PROJECT REPORT

EXECUTIVE SUMMARY

This project is about the analysis of gain per flight for United Airlines (UA). The main goals of this project are to comprehend the effects of departure delays, investigate the five most popular airports for destinations, calculating the gain per hour and to investigate whether the average gain per hour differs for longer flights and shorter flights. Based on the circumstances some variables are added in the table like net_gain, gain per hour, late, very_late and flight_duration. The net_gain is calculated by subtracting arrival delay from departure delay.

The goal is to provide practical insights for optimizing flight efficiency and improving customer satisfaction by performing confidence intervals and hypothesis test. Alongside Exploratory Data Analysis (EDA) is also used to perform analysis and learn insights from the data

After performing hypothesis test for all cases it is found that there is difference of values in average gain of departure delays, difference of values in average gain per hour of departure delays, difference of values in net gain of longer and shorter flights. The five most common destinations for United Airlines found out be “IAH”, “ORD”, “SFO”, “LEX”, “DEN” and the distribution among all these airports is uniformly distributed and are normal.

INTRODUCTION

The aim of this project is to analyze and investigate gain per flight for United Airlines(UA). The dataset flights is retrieved from nycflights13 package. We filter the entire dataset to only for United Airlines(UA) flights carrier and if there are any missing or “NA” value it will be omitted. To calculate net gain we subtract arrival delay (arr_delay) from departure delay (dep_delay). To measure net gain we add this column to the dataset.

We perform Exploratory Data Analysis (EDA) , conduct confidence intervals and hypothesis test and if possible we also add new variables based on the analysis we will be performing.

We will be analyzing based on the following :

- Whether average gain differ for flights that departed late versus those that did not and also about the flights that departed for more than 30 minutes late.
- The five most common destination airports for United Airlines flights from New York City and to describe the distribution and the average gain for each of these five airports.
- Is the net gain relative to the duration of the flight and calculating the gain per hour by dividing the total gain by the duration in hours of each flight. Whether the average gain per hour differ for flights that departed late versus those that did not and about the flights that departed more than 30 minutes late.
- Whether the average gain per hour differ for longer flights versus shorter flights.

From the above circumstances we will be adding some new variables other than net gain. Those new variables are described below :

1. late : late variable is constructed , if the departure delay(dep_delay) is greater than 0 it is TRUE and otherwise it is FALSE.

2. very_late : very_late variable is constructed, if the departure delay(dep_delay) was greater than 30 minutes and otherwise it is FALSE.

3. gain_per_hour : This variable is constructed by dividing the total gain (net_gain) by hours of each flight duration(hour).

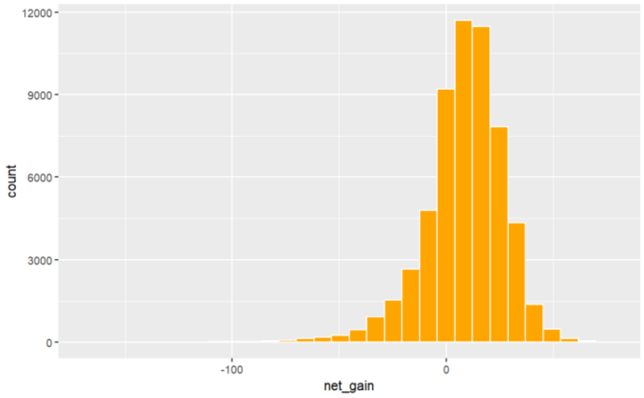
4. flight_duration : This variable is added to divide the duration of flight into short and longer. If hour is less than 6 it is considered as short duration and else if hour is greater than or equal to 6 it is considered as longer duration.

EXPLORATORY DATA ANALYSIS (EDA)

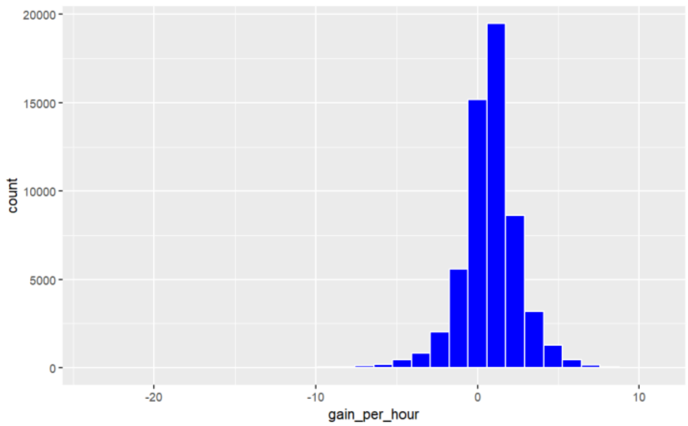
Summary Table

	dep_delay <dbl>	**net_gain** <dbl>	**gain_per_hour** <dbl>
Min.	-20.00000	-165.000000	-23.0000000
1st Qu.	-4.00000	-1.000000	-0.0625000
Median	0.00000	10.000000	0.8235294
Mean	12.01691	8.458897	0.7885025
3rd Qu.	11.00000	20.000000	1.7142857
Max.	483.00000	74.000000	10.8333333

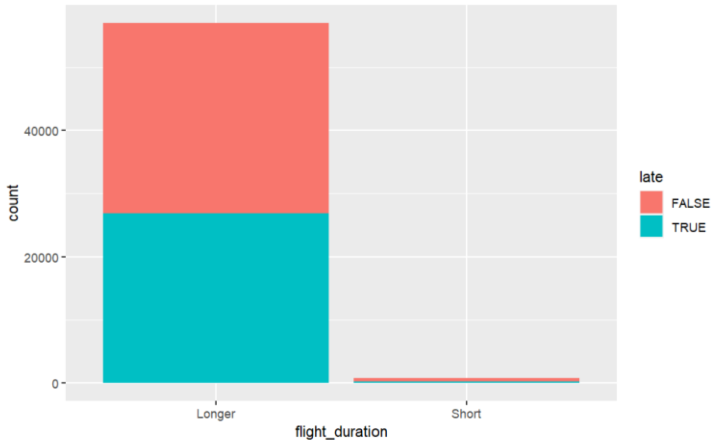
Histogram of net_gain



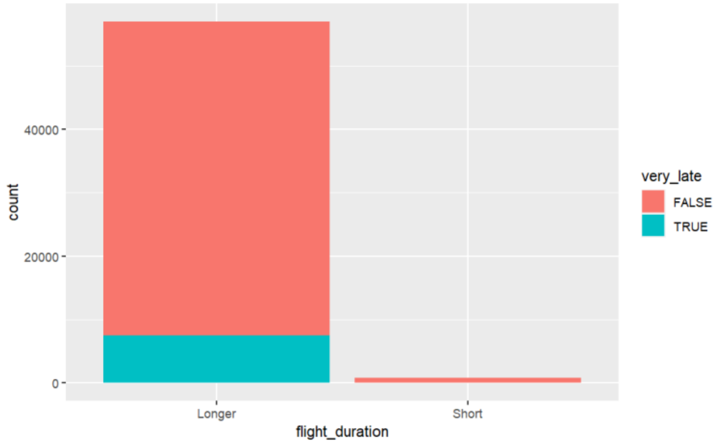
Histogram plot of gain_per_hour



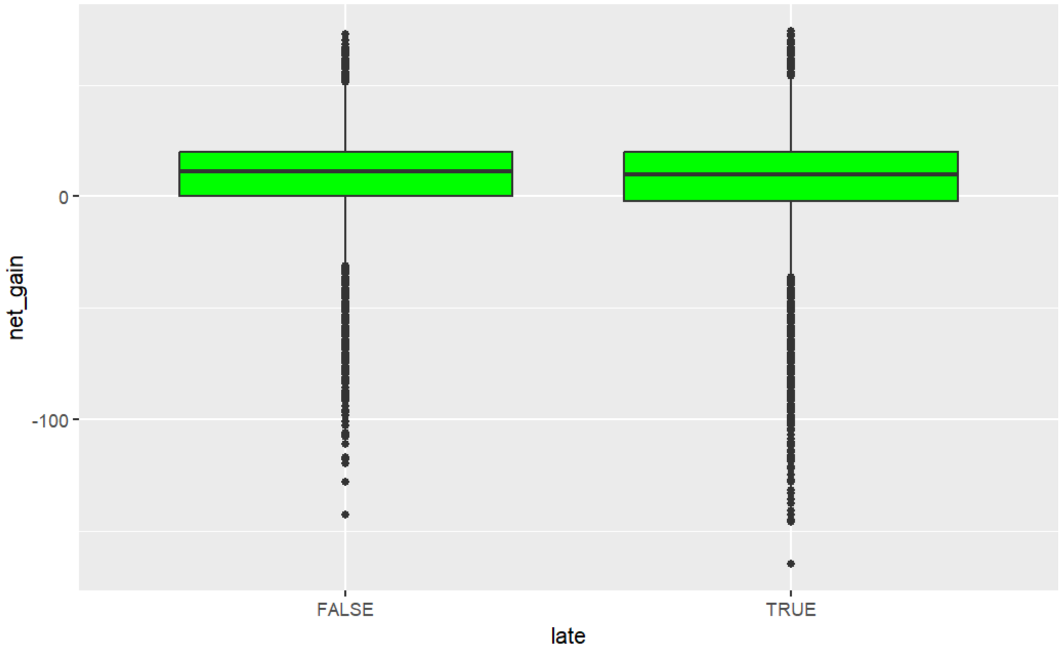
Relationship between flight_duration and late (Barplot)



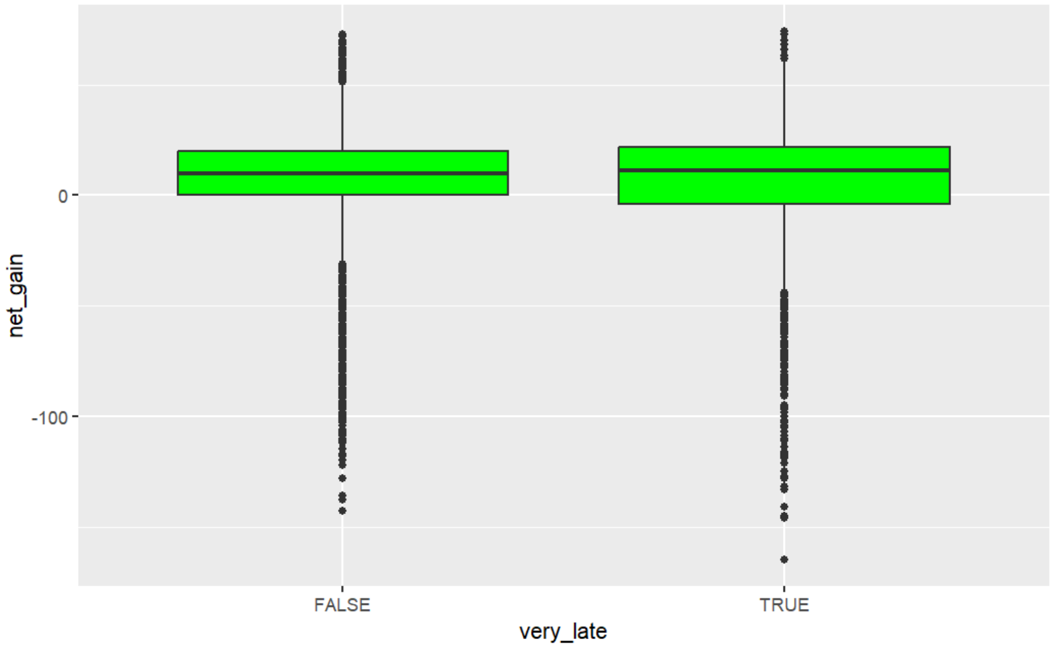
Relationship between flight_duration and very_late (Bar plot)



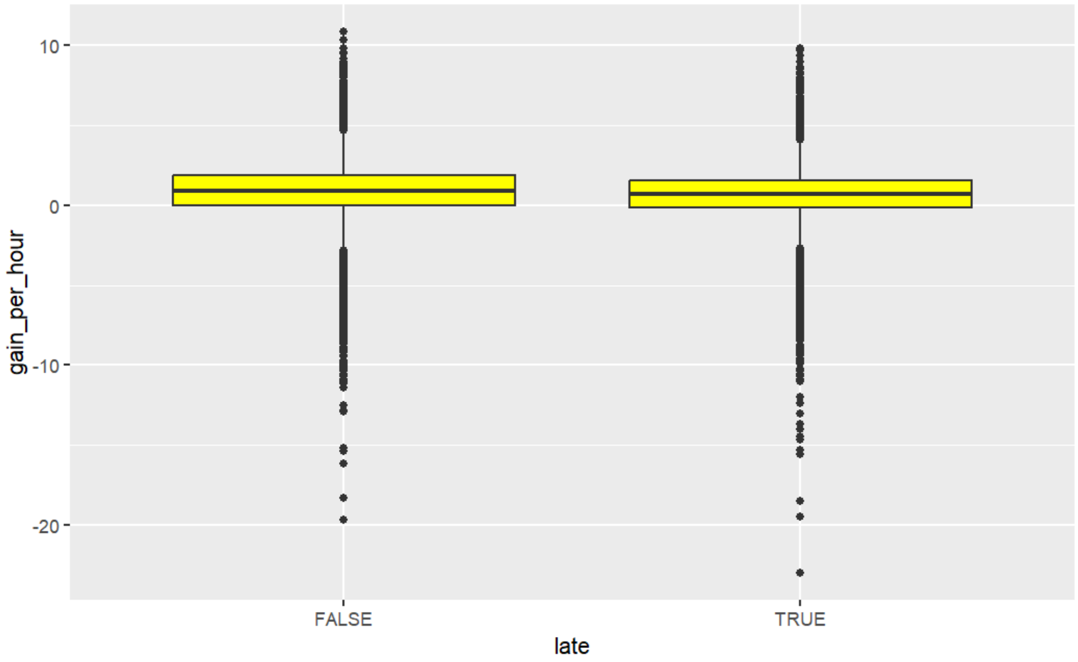
Relationship between late and net_gain



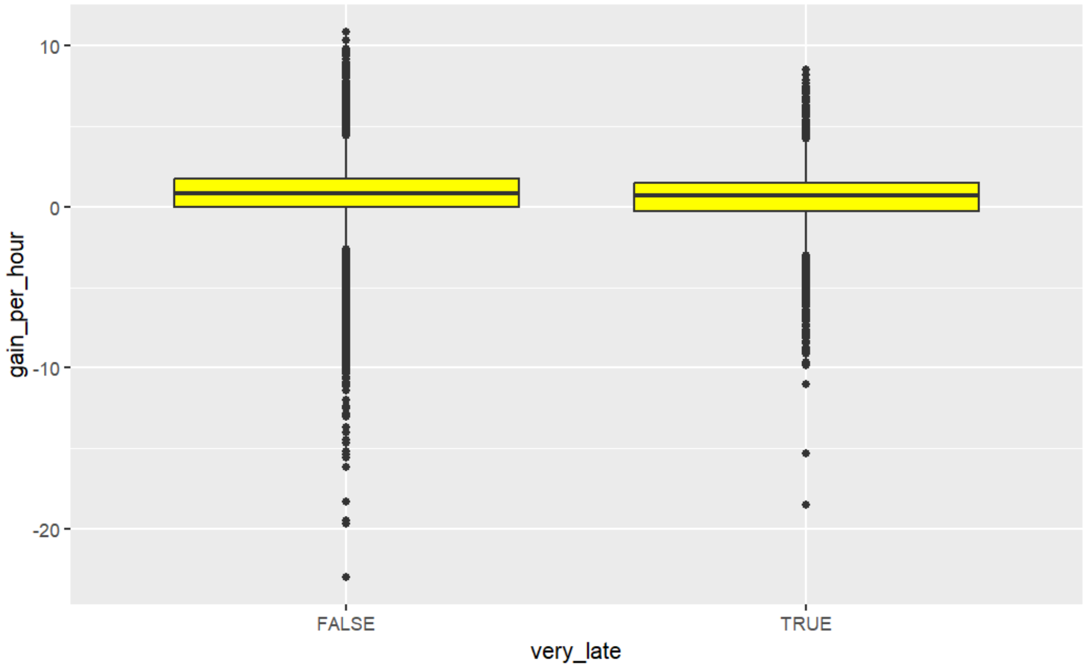
Relationship between very_late and net_gain



Relationship between late and gain_per_hour



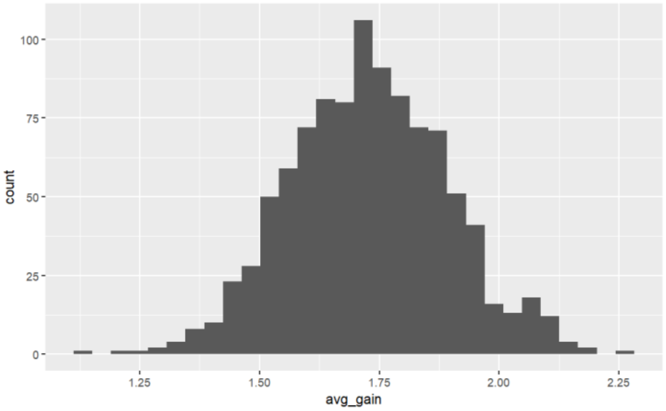
Relationship between very_late and gain_per_hour



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?

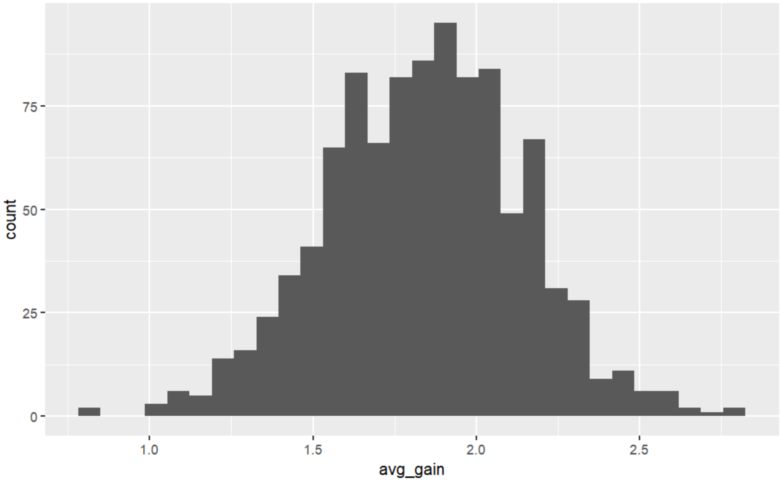
    a) After performing confidence interval and hypothesis test we conclude that there is a statistical significance difference between departed late flights and not late flights. Therefore, the average net_gain for not late departed flights have more values than late flights.
    The confidence interval result is (1.401929, 2.030207)

    Histogram for the average gain differ for the flights that departed late vs not



    b) After conducting confidence interval and hypothesis test it is concluded that there is statistical significance difference in average gain for very_late departed flights and not very_late flights. Therefore, the average net_gain for not very_late flights have more values than very_late. The confidence interval result is (1.276680, 2.377790)

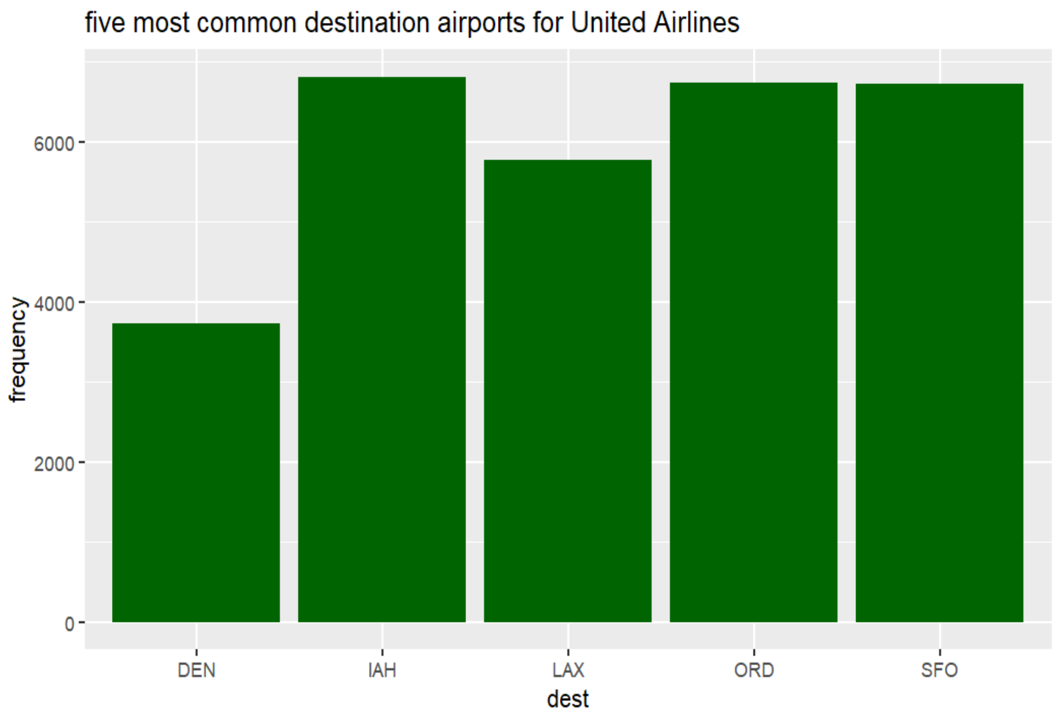
    Histogram for the average gain differ for the flights that departed very late or not

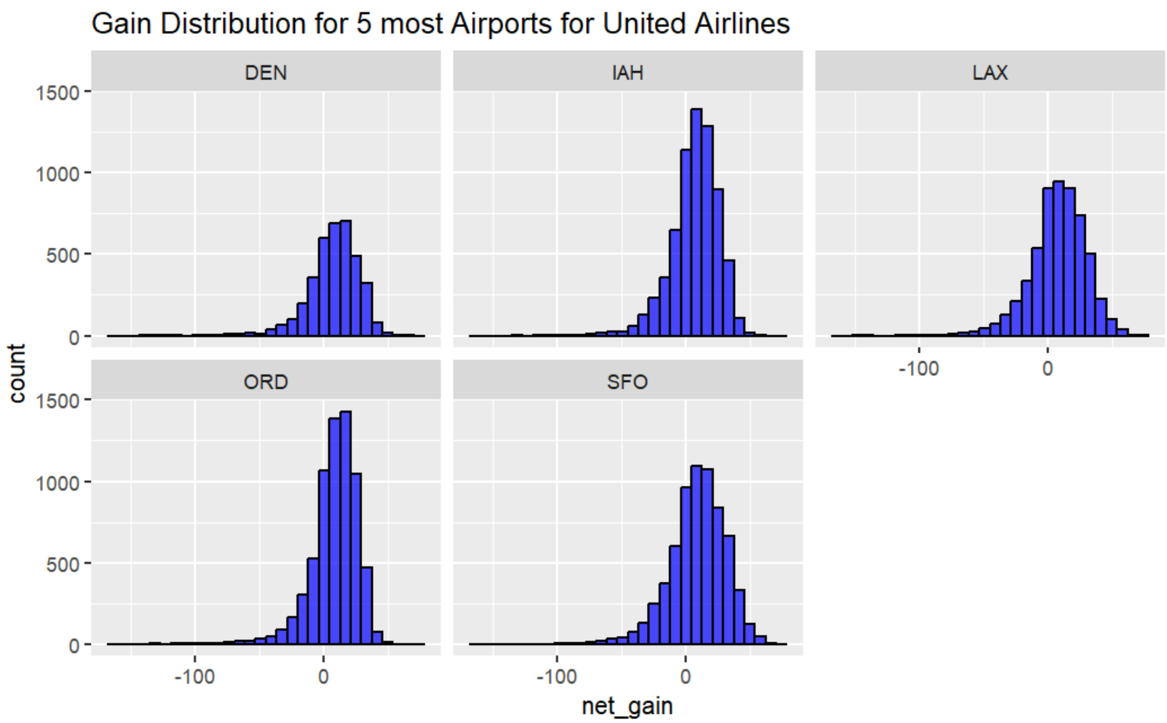


2. What are the five most common destination airports for United Airlines flights from New York City? Describe the distribution and the average gain for each of these five airports.

   The five most common destination airports for United Airlines flights from New York City are tabulated below :

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





The distribution for five most common airports appears to be uniformly distributed and are normal.

The average gain and confidence intervals for each of these are :

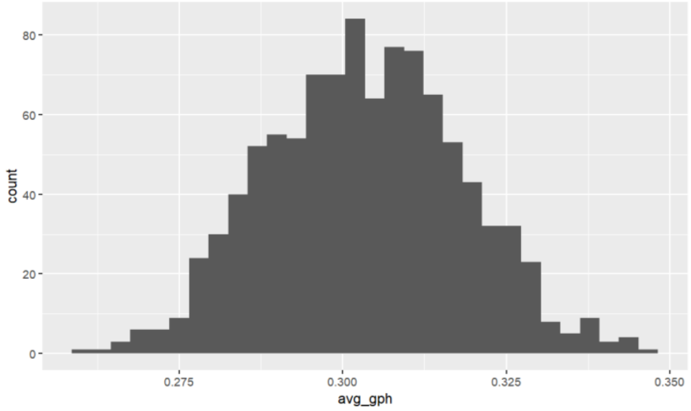
```
"Average Gain for IAH : 6.86175520986205"
"Confidence interval for  IAH : 6.42381974530183 7.29969067442227"

"Average Gain for ORD : 7.77743179122183"
"Confidence interval for  ORD : 7.32013459022188 8.23472899222177"

"Average Gain for SFO : 8.69500594530321"
"Confidence interval for  SFO : 8.15947541162006 9.23053647898636"

"Average Gain for LAX : 7.82530329289428"
"Confidence interval for  LAX : 7.25968142356738 8.39092516222118"

"Average Gain for DEN : 7.3023815895103"
"Confidence interval for  DEN : 6.65934839135574 7.94541478766487"
```

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?

   The gain per hour value is calculated.

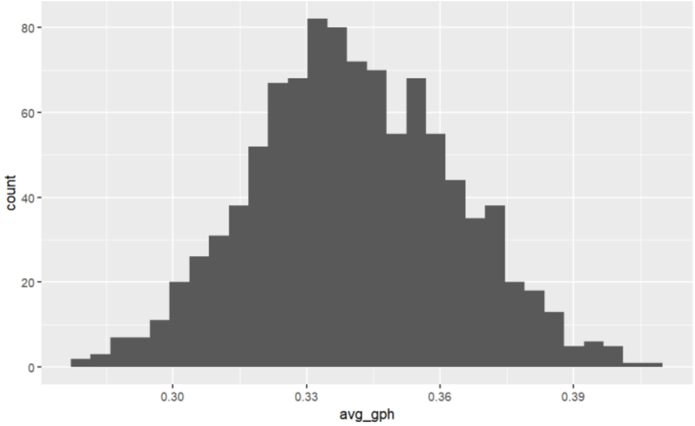
   a) After calculating confidence interval and conducting hypothesis test it is concluded that there is a statistical difference of gain per hour for flights that departed late and not late. Therefore, the average gain per hour for not late flights have more values than late flights. The result of confidence interval is (0.274,0.332)

   Histogram plot for average gain per hour differ for flights that departed late vs those that did not



   b) After calculating confidence interval and conducting hypothesis test it is concluded that there is a statistical significance difference of average gain per hour for flights that departed very late. Therefore, the average gain per hour for not very_late flights have more values than very_late flights. The confidence interval result is (0.294, 0.384)

   Histogram plot for average gain per hour differ for flights that departed very_late vs those that did not

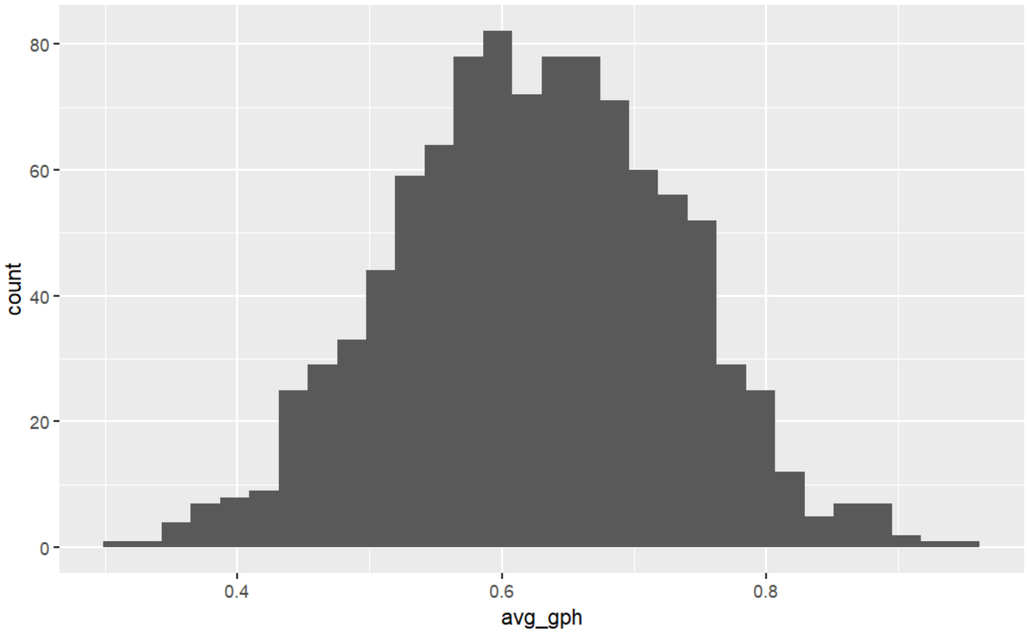


4. Does the average gain per hour differ for longer flights versus shorter flights?

   After calculating the confidence interval and conducting the hypothesis test it is concluded that there is a statistical significance difference of average gain per hour between longer flights and shorter flights. Therefore, average gain per hour for longer flights are having large value than shorter flights.

   The confidence interval is calculated as (0.4081, 0.8226)

   Histogram plot for average gain per hour differ for longer flights vs shorter flights



APPENDIX

~~~
```{r}
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
library(nycflights13)
```
~~~

~~~
```{r}
data("flights")
```


```{r}
UA_flights <- flights %>%
  filter(carrier=="UA")
```


```{r}
UA_flights_new <- na.omit(UA_flights)
```


```{r}
UA_flights <- UA_flights_new %>%
  mutate(net_gain = dep_delay - arr_delay,
         late = dep_delay > 0 ,
    very_late = dep_delay > 30,
    flight_duration = case_when(
    hour < 6 ~ "Short",
    hour >= 6 ~ "Longer"),
    gain_per_hour = net_gain/hour
    )
```


```{r}
table <- cbind(
summary(UA_flights$dep_delay),
~~~

~~~
summary(UA_flights$net_gain),
summary(UA_flights$gain_per_hour))
columns <- c("dep_delay", "net_gain", "gain_per_hour")
colnames(table)<- columns
data.frame(table)
```
```{r}
ggplot(data = UA_flights, mapping=aes(x=net_gain))+
  geom_histogram(color="white", fill="orange")
```
```{r}
ggplot(data = UA_flights, mapping=aes(x=gain_per_hour))+
  geom_histogram(color="white", fill="blue")
```
```{r}
ggplot(data = UA_flights, mapping = aes(x = flight_duration, fill = late)) +
  geom_bar()

ggplot(data = UA_flights, mapping = aes(x = flight_duration, fill = very_late)) +
  geom_bar()
```
```{r}
ggplot(data = UA_flights, mapping = aes(x = late, y = net_gain)) +
  geom_boxplot(fill="green")

ggplot(data = UA_flights, mapping = aes(x = very_late, y = net_gain)) +
  geom_boxplot(fill="green")
```
~~~

~~~
```{r}
ggplot(data = UA_flights, mapping = aes(x = late, y = gain_per_hour)) +
  geom_boxplot(fill="yellow")

ggplot(data = UA_flights, mapping = aes(x = very_late, y = gain_per_hour)) +
  geom_boxplot(fill="yellow")
```

```{r}
UA_flight.not_late <- UA_flights$net_gain[UA_flights$late == "FALSE"]
UA_flight.late <- UA_flights$net_gain[UA_flights$late == "TRUE"]

n.late <- length(UA_flight.late)
n.not_late <- length(UA_flight.not_late)

avg_gain <- numeric(1000)
for(i in 1:1000)
{
  sample.late <- sample(UA_flight.late, size = n.late, replace = TRUE)
  sample.not_late <- sample(UA_flight.not_late, size = n.not_late, replace = TRUE)
  avg_gain[i] <- mean(sample.not_late) - mean(sample.late)
}
ggplot(data=tibble(avg_gain), mapping = aes(x = avg_gain)) +
  geom_histogram()

quantile(avg_gain, c(.025, .975))
```
~~~

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


```{r}
UA_flight.not_verylate <- UA_flights$net_gain[UA_flights$very_late == "FALSE"]
UA_flight.very_late <- UA_flights$net_gain[UA_flights$very_late == "TRUE"]

n.very_late <- length(UA_flight.very_late)
n.not_verylate <- length(UA_flight.not_verylate)
avg_gain <- numeric(1000)
for(i in 1:1000)
{
  sample.very_late <- sample(UA_flight.very_late, size = n.very_late, replace = TRUE)
  sample.not_verylate <- sample(UA_flight.not_verylate, size = n.not_verylate, replace = TRUE)
  avg_gain[i] <- mean(sample.not_verylate) - mean(sample.very_late)
}

ggplot(data=tibble(avg_gain), mapping = aes(x = avg_gain)) +
  geom_histogram()

quantile(avg_gain, c(.025, .975))
```


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

~~~~
```{r}
top_airports <- UA_flights %>%
  group_by(dest) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency)) %>%
  head(5)
print(top_airports)
ggplot(data = top_airports, aes(x = dest, y = frequency)) +
  geom_bar(stat = "identity", fill="darkgreen") +
  ggtitle("five most common destination airports for United Airlines")
```


```{r}
airport_data <- UA_flights %>%
  filter(dest %in% top_airports$dest)
ggplot(data = airport_data, aes(x = net_gain)) +
  geom_histogram(position = "identity", alpha = 0.7, fill = "blue",color="black") +
  ggtitle("Gain Distribution for 5 most Airports for United Airlines") +
  facet_wrap(~ dest, ncol = 3)

for (airport in top_airports$dest) {
  airport_data <- UA_flights %>%
    filter(dest == airport)
  avg_gain <- mean(airport_data$net_gain)
  conf_int <- t.test(airport_data$net_gain)$conf.int
  print(paste("Average Gain for", airport, ":", avg_gain))
  print(paste("Confidence interval for ", airport, ":", conf_int[1],conf_int[2]))
}```
~~~~

~~~
```{r}
UA_flight.not_late <- UA_flights$gain_per_hour[UA_flights$late == "FALSE"]
UA_flight.late <- UA_flights$gain_per_hour[UA_flights$late == "TRUE"]

n.late <- length(UA_flight.late)
n.not_late <- length(UA_flight.not_late)

avg_gph <- numeric(1000)
for(i in 1:1000)
{
  sample.late <- sample(UA_flight.late, size = n.late, replace = TRUE)
  sample.not_late <- sample(UA_flight.not_late, size = n.not_late, replace = TRUE)
  avg_gph[i] <- mean(sample.not_late) - mean(sample.late)
}
ggplot(data=tibble(avg_gph), mapping = aes(x = avg_gph)) +
  geom_histogram()

quantile(avg_gph, c(.025, .975))
```

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

```{r}
UA_flight.not_verylate <- UA_flights$gain_per_hour[UA_flights$very_late == "FALSE"]
UA_flight.very_late <- UA_flights$gain_per_hour[UA_flights$very_late == "TRUE"]
~~~

~~~
n.very_late <- length(UA_flight.very_late)
n.not_verylate <- length(UA_flight.not_verylate)

avg_gph <- numeric(1000)
for(i in 1:1000)
{
  sample.very_late <- sample(UA_flight.very_late, size = n.very_late, replace = TRUE)
  sample.not_verylate <- sample(UA_flight.not_verylate, size = n.not_verylate, replace = TRUE)
  avg_gph[i] <- mean(sample.not_verylate) - mean(sample.very_late)
}
ggplot(data=tibble(avg_gph), mapping = aes(x = avg_gph)) +
  geom_histogram()

quantile(avg_gph, c(.025, .975))
```

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

```{r}
UA_flight.shorter <- UA_flights$gain_per_hour[UA_flights$flight_duration == "Short"]
UA_flight.longer <- UA_flights$gain_per_hour[UA_flights$flight_duration == "Longer"]

n.shorter <- length(UA_flight.shorter)
n.longer <- length(UA_flight.longer)

avg_gph <- numeric(1000)
~~~

~~~
for(i in 1:1000)
{
  sample.shorter <- sample(UA_flight.shorter, size = n.shorter, replace = TRUE)
  sample.longer <- sample(UA_flight.longer, size = n.longer, replace = TRUE)
  avg_gph[i] <- mean(sample.shorter) - mean(sample.longer)
}
ggplot(data=tibble(avg_gph), mapping = aes(x = avg_gph)) +
  geom_histogram()

quantile(avg_gph, c(.025, .975))
```


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