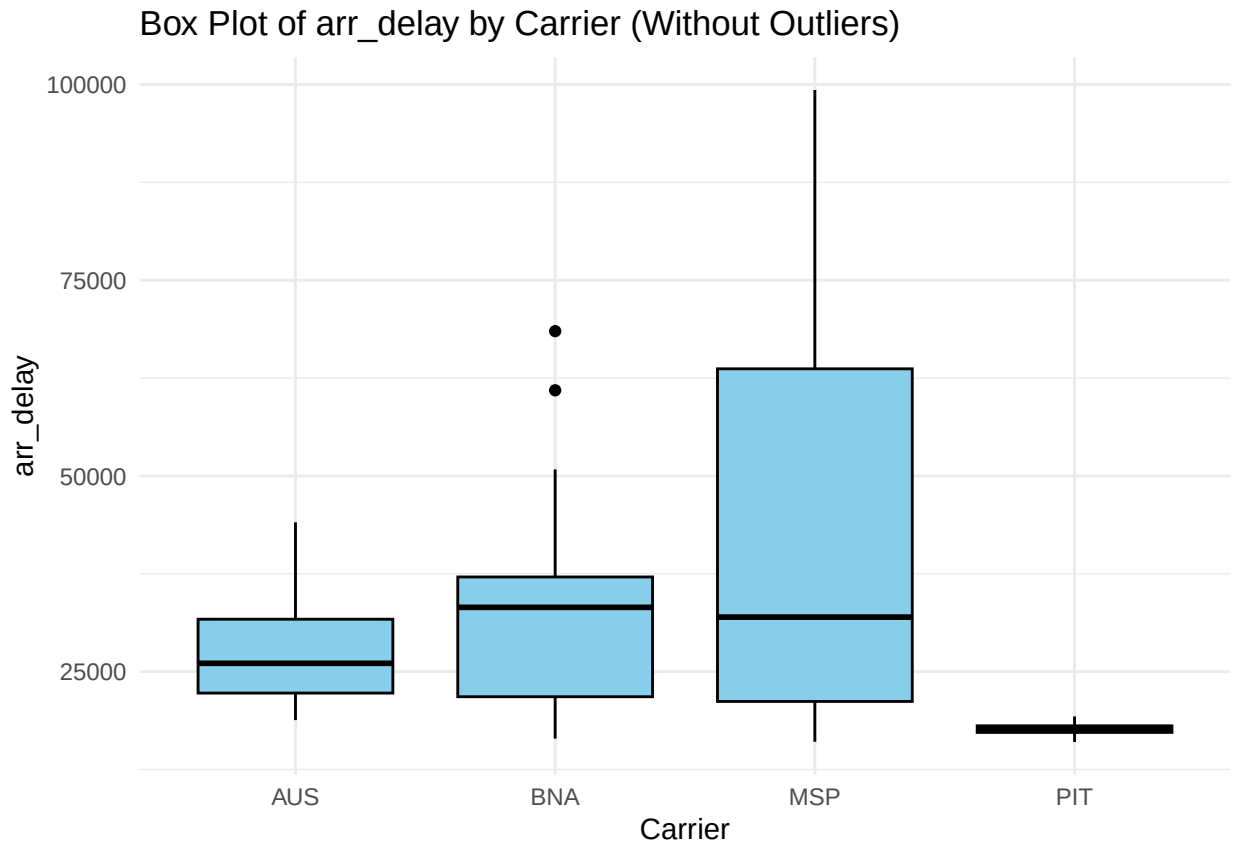
b) We are eager to know what effect if we remove outliers So, after removing outlier, we performed same analysis again.

```r
# Load the required libraries
library(ggplot2)

# Function to remove outliers
remove_outliers <- function(x) {
  q <- quantile(x, c(0.25, 0.75))
  iqr <- q[2] - q[1]
  lower_bound <- q[1] - 1.5 * iqr
  upper_bound <- q[2] + 1.5 * iqr
  return(x[x >= lower_bound & x <= upper_bound])
}

# Remove outliers from 'arr_delay'
flight_data_no_outliers <- raw_top4_airport %>%
  filter(!arr_delay %in% remove_outliers(raw_top4_airport$arr_delay))

# Create a box plot without outliers
ggplot(flight_data_no_outliers, aes(x = airport, y = arr_delay)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Box Plot of arr_delay by Carrier (Without Outliers)",
       x = "Carrier",
       y = "arr_delay") +
  theme_minimal()
```




```r
# Load the required libraries
library(tidyverse)
```

```r
# Assuming your data is stored in a dataframe named flight_data_no_outliers
# Perform one-way ANOVA
anova_result_no_outliers <- aov(arr_delay ~ airport, data = flight_data_no_outliers)

# Display ANOVA results
print(summary(anova_result_no_outliers))
```

```
##              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
```

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

Tukey Multiple Comparisons:

```r
# Perform Tukey's post-hoc test
tukey_result <- TukeyHSD(anova_result_no_outliers)

# Display Tukey's post-hoc results
print(tukey_result)
```

```
##   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
```

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