PROJECT – 2

EXECUTIVE SUMMARY:

This report gives us insights for United Airlines (UA) departing from New York City, specifically focusing on the gain per flight. To elaborate, we are focusing on how much quicker the flights ended up being. In this analysis, we are using confidence intervals and hypothesis tests either of them in each case, and also using exploratory data analysis to answer the problem questions and get an insight on what shall be worked on for improvement.

Key Take-aways:

1. **Average Gain for flights that departed late versus that departed on-time:**
   This analysis gives us difference in the average gain for flights that departed late compared to those that departed on time. Here, flights are being counted as late when they are late even by a minute.
   Average Gain for flights that departed very late versus not very late:
   This analysis gives us difference in the average gain for flights that departed very late compared to those that were not very late. Here, flights are being counted as very late when they are late by 30 minutes.
2. **Top Five Destination Airports:**
   The five most common destination airports for UA flights from New York City are being considered. For each airport, the report gives the distribution of gains and also gives the average gain.
3. **Gain Per Hour for flights that departed late versus that departed on-time:**
   Gain per hour is related to the flight duration. It is calculated as net time divided by flight duration. This analysis gives us the average gain per hour for flights that departed late versus on time.
   Gain Per Hour for flights that departed very late versus not very late:
   This analysis gives us the average gain per hour for flights that departed very late versus not very late.
4. **Gain Per Hour for Longer versus Shorter Flights:**
   We are dividing the flights to longer and shorter flights by the median of flight duration. This analysis gives us the average gain per hour for longer flights compared to shorter flights.

Results:
Every question is discussed with the help of tests, graphs and tables for a better understanding and clear results.

INTRODUCTION:

The goal of this analysis is to give the gain per flight for United Airlines (UA) departing from New York City. By employing statistical methods such as confidence intervals and hypothesis tests, complemented by exploratory data analysis, we give a report for efficient management of United Airlines.

Objectives:

1. Late Departures and Average Gain: This information helps us understand the overall efficiency and performance of UA flights.
2. Top Destination Airports: This information helps us with the destination allocation.
3. Gain Per Hour Analysis for late and very late flights: This information helps us understand the efficiency of UA flights in terms of time.
4. Gain Per Hour Analysis for shorter and longer flights: This information would help us regarding scheduling and resource utilization.

Data and Variables:

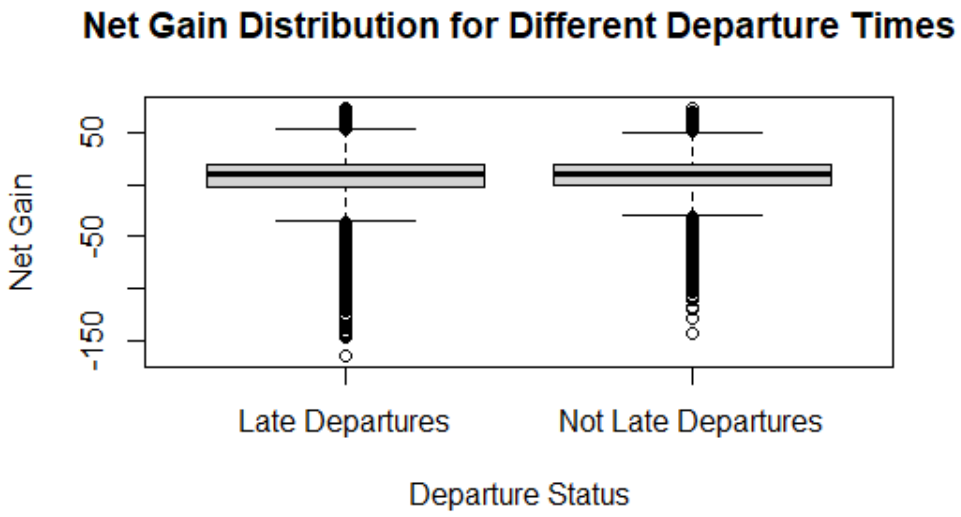
The analysis utilizes a comprehensive dataset containing information on departure delays(dep_delay), arrival delays(arr_delay), flight durations(air_time), and destination airports for United Airlines flights departing from New York City.

The variables we make and add include:

1. **Net Gain(net_gain):** Calculated by subtracting the arrival delay from the departure delay, representing the overall gain or loss in time for each flight.
2. **Most Common Destination Airports(top_destinations):** The airports where United Airlines flights from New York City commonly arrive.
3. **Gain Per Hour(gain_per_hour)*:*** Calculated by dividing the total gain by the duration in hours of each flight.

ANALYSIS:

1. **Average Gain for flights that departed late versus that departed on-time:**
   Graph:



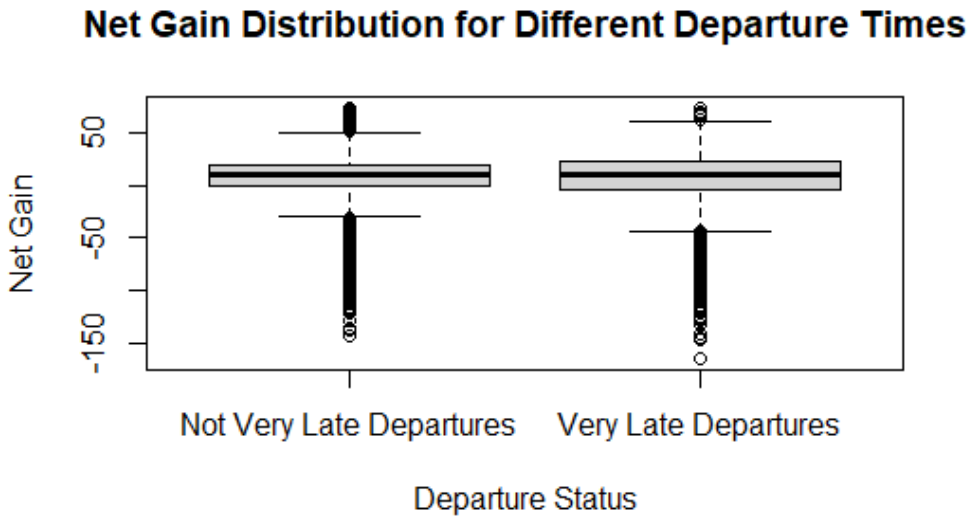
Net Gain for main part of the data for late departures and not late departures lie between a little less than 0 and somewhere around 20. There are many outliers as well. Although both of the variables look similar, late departures has relatively higher net gain than on time departures.

Numerical Results:

P-value: < 2.2e-16 (extremely small)

The p-value says that the observed difference in average gain between flights that departed late and those that departed on time is highly unlikely to be due to random chance.
We reject the null hypothesis and conclude that there is a statistically significant difference in average gain between late and on time flights.
Confidence Interval: (1.411308, 2.040805)
Mean of Not Late Departures: 9.269172
Mean of Late Departures: 7.543115
Confidence Intervals or Hypothesis test:
Confidence Interval: (1.411308, 2.040805)
This interval suggests that we are 95% confident that the true difference in the average gain between flights that departed late and those that departed on time lies between 1.411308 and 2.040805.
As the interval does not include zero, it says that there is a statistically significant difference in average gain between the two groups.
Conclusion:
The analysis shows a statistically significant difference in average gain between flights that departed late and those that departed on time. On average, late departures have a net gain that is significantly different from on-time departures.

Average Gain for flights that departed very late versus not very late:
Graph:



Net Gain for main part of the data for very late departures and not very late departures lie between a little less than 0 and somewhere around 20. There are many outliers as well. The very late departures seem to have high net gain when compared to not very late departures.
Numerical Results:
P-value: 3.215e-10 (extremely small)
The p-value is against the null hypothesis. It says that the observed difference in average gain between flights that depart very late and those that do not is highly unlikely to occur by random chance.

The null hypothesis is rejected, and there is a statistically significant difference in average gain between the two groups.
Confidence Interval: (1.268195, 2.415112)
Mean of Not Very Late Departures: 8.699534
Mean of Very Late Departures: 6.857881

Confidence Intervals or Hypothesis test:

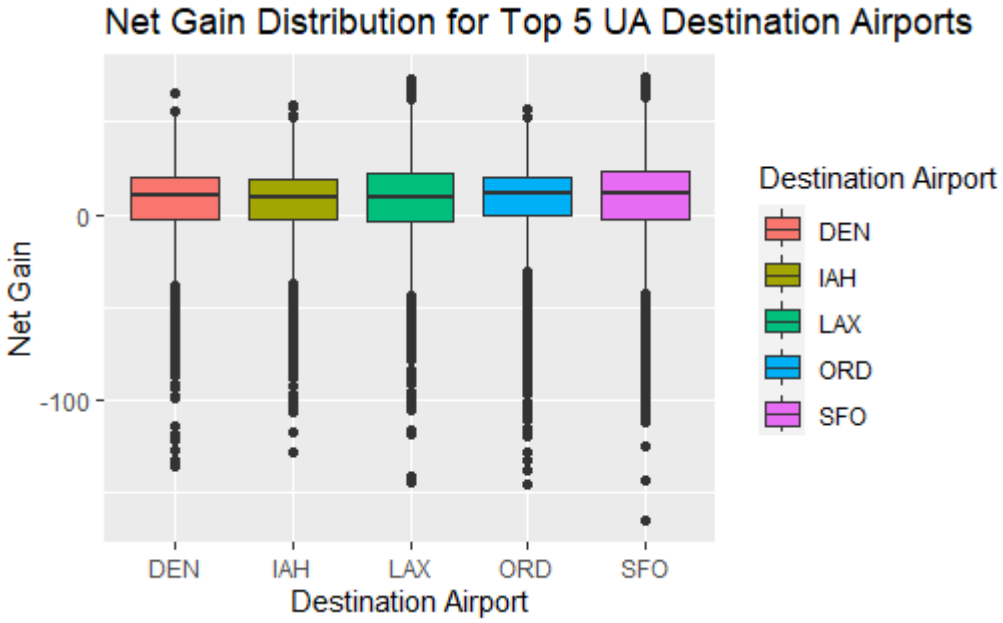
Confidence Interval: (1.268195, 2.415112)
This interval indicates that we are 95% confident that the true difference in average gain between flights that depart very late and those that do not depart very late lies between 1.268195 and 2.415112.
As the interval does not include zero, it says that there is a statistically significant difference in average gain between the two groups.

Conclusion:

There is a statistically significant difference in average gain between flights that depart very late and those that do not depart very late. Flights departing very late have a net gain that is significantly different from those departing not very late.

2. **Top Five Destination Airports:**

Graph:



The graph shows the distribution of each airport. The net gain seems to be 0 and above. There are a lot of outliers for each airport.

Numerical Results:

A tibble: 5 × 3

dest <chr>	count <int>	avg_gain <dbl>
IAH	6814	6.861755
ORD	6744	7.777432
SFO	6728	8.695006
LAX	5770	7.825303
DEN	3737	7.302382

5 rows

This table clearly shows us the average gain for each airport.

Confidence Intervals or Hypothesis test:

Confidence Interval for IAH (6.423820, 7.299691)
Confidence Interval for ORD (7.320135, 8.234729)
Confidence Interval for SFO (8.159475, 9.230536)
Confidence Interval for LAX (7.259681, 8.390925)
Confidence Interval for DEN (6.659348, 7.945415)

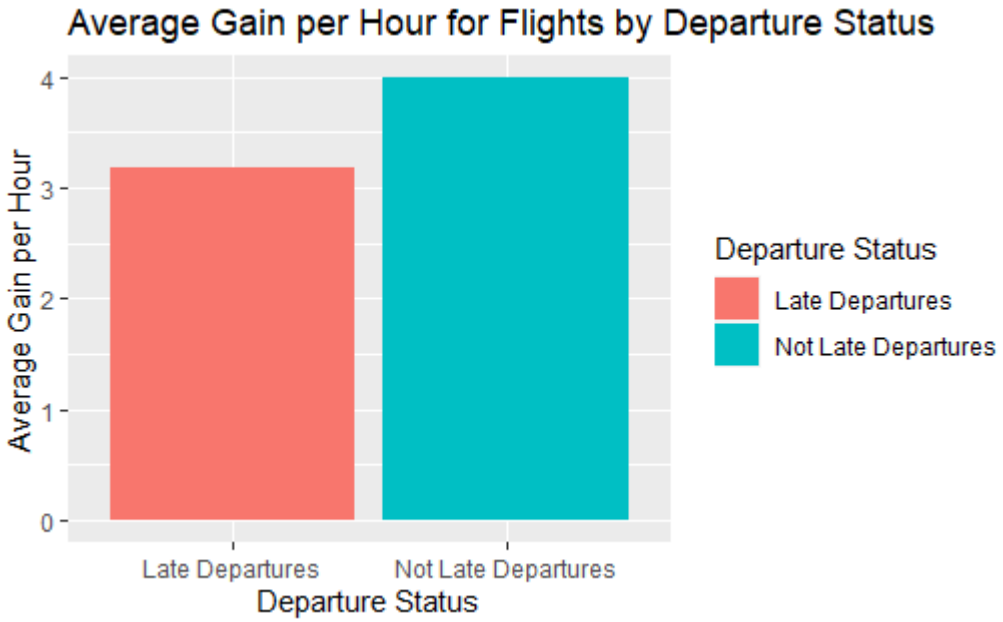
The intervals seem to overlap, saying that that the average gains for these airports are not significantly different at 95% confidence level.

Conclusion:

As the confidence intervals overlap, there are no statistically significant differences in average gains within these airports.

3. **Gain Per Hour for flights that departed late versus that departed on-time:**

Graph:



The graph suggests that late departures give lower average gain per hour than not late departures.

Numerical Results:

P-value: < 2.2e-16 (extremely small)
The p-value gives strong evidence against null hypothesis.
Confidence Interval: (0.6662688, 0.9463657)
Mean of Not Late Departures: 3.990898
Mean of Late Departures: 3.184581

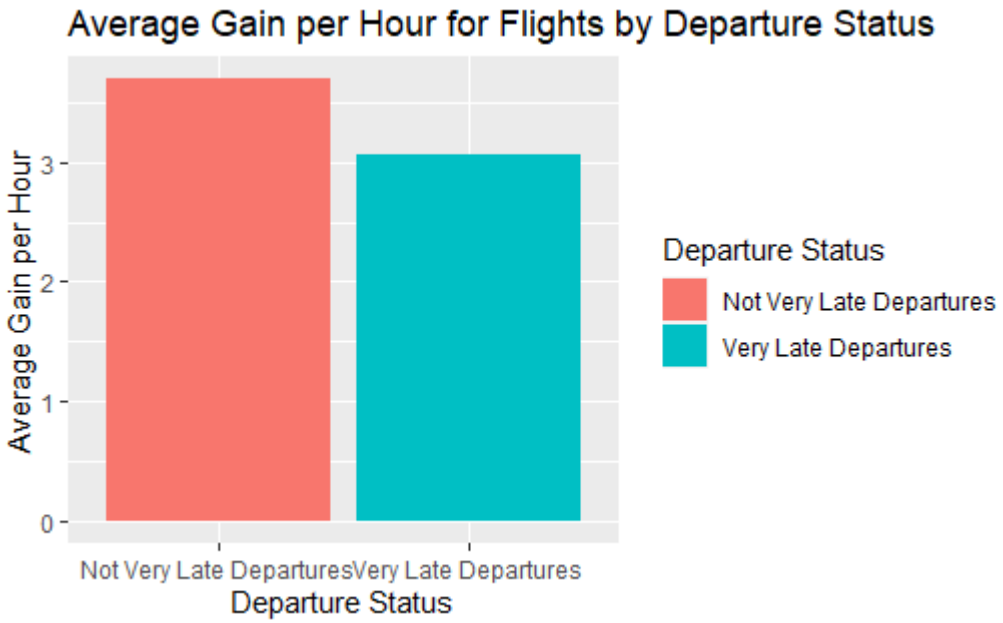
Confidence Intervals or Hypothesis test:

The 95% confidence interval for the true difference in means is (0.6662688, 0.9463657). This interval does not include zero, indicating a statistically significant difference in average gain per hour between the late and not late departures.

Conclusion:

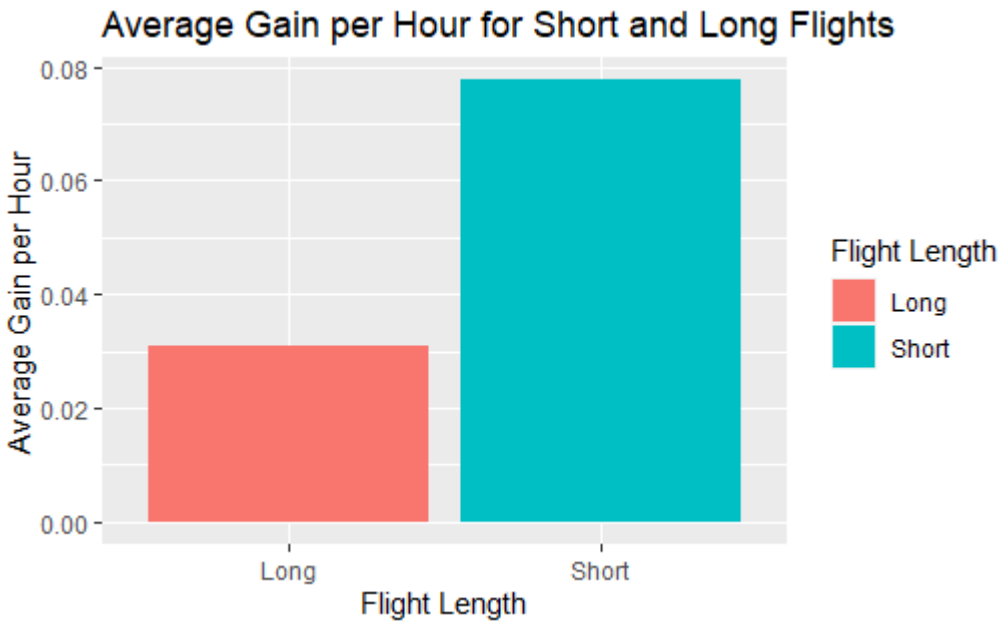
The average gain per hour differs significantly between flights that departed late and those that departed on time. The positive mean difference and the lack of overlap with zero in the confidence interval indicate that, on average, not late departures result in a higher gain per hour.

Gain Per Hour for flights that departed very late versus not very late:
Graph:



The graph suggests that very late departures give lower average gain per hour than not very late departures.
Numerical Results:
P-value: 1.372e-06 (extremely small)
The p-value indicates strong evidence against the null hypothesis.
Confidence Interval: (0.3745605, 0.8858401)
Mean of Not Very Late Departures (x): 3.694727
Mean of Very Late Departures (y): 3.064527
Confidence Intervals or Hypothesis test:
The 95% confidence interval for the true difference in means is (0.3745605, 0.8858401). This interval does not include zero, indicating a statistically significant difference in average gain per hour between the two groups.
The positive values in the interval suggest that, on average, flights departing not very late have a higher gain per hour compared to those that depart very late.
Conclusion:
The small p-value and the confidence interval excluding zero strongly support the conclusion that not very late departures result in a higher average gain per hour.

4. **Gain Per Hour for Longer versus Shorter Flights:**
   Graph:



The graph suggests that average gain per hour is very high for short flights than long flights.

Numerical Results:

P-value: < 2.2e-16 (extremely small)

The p-value indicates strong evidence against the null hypothesis.

Confidence Interval: (2.725528, 2.877603)

Mean of Short Flights (x): 4.659777

Mean of Long Flights (y): 1.858212

Confidence Intervals or Hypothesis test:

The 95% confidence interval for the true difference in means is (2.725528, 2.877603). This interval does not include zero, indicating a statistically significant difference in average gain per hour between short and long flights.

The positive values in the interval suggest that, on average, short flights have a higher gain per hour compared to long flights.

Conclusion:

The p-value and the confidence interval excluding zero strongly support the conclusion that short flights, on average, have a higher gain per hour than long flights.

APPENDIX:

---

title: 'Project #2'

author: "Tejaswi Neelapu"

output: html_document

date: "2023-11-24"

Selecting United Airlines flights

```{r}
library(nycflights13)
library(tidyverse)

United_Airlines <- flights %>% filter(carrier == "UA") %>% mutate(net_gain = dep_delay - arr_delay)
glimpse(United_Airlines)
```

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}
late_departures <- United_Airlines %>% filter(dep_delay > 0)
not_late_departures <- United_Airlines %>% filter(dep_delay <= 0)

t.test(not_late_departures$net_gain, late_departures$net_gain)

United_Airlines <- United_Airlines %>%
  mutate(dep_status = case_when(
    dep_delay <= 0 ~ "Not Late Departures",
    dep_delay > 0 ~ "Late Departures",
  ))

United_Airlines <- na.omit(United_Airlines)
boxplot(net_gain ~ dep_status, data = United_Airlines,
        main = "Net Gain Distribution for Different Departure Times",
        xlab = "Departure Status",
        ylab = "Net Gain")
```

```{r}
```

```
very_late_departures <- United_Airlines %>% filter(dep_delay > 30)
not_very_late_departures <- United_Airlines %>% filter(dep_delay <= 30)

t.test(not_very_late_departures$net_gain, very_late_departures$net_gain)

United_Airlines <- United_Airlines %>%
  mutate(dep_status = case_when(
    dep_delay <= 30 ~ "Not Very Late Departures",
    dep_delay > 30 ~ "Very Late Departures",
  ))

United_Airlines <- na.omit(United_Airlines)
boxplot(net_gain ~ dep_status, data = United_Airlines,
        main = "Net Gain Distribution for Different Departure Times",
        xlab = "Departure Status",
        ylab = "Net Gain")
```

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}
top_destinations <- United_Airlines %>%
  group_by(dest) %>%
  summarize(count = n(), avg_gain = mean(net_gain)) %>%
  arrange(desc(count)) %>%
  head(5)

top_dest_flights <- United_Airlines %>% filter(dest %in% top_destinations$dest)
top_dest_airports <- top_destinations$dest
other_dest_flights <- United_Airlines %>% filter(!dest %in% top_dest_airports)
```

```
ggplot(top_dest_flights, aes(x = dest, y = net_gain, fill = dest)) + geom_boxplot() + labs(title = "Net Gain Distribution for Top 5 UA Destination Airports", x = "Destination Airport", y = "Net Gain", fill = "Destination Airport")

top_destinations

top_dest_flights1 <- top_dest_flights %>% filter(dest == "IAH")

top_dest_flights2 <- top_dest_flights %>% filter(dest == "ORD")

top_dest_flights3 <- top_dest_flights %>% filter(dest == "SFO")

top_dest_flights4 <- top_dest_flights %>% filter(dest == "LAX")

top_dest_flights5 <- top_dest_flights %>% filter(dest == "DEN")

t.test(top_dest_flights1$net_gain)$conf

t.test(top_dest_flights2$net_gain)$conf

t.test(top_dest_flights3$net_gain)$conf

t.test(top_dest_flights4$net_gain)$conf

t.test(top_dest_flights5$net_gain)$conf
```

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}
United_Airlines <- United_Airlines %>%
  mutate(gain_per_hour = net_gain / (air_time / 60))


t.test(not_late_departures$gain_per_hour, late_departures$gain_per_hour)


United_Airlines <- United_Airlines %>%
  mutate(dep_status = case_when(
    dep_delay <= 0 ~ "Not Late Departures",
    dep_delay > 0 ~ "Late Departures",
```

```
  ))

ggplot(United_Airlines, aes(x = dep_status, y = net_gain / (air_time / 60), fill = dep_status)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Gain per Hour for Flights by Departure Status",
       x = "Departure Status",
       y = "Average Gain per Hour",
       fill = "Departure Status")
```

```{r}
t.test(not_very_late_departures$gain_per_hour, very_late_departures$gain_per_hour)
```

```{r}

United_Airlines <- United_Airlines %>%
  mutate(dep_status = case_when(
    dep_delay <= 30 ~ "Not Very Late Departures",
    dep_delay > 30 ~ "Very Late Departures",
  ))


ggplot(United_Airlines, aes(x = dep_status, y = net_gain / (air_time / 60), fill = dep_status)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Gain per Hour for Flights by Departure Status",
       x = "Departure Status",
       y = "Average Gain per Hour",
       fill = "Departure Status")
```

```

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

```{r}
United_Airlines <- United_Airlines %>%
  mutate(flight_length = ifelse(air_time <= median(air_time), 'Short', 'Long'))

short_flights <- flights %>% filter(flight_length == 'Short')
long_flights <- flights %>% filter(flight_length == 'Long')

t.test(short_flights$gain_per_hour, long_flights$gain_per_hour)

ggplot(flights, aes(x = flight_length, y = net_gain / air_time, fill = flight_length)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Gain per Hour for Short and Long Flights",
       x = "Flight Length",
       y = "Average Gain per Hour",
       fill = "Flight Length")
```
```