Gain Per Flight Analysis

DATA 5300, Group Project #2

By: Alyssa Zukas, Prince Newman, Badamgarav Battushig, Edwin Okwor

Agenda

Introduction

- **Domain Problems**
- Imports + Technique
- Data Prepping
- Timeline of Analysis

Problem 1

- Methodology
- Analysis
- Conclusion

Problem 2

- Methodology
- Analysis
- Conclusion

Problem 3

- Methodology
- Analysis
- Conclusion

Problem 4

- Methodology
- Analysis
- Conclusion

Link to Google Colab Notebook: gain_per_flight_analysis.ipynb

Domain Problems

- **Problem 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?
- **Problem 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.
- **Problem 3**
  - 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?
- **Problem 4**
  - Does the average gain per hour differ for longer flights versus shorter flights?

Link to Google Colab Notebook: gain_per_flight_analysis.ipynb

Imports

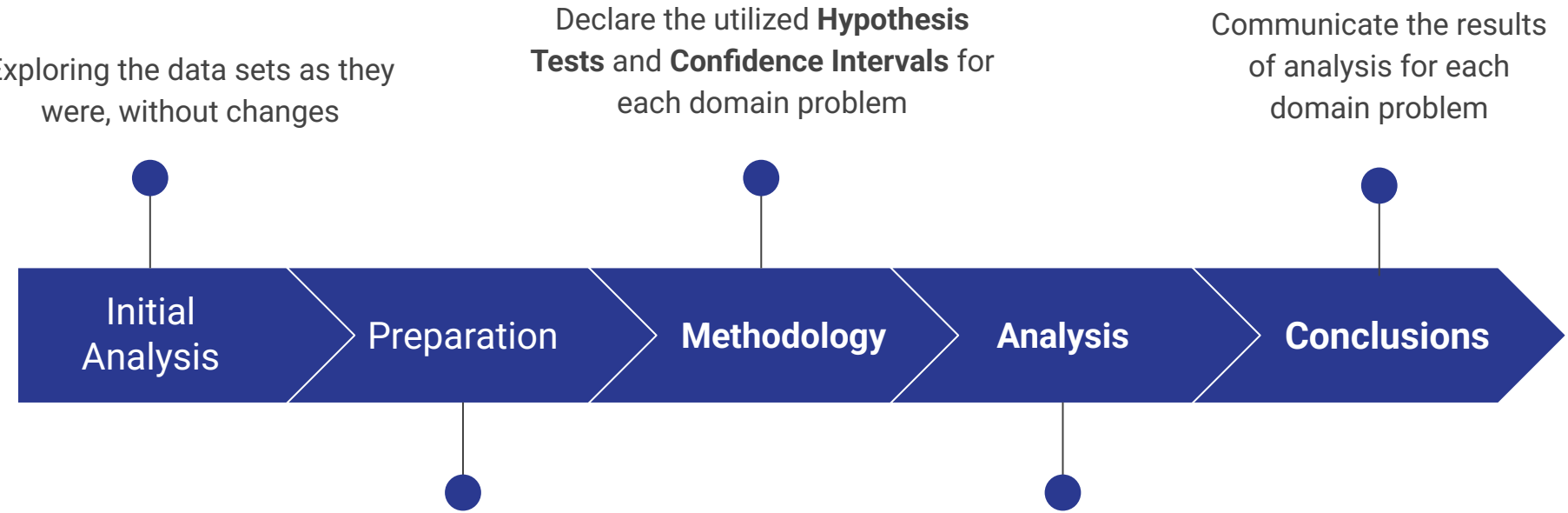
Packages

- **‘nycflights13’**
  - *Airports*
  - *Flights*

Libraries

- **Ggplot2**: Imported for data visualization
- **Dplyr**: Used for data manipulation in R
  - (such as the filter, select, muttate, arrange, and summarize functions)
- **Tidyr**: Used for tidying and reshaping data

Timeline of Analysis



Data Prepping

- Data Cleaning
  - Removed flights with missing values for departures delays, airtimes and arrival delays times
  - Excluded cancelled flights to ensure accuracy
- New Variables Created:
  - **Net gain:** departure delay - arrival delay
  - **Gain per hour:** net gain / (airtime/60)
  - Flight departure delay **status:** whether flights were late, very late or on-time.
- Data Merging + Filtering
  - Joined the flight and airport dataset to obtain airport names for each destination
  - Filter dataset to include only columns of interest

Problem 1

1. Does the average gain differ for flights that departed late versus those that did not?
2. Does the average gain differ for flights that departed very late versus those that did not?

Statistical Methods

Hypothesis Testing: Two-sample t-test to compare means between groups

Confidence Intervals: 95% CI calculated for mean gains to quantify uncertainty

Analysis - Part 1

Does the average gain differ for flights that departed late versus those that did not?

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

μ_1 – mean gain for flights that departed late

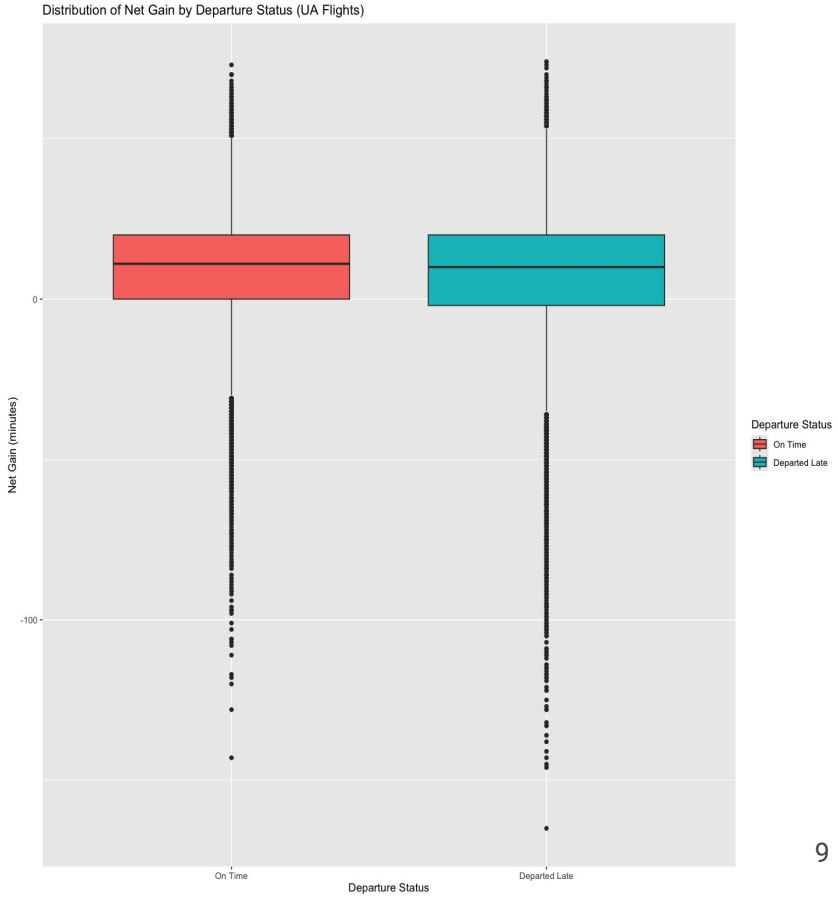
μ_2 – mean gain for flights that did not depart late

Mean Gain:

- Departed Late: ~7.54 average gain minutes
- Did not depart late(Ontime): ~9.27 average gain minutes

Welch Two Sample T-Test

- t-statistic = -10.749
- p-value < 2.2e-16
- 95% confidence interval for the difference = [-2.04, -1.411]



Analysis - Part 2

Does the average gain differ for flights that departed very late versus those that did not?

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

μ_1 – mean gain for flights that departed very late(>30 minutes)

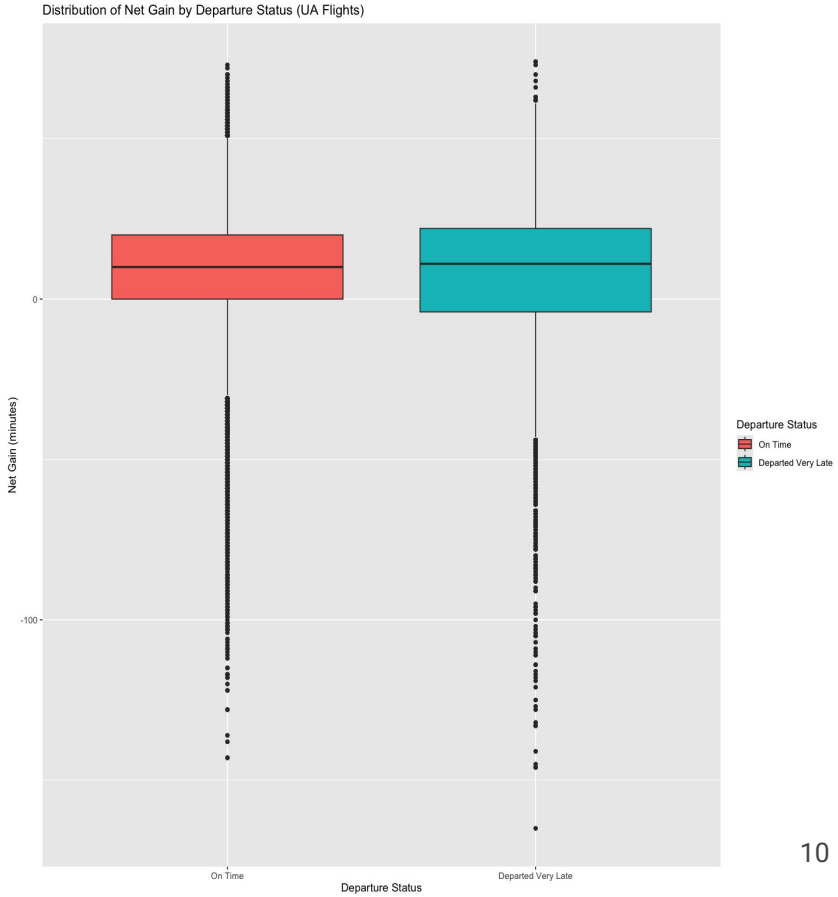
μ_2 – mean gain for flights that did not depart very late

Mean Net Gain:

- Departed Late: ~6.86 average gain minutes
- Did not depart late(Ontime): ~8.69 average gain minutes

Welch Two Sample T-Test

- t-statistic = -6.2953
- p-value < 3.215e-10
- 95% confidence interval for the difference = [-2.42, -1.268]



Conclusion- Part 1 & Part 2

For part 1:

- Given the p-value is very small(<0.05), we reject the null hypothesis
- Late flights gain on average ~1.41 - 2.04 minutes less than flights that do not depart late.
- The difference is statistically significant.

For part 2:

- P-value = 3.215e-10 is very small(<0.05), reject the null hypothesis.
- Very Late flights gain on average ~1.27 to 2.42 minutes less than flights that do not depart very late.
- The difference is statistically significant.

Problem 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.

Analysis - Part 1

What are the five most common destination airports for United Airlines flights from New York City?

Check count of UA flights with origin from NYC

A tibble: 3 × 2

origin	count
<chr>	<int>
EWR	45501
JFK	4478
LGA	7803

Check five (5) most common destination airports

top_destinations

A tibble: 5 × 2

dest	count
<chr>	<int>
IAH	6814
ORD	6744
SFO	6728
LAX	5770
DEN	3737

Analysis - Part 2

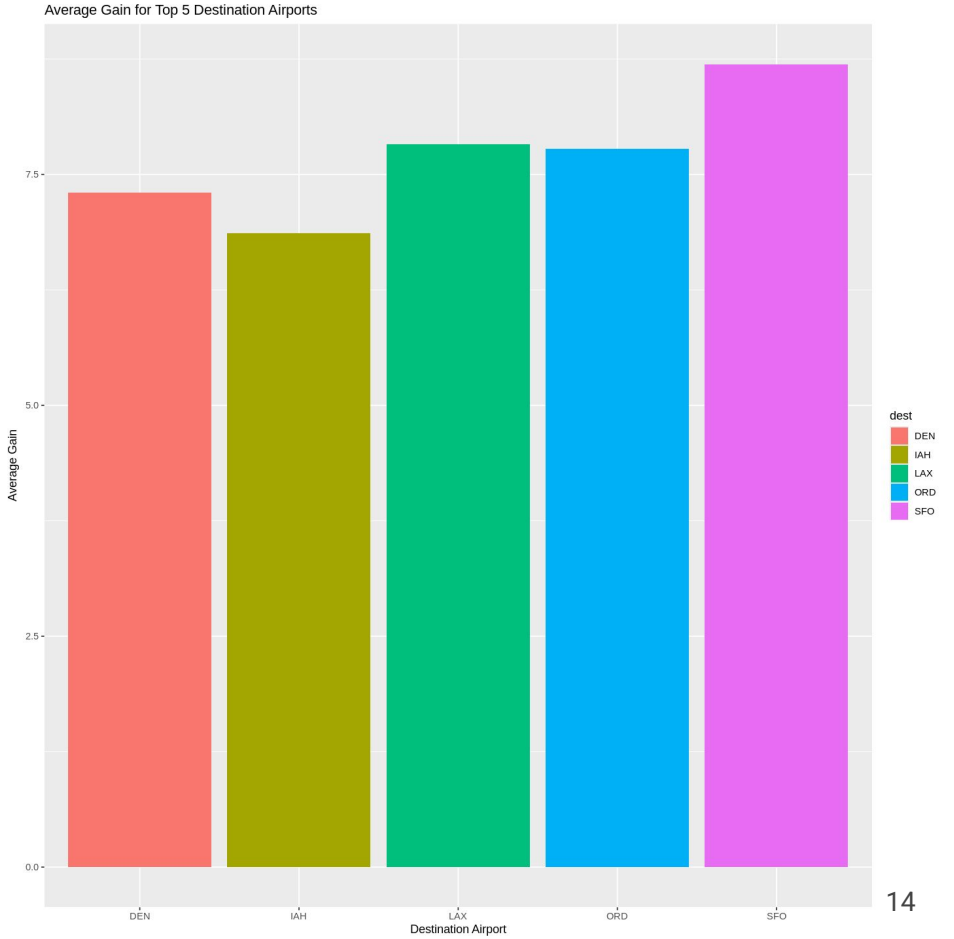
Describe the distribution and the average gain for each of these five airports.

- Average gain for top 5 destination airports (Histogram distribution)

average_gain_table

A tibble: 5 × 2

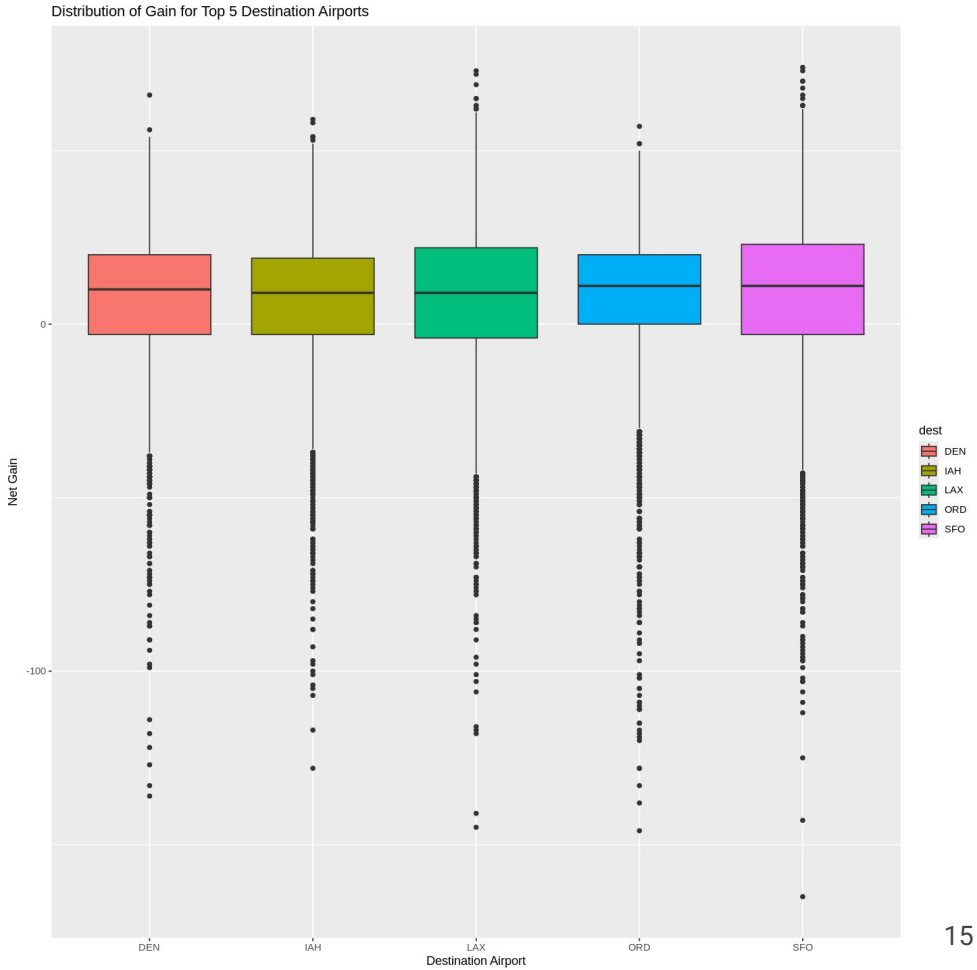
dest	average_gain_minutes
<chr>	<dbl>
DEN	7.302382
IAH	6.861755
LAX	7.825303
ORD	7.777432
SFO	8.695006



Analysis - Part 2

Describe the distribution and the average gain for each of these five airports.

- Average gain for top 5 destination airports (Boxplot distribution)



Conclusion - Part 1

Part 1 Summary: Most Common Destination Airports

The five most common destination airports in order of flight count are:

- IAH (Houston George Bush Intercontinental): 6,814 flights
- ORD (Chicago O'Hare International): 6,744 flights
- SFO (San Francisco International): 6,728 flights
- LAX (Los Angeles International): 5,770 flights
- DEN (Denver International): 3,737 flights

All five destination airports show a positive average net gain, meaning that on average, United Airlines flights from NYC to these locations are making up time while airborne. SFO (San Francisco) demonstrates the highest average time recovery at approximately 8.70 minutes. This suggests the SFO route is the most efficient at compensating for ground delays. IAH (Houston) shows the lowest average time recovery at approximately 6.86 minutes among the top five.

The data shows a strong concentration of flights to major airport hubs across the United States, with a relatively small difference in flight count among the top three destinations (IAH, ORD, and SFO).

Conclusion - Part 2

Part 2 Summary: Distribution and the average gain for each of these five airports.

SFO (San Francisco): SFO shows the strongest time recovery, with flights gaining an average of 8.7 minutes. Its distribution is skewed toward positive gains, featuring many flights that recover over 100 minutes. Despite these extreme outliers, most flights cluster tightly around small gains or losses, indicating generally stable performance with occasional large recoveries.

LAX (Los Angeles): LAX recovers an average of 7.8 minutes, making it the second-best performer. Its distribution resembles SFO and ORD, with several major time-recovery outliers but also some significant time-loss outliers. This suggests that while the route typically makes up time, a notable subset of flights still encounters delays in the air.

ORD (Chicago): ORD averages a 7.8-minute recovery and displays a distribution tightly centered near zero, reflecting consistent performance. Although both time-gaining and time-losing outliers appear, the positive extremes are strong enough to pull the overall average up, showing a reliable but occasionally variable route.

DEN (Denver): DEN flights gain an average of 7.3 minutes, slightly below the top three destinations. Its mid-range variability is somewhat tighter, but the route still experiences substantial positive and negative outliers, characteristic of long-haul flights that can either recover significant time or encounter meaningful setbacks.

IAH (Houston): IAH has the lowest average recovery at 6.9 minutes, indicating less flexibility for making up delays. The distribution shows the smallest overall spread, suggesting consistently scheduled performance, though a cluster of flights losing time in the air contributes to its comparatively lower average gain.

Problem 3

Does the average gain per hour differ for flights that departed late versus those that did not?

Does the average gain per hour differ for flights that departed late versus those that departed more than 30 minutes late?

Statistical Methods

Hypothesis Testing:
Two-sample t-tests were used to compare the mean gain per hour between the defined departure groups (late vs. on-time, and very late vs. not very late).

Confidence Intervals:
95% confidence intervals were computed for the differences in mean gain per hour to quantify the uncertainty around the estimated group differences.

Analysis - Part 1

Does the average gain per hour differ for flights that departed late versus those that did not?

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

μ_1 – mean gain per hour for flights that departed late

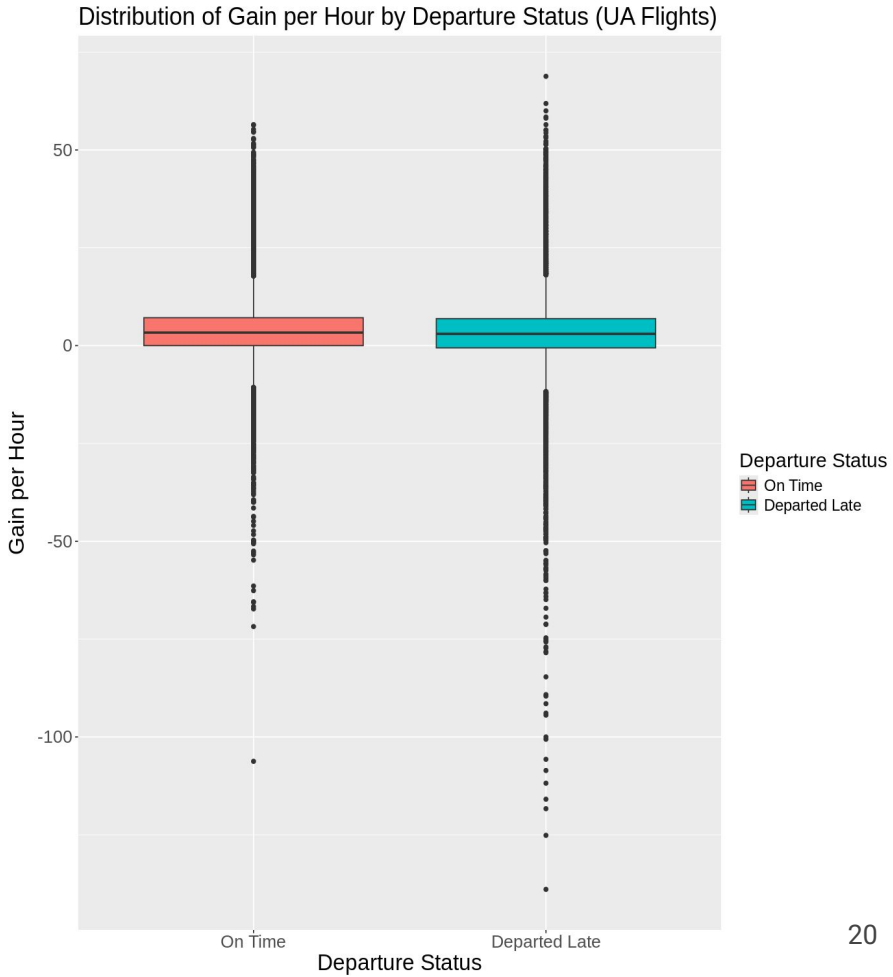
μ_2 – mean gain per hour for flights that did not depart late

Mean Net Gain:

- Departed Late: ~3.18 average gain minutes
- Did not depart late(Ontime): ~3.99 average gain minutes

Welch Two Sample T-Test

- t-statistic = -11.285
- p-value < 2.2e-16
- 95% confidence interval for the difference = [-0.95, -0.67]



Analysis - Part 2

Does the average gain per hour differ for flights that departed late versus those that departed more than 30 minutes late?

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

μ_1 – mean gain per hour for flights that departed more than 30 minutes late

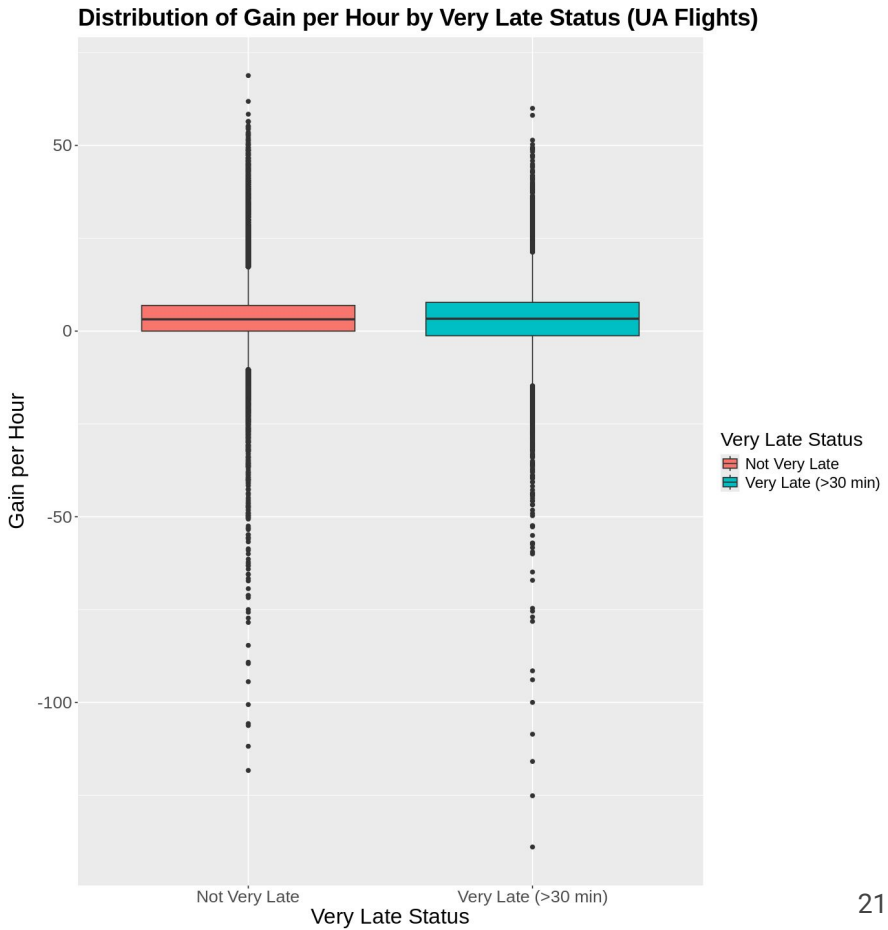
μ_2 – mean gain per hour for flights that did not depart more than 30 minutes late

Mean Net Gain:

- Departure very late (>30min) : ~3.06 average gain minutes
- Departed Late (≤30 min) : ~3.69 average gain minutes

Welch Two Sample T-Test

- t-statistic = -4.8323
- p-value < 1.372e-06
- 95% confidence interval for the difference = [-0.89, -0.37]



Conclusion- Part 1 & Part 2

For part 1:

- Late flights make less gain per hour than on-time flights.
- The difference is statistically significant.

For part 2:

- Flights that depart more than 30 minutes late also gain less time per hour than the not-very-late group.
- This difference is also statistically significant.

Problem 4

Does the Average Gain Per Hour differ for Long Flights vs Shorter Flights?

Methodology

Null Hypothesis (H_0):

non-technically written: Mean GPH of Long Flights *equals* Mean GPH of Short Flights

technically written: μlong = μshort

Alternative Hypothesis (H_1):

non-technically written: Mean GPH of Long Flights *does not equal* Mean GPH of Short Flights

technically written: μlong != μshort

Analysis

Median Flight Times:

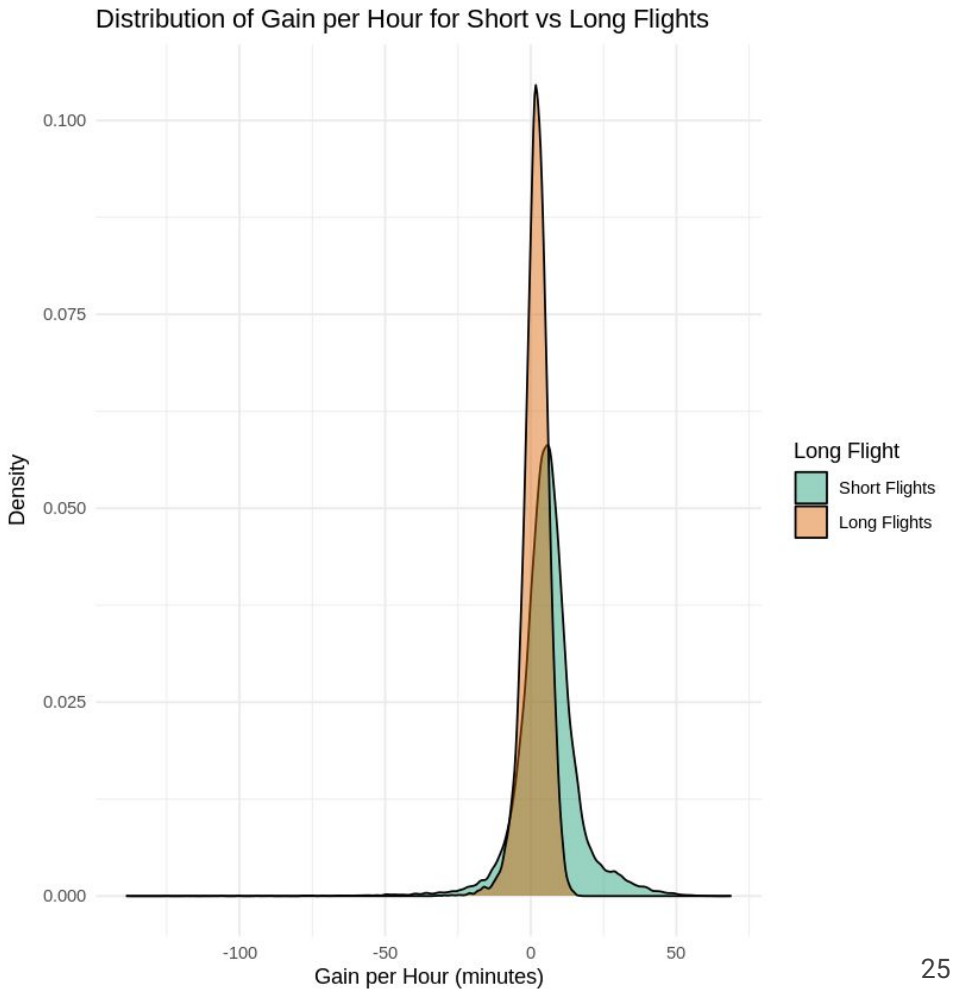
- Long flights: airtime ≥ median of 197 minutes
- Short flights: airtime < median of 197 minutes

Mean Gain Per Hour (GPH):

- Long flights: 1.64 average gained minutes
- Short flights: 5.59 average gained minutes

Welch Two Sample T-Test:

- t-statistic = 57.445
- p-value < 2.2e-16
- 95% confidence interval for the difference = [3.822, 4.0925]



Conclusions

Group Means:

- Short flights: mean gain per hour = **5.599 minutes**
- Long flights: mean gain per hour = **1.642 minutes**

Short flights gain much more time per hour than long flights – about 4 minutes more per hour, on average.

Confidence Interval:

- The 95% confidence interval for the true difference in **mean gain per hour** is:

  [3.822, 4.093]

*We are 95% confident that short flights gain between **3.8** and **4.1** more minutes per hour than long flights.*

Because the entire CI is positive, the difference is statistically meaningful.

Statistical Conclusion

- **t = 57.445**
- **p-value < 2.2e-16**

Hypothesis Testing:

Null Hypothesis (H_0): Mean GPH of Long Flights *equals* Mean GPH of Short Flights

Alternative Hypothesis (H_1): Mean GPH of Long Flights *does not equal* Mean GPH of Short Flights

H_0: $\mu long = \mu short$

H_1: $\mu long \neq \mu short$

*Since the **p<0.05** we **Reject the null hypothesis**:*

There is extremely strong evidence that short flights and long flights differ in their average gain per hour.

Thank You!