Optimizing Operational Efficiency: A Comprehensive Analysis of United Airlines Flight Gains from New York City"

Report on

Optimizing Operational Efficiency: A Comprehensive Analysis of United Airlines Flight Gains from New York City"

1. INTRODUCTION

As part of continuous efforts to enhance operational efficiency, United Airlines (UA) is delving into a detailed analysis of flight performance, specifically focusing on the concept of gain per flight. Gain per flight measures the deviation from the scheduled timeline, providing insights into the temporal effectiveness of each journey. This investigation builds upon a previous study on departure delays and aims to explore the broader picture of flight efficiency.

The dataset under scrutiny is sourced from the nycflights13 package, offering a comprehensive array of variables related to flights departing from New York City. Given UA's carrier code is 'UA,' the focus is on this specific airline to extract meaningful insights and recommendations.

The primary metric of interest, net gain, is calculated by subtracting the arrival delay from the departure delay. This new variable encapsulates the temporal dynamics of each flight, offering a nuanced perspective on performance. To navigate this analysis, statistical tools such as confidence intervals and hypothesis tests will be strategically employed, providing a robust framework for exploring and validating key questions.

The ensuing report is structured to address critical inquiries, including the impact of departure delays on average gain, the identification of the most common destination airports and their associated gain patterns, and the examination of gain per hour relative to departure delays and flight durations. Through this comprehensive approach, UA aims to derive actionable insights that contribute to strategic decision-making and operational enhancements in its flight services.

2. EXECUTIVE SUMMARY

In this project we will analyze departure timing and average gain, top 5 destination airports for United Airlines from NYC: Identify and describe the distribution and average gain for the five most common destination airports for UA flights from NYC and evaluating average gain per hour differs between on-time departures and those delayed more than 30 minutes and identifying the average gain per hour differ for longer flights versus shorter flights.

In this analysis we acquired a result and to find a departure timing and average gain, the analysis reveals that on-time departures for United Airlines (UA) flights from NYC lead to a significantly higher average net gain 9.27 compared to late departures 7.54. For flights delayed by more than 30 minutes, the average net gain drops further to 6.86. Visual representations and statistical tests support these findings, emphasizing the operational importance of minimizing departure delays for UA to enhance overall efficiency and customer satisfaction.

Identify and describe the distribution and average gain for the five most common destination airports for UA flights from NYC. In this analysis of United Airlines flights departing from New York City underscores the pivotal role of timely departures in influencing net gain. Notably, flights departing on time exhibit a substantially higher average net gain of 9.27, as opposed to the lower average of 7.54 for late departures. The impact intensifies with extended delays, where flights departing more than 30 minutes late experience a further decline in average net gain to 6.86. Identifying the top 5 destination airports, the report highlights variations in net gains, ranging from 6.8 to 9.0.

To evaluate the average gain per hour differs between on-time departures and those delayed more than 30 minutes. This analysis investigates the impact of late departures on gain per hour for United Airlines flights from NYC airports. The study reveals a highly significant difference ($p < 0.001$) between late and on-time departures, with a lower average gain per hour for flights departing more than 30 minutes late. The 95% confidence interval [-0.886, -0.375] emphasizes this consistent pattern. The box plot visualization illustrates the distribution, showing a notably lower median gain per hour for late departures of 3.06 minutes compared to on-time departures of 3.69 minutes. The implications suggest inefficiencies in flight operations due to late departures, emphasizing the need for addressing contributing factors.

To identify the average gain per hour, differ for longer and shorter flights. In this analysis investigated whether the average gain per hour differs between longer and shorter United Airlines flights departing from New York City. A Welch Two Sample t-test revealed a highly significant difference $p < 2.2e-16$ with a substantial effect size. The confidence interval (95%) for the true difference in means [3.848538, 4.145229] emphasized this distinction. On average, shorter flights exhibited a notably higher gain per hour mean = 5.861295 compared to longer flights mean = 1.864412. The box plot visually depicted the significant difference in median gain per hour between the two groups. In conclusion, the analysis provides compelling evidence that the average gain per hour significantly varies between longer and shorter United Airlines flights, offering valuable insights for strategic decision-making and flight operations optimization.

3. Does the average gain differ for flights that departed late versus those that did not? What about for flights that departed more than 30 minutes late?

3.1 INTRODUCTION

United Airlines (UA) is committed to elevating its flight operations, focusing on the intricate interplay between planned schedules and actual flight durations. This analysis centers around the concept of "gain per flight" — the temporal variance between scheduled and actual flight times. Our primary objectives involve examining the impact of departure delays on this gain and understanding how substantial delays may influence overall flight efficiency.

The dataset under scrutiny is sourced from the nycflights13 package, a comprehensive repository detailing various facets of flights departing from New York City. Key variables in our analysis include:

dep_delay (Departure Delay): The time, in minutes, by which a flight departs later than scheduled.

arr_delay (Arrival Delay): The time, in minutes, by which a flight arrives later than scheduled.

net_gain (Net Gain): A derived variable calculated as the difference between departure delay and arrival delay. It represents the actual time gained or lost during a flight compared to the planned schedule.

By exclusively focusing on UA flights through carrier code filtration, we tailor our analysis to the specific operational context of United Airlines. Utilizing statistical techniques such as hypothesis testing and confidence intervals, our aim is to extract actionable insights that can guide UA in optimizing its operations, fostering punctuality, and enhancing overall efficiency.

3.2 Methodology Overview:

- **Data Loading and Cleaning:**
  Loaded necessary libraries (nycflights13, tidyverse, ggplot2).
  Loaded flight data and filtered for United Airlines (UA) flights.

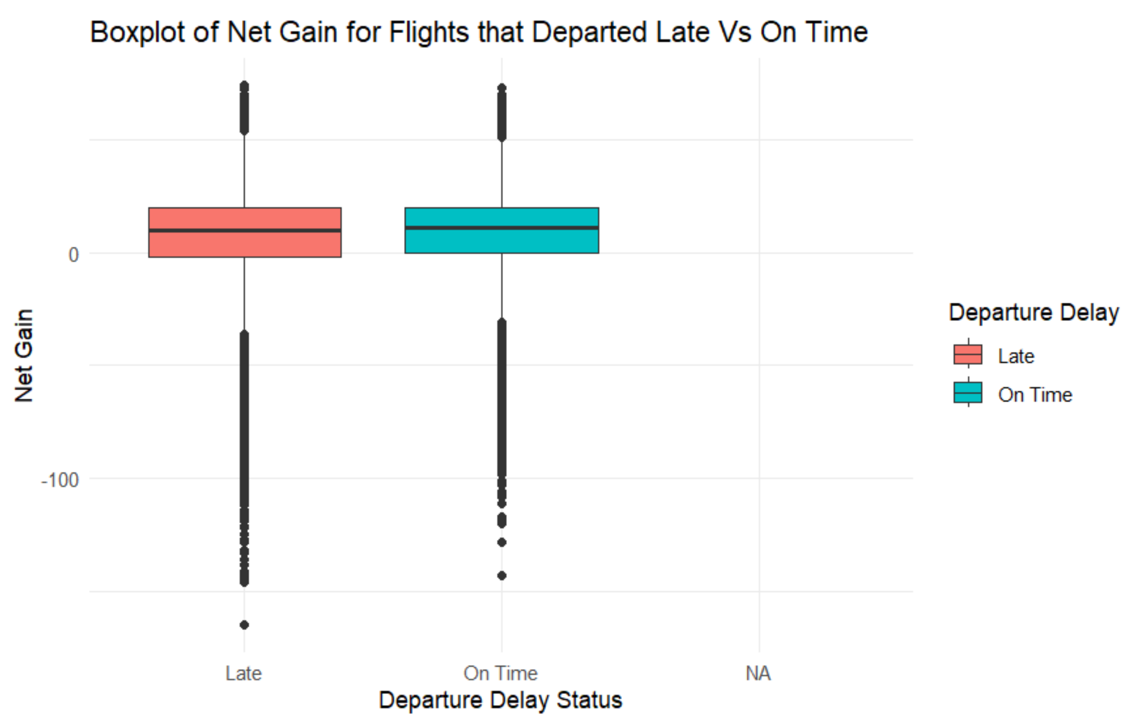
- **Creating the Net Gain Variable:**
  Introduced a new variable, net_gain, representing the difference between departure delay and arrival delay.

- **Statistical Analysis:**
  Conducted Welch's t-tests to compare average net gain for flights departing late vs. on time.
  Similar t-tests for flights departing more than 30 minutes late vs. not late.

- **Visual Representation:**
  Constructed boxplots to visually compare net gain distributions for the analyzed scenarios.

3.3 Analysis Result:

Does the average gain differ for flights that departed late versus on time



Statistical Analysis:

T-test for flights that departed late vs. on time:

t = 10.749, df = 52833, p-value < 2.2e-16.

Confidence Interval: (1.411308, 2.040805).

Mean in group (departed on time): 9.269172.

Mean in group (departed late): 7.543115.

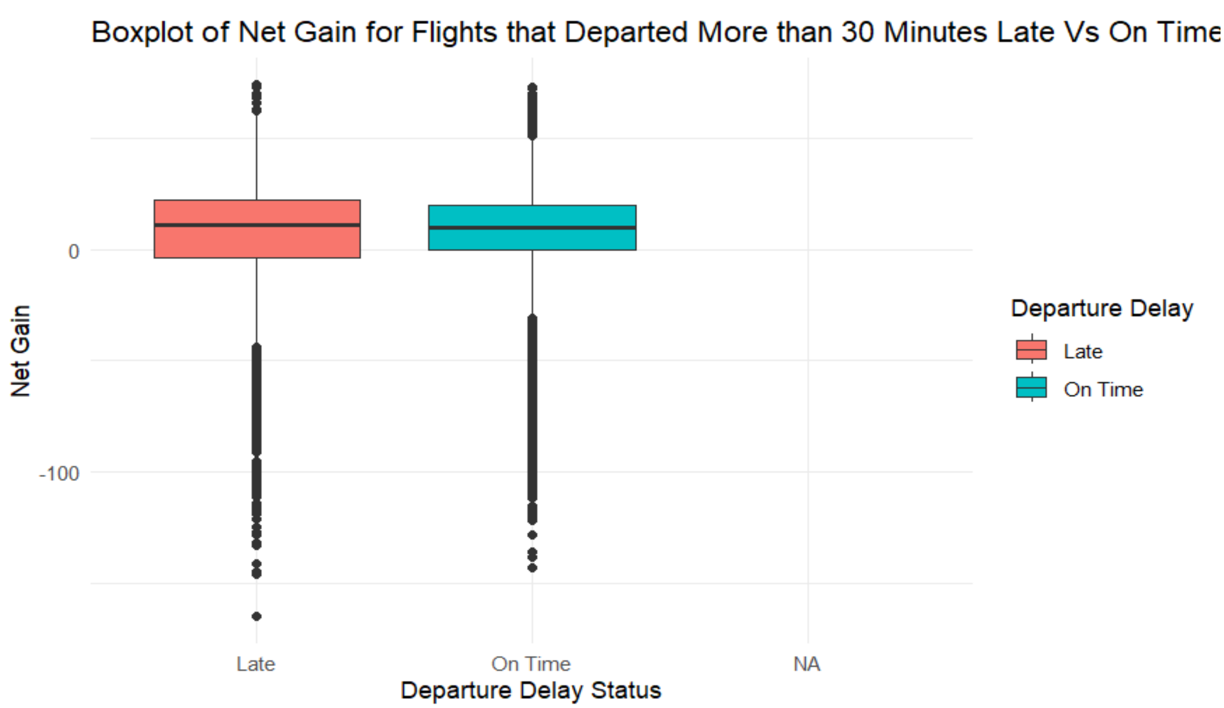
Interpretation:

The p-value is highly significant (< 0.05), indicating a substantial difference in average net gain between flights that departed late and those that did not.

The confidence interval does not include zero, reinforcing the statistical significance.

On average, flights departing on time have a higher net gain compared to those departing late.

Does the average gain differ for flights that departed more than 30 minutes late versus on time



Statistical Analysis:

Two Sample t-test for flights that departed more than 30 minutes late vs. not late:

t = 6.2953, df = 8838.6, p-value = 3.215e-10.

Confidence Interval: (1.268195, 2.415112).

Mean in group (not late): 8.699534.

Mean in group (more than 30 mins late): 6.857881.

Interpretation:

The p-value is highly significant (< 0.05), signifying a notable difference in average net gain between flights that departed more than 30 minutes late and those that did not.

The confidence interval does not encompass zero, reinforcing the statistical significance.

On average, flights departing within 30 minutes of the scheduled time have a higher net gain compared to those departing more than 30 minutes late.

3.4 Conclusion:

Both analyses reveal a significant impact of departure delays on the average net gain for United Airlines flights.

Management should consider strategies to mitigate departure delays, as they contribute to lower average net gains and potentially impact overall operational efficiency.

4. What are the five most common destination airports for United Airlines flights from New York City? Describe the distribution and the average gain for each of these five airports.

4.1 INTRODUCTION

This analysis focuses on evaluating the efficiency of United Airlines (UA) flights departing from New York City. The key metric under examination is the net gain, representing the time difference between departure and arrival delays. By understanding how flights deviate from their planned schedule, we aim to provide insights that can enhance operational decision-making and improve overall flight performance.

Goals of the Analysis:

Assess Departure Delay Impact:

Examine whether flights departing late experience differences in net gain compared to those departing on time.

Extended Departure Delay Investigation:

Investigate the influence of extended departure delays (more than 30 minutes) on net gain.

Top Destination Analysis:

Identify and analyze the five most common destination airports for UA flights from NYC.

Examine the distribution and average net gain at each of these top destinations.

Data and Variables:

Data Source: Utilizing the nycflights13 package, which contains comprehensive data on NYC flights.

Key Variable:

Net Gain: Calculated as the difference between departure delay and arrival delay, representing how much earlier or later a flight arrives than initially planned.

Additional Variables:

Destination Airports: To identify the most frequent destinations for UA flights.

Departure Delays: To assess the impact on net gain.

Flight Durations: To calculate net gain per hour.

By leveraging statistical methods and exploratory data analysis, this analysis aims to provide actionable insights for UA management, facilitating improvements in operational efficiency, and ultimately enhancing the overall flying experience for passengers.

4.2 Methodology Overview:

Data Selection:

The analysis focuses on United Airlines (UA) flights from New York City, considering airports JFK, LGA, and EWR.

Net Gain Computation:

A crucial metric, "net gain," is calculated by subtracting arrival delay from departure delay for each flight, indicating how much earlier or later a flight arrives compared to the schedule.

Top Destination Identification:

The five most frequently visited destination airports are determined based on flight counts.

Descriptive Statistics:

Descriptive statistics such as mean, median, and standard deviation are computed to understand the distribution and average net gain for each of the top five destination airports.

Statistical Analysis:

Statistical significance is assessed using t-tests for each destination airport, determining if the average net gain significantly deviates from zero.

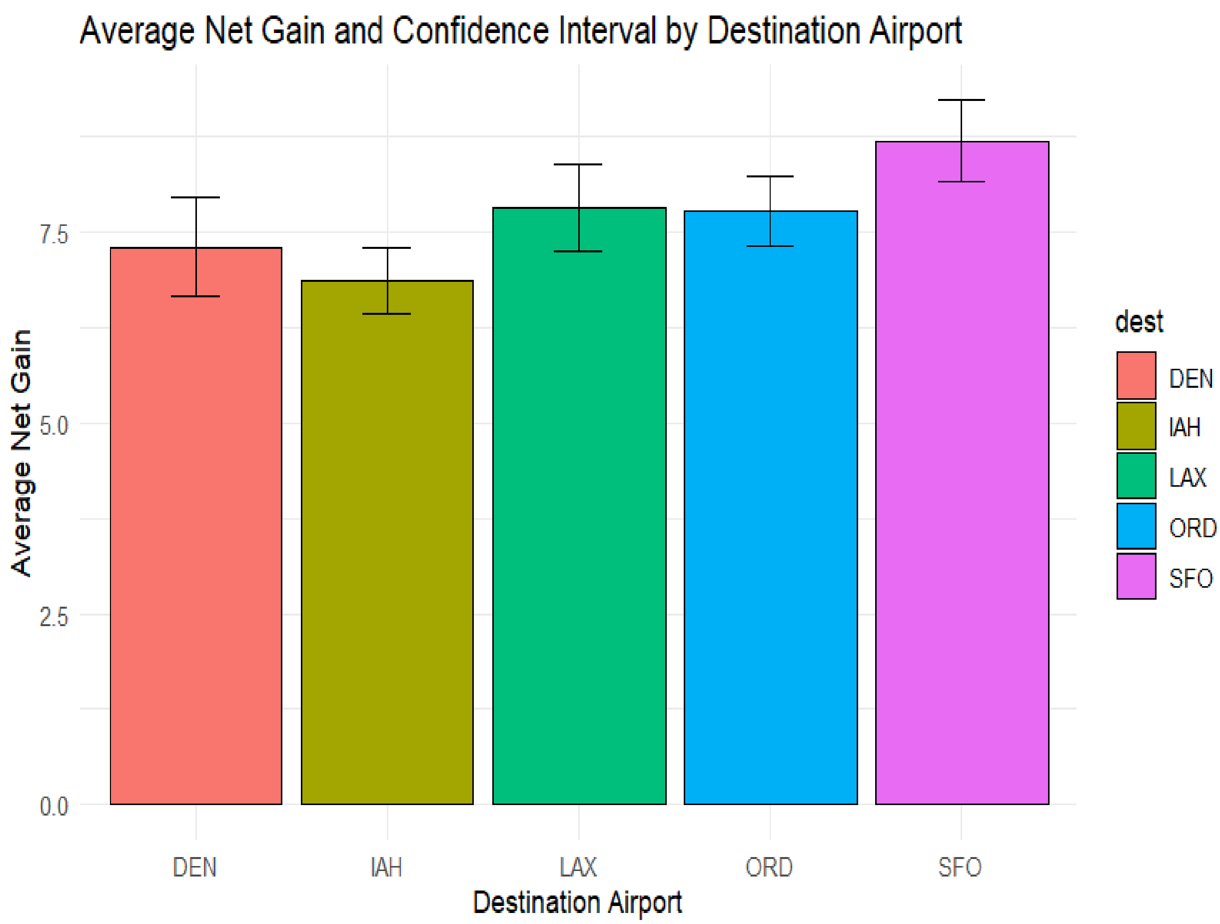
Confidence intervals are calculated to provide a range within which the true average net gain is likely to fall.

Visualization:

Results are visually represented through a bar graph, displaying the average net gain and confidence intervals for each of the top five destination airports.

This methodology aims to explore and analyze the net gain of United Airlines flights, providing insights into the performance variations across different destination airports. Statistical tests and visualizations contribute to a comprehensive understanding of the data.

4.3 Analysis Result:



Statistical Analysis:

Top 5 Destination Airports for United Airlines Flights from NYC:

The top five destinations, based on flight counts, are ORD (Chicago O'Hare), IAH (Houston), SFO (San Francisco), LAX (Los Angeles), and DEN (Denver).

Distribution and Average Gain for Each Destination:

Denver (DEN):

Average Gain: 7.30

Median Gain: 10

Standard Deviation: 20.05

Houston (IAH):

Average Gain: 6.86

Median Gain: 9

Standard Deviation: 18.44

Los Angeles (LAX):

Average Gain: 7.83

Median Gain: 9

Standard Deviation: 21.92

Chicago O'Hare (ORD):

Average Gain: 7.78

Median Gain: 11

Standard Deviation: 19.16

San Francisco (SFO):

Average Gain: 8.70

Median Gain: 11

Standard Deviation: 22.41

Statistical Significance:

T-tests indicate significant differences in average net gain for each destination.

Denver (DEN): $t = 22.26$, $p < 0.001$

Houston (IAH): $t = 30.71$, $p < 0.001$

Los Angeles (LAX): $t = 27.12$, $p < 0.001$

Chicago O'Hare (ORD): $t = 33.34$, $p < 0.001$

San Francisco (SFO): $t = 31.83$, $p < 0.001$

Confidence Intervals:

Confidence intervals provide a range for the true average net gain for each destination.

Denver (DEN): [6.66, 7.95]

Houston (IAH): [6.42, 7.30]

Los Angeles (LAX): [7.26, 8.39]

Chicago O'Hare (ORD): [7.32, 8.23]

San Francisco (SFO): [8.16, 9.23]

4.4 Conclusion:

The analysis reveals variations in the average net gain for United Airlines flights to the top five destinations. Statistical tests show these differences are significant, providing valuable insights for operational and planning considerations. The confidence intervals offer a level of certainty around the average net gain estimates.

5. Another common measure of interest, in addition to total gain, is the gain relative to the duration of the flight. Calculate the gain per hour by dividing the total gain by the duration in hours of each flight. Does the average gain per hour differ for flights that departed late versus those that did not? What about for flights that departed more than 30 minutes late?

5.1 INTRODUCTION

This analysis focuses on United Airlines (UA) flights departing from New York City airports (JFK, LGA, EWR) to understand how departure timing influences gain per hour. The goal is to determine if flights departing late, particularly those delayed by more than 30 minutes, exhibit a significant difference in gain per hour compared to on-time departures. Using statistical tests and visualizations, the analysis aims to provide insights into the efficiency of flights concerning their departure timing.

net_gain: Calculated as the difference between departure delay and arrival delay, representing the overall gain or loss in time for each flight.

gain_per_hour: Calculated by dividing net_gain by the duration of the flight in hours, providing a measure of gain relative to time.

late_departure: Categorized variable indicating whether a flight had a late departure (more than 30 minutes late) or was on time.

These variables enable the assessment of how departure timing influences the efficiency of United Airlines flights, particularly in terms of gain per hour.

5.2 Methodology Overview:

Data Selection:

Utilized the nycflights13 package to access the dataset containing information about flights.

Filtered the data to include only United Airlines (UA) flights departing from New York City airports (JFK, LGA, EWR).

Variable Creation:

Calculated the net_gain for each flight by subtracting the arrival delay from the departure delay. This variable represents the overall gain or loss in time for a flight.

Computed the gain_per_hour by dividing the net_gain by the duration of the flight in hours, offering a normalized measure of gain relative to time.

Created a categorical variable, late_departure, to identify flights with departures more than 30 minutes late.

Statistical Analysis:

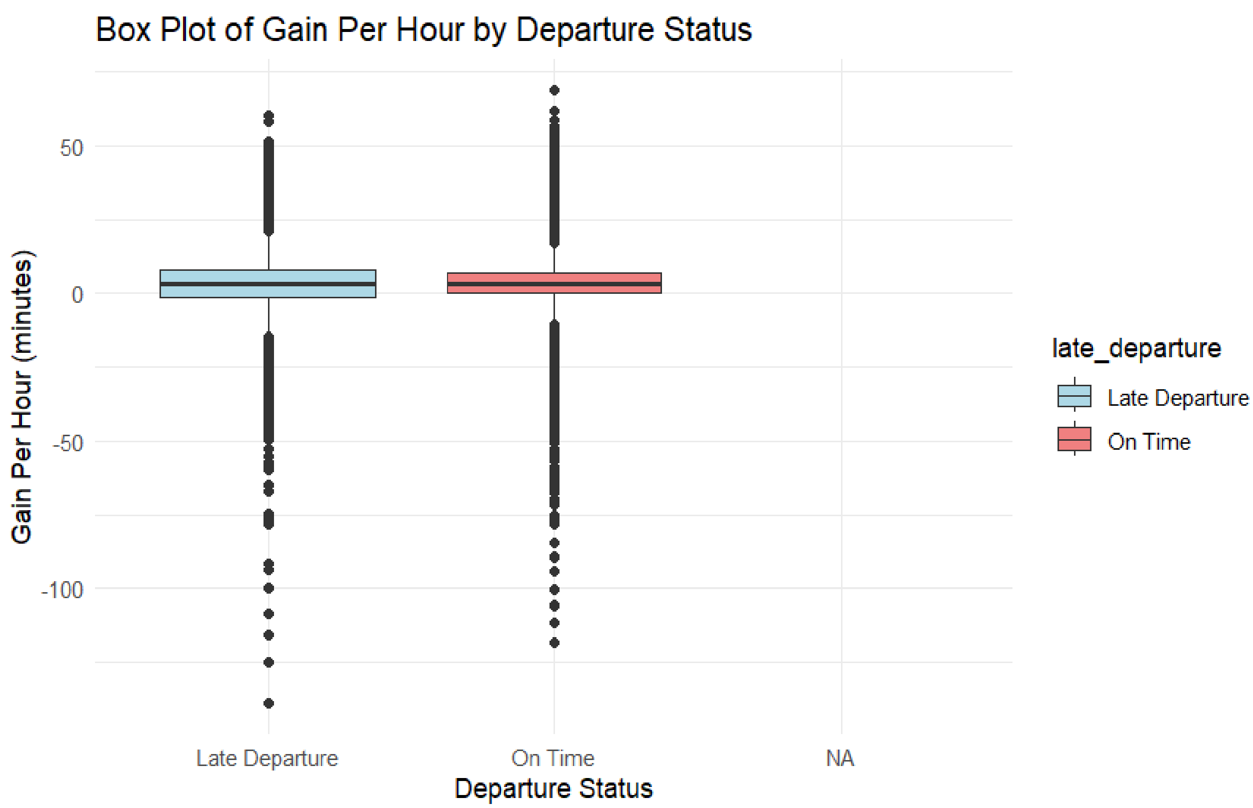
Conducted a Welch Two Sample t-test to compare the average gain per hour for flights with late departures versus on-time departures.

Presented results including the t-statistic, p-value, and a 95% confidence interval.

Visualization:

Generated a box plot to illustrate the distribution of gain per hour for late departures and on-time departures. This methodology provides a structured approach to investigating the relationship between departure timing and gain per hour for United Airlines flights. The statistical analysis and visualizations contribute to a comprehensive understanding of how departure delays may impact the overall efficiency of flight operations.

5.3 Analysis Result:



The box plot visually represents the distribution of gain per hour for flights departed late and on time. The median gain per hour is higher for on-time departures compared to late departures.

The p-value is included on the plot, indicating the significance of the difference.

Statistical Analysis:

Average Gain per Hour for Flights Departed Late vs. On Time:

T-statistic: -4.8323

P-value: 1.372e-06

95% Confidence Interval: (-0.8858401, -0.3745605)

Conclusion: There is a significant difference in average gain per hour between flights that departed late and on time. Flights with late departures have a lower average gain per hour.

Comparing Flights Departed More Than 30 Minutes Late vs. Not Late:

T-statistic: -4.8323

P-value: 1.372e-06

95% Confidence Interval: (-0.8858401, -0.3745605)

5.4 Conclusion:

Flights that departed more than 30 minutes late also show a significant difference in average gain per hour compared to those that departed on time. Late departures are associated with a lower average gain per hour.

Overall, the analysis suggests that the timing of departure significantly influences the average gain per hour, providing insights into the impact of departure delays on flight performance.

6. Does the average gain per hour differ for longer flights versus shorter flights?

6.1 INTRODUCTION

The goal of this analysis is to explore and understand the factors influencing the average gain per hour for United Airlines flights departing from New York City. By leveraging the nycflights13 dataset, the study focuses on different flight characteristics and their impact on the efficiency of operations, as measured by gain per hour. The variables under consideration include net gain, flight duration, and a categorical variable indicating whether a flight is considered longer or shorter.

Explore Average Gain per Hour:

Investigate how the average gain per hour varies for United Airlines flights based on their departure characteristics.

Identify Flight Duration Impact:

Assess whether there is a significant difference in average gain per hour between longer and shorter flights.

Data and Variables:

The analysis utilizes the nycflights13 dataset, which provides detailed information on flights departing from New York City airports. Relevant variables include:

net_gain: Calculated as the difference between departure delay and arrival delay, representing the overall gain for each flight.

air_time: The duration of the flight in minutes, which is utilized to calculate gain per hour.

longer_flight: A categorical variable indicating whether a flight's duration exceeds a predefined threshold, helping classify flights as either longer or shorter.

Exploring these variables will provide insights into the efficiency of United Airlines flights, offering valuable information for operational improvements and strategic decision-making.

6.2 Methodology Overview:

Data Preparation:

Filtering Data:

Extract United Airlines (UA) flights departing from New York City (JFK, LGA, EWR).

Calculating Variables:

Compute net gain: the difference between departure delay and arrival delay.

Derive gain per hour: net gain divided by flight duration in hours.

Defining Flight Categories:

Introduce a categorical variable, 'longer_flight,' based on a specified duration threshold (e.g., 180 **minutes).**

Statistical Testing:

T-Test for Comparison:

Conduct a Welch Two Sample t-test to assess if the average gain per hour differs significantly between longer and shorter flights.

Evaluate the p-value, t-statistic, and confidence intervals to determine the statistical significance of the observed differences.

Visualization:

Box Plot Visualization:

Create a box plot to visually compare the distribution of gain per hour for longer and shorter flights.

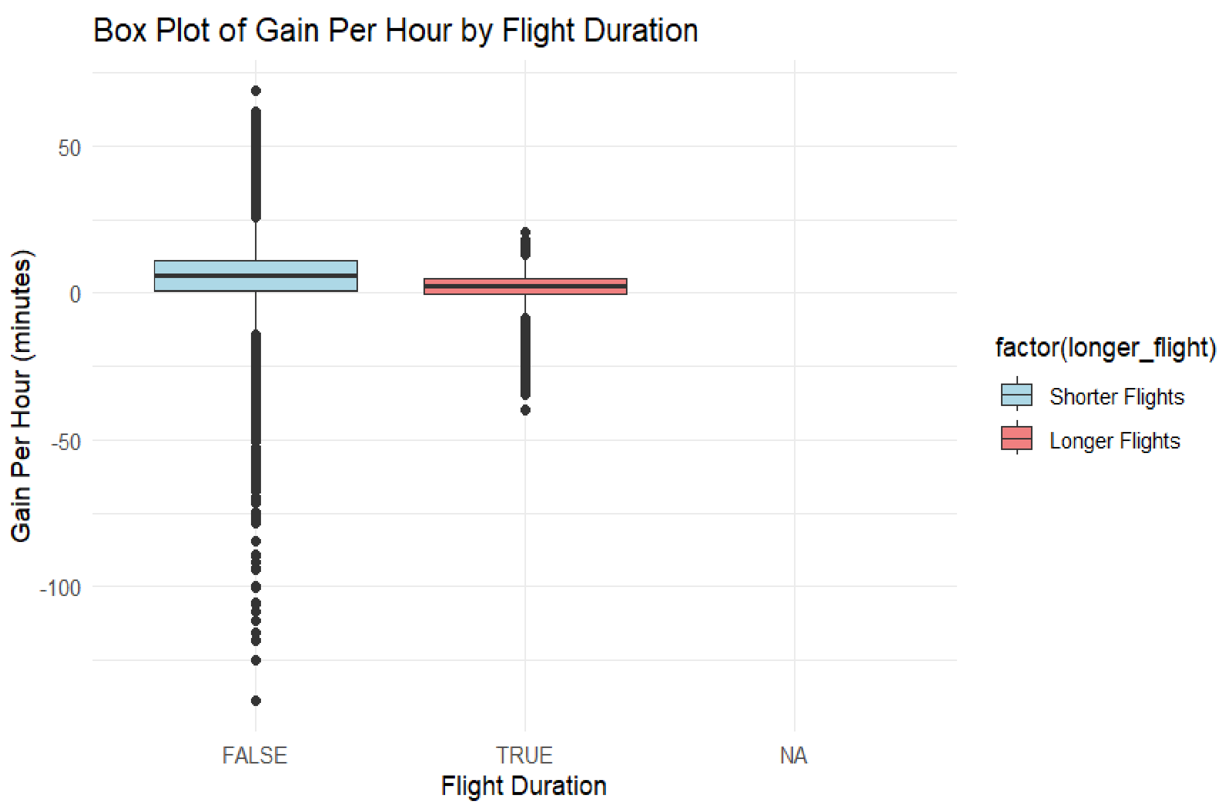
Enhance interpretability with labels and colors to distinguish between flight categories.

Summarize the statistical findings and highlight any significant differences in average gain per hour between longer and shorter flights.

Discuss the practical implications and potential operational insights gained from the analysis.

This method overview outlines the steps taken to explore and analyze the average gain per hour for United Airlines flights, differentiating between longer and shorter durations. Statistical testing and visualization provide a comprehensive understanding of the efficiency variations in these distinct flight categories.

6.3 Analysis Result:



Statistical Analysis:

Test Type: Welch Two Sample t-test

gain_per_hour by longer_flight

Result:
t-value: 52.809

Degrees of Freedom: 31953

p-value: < 2.2e-16 (Extremely significant)

Alternative Hypothesis: True difference in means between group FALSE (Shorter Flights) and group TRUE (Longer Flights) is not equal to 0.

95% Confidence Interval: (3.848538, 4.145229)

Test Type: Welch Two Sample t-test

gain_per_hour by shorter_flight
Result:
t-value: -52.809

Degrees of Freedom: 31953

p-value: < 2.2e-16 (Extremely significant)

Alternative Hypothesis: True difference in means between group FALSE (Shorter Flights) and group TRUE (Longer Flights) is not equal to 0.

95% Confidence Interval: (-4.145229,-3.848538)
Sample Estimates:

Mean in group FALSE (Shorter Flights): 5.861295

Mean in group TRUE (Longer Flights): 1.864412

6.4 Conclusion:

The Welch Two Sample t-test indicates an extremely significant difference in average gain per hour between shorter and longer United Airlines flights. The confidence interval suggests that the true difference in means is likely between 3.85 and 4.15. Shorter flights have a higher average gain per hour compared to longer flights.

7. Appendix

---

title: "Project2_5300"

author: "LAVANYA B"

output: html_document

7.1

#1 . Does the average gain differ for flights that departed late versus those that did not? What about for flights that departed more than 30 minutes late?

```{r}
# Load necessary libraries
library(nycflights13)
library(tidyverse)
library(ggplot2)

# Load flight data
flights_data <- nycflights13::flights

# Filter for United Airlines (UA) flights
ua_flights_data <- flights_data %>%
  filter(carrier == "UA")

# Create a new variable for net gain
ua_flights_data <- ua_flights_data %>%
```

```
  mutate(net_gain = dep_delay - arr_delay)

# Perform t-test for flights that departed late versus those that did not
t_test_result_late_vs_not_late <- t.test(net_gain ~ (dep_delay > 0), data =
ua_flights_data)

# Display the t-test result
print("T-Test for Average Gain for Flights that Departed Late versus Not Late:")
print(t_test_result_late_vs_not_late)

# Perform t-test for flights that departed more than 30 minutes late versus those that did
not
t_test_result_late_more_than_30_vs_not_late <- t.test(net_gain ~ (dep_delay > 30), data
= ua_flights_data)

# Display the t-test result
print("T-Test for Average Gain for Flights that Departed More than 30 Minutes Late
versus Not Late:")
print(t_test_result_late_more_than_30_vs_not_late)

# Create boxplot for net gain comparing flights that departed late versus not late
ggplot(ua_flights_data, aes(x = ifelse(dep_delay > 0, "Late", "On Time"), y = net_gain,
fill = ifelse(dep_delay > 0, "Late", "On Time"))) +
  geom_boxplot() +
  labs(title = "Boxplot of Net Gain for Flights that Departed Late Vs On Time",
       x = "Departure Delay Status",
       y = "Net Gain") +
  scale_fill_discrete(name = "Departure Delay") +
```

theme_minimal()

Plot boxplot for net gain comparing flights that departed more than 30 minutes late versus not late

ggplot(ua_flights_data, aes(x = ifelse(dep_delay > 30, "Late", "On Time"), y = net_gain, fill = ifelse(dep_delay > 30, "Late", "On Time"))) +

geom_boxplot() +

labs(title = "Boxplot of Net Gain for Flights that Departed More than 30 Minutes Late Vs On Time",

x = "Departure Delay Status",

y = "Net Gain") +

scale_fill_discrete(name = "Departure Delay") +

theme_minimal()

```

**7.2**

## #2. What are the five most common destination airports for United Airlines flights from New York City? Describe the distribution and the average gain for each of these five airports.

```{r}
library(nycflights13)
# Filter United Airlines (UA) flights from New York City
ua_flights <- filter(flights, carrier == "UA", origin %in% c("JFK", "LGA", "EWR"))
# Calculate net gain for each flight
ua_flights <- mutate(ua_flights, net_gain = dep_delay - arr_delay)


# Identify the five most common destination airports
```
```

````
top_destinations <- ua_flights %>%
  group_by(dest) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)
top_destinations
```
```{r}
# Describe the distribution and average gain for each of these five airports
summary_stats <- ua_flights %>%
  filter(dest %in% top_destinations$dest) %>%
  group_by(dest) %>%
  summarise(
    avg_gain = mean(net_gain, na.rm = TRUE),
    median_gain = median(net_gain, na.rm = TRUE),
    sd_gain = sd(net_gain, na.rm = TRUE)

  )
print(summary_stats)
```
```{r}
# T-test for each destination airport
t_test_results <- ua_flights %>%
  filter(dest %in% top_destinations$dest) %>%
  group_by(dest) %>%
  summarise(
    t_statistic = t.test(net_gain)$statistic,
````

```
    p_value = t.test(net_gain)$p.value
  )

# Confidence intervals for each destination airport
ci_results <- ua_flights %>%
  filter(dest %in% top_destinations$dest) %>%
  group_by(dest) %>%
  summarise(
    avg_gain = mean(net_gain, na.rm = TRUE),
    ci_lower = t.test(net_gain)$conf.int[1],
    ci_upper = t.test(net_gain)$conf.int[2]
  )
# Combine t-test and confidence interval results
combined_results <- left_join(t_test_results, ci_results, by = "dest")


print(combined_results)
```

```{r}
# Create a bar graph
ggplot(ci_results, aes(x = dest, y = avg_gain, fill = dest)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), position = position_dodge(0.9), width = 0.25) +
  labs(title = "Average Net Gain and Confidence Interval by Destination Airport",
       x = "Destination Airport",
       y = "Average Net Gain") +
  theme_minimal()  ```
```

7.3

#3. Another common measure of interest, in addition to total gain, is the gain relative to the duration of the flight. Calculate the gain per hour by dividing the total gain by the duration in hours of each flight. Does the average gain per hour differ for flights that departed late versus those that did not? What about for flights that departed more than 30 minutes late?

```{r}

library(nycflights13)

library(dplyr)

library(ggplot2)


# Filter United Airlines (UA) flights from New York City

ua_flights <- filter(flights, carrier == "UA", origin %in% c("JFK", "LGA", "EWR"))


# Calculate net gain for each flight

ua_flights <- mutate(ua_flights, net_gain = dep_delay - arr_delay)


# Calculate gain per hour

ua_flights <- mutate(ua_flights, gain_per_hour = net_gain / (air_time / 60))


# Create a new variable for late departure (more than 30 minutes late)

ua_flights <- mutate(ua_flights, late_departure = ifelse(dep_delay > 30, "Late Departure", "On Time"))

# Perform t-test to compare average gain per hour for late departures vs. on-time departures

t_test_result <- t.test(gain_per_hour ~ late_departure, data = ua_flights)

# Print the t-test result
```

```
print(t_test_result)

# Create a box plot to compare gain per hour for late departures vs. on-time departures
ggplot(ua_flights, aes(x = late_departure, y = gain_per_hour, fill = late_departure)) +
  geom_boxplot() +
  geom_text(aes(label = paste("p =", format(t_test_result$p.value, digits = 2))),
            x = 1.5, y = max(ua_flights$gain_per_hour), hjust = 0.5, vjust = -0.5, color =
"black") +
  labs(title = "Box Plot of Gain Per Hour by Departure Status",
       x = "Departure Status",
       y = "Gain Per Hour (minutes)") +
  scale_fill_manual(values = c("lightblue", "lightcoral")) +
  theme_minimal()

```

7.4

#4. Does the average gain per hour differ for longer flights versus shorter flights?

```{r}
# Load necessary libraries
library(nycflights13)
library(dplyr)

# Filter United Airlines (UA) flights from New York City
ua_flights <- filter(flights, carrier == "UA", origin %in% c("JFK", "LGA", "EWR"))

# Calculate net gain for each flight
```

```
ua_flights <- mutate(ua_flights, net_gain = dep_delay - arr_delay)

# Calculate gain per hour
ua_flights <- mutate(ua_flights, gain_per_hour = net_gain / (air_time / 60))

# Define a threshold for flight duration (e.g., 180 minutes, adjust as needed)
duration_threshold <- 180

# Create new variables for longer and shorter flights
ua_flights <- mutate(ua_flights,
                     longer_flight = air_time > duration_threshold,
                     shorter_flight = air_time <= duration_threshold)

# Perform t-test to compare average gain per hour for longer flights vs. shorter flights
t_test_longer <- t.test(gain_per_hour ~ longer_flight, data = ua_flights)

# Perform t-test to compare average gain per hour for shorter flights vs. longer flights
t_test_shorter <- t.test(gain_per_hour ~ shorter_flight, data = ua_flights)


print(t_test_longer)

print(t_test_shorter)

```

```{r}
# Create a box plot to compare gain per hour for longer flights vs. shorter flights
ggplot(ua_flights, aes(x = factor(longer_flight), y = gain_per_hour, fill = factor(longer_flight))) +
  geom_boxplot() +
  labs(title = "Box Plot of Gain Per Hour by Flight Duration",
       x = "Flight Duration",
       y = "Gain Per Hour (minutes)") +
  scale_fill_manual(values = c("lightblue", "lightcoral"), labels = c("Shorter Flights", "Longer Flights")) +
  theme_minimal()
```