Project 2

Ting-Yu, Lin

Executive Summary

This report analyzes United Airlines' profitability using nycflights13 dataset. On-time departures correlate with higher gain, emphasizing operational efficiency. Flights to SFO (most common destination airport) are more profitable than IAH (fifth common destination airport), highlighting the impact of popular routes. Shorter flights yield higher gain per hour, emphasizing the profitability of time-efficient operations. Actionable insights include prioritizing punctuality, optimizing popular routes, and considering flight duration for enhanced profitability.

Introduction

This report analyzes the gain per flight for United Airlines(carrier code UA), using data from the nycflights13 package. The focus is on understanding how much quicker flights end up being than planned, measured as the net gain (departure delay minus arrival delay) and gain per hour(dividing the total gain by the duration in hours of each flight). The analysis employs a combination of exploratory data analysis, confidence intervals, and hypothesis tests to address key questions. The questions include investigating average gain differences for flights departing late, five common destination airports, gain per hour, and differences in gain per hour for longer versus shorter flights.

Analysis

1. Average Gain for Late Departures

In our analysis, we aim to investigate whether there is a difference in the average gain for flights based on their departure punctuality.

Three new columns have been created in the dataset. The first column, labeled 'late,' is assigned a value of True when there is a departure delay and False otherwise. The second column, 'very_late,' is marked as True when the departure delay exceeds 30 minutes and False otherwise. Lastly, the third column, 'net_gain,' is calculated as the difference between departure delay and arrival delay.

Figure 1 displays the distribution of net gain, revealing a slightly left-skewed pattern. The majority of net gains are positive.

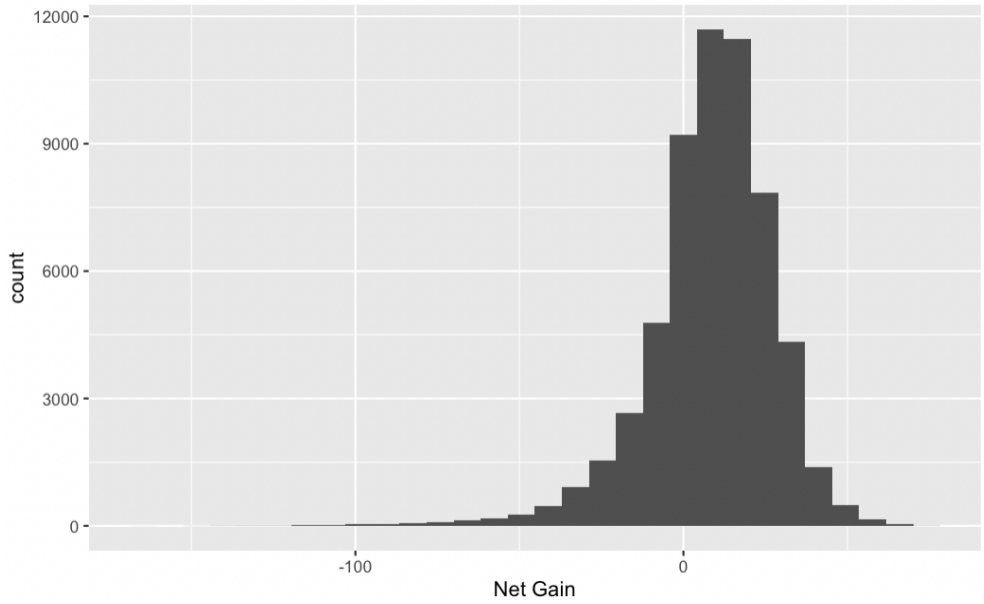

Figure 1

We conducted a hypothesis test using a t-test to assess the difference in means of the net_gain variable between flights with departure delays and those without. The p-value, smaller than 2.2e-16, provides robust evidence indicating a significant disparity in average gain between these two groups. Specifically, flights that departed on time exhibited a higher average gain compared to those that departed late. Moreover, with a 95% confidence level, we estimate that the difference in means lies within the range of 1.411 to 2.041. This confidence interval reinforces our conclusion, affirming a statistically significant distinction in the average gain for flights based on their departure punctuality.

Subsequently, we conducted a hypothesis test using a t-test to explore the variance in means of the net_gain variable between flights with departure delays exceeding 30 minutes and those without such delays. The resulting p-value, equal to 3.215e-10, underscores compelling evidence of a significant difference in average gain between these two groups. Specifically, flights that did not experience much delays exhibited a higher average gain compared to those that departed very late. Furthermore, with 95% confidence, we estimate that the difference in means falls within the range of 1.268 to 2.415.

Based on the aforementioned results, we can conclude that flights departing on time tend to yield higher profits, as indicated by the net gain.

2. Five Common Destination Airports

In Table 1, we observe that ORD, IAH, SFO, LAX, and DEN are identified as the five most common destination airports for United Airlines flights originating from New York City.

dest <chr>	count <int>	mean_net_gain <dbl>
ORD	6984	7.77743179
IAH	6924	6.86175521
SFO	6819	8.69500595
LAX	5823	7.82530329
DEN	3796	7.30238159
BOS	3342	9.38064907
MCO	3217	9.38389220
FLL	2407	7.37205387
LAS	2010	11.91959799
TPA	1968	7.44159836

Table 1

In Figure 2, we observe that the mean gain is relatively consistent across these five destination airports. However, it is noteworthy that each of these airports exhibits numerous outliers with negative net gains.

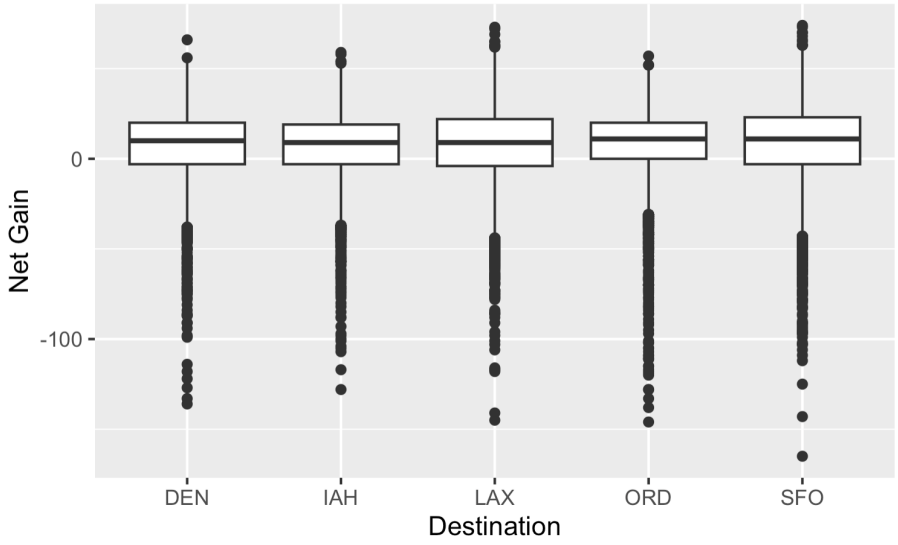

Figure 2

We conducted a hypothesis test using a t-test to evaluate the difference in means of the net_gain variable between the most common destination airport (SFO) from New York

City and the fifth common destination airport (IAH). The resulting p-value, equal to 2.08e-07, provides evidence of a disparity in average net gain between these two airports. Specifically, SFO airport exhibited a higher average net gain compared to IAH.

Furthermore, with a 95% confidence level, we estimate that the difference in mean gain between SFO and IAH falls within the range of 1.142 to 2.525.

In conclusion, we find that SFO destination airport tends to generate more profit than IAH based on net gain.

3. Gain per Hour

Another common measure of an airline company's profit is the gain relative to the duration of the flight. We created a new column by calculating the gain per hour, obtained by dividing the total gain by the duration in hours of each flight.

Figure 3 illustrates the distribution of gain per hour, showing a pattern resembling a normal distribution. The graph indicates that the majority of gain per hour values are positive.

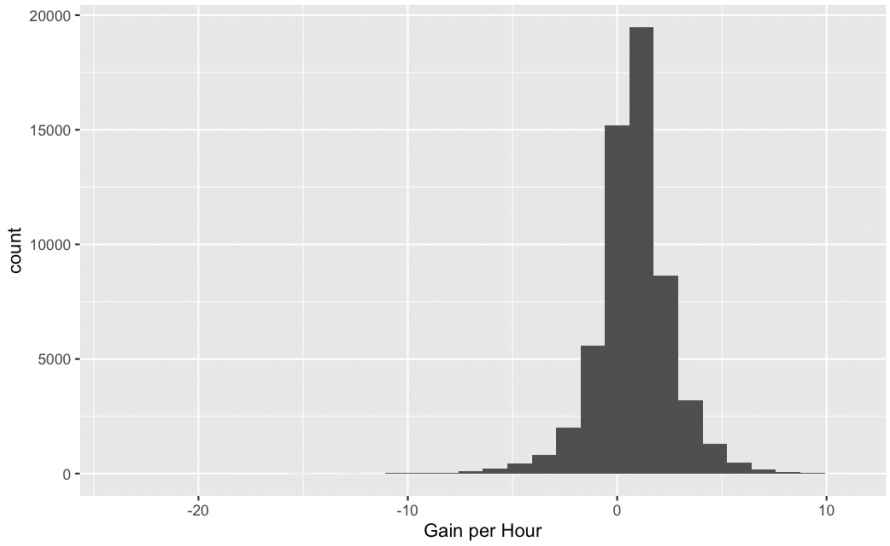

Figure 3

We conducted a hypothesis test using a t-test to determine whether the average gain per hour differs between flights that departed late and those that did not. Additionally, we explored differences in the average gain per hour for flights with delays exceeding 30 minutes compared to those without.

The resulting p-value (< 2.2e-16) in the hypothesis test indicates evidence of a significant disparity in average gain per hour between flights that departed on time and those that departed late. Specifically, flights that departed on time exhibited a higher average gain per hour than those that departed late. Furthermore, with 95% confidence, we estimate that the difference in mean between these two groups lies within the range of 0.274 to 0.333.

Similarly, the obtained p-value (< 2.2e-16) suggests significant evidence of a difference in average gain per hour between flights with delays under 30 minutes and those with delays exceeding 30 minutes. Specifically, flights with delays shorter than 30 minutes showed a higher average gain per hour compared to those with delays over 30 minutes. With 95% confidence, the estimated difference in mean between these two groups falls within the range of 0.296 to 0.385.

In conclusion, we find that flights that departed on time tend to generate more profit based on gain per hour.

4. Gain per Hour for Longer vs. Shorter Flights

Figure 4 illustrates the gain per hour across flights of varying durations. All means fall within the range of 0 to 5. Flights with shorter durations exhibit more outliers, and the range of these outliers is larger. This observation suggests a higher level of uncertainty associated with flights of shorter duration.

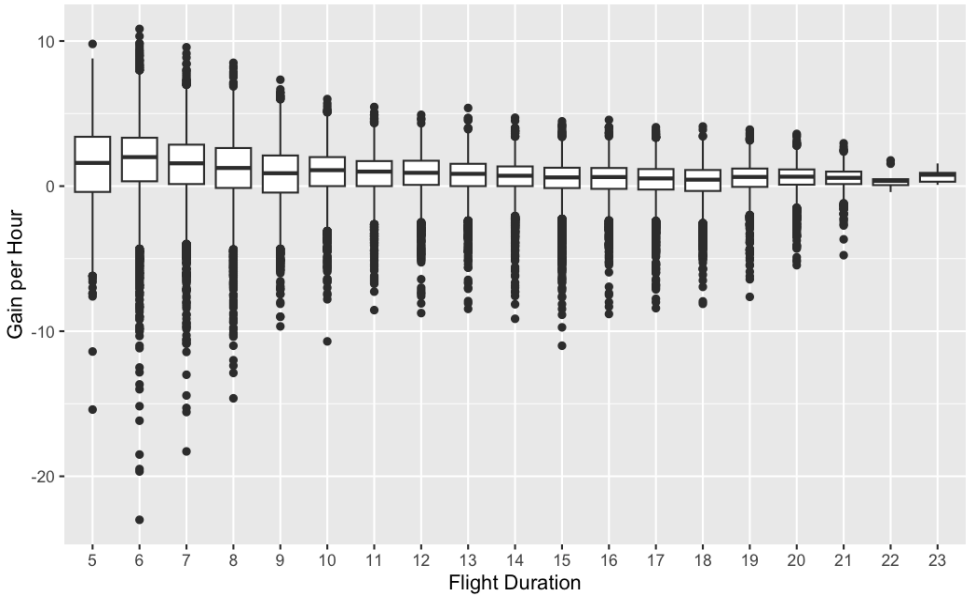

Figure 4

We defined flights with durations between 5 and 13 hours as shorter flights and those with durations between 14 and 23 hours as longer flights. We then investigated whether the average gain per hour differs between longer and shorter flights. The resulting p-value, smaller than 2.2e-16, provides substantial evidence of a significant disparity in average gain per hour between the two groups. Specifically, shorter flights exhibited a higher average gain per hour compared to longer flights. Moreover, with 95% confidence, we estimate that the difference in mean between these two groups falls within the range of 0.657 to 0.599.

In conclusion, we find that shorter flights tend to generate more profit based on gain per hour.

Appendix

The dataset only includes carrier UA and create three new column: late, very_late and net_gain

```{r}
ua_flights <- flights%>%
  filter(carrier=='UA')%>%
  mutate(late = dep_delay > 0,
         very_late = dep_delay > 30,
         net_gain =  dep_delay - arr_delay)
ua_flights
```

Figure 1

```{r}
ggplot(data=ua_flights, aes(x= net_gain))+
  geom_histogram()+
  xlab('Net Gain')
```

A t-test to compare the means of the net_gain variable between two groups defined by the late variable.

```{r}
t.test(net_gain~late,data=ua_flights, alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  net_gain by late
t = 10.749, df = 52833, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 1.411308 2.040805
sample estimates:
mean in group FALSE  mean in group TRUE
           9.269172            7.543115
```

A t-test to compare the means of the net_gain variable between two groups defined by the very_late variable.

```{r}
t.test(net_gain~very_late,data=ua_flights, alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  net_gain by very_late
t = 6.2953, df = 8838.6, p-value = 3.215e-10
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 1.268195 2.415112
sample estimates:
mean in group FALSE  mean in group TRUE
           8.699534            6.857881
```

Table1

```{r}
by_dest <- ua_flights %>%
  group_by(dest) %>%
  summarize(count = n(), mean_net_gain = mean(net_gain, na.rm = TRUE)) %>%
  arrange(desc(count))
by_dest
```

Figure2

```{r}
five_dest <- ua_flights%>%
  filter(dest %in% c('ORD', 'IAH', 'SFO', 'LAX', 'DEN'))

ggplot(five_dest, aes(x=dest, y=net_gain))+
  geom_boxplot()+
  xlab('Destination')+
  ylab('Net Gain')
```

A t-test to compare the means of the net_gain variable between SFO and IAH.

```{r}
t.test(ua_flights$net_gain[ua_flights$dest=='SFO'],ua_flights$net_gain[ua_flights$dest==
='IAH'], alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  ua_flights$net_gain[ua_flights$dest == "SFO"] and ua_flights$net_gain[ua_flights$dest == "IAH"]
t = 5.1948, df = 12995, p-value = 2.08e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.141515 2.524986
sample estimates:
mean of x mean of y
 8.695006  6.861755
```

Create gain_per_hour variable

```{r}
ua_flights <- ua_flights %>%
  mutate(gain_per_hour = net_gain / hour)
ua_flights
```

A t-test to compare the means of the gain_per_hour variable between two groups defined by the late variable.

```{r}
t.test(gain_per_hour~late,data=ua_flights, alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  gain_per_hour by late
t = 20.056, df = 57473, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 0.2739012 0.3332350
sample estimates:
mean in group FALSE  mean in group TRUE
          0.9310086           0.6274405
```

A t-test to compare the means of the gain_per_hour variable between two groups defined by the very_late variable.

```{r}
t.test(gain_per_hour~very_late,data=ua_flights, alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  gain_per_hour by very_late
t = 14.971, df = 9882.8, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 0.2960713 0.3852843
sample estimates:
mean in group FALSE  mean in group TRUE
          0.8330167           0.4923389
```

Figure 3

```{r}
ggplot(data=ua_flights, aes(x= gain_per_hour))+
  geom_histogram()+
  xlab('Gain per Hour')
```

Figure4

```{r}
ggplot(data=ua_flights, aes(x= factor(hour), y= gain_per_hour))+
  geom_boxplot()+
  xlab('Flight Duration')+
  ylab('Gain per Hour')
```

Create duration variable

```{r}
ua_flights <- ua_flights %>%
  mutate(duration = ifelse(hour >= 14, "long", "short"))
ua_flights
```

A t-test to compare the means of the gain_per_hour variable between two groups defined by the duration variable.

```{r}
t.test(gain_per_hour~duration,data=ua_flights, alternative = "two.sided")
```

```
        Welch Two Sample t-test

data:  gain_per_hour by duration
t = -42.746, df = 47666, p-value < 2.2e-16
alternative hypothesis: true difference in means between group long and group short is not equal to 0
95 percent confidence interval:
 -0.6566907 -0.5991095
sample estimates:
 mean in group long mean in group short
          0.4624252           1.0903253
```