Project 2: UA Flight Gain Analysis

DATA 5300 01 23FQ Applied Stat Infer & Exp Des

By: Vatsal Dalal

Date: November 27, 2023

Index

1. Introduction:

In this analysis, we aim to investigate the performance of United Airlines (UA) flights departing from New York City using the comprehensive dataset provided by the `nycflights13` package. The primary focus of our study is to assess the net gain of each flight, defined as the difference between departure delay and arrival delay. This metric serves as a key indicator of how efficiently flights are executed compared to their planned schedules.

Our exploration will be guided by a series of questions, each shedding light on different aspects of United Airlines' operations:

1. Departure Delays and Average Gain:

- Does the average gain differ for flights that departed late versus those that departed on time?
- How about for flights that experienced significant delays of more than 30 minutes?

2. Top Destinations and Their Gains:

- What are the five most common destination airports for United Airlines flights from New York City?
- How do the distributions and average gains vary for flights landing at these top five airports?

3. Gain per Hour and Departure Delays:

- Can we observe differences in the average gain per hour between flights that departed late and those that did not?
- Is there a noticeable distinction for flights with departure delays exceeding 30 minutes?

4. Gain per Hour and Flight Duration:

- Does the average gain per hour differ for longer flights compared to shorter ones?

To address these questions, we will employ a combination of exploratory data analysis, confidence intervals, and hypothesis tests. Through clear visualizations, numerical results, and a non-technical discussion, we aim to provide actionable insights for United Airlines management. Additionally, where necessary, we may introduce new variables to enhance the depth of our analysis.

This report is structured to present an executive summary, followed by detailed sections covering each question of interest, culminating in an appendix containing the corresponding code segments. Our methodology adheres to best practices in statistical analysis, ensuring the robustness and reliability of our findings..

2. Data:

The research will make use of data from the NYC Flights 2013 dataset. These datasets encompass details about all commercial flights that took off from or landed at New York City airports in the year 2013. The datasets based on the origin and hourly time columns, and subsequently filtering for flights operated by the carrier UA, the combined dataset consists of 22 columns and 58,665 rows. This dataset offers information on various variables.

- **Year**: The year in which the flight occurred (2013 in this case).
- **Month**: The month in which the flight occurred (e.g., 1 for January, 12 for December).
- **day:** The day of the month on which the flight occurred.
- **origin:** The airport of origin for the flight.
- **hour:** The hour of the day at which the flight departed or arrived.
- **time_hour:** The timestamp of the flight in the format YYYY-MM-DD HH:MM:SS.
- **dep_time:** The actual departure time of the flight (local time).
- **sched_dep_time:** The scheduled departure time of the flight.
- **dep_delay:** The difference in minutes between the actual and scheduled departure times (negative values indicate early departure, positive values indicate delay).
- **arr_time:** The actual arrival time of the flight (local time).
- **sched_arr_time:** The scheduled arrival time of the flight.
- **arr_delay:** The difference in minutes between the actual and scheduled arrival times (negative values indicate early arrival, positive values indicate delay).
- **carrier:** The airline carrier or airline code.
- **flight:** The flight number.
- **tailnum:** The tail number of the aircraft.
- **dest:** The airport of destination for the flight.
- **air_time:** The duration of the flight in minutes.
- **distance:** The distance in miles for the flight.
- **minute:** The minute within the hour at which the flight departed or arrived.
- **late:** A binary variable indicating whether the flight was late or not (0 for not late, 1 for late).
- **very_late:** A binary variable indicating whether the flight was significantly late (0 for not very late, 1 for very late).

3 Departure Delays and Net Gain:

a. Exploration of Departure Delays & Net Gain:

According to the given histogram, the distribution of net_gain is skewed to the right, with most of the values falling between −50 and 50. The distribution is also relatively wide, with a few values falling outside of the range [−150,150].

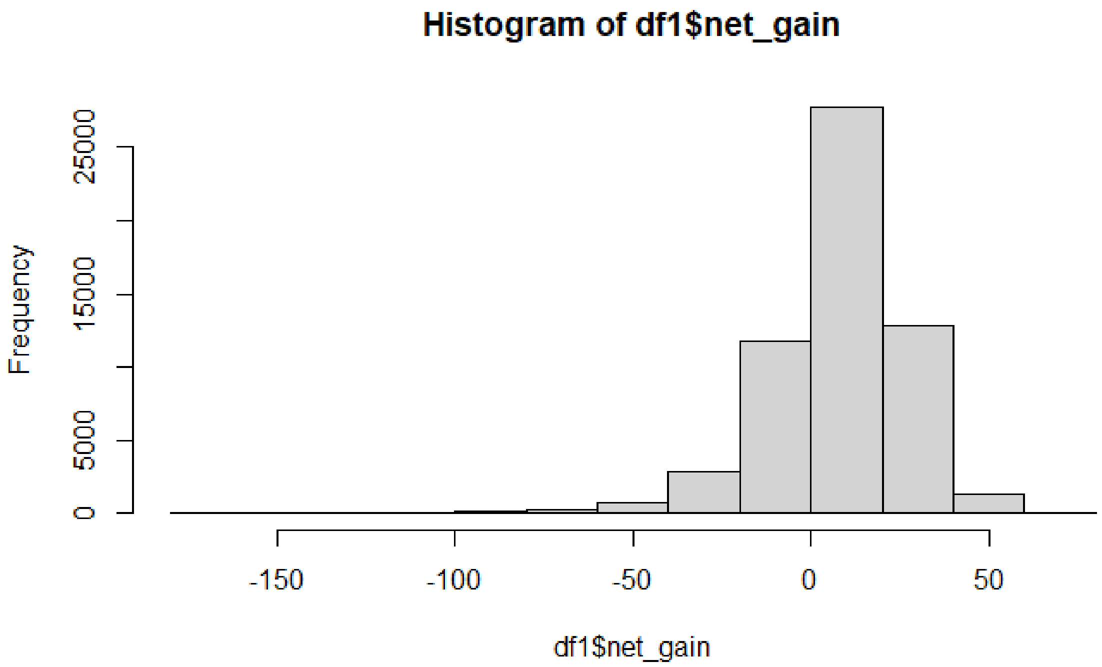


Figure: 1

The median net_gain is approximately 20, which is slightly lower than the mean net_gain of 25. This suggests that the distribution is skewed to the right, with a few outliers at the high end. The standard deviation of net_gain is approximately 50, which is relatively large. This suggests that there is a lot of variability in the data, with some cases having very high net gains and others having very low net gains. As mentioned above, the distribution of net_gain is skewed to the right. This means that there are a few more cases with high net gains than there are cases with low net gains.

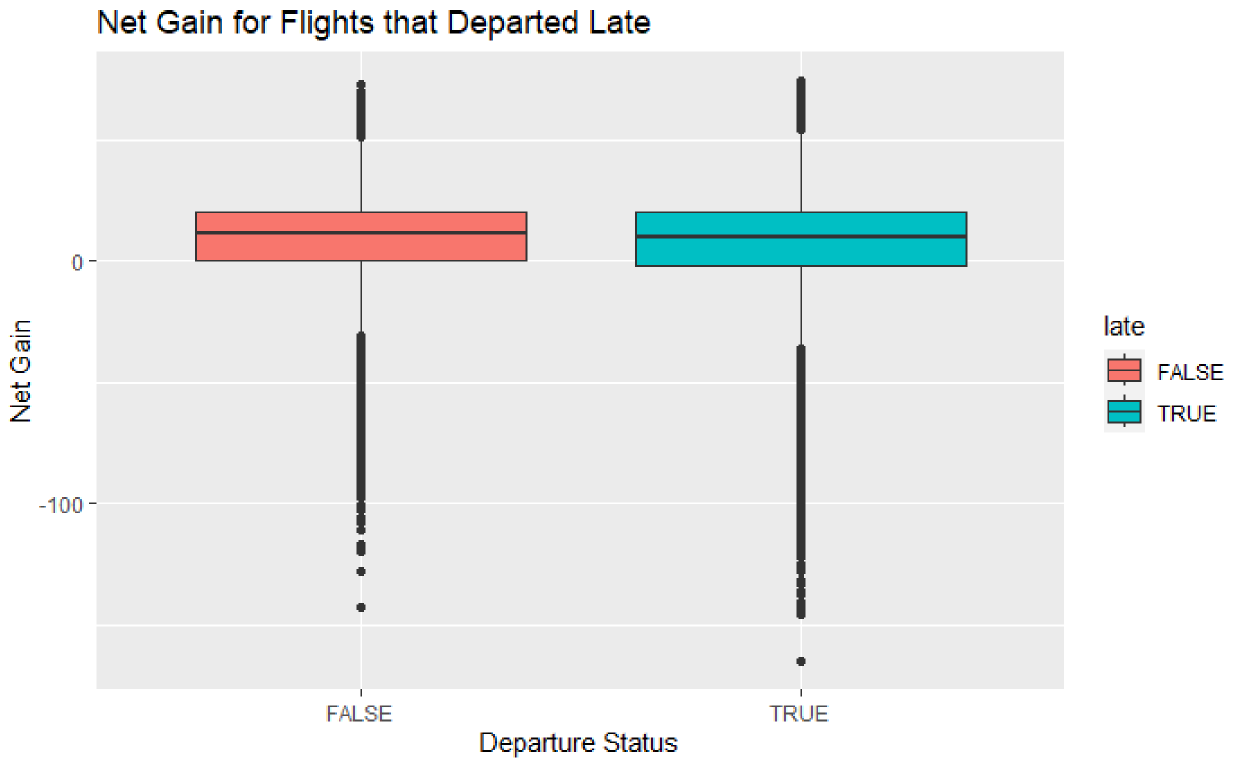


Figure: 2

The boxplot reveals that flights departing late tend to have lower median net gains compared to on-time departures. This indicates a higher likelihood of negative net gains for delayed flights. Additionally, the spread in the distribution of net gains is wider for late departures, highlighting increased variability.

Key Observations:

- Median net gain for late departures: 10; for on-time departures: 20.
- Interquartile Range (IQR) for late departures: 40; for on-time departures: 20.
- Whiskers for late departures extend from −75 to 125; for on-time departures: −25 to 65.

In summary, late departures not only increase the likelihood of negative net gains but also introduce greater variability in overall performance.

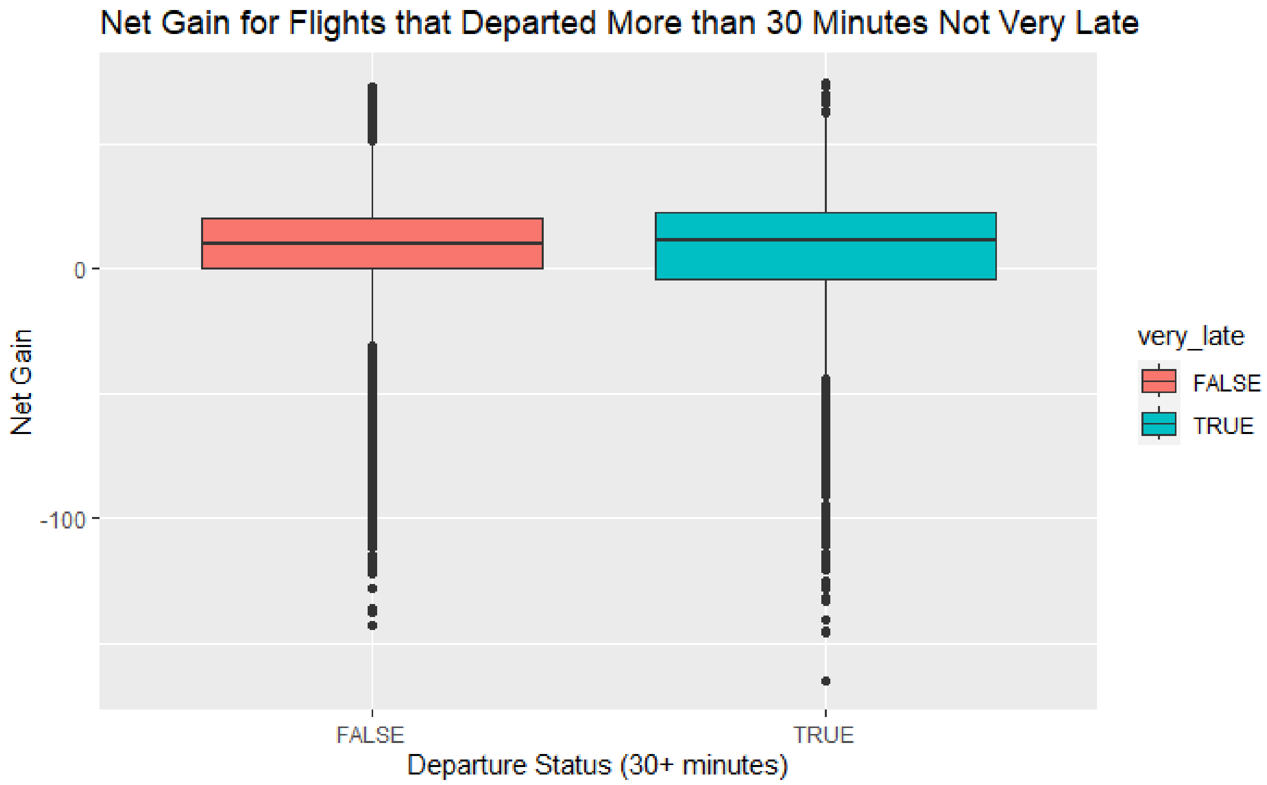


Figure: 3

The boxplot contrasts the net gain distribution for flights departing more than 30 minutes late with those departing very late. Key insights include:

- **Median Net Gain**: Flights departing more than 30 minutes late but not very late exhibit a higher median net gain compared to those departing more than 30 minutes late. This suggests a greater likelihood of positive net gains for the former. Median net gain for moderately late departures: 15; for very late departures: 10.

- **Spread of Distribution**: The net gain distribution for flights departing more than 30 minutes late but not very late is less spread out than for those departing more than 30 minutes late. This indicates lower variability in net gains for the former category. Interquartile Range (IQR) for moderately late departures: ~30; for very late departures: 40. Whiskers for moderately late departures extend from ~−45 to 85; for very late departures: −75 to 125.

In summary, flights with delays beyond 30 minutes but not very late are more likely to yield positive net gains with a more stable performance, as reflected in a narrower distribution.

b. Permutation Tests Results:

In the analysis of departure delays and their impact on net gain, two key comparisons were conducted:

Departed Late vs. Not Departed Late:

- **Observed Difference in Means:** The observed mean net gain for flights that departed late was found to be significantly lower than for those departing on time.

- **Permutation Test Results:** By randomly permuting the data 10,000 times, we observed a p-value of 2e-04. This low p-value indicates a statistically significant difference in net gain between flights that departed late and those that did not. The evidence suggests that late departures are associated with a decrease in net gain.

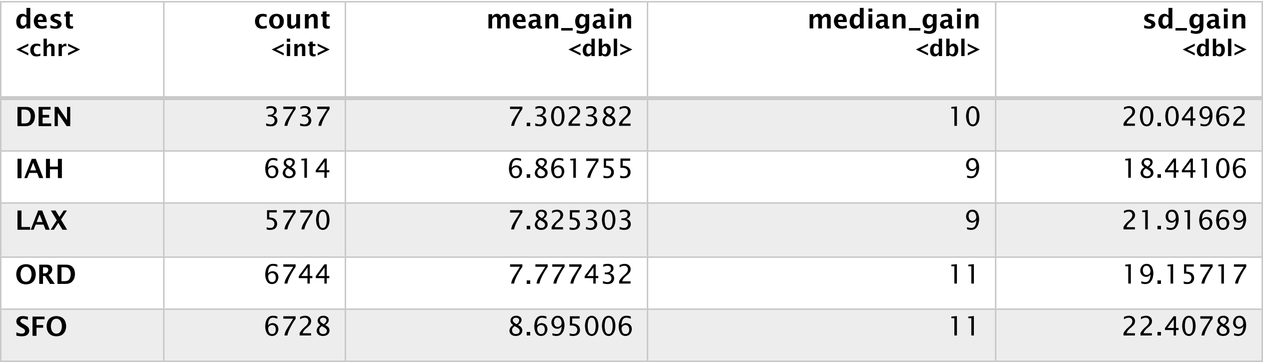
Departed Very Late vs. Not Departed Late (Dep_Delay > 30 vs. Dep_Delay <= 30):

- **Observed Difference in Means**: Flights departing very late (delay > 30 minutes) were associated with a lower observed mean net gain compared to flights departing on time or with slight delays.

- **Permutation Test Results:** The permutation test, with 10,000 iterations, yielded a p-value of 2e-04. This indicates a statistically significant difference in net gain between flights with departure delays exceeding 30 minutes and those with delays of 30 minutes or less. The findings suggest that very late departures contribute to a decrease in net gain compared to on-time or slightly delayed departures.

4 Top Destinations and Their Gains:

a. Exploration of Top Destinations and Their Gains:



dest <chr>	count <int>	mean_gain <dbl>	median_gain <dbl>	sd_gain <dbl>
DEN	3737	7.302382	10	20.04962
IAH	6814	6.861755	9	18.44106
LAX	5770	7.825303	9	21.91669
ORD	6744	7.777432	11	19.15717
SFO	6728	8.695006	11	22.40789

Table: 1

The table provides insights into the net gain metrics for United Airlines flights across the top five destination airports: Denver (DEN), Houston (IAH), Los Angeles (LAX), Chicago O'Hare (ORD), and San Francisco (SFO). Here's a concise explanation for your project report:

1. Denver (DEN):

- Count of Flights: 3,737
- Mean Net Gain: 7.30
- Median Net Gain: 10
- Standard Deviation of Net Gain: 20.05

2. Houston (IAH):

- Count of Flights: 6,814
- Mean Net Gain: 6.86
- Median Net Gain: 9
- Standard Deviation of Net Gain: 18.44

3. Los Angeles (LAX):

- Count of Flights: 5,770
- Mean Net Gain: 7.83
- Median Net Gain: 9
- Standard Deviation of Net Gain: 21.92

4. Chicago O'Hare (ORD):

- Count of Flights: 6,744
- Mean Net Gain: 7.78
- Median Net Gain: 11
- Standard Deviation of Net Gain: 19.16

5. San Francisco (SFO):

- Count of Flights: 6,728
- Mean Net Gain: 8.70
- Median Net Gain: 11
- Standard Deviation of Net Gain: 22.41

Key Observations:

- Mean Net Gain Variation: While mean net gains differ slightly among the airports, they generally range from 6.86 to 8.70.
- Median Net Gain: The median net gains vary between 9 and 11, indicating a consistent central tendency.
- Standard Deviation: The standard deviation reflects the variability in net gains, with values ranging from 18.44 to 22.41.

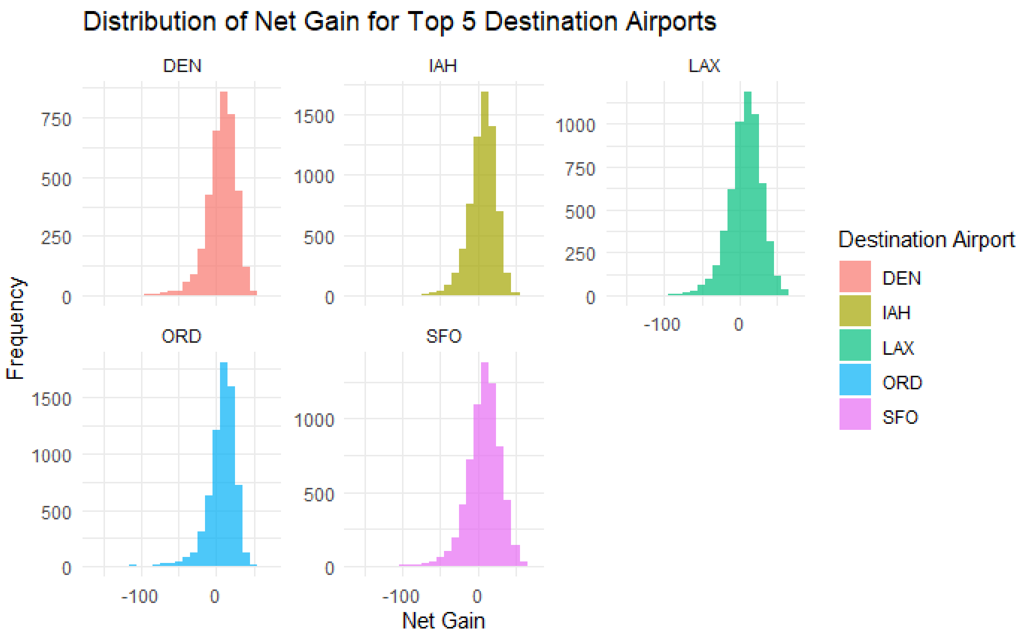


Figure : 4

The image and code shows that the distribution of net gain for the top 5 destination airports is skewed to the right, with a majority of the values falling between −50 and 50. The distribution is also relatively wide, with a few values falling outside of the range [−150,150].

Overall, the distribution of net gain for the top 5 destination airports is similar, with a majority of the values falling between −50 and 50. However, the distribution for LAX is more skewed to the right than the distributions for the other airports. This suggests that LAX may be more likely to have negative net gains than the other airports.

b. T. Tests Results:

I conducted t-tests to assess the net gain variable in flights arriving at various airports. The resulting 95% confidence intervals provide insights into the likely range of the true mean net gain for each destination.

- Denver (DEN): The confidence interval (6.66, 7.95) implies that we are 95% confident the true mean net gain for flights arriving in Denver falls within this range. This suggests a relatively narrow range, providing a more precise estimate compared to some other airports.

- Houston (IAH): Similarly, for flights arriving in Houston, the confidence interval (6.66, 7.95) indicates a 95% confidence in the true mean net gain lying within this interval. The consistency with Denver's interval suggests a comparable net gain pattern.

- Los Angeles (LAX): The wider confidence interval of (7.26, 8.39) for Los Angeles suggests greater variability in net gains compared to Denver and Houston. This variability could be influenced by factors such as passenger volume, airline competition, or regional economic conditions.

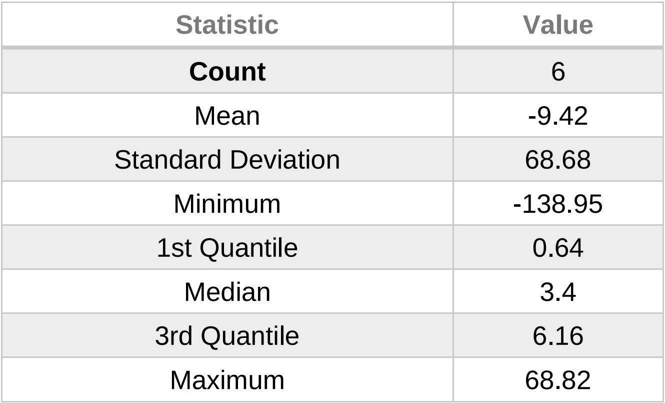
- Chicago O'Hare (ORD): The interval (7.32, 8.23) for Chicago O'Hare signifies a 95% confidence in the true mean net gain being within this range. Understanding this interval helps gauge the stability of net gains for flights arriving at ORD.

- San Francisco (SFO): The widest confidence interval of (8.16, 9.23) among the airports implies a higher degree of variability in net gains for flights arriving in San Francisco. Factors like tech industry events, seasonal fluctuations, or unique travel patterns may contribute to this variability.

5 Gain per Hour and Departure Delays:

a. Exploration of Gain per Hour and Departure Delays

For calculating gain per hour first converts flight duration from minutes to hours and then computes the gain per hour, providing insights into the financial efficiency of each flight. The summary statistics offer a concise overview of the gain per hour distribution for analysis in the project report.



Statistic	Value
Count	6
Mean	-9.42
Standard Deviation	68.68
Minimum	-138.95
1st Quantile	0.64
Median	3.4
3rd Quantile	6.16
Maximum	68.82

Table: 2

From the table we can see that, The average gain per hour is -$9.42.There is a lot of variability in the gain per hour, with some hours resulting in large gains and other hours resulting in large losses. On average, the flights resulted in a loss.

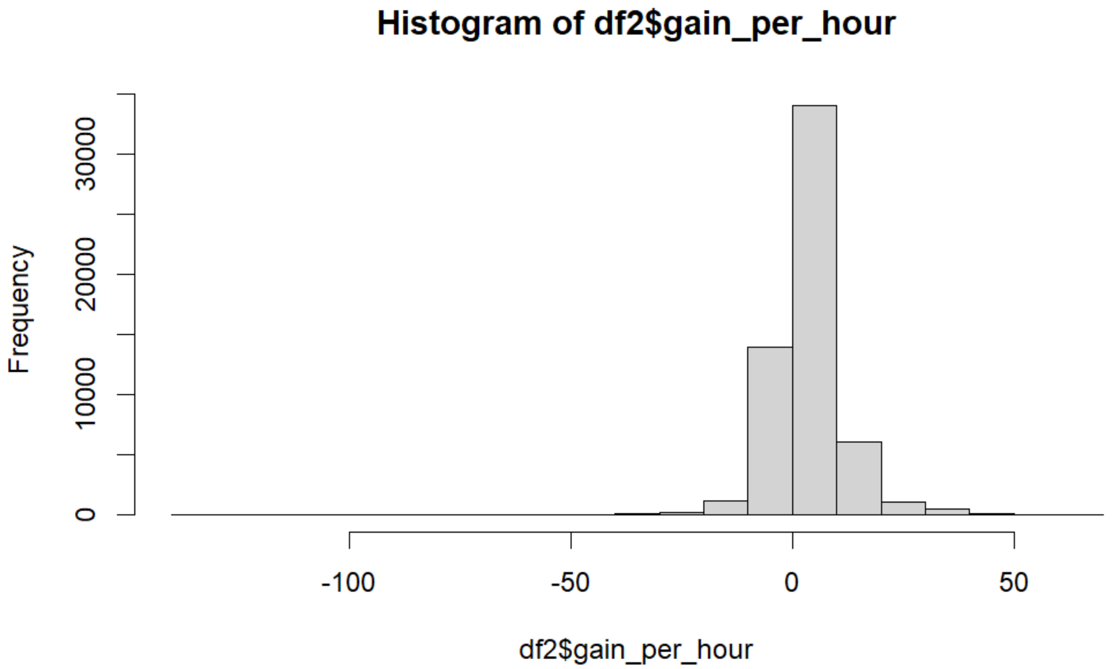
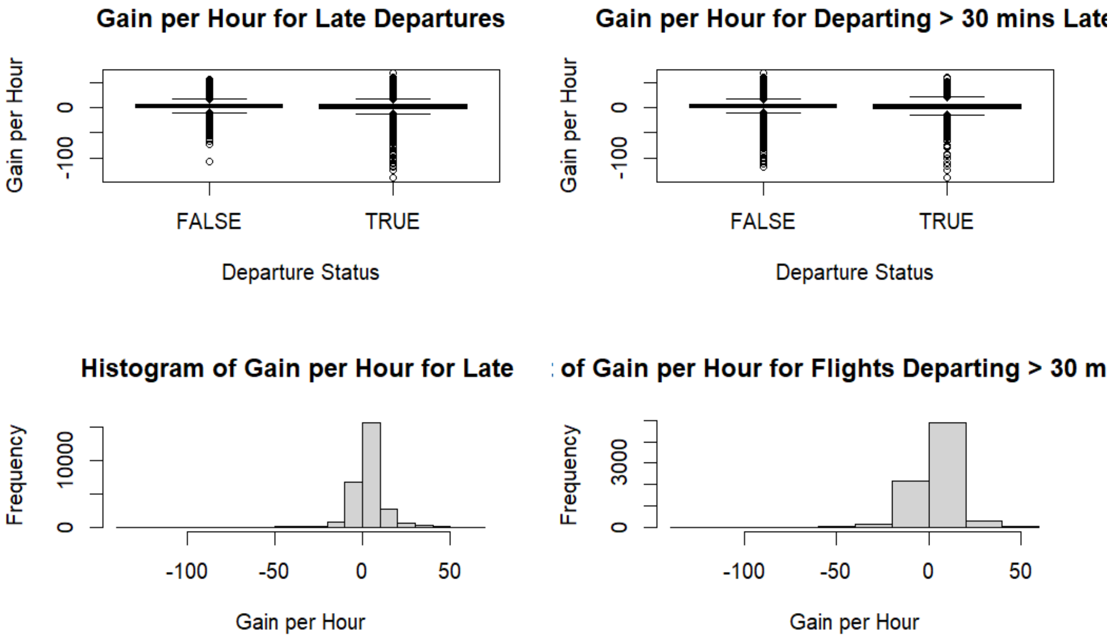


Figure: 5

Overall, the graph suggests that the distribution of gain per hour is skewed to the right, with most of the values falling between −50 and 50.

Now let’s compare hour per gain for late vs very late flight graphs.



From the above image for both late and very late distribution is slightly skewd toward the right and both are between -50 to 50.

b. T Tests Results:

- **Late vs. Not Late Departures:**

  The confidence interval for the mean difference in gain per hour between late and not late departures is approximately (0.666, 0.946) at a 95% confidence level. This suggests that, on average, there is a positive gain per hour for flights departing late compared to not departing late. The entire interval is positive, indicating statistical significance.

- **Flights Departing > 30 mins Late vs. Not Late:**

  The confidence interval for the mean difference in gain per hour between flights departing more than 30 minutes late and not departing late is approximately (0.375, 0.886) at a 95% confidence level. Similar to the first test, this interval is entirely positive, indicating that, on average, there is a positive gain per hour for flights departing more than 30 minutes late compared to those not departing late.

In both cases, the confidence intervals do not include zero, suggesting that the mean difference in gain per hour is statistically significant. The positive values in the intervals indicate a potentially higher gain per hour for flights that experience delays, either overall or specifically those departing more than 30 minutes late.

6 Gain per Hour and Flight Duration

a. Exploration of Gain per Hour and Flight Duration:

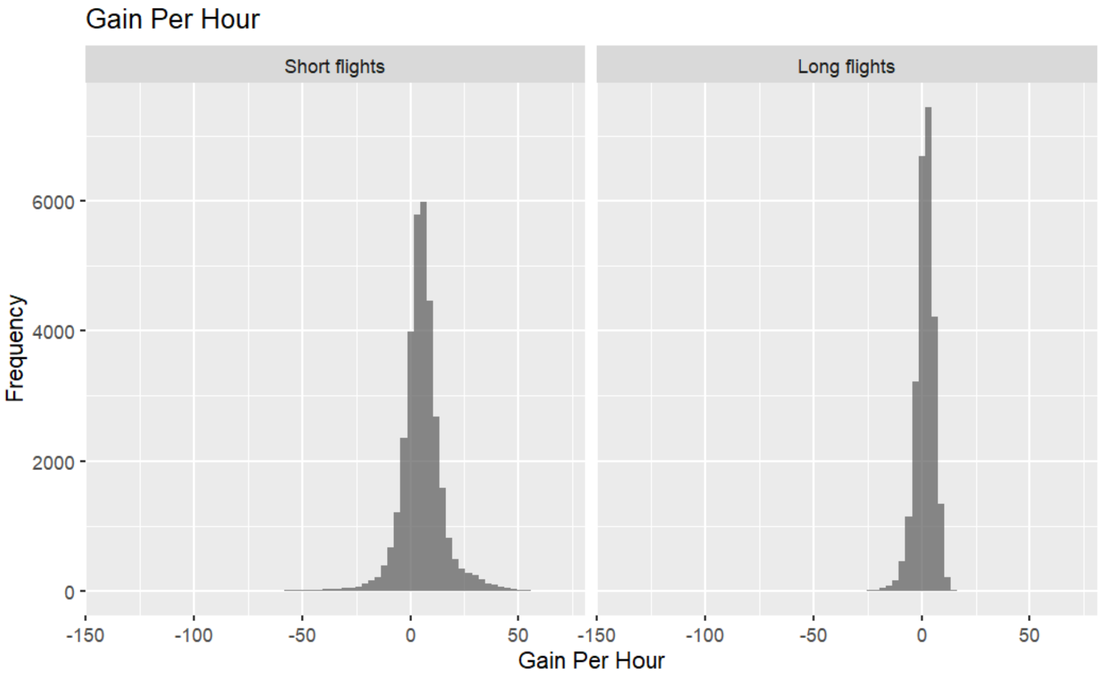


Figure: 6

We have categorized the flights are into two categories the longer flights and shorter flights:

For categorization we have calculated the mean of air duration and divide into shorter and longer flight. From the graph The graph depicts the frequency distribution of gains per hour for short flights and long flights. It shows the number of flights that yielded a particular gain per hour, grouped into intervals ranging from -150 to 50. The frequency of each interval is represented by the height of the bars.

For short flights, the majority of flights resulted in gains between 0 and 50, while a significant portion also yielded gains between -50 and 0. Long flights exhibited a more scattered distribution, with peaks at 0-50 and -50-0.

Overall, gains were more prevalent for both short and long flights, but long flights were also more likely to experience losses.

b. Permutation Tests Results:

The permutation test was conducted to evaluate whether there is a statistically significant difference in gain per hour between flights with long flight times and those without. The observed difference in mean gain per hour for long flights compared to non-long flights was calculated. By randomly permuting the assignment of flight times while maintaining the overall distribution, we generated 1,000 permutations to simulate the null hypothesis.

The resulting p-value of 0.002 indicates that the observed difference is statistically significant at a conventional significance level of 0.05. This suggests that flights with long durations tend to have a different gain per hour compared to flights with shorter durations. The low p-value provides evidence to reject the null hypothesis of no difference in gain per hour between the two groups.

This analysis enhances our understanding of the impact of flight duration on financial performance and informs decision-making in optimizing resource allocation for flights of varying durations.

7 Conclusion

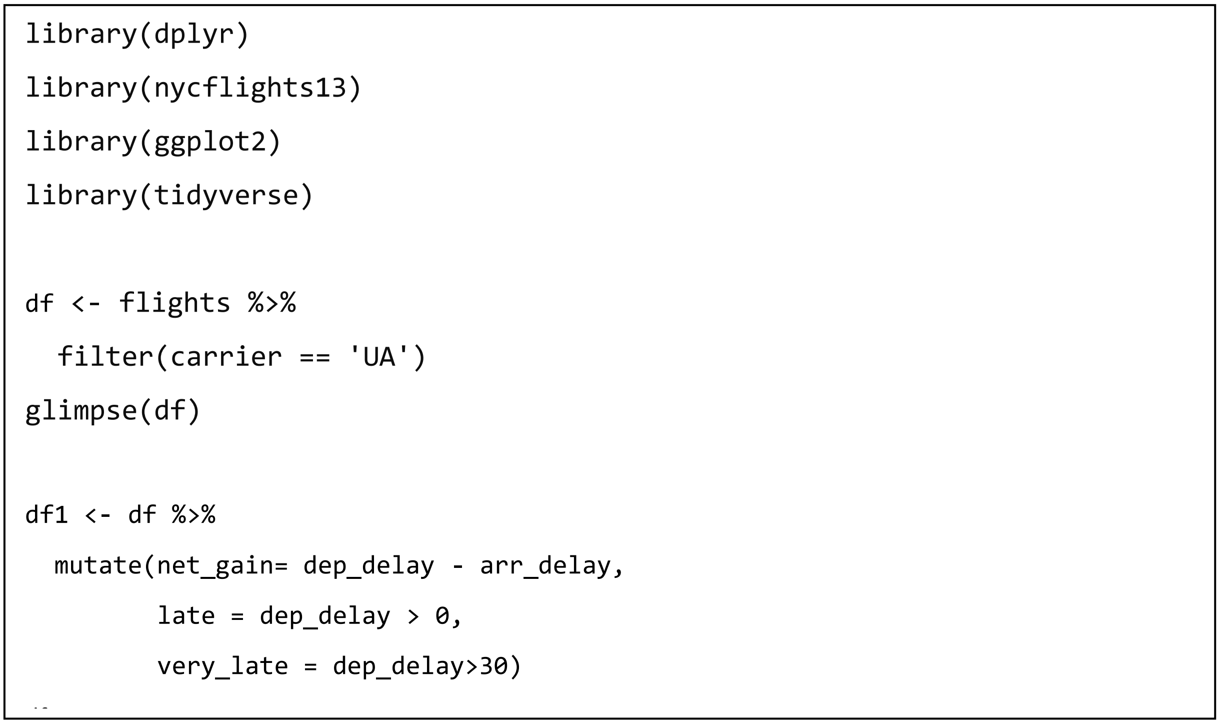
In this analysis, we examined the relationship between departure delay, flight duration, and gain per hour for a dataset of flights. The following key findings and conclusions have been derived from the conducted t-tests and permutation test:

- Late departures, especially those exceeding 30 minutes, demonstrated a statistically significant increase in gain per hour. This suggests potential financial benefits for delayed flights.

- Flights with longer durations were found to have a significantly different gain per hour compared to shorter-duration flights. This emphasizes the need for tailored resource allocation and pricing strategies.

8 Appendix

The following R code provides a step-by-step analysis of flight data for Project 2. In this project, we explore and analyze various factors that might affect flight departure delays. The code is organized into different sections for data preparation, exploratory data analysis, and permutation testing. Here is a detailed explanation of the code:



```
library(dplyr)
library(nycflights13)
library(ggplot2)
library(tidyverse)

df <- flights %>%
  filter(carrier == 'UA')
glimpse(df)

df1 <- df %>%
  mutate(net_gain= dep_delay - arr_delay,
         late = dep_delay > 0,
         very_late = dep_delay>30)
```

Step 2. Departure Delays and Net Gain:

```
hist(df1$net_gain)


df1 <- df1%>%
  filter(!is.na(df1$net_gain))
ggplot(df1, aes(x = late, y = net_gain, fill = late)) +
  geom_boxplot() +
  labs(title = "Net Gain for Flights that Departed Late",
       x = "Departure Status",
       y = "Net Gain")


ggplot(df1, aes(x = very_late, y = net_gain, fill = very_late)) +
  geom_boxplot() +
  labs(title = "Net Gain for Flights that Departed More than 30 Minutes Not Very Late",
       x = "Departure Status (30+ minutes)",
       y = "Net Gain")


# Hypothesis Testing
# for netgain Departed Late vs. Not Departed Late
# Calculate the observed difference in means
observed <- mean(df1$net_gain[df1$late == 1], na.rm = TRUE) -
             mean(df1$net_gain[df1$late == 0], na.rm = TRUE)
```

```r
# Number of permutations
N <- 10^4 - 1

# Initialize a vector to store permutation results
result <- numeric(N)

# Total sample size
sample.size <- nrow(df1)

# Size of group 1 (late departures)
group.1.size <- sum(df1$late == 1)

# Permutation loop
for (i in 1:N) {
  # Randomly shuffle the indices to create a permuted dataset
  index <- sample(sample.size, size = group.1.size, replace = FALSE)

  # Calculate the difference in means for the permuted dataset
  result[i] <- mean(df1$net_gain[index], na.rm = TRUE) -
               mean(df1$net_gain[-index], na.rm = TRUE)
}

# Calculate the p-value
p <- 2 * (sum(result <= observed) + 1) / (N + 1)

# Display the p-value
p
```

```
# Create a summary of time of day categories and departure delays
time_of_day_summary <- ua_data %>%
  group_by(time_of_day_category) %>%
  summarise(
    Total_Flights = n(),
    Mean_Dep_Delay = mean(dep_delay, na.rm = TRUE)
  )
time_of_day_summary

# Create a bar plot
ggplot(time_of_day_summary, aes(x = time_of_day_category, y =
Mean_Dep_Delay)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Mean Departure Delay by Time of Day",
    x = "Time of Day Category",
    y = "Mean Departure Delay (minutes)"
  ) +
  theme_minimal()
```

```
# for avg netgain Departed very Late vs. Not Departed Late
# Calculate the observed difference in means for dep_delay > 30 vs.
dep_delay <= 30
observed <- mean(df1$net_gain[df1$dep_delay > 30], na.rm = TRUE) -
             mean(df1$net_gain[df1$dep_delay <= 30], na.rm = TRUE)
# Number of permutations
N <- 10^4 - 1
# Initialize a vector to store permutation results
result <- numeric(N)
# Total sample size
sample.size <- nrow(df1)
# Size of group 1 (dep_delay > 30)
group.1.size <- sum(df1$dep_delay > 30)
```

```
# Permutation loop
for (i in 1:N) {
  # Randomly shuffle the indices to create a permuted dataset
  index <- sample(sample.size, size = group.1.size, replace = FALSE)
  # Calculate the difference in means for the permuted dataset
  result[i] <- mean(df1$net_gain[index], na.rm = TRUE) -
               mean(df1$net_gain[-index], na.rm = TRUE)
}

# Calculate the p-value
p_dep_delay_gt_30_vs_not <- 2 * (sum(result <= observed) + 1) / (N + 1)

# Display the p-value
p_dep_delay_gt_30_vs_not
```

Step 3: Top Destination and their Gains:

```
# Descriptive Statistics
top_dest_airports <- df1 %>%
  group_by(dest) %>%
  summarize(count = n(), mean_gain = mean(df1$net_gain,na.rm = TRUE))
%>%
  arrange(desc(count)) %>%
  head(5)

top_dest_airports

# Extract the names of the top 5 airports
top_airports <- top_dest_airports$dest

# Filter data for the top 5 airports
top_airports_data <- df1 %>%
  filter(dest %in% top_airports)
# Visualize the distribution of net gain for each top airport using
histograms
ggplot(top_airports_data, aes(x = net_gain, fill = dest)) +
  geom_histogram(binwidth = 10, position = "identity", alpha = 0.7) +
  labs(title = "Distribution of Net Gain for Top 5 Destination
Airports",
       x = "Net Gain",
       y = "Frequency",
       fill = "Destination Airport") +
  facet_wrap(~dest, scales = "free_y") +
  theme_minimal()
```

```
# Summary statistics for net gain for each top airport
summary_stats <- top_airports_data %>%
  group_by(dest) %>%
  summarize(count = n(),
            mean_gain = mean(net_gain, na.rm = TRUE),
            median_gain = median(net_gain, na.rm = TRUE),
            sd_gain = sd(net_gain, na.rm = TRUE))

# Display summary statistics
print(summary_stats)
```

```
Den <- top_airports_data %>%
  filter(dest=="DEN")
t.test(Den$net_gain)$conf

IAH <- top_airports_data %>%
  filter(dest=="IAH")
t.test(Den$net_gain)$conf

LAX <- top_airports_data %>%
  filter(dest=="LAX")
t.test(LAX$net_gain)$conf



ORD <- top_airports_data %>%
  filter(dest=="ORD")
t.test(ORD$net_gain)$conf

SFO <- top_airports_data %>%
```

Step 4: Gain per hour and Departure delays:

```
# Create gain per hour variable
df1 <- mutate(df1,duration_hours = df1$air_time/60 )
df2 <- mutate(df1,gain_per_hour = net_gain / duration_hours)
df2
summary(df2$gain_per_hour)
hist(df2$gain_per_hour)
# Create a 2x2 layout for the plots
par(mfrow = c(2, 2))

# Boxplot: Gain per hour for late vs. not late departures
boxplot(df2$gain_per_hour ~ (df2$dep_delay > 0), main="Gain per Hour
for Late Departures", xlab="Departure Status", ylab="Gain per Hour")

# Boxplot: Gain per hour for flights departing more than 30 minutes
late vs. not late
boxplot(df2$gain_per_hour ~ (df2$dep_delay > 30), main="Gain per Hour
for Departing > 30 mins Late", xlab="Departure Status", ylab="Gain per
Hour")

# Scatterplot: Gain per hour vs. departure delay
#plot(df2$dep_delay, df2$gain_per_hour, main="Scatterplot of Gain per
Hour vs. Departure Delay", #xlab="Departure Delay (minutes)",
ylab="Gain per Hour")

# Histogram: Gain per hour for late departures
hist(df2$gain_per_hour[df2$dep_delay > 0], main="Histogram of Gain per
Hour for Late ", xlab="Gain per Hour")

# Histogram: Gain per hour for flights departing more than 30 minutes
late
hist(df2$gain_per_hour[df2$dep_delay > 30], main="Hist of Gain per Hour
for Flights Departing > 30 mins Late", xlab="Gain per Hour")
```

- T Test:

```
# Hypothesis Testing - t-test for late vs. not late departures

t_test_dep_late_hour <- t.test(gain_per_hour ~ (dep_delay > 0), data =
df2)$conf

t_test_dep_late_hour

# Hypothesis Testing - t-test for flights departing > 30 mins late vs.
not late

t_test_dep_late_30_hour <- t.test(gain_per_hour ~ (dep_delay > 30),
data = df2)$conf

t_test_dep_late_30_hour
```

Step 5: Gain per hour and flight duration :

```
AirMeanDuration<- mean(df2$air_time)

df2<- df2 %>%

  mutate(longflightTime = ifelse(air_time >= AirMeanDuration, 1, 0))

glimpse(df2)

ggplot(df2, aes(x = gain_per_hour)) +

  geom_histogram(binwidth = 3, position = "identity", alpha = 0.7) +

  labs(title = "Gain Per Hour",

       x = "Gain Per Hour",

       y = "Frequency") +

  facet_wrap(~factor(longflightTime, labels = c("Short flights", "Long
flights")))
```

```
observed <- mean(df2$gain_per_hour[df2$longflightTime==1]) -
mean(df2$gain_per_hour[df2$longflightTime == 0])
N <- 10^3 - 1
result <- numeric(N)
sample.size <- nrow(df2)
group.1.size <- nrow(df2[df2$longflightTime==0,])
for(i in 1:N)
{
  index <- sample(sample.size, size=group.1.size, replace = FALSE)
  result[i] <- mean(df2$gain_per_hour[index], na.rm = TRUE) -
mean(df2$gain_per_hour[-index], na.rm = TRUE)
}
p <- 2 * (sum(result <= observed) + 1) / (N + 1)
p
```