1. PROJECT STATEMENT:

This project will use the data included in the nycflights13 package.

Again, suppose you work for United Airlines (carrier code UA). After having previously studied departure delays, you will now be investigating gain per flight - that is, how much quicker the flight ended up being than planned. We can find the net gain by subtracting the arrival delay from the departure delay. Create a new variable to measure the net gain.

Prepare a report that utilizes confidence intervals and hypothesis tests, alongside appropriate exploratory data analysis, to analyze flight gains. You can choose where you want to use confidence intervals and where you want to use hypothesis tests, but one or the other should be used to address each question.

Your report should address the following questions:

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

2. EXECUTIVE SUMMARY

In this project, we are importing dataset of nycflights and then filtering the data in order to find the flights of UA carrier. Then, we can add the following new variables which are useful in our analysis as follows:

late = TRUE if delay > 0 else FALSE.

very_late = TRUE if delay > 30 else FALSE.

gain = departure delay – arrival delay.

gain per hour = gain / hour.

Now, to answer the above four questions mentioned in problem statement, I performed some exploratory data analysis using some graphs. And also some permutation tests and t-test in order to analyze and support the answers as mentioned in the following report.

3. AVERAGE GAIN FOR DELAYED FLIGHTS

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?

The net gain can be calculated using **Gain = Departure delay – Arrival delay**.

The flights which are delayed i.e., the flights with departure delay > 0 are the flights which are late. And, the flights whole delay <= 0 are not late.

The flights which are delayed30 i.e., the flights with departure delay > 30 are the flights which are very late. And, the flights whole delay <= 30 are not very late.

3.1 Late Vs Not Late

A tibble: 2 × 2

late <lgl>	avg_gain <dbl>
FALSE	9.269172
TRUE	7.543115

2 rows

Fig 3.1 Table having average gain for late and not late flights.

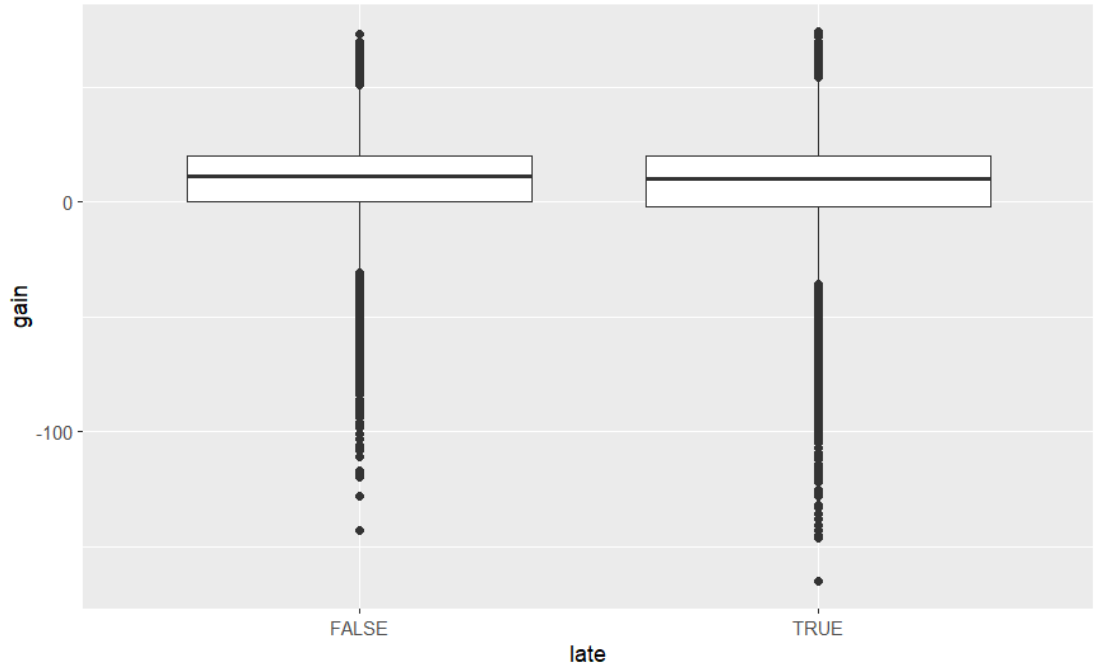

Fig 3.2 Boxplot of gain for late and not late flights.

The boxplots of late and not late flights look almost similar. Hence, to further analyze if there is a difference in average gain between late and not late flights, we can perform permutation tests as follows.

Null Hypothesis H0: Mean of net gain for not late flights = Mean of net gain for late flights.

Alternate Hypothesis H1: Mean of net gain for not late flights > Mean of net gain for late flights.

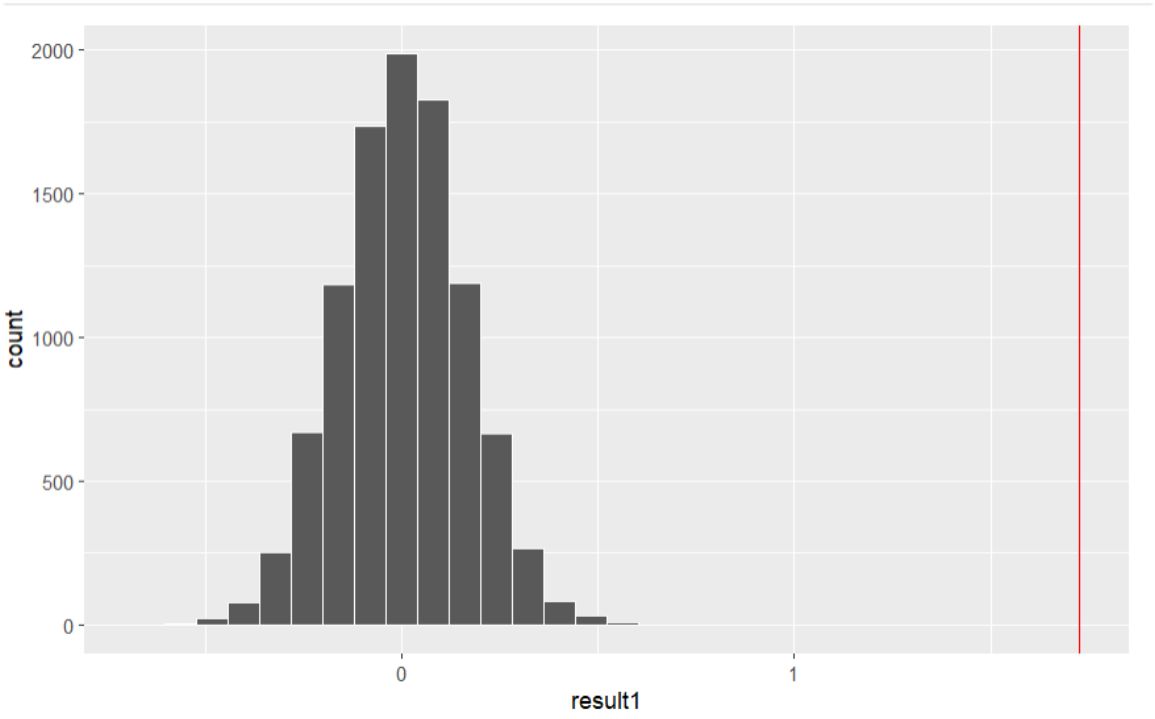

Fig 3.3 Graph for differences in gain for late and not late flights.

From the permutation test the resultant p-value is 1e-04. As it is less than 0.05, there is evidence for significant difference in net gain between late and not late flights. So, we can support our alternate hypothesis H1 i.e., mean of net gain for not late flights > mean of net gain for late flights.

```
        Welch Two Sample t-test

data:  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 greater than 0
95 percent confidence interval:
 1.461913      Inf
sample estimates:
mean in group FALSE  mean in group TRUE
           9.269172            7.543115
```

Fig 3.4 T-test for Gain between late and not late flights.

From Fig 3.3 and Fig 3.4, we can conclude that true difference in means between group FALSE and group TRUE is greater than 0. So, the average gain of flights that are not delayed is greater than flights that are delayed.

3.2 Very Late Vs Not Very Late

A tibble: 2 × 2

very_late <lgl>	avg_gain <dbl>
FALSE	8.699534
TRUE	6.857881

2 rows

Fig 3.5 Table of average gain for very_late and not very_late flights.

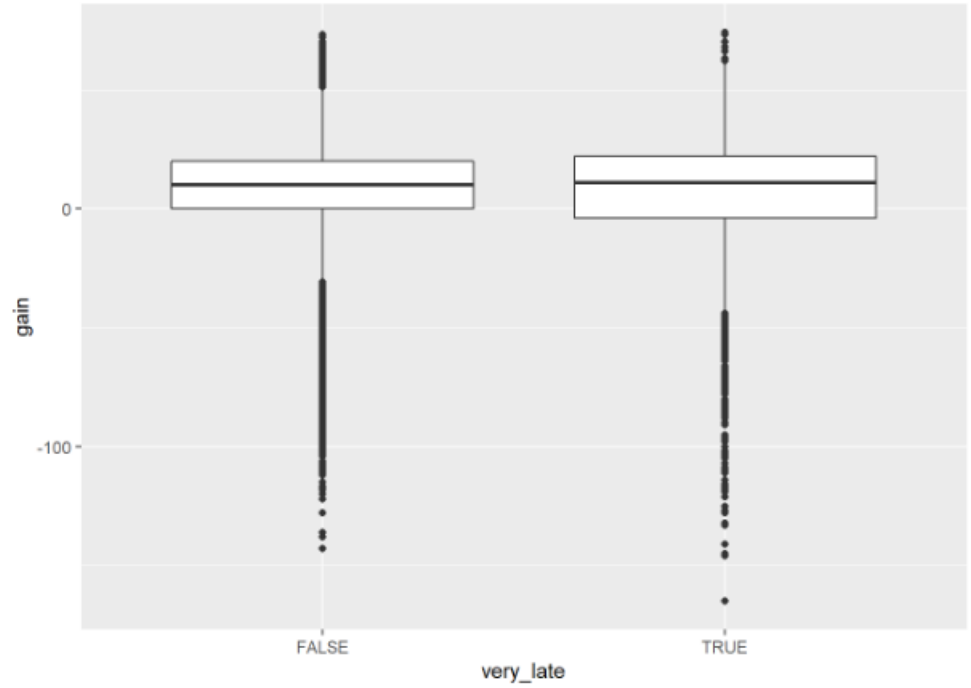

Fig 3.6 Boxplot of gain for very late and not very late flights.

The boxplots of very late and not very late flights do not give much information about the average gain. Hence, to further analyze if there is a difference in average gain between very late and not very late flights, we can perform permutation tests as follows.

Null Hypothesis H0: Mean of net gain for not very late flights = Mean of net gain for very late flights.

Alternate Hypothesis H1: Mean of net gain for not very late flights > Mean of net gain for very late flights.

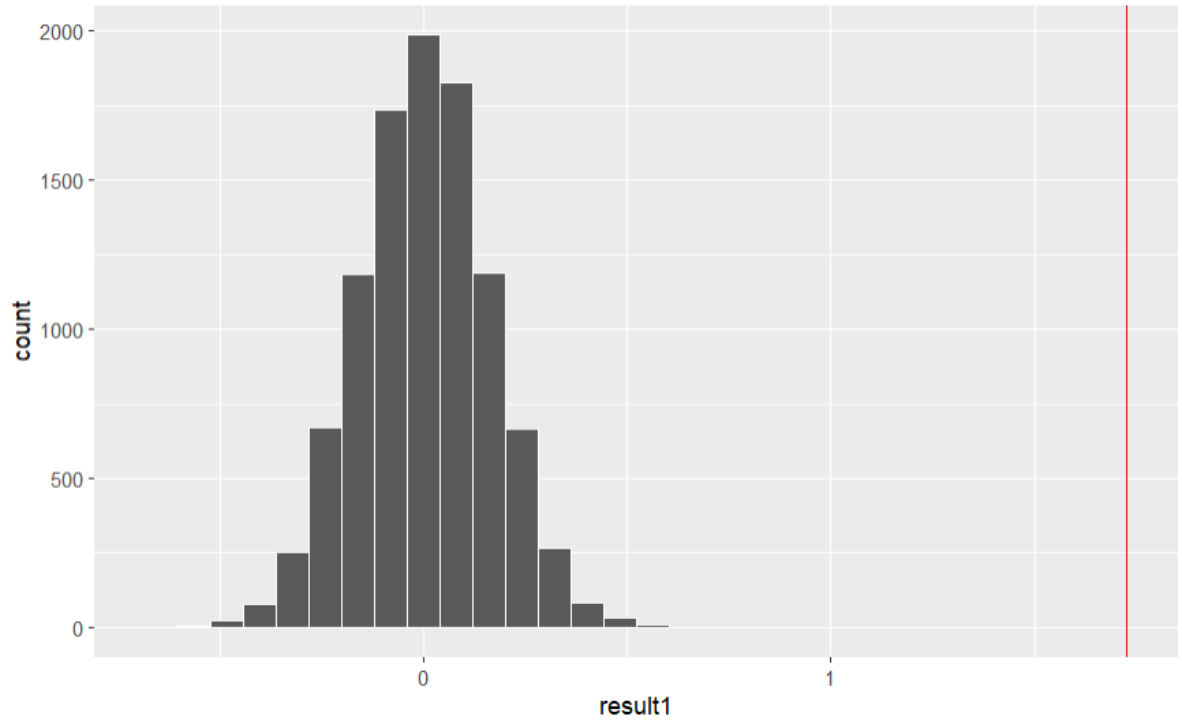

Fig 3.7 Graph for differences in Gain for very late and not very late flights.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in net gain for very late and not very late flights. So, we can support our alternate hypothesis H1 i.e., mean of net gain for not very late flights > mean of net gain for very late flights.

```
        Welch Two Sample t-test

data:  gain by very_late
t = 6.2953, df = 8838.6, p-value = 1.607e-10
alternative hypothesis: true difference in means between group FALSE and group TRUE is greater than 0
95 percent confidence interval:
 1.360407      Inf
sample estimates:
mean in group FALSE  mean in group TRUE
           8.699534            6.857881
```

Fig 3.8 T-test for Gain between very late and not very late flights.

From Fig 3.7 and Fig 3.8, we can conclude that true difference in means between group FALSE and group TRUE is greater than 0. So, the average gain of flights which are not delayed by 30 minutes is greater than flights that are delayed by 30 minutes.

4. MOST COMMON DESTINATION AIRPORTS

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

A tibble: 5 × 2 Groups: dest [5]

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

5 rows

Fig 4.1 Table of 5 most common destination airports.

The five most common destination airports for United Airlines from New York City are ORD, IAH, SFO, LAX, DEN.

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

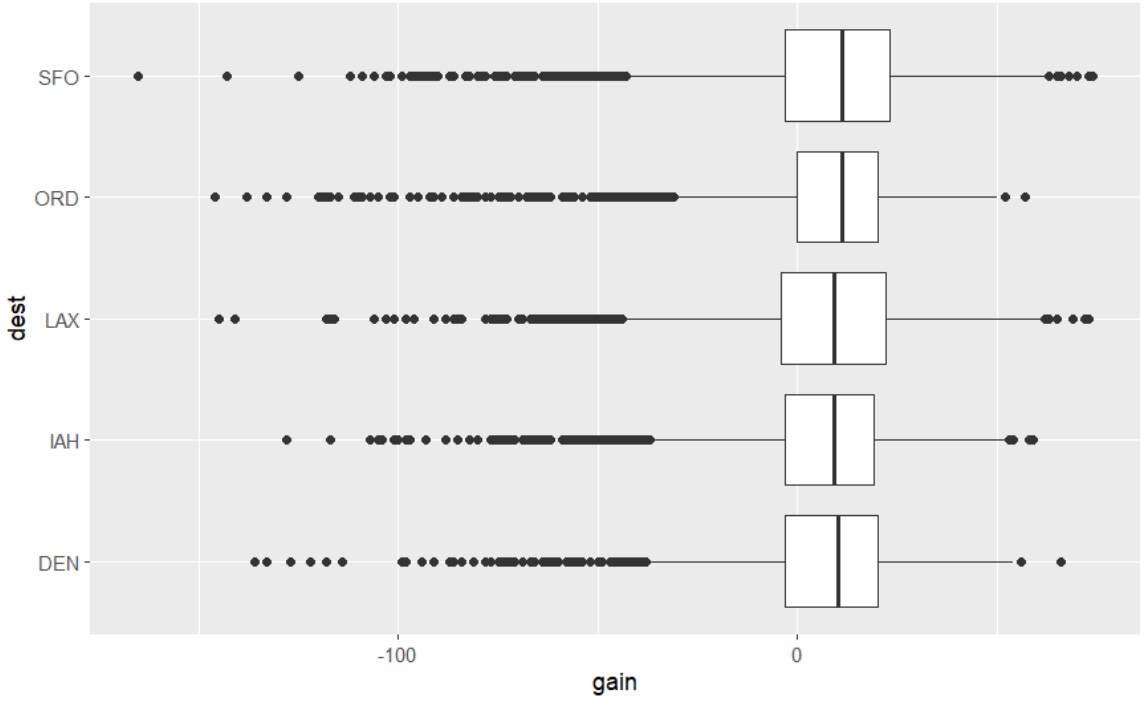

Fig 4.2 Boxplot of gain for 5 most common destination airports.

A tibble: 5 × 2

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

5 rows

Fig 4.3 Table of average gain for 5 most common destination airports.

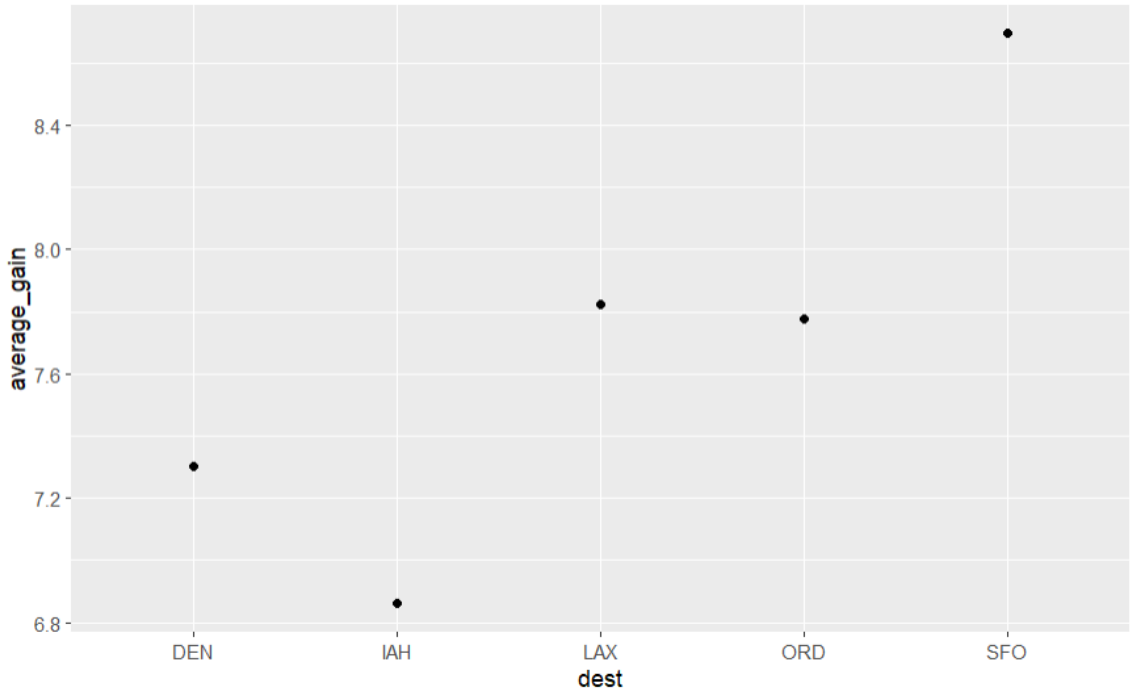

Fig 4.4 Scatter plot of average gain for 5 most common destination airports.

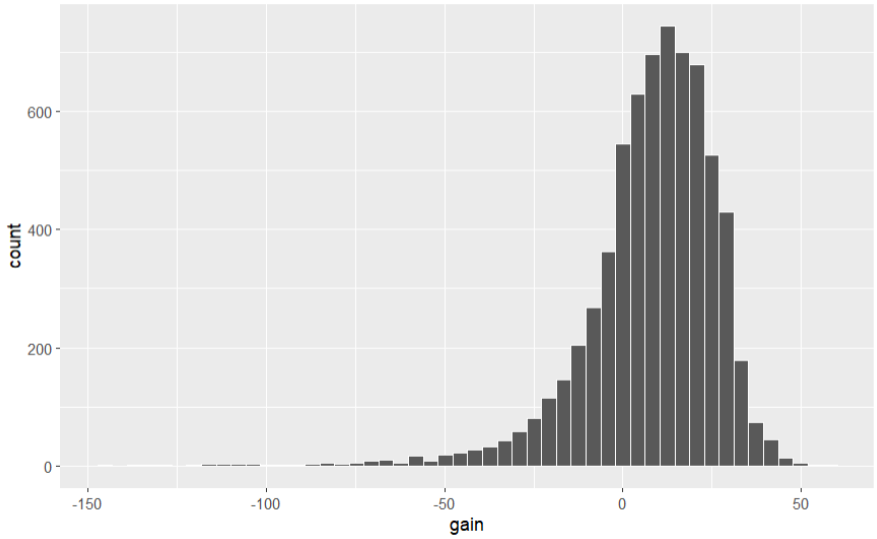

Fig 4.5 Histogram of net gain for the destination airport 'ORD'

From the above histogram, the distribution follows normal distribution and the graph looks left-skewed.

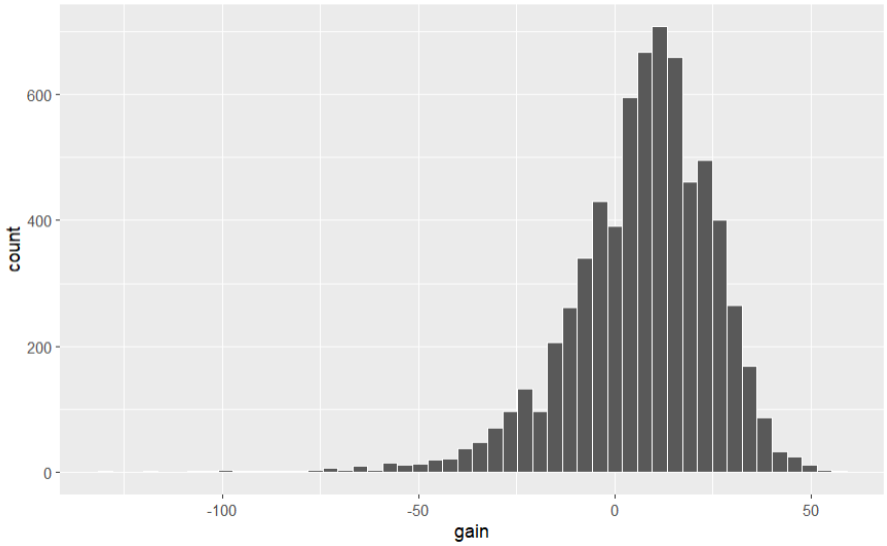

Fig 4.6 Histogram of net gain for the destination airport 'IAH'

From the above histogram, the distribution follows normal distribution and the graph looks left-skewed.

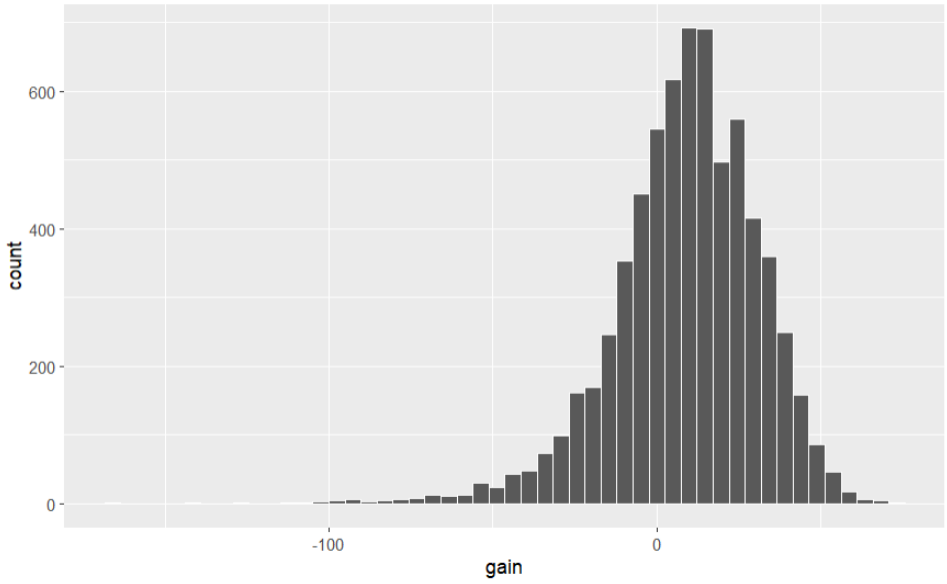

Fig 4.7 Histogram of net gain for the destination airport ‘SFO’

From the above histogram, the distribution follows normal distribution and the graph looks left-skewed.

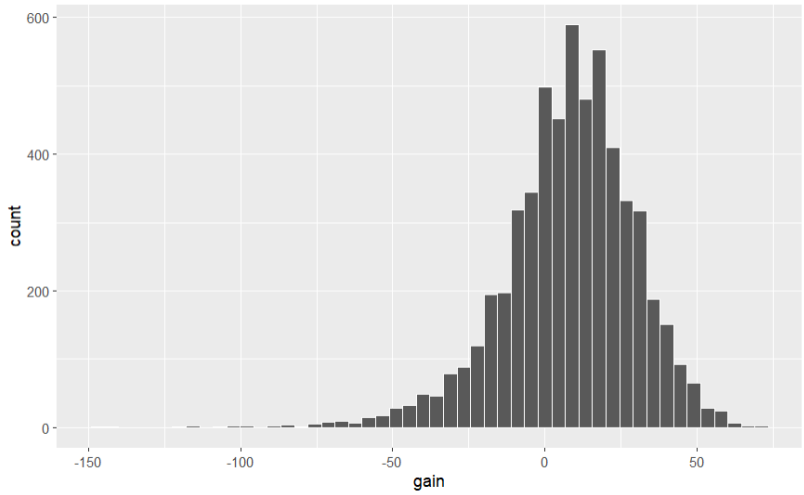

Fig 4.8 Histogram of net gain for the destination airport ‘LAX’

From the above histogram, the distribution follows normal distribution and the graph looks left-skewed.

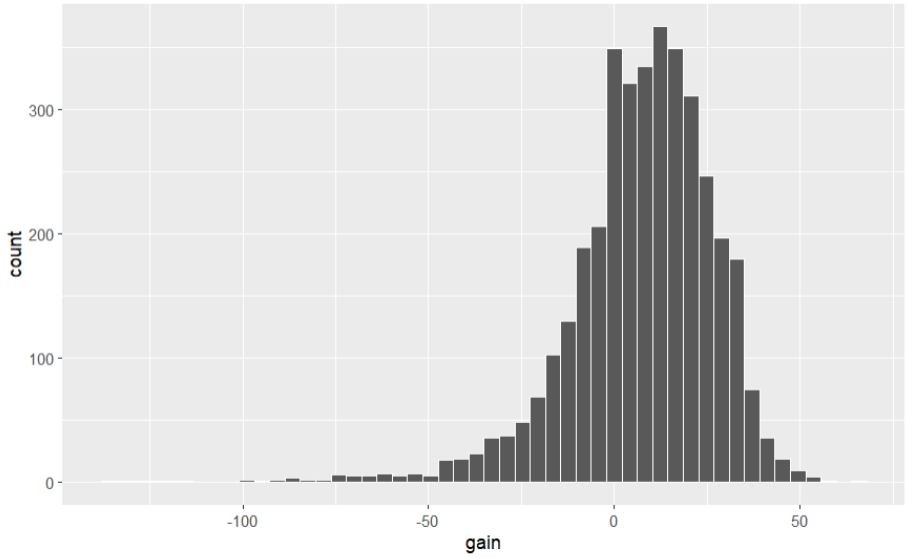

Fig 4.9 Histogram of net gain for the destination airport ‘DEN’

From the above histogram, the distribution follows normal distribution and the graph looks left-skewed.

5. AVERAGE GAIN PER HOUR FOR DELAYED FLIGHTS

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 is calculated by using the formula

Gain per hour = Gain / hour.

The flights which are delayed i.e., the flights with departure delay > 0 are the flights which are late. And, the flights whole delay <= 0 are not late.

The flights which are delayed30 i.e., the flights with departure delay > 30 are the flights which are very late. And, the flights whole delay <= 30 are not very late.

5.1 Late Vs Not Late

A tibble: 2 × 2

late <lgl>	avg_gain <dbl>
FALSE	0.9310086
TRUE	0.6274405

2 rows

Fig 5.1 Table having average gain per hour for late and not late flights.

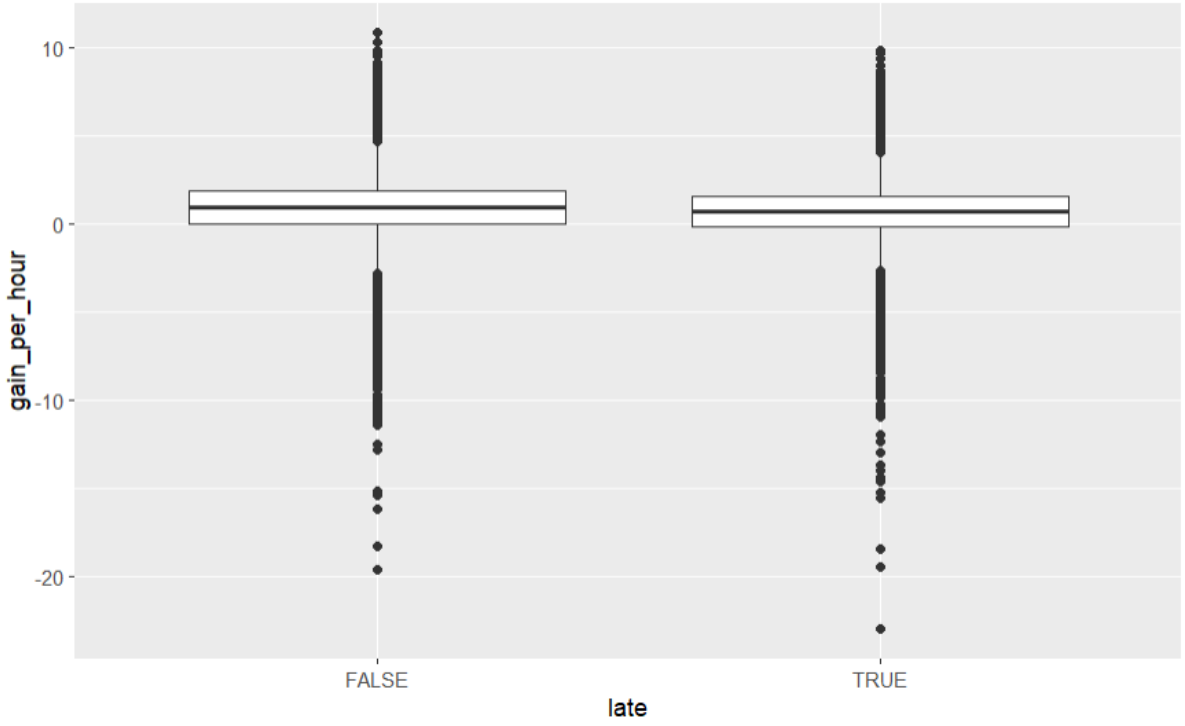

Fig 5.2 Boxplot of gain per hour for late and not late flights.

As the boxplots for both late and not late flights are almost similar, to further analyze if there is any difference in average gain per hour between late and not late flights, we can perform the permutation tests as follows.

Null Hypothesis H0: Mean of gain per hour for not late flights = Mean of gain per hour for late flights.

Alternate Hypothesis H1: Mean of gain per hour for not late flights > Mean of gain per hour for late flights.

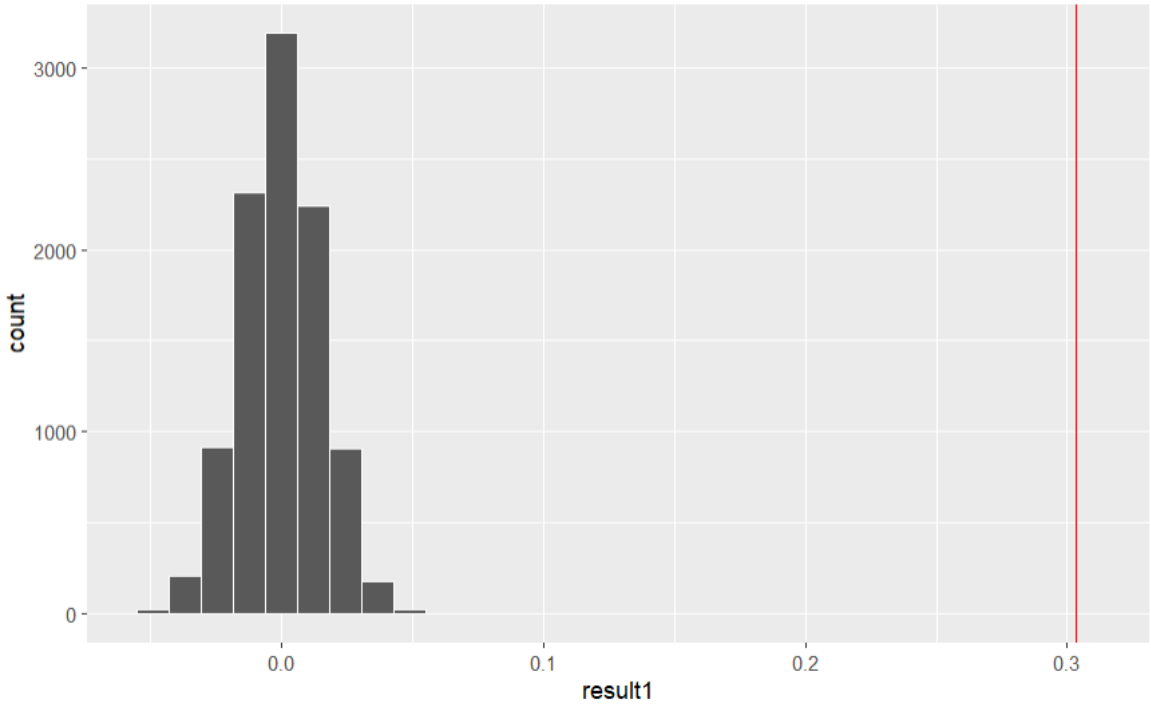

Fig 5.3 Graph for differences in Gain per hour for late and not late flights.

From the permutation test the resultant p-value is 1e-04. As it is less than 0.05, there is evidence for significant difference in gain per hour between late and not late flights. So, we can support our alternate hypothesis H1 i.e., mean of gain per hour for not late flights > mean of gain per hour for late flights.

```
        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 greater than 0
95 percent confidence interval:
 0.278671      Inf
sample estimates:
mean in group FALSE  mean in group TRUE
          0.9310086           0.6274405
```

Fig 5.4 T-test for Gain per hour between late and not late flights.

From Fig 5.3 and Fig 5.4, we can conclude that true difference in means between group FALSE and group TRUE is greater than 0. So, the average gain per hour of flights that are not delayed is greater than flights that are delayed.

5.2 Very Late Vs Not Very Late

A tibble: 2 × 2

very_late <lgl>	avg_gain <dbl>
FALSE	0.8330167
TRUE	0.4923389

2 rows

Fig 5.5 Table having average gain per hour for very late and not very late flights.

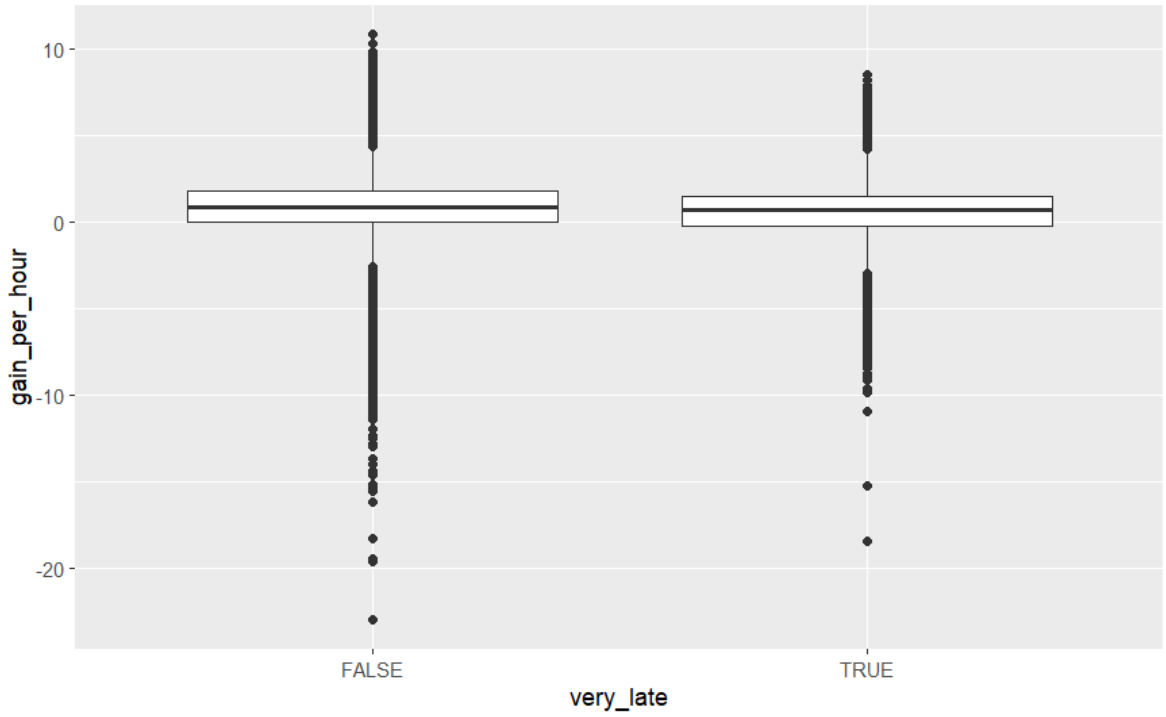

Fig 5.6 Boxplot of gain per hour for very late and not very late flights.

To further understand the difference of gain per hour between late and not late flights, we can perform permutation tests as follows.

Null Hypothesis H0: Mean of gain per hour for not very late flights = Mean of gain per hour for very late flights.

Alternate Hypothesis H1: Mean of gain per hour for not very late flights > Mean of gain per hour for very late flights.

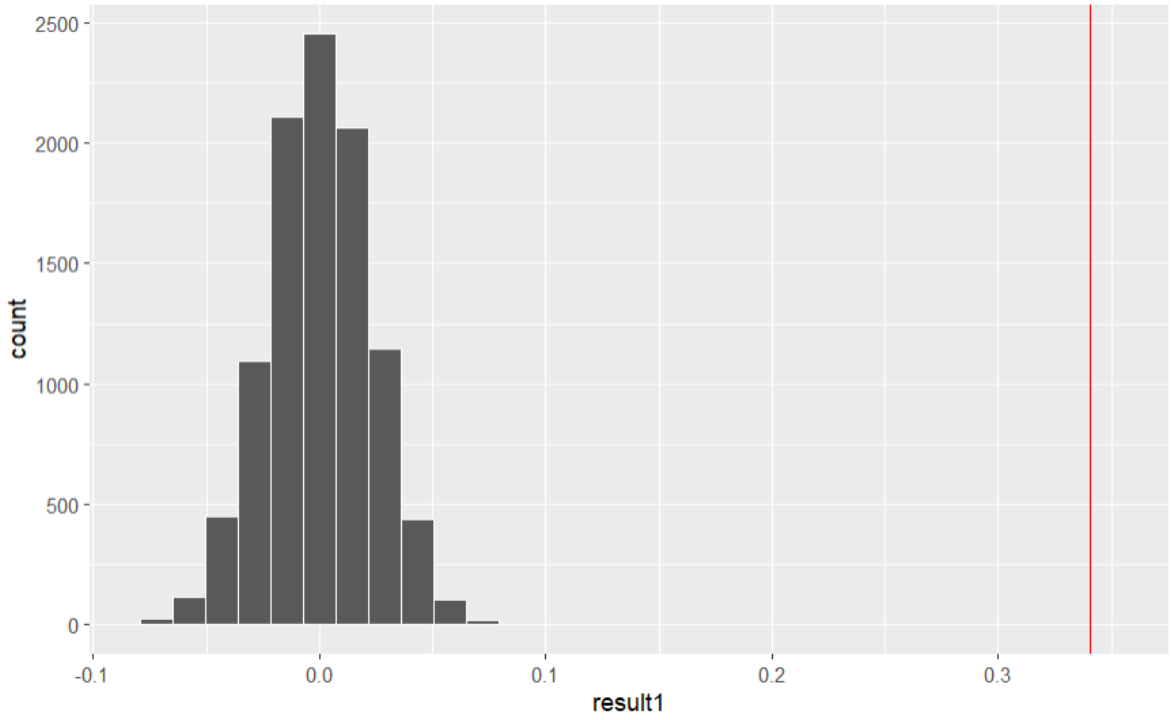

Fig 5.7 Graph for differences in Gain per hour for very late and not very late flights.

From the permutation test the resultant p-value is 1e-04. As it is less than 0.05, there is evidence for significant difference in net gain for very late and not very late flights. So, we can support our alternate hypothesis H1 i.e., mean of gain per hour for not very late flights > mean of gain per hour for very late flights.

```
        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 greater than 0
95 percent confidence interval:
 0.303244      Inf
sample estimates:
mean in group FALSE  mean in group TRUE
          0.8330167           0.4923389
```

Fig 5.8 T-test for Gain per hour between very late and not very late flights.

From Fig 5.7 and Fig 5.8, we can conclude that true difference in means between group FALSE and group TRUE is greater than 0. So, the average gain per hour of flights which are not delayed by 30 minutes is greater than flights that are delayed by 30 minutes.

6. GAIN PER HOUR FOR LONGER AND SHORTER FLIGHTS

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

A tibble: 2 × 2

air_time <= m <lgl>	avg_gain <dbl>
FALSE	0.7053551
TRUE	0.8522031

2 rows

Fig 6.1 Table having average gain per hour for shorter and longer flights based on threshold of mean m of air time.

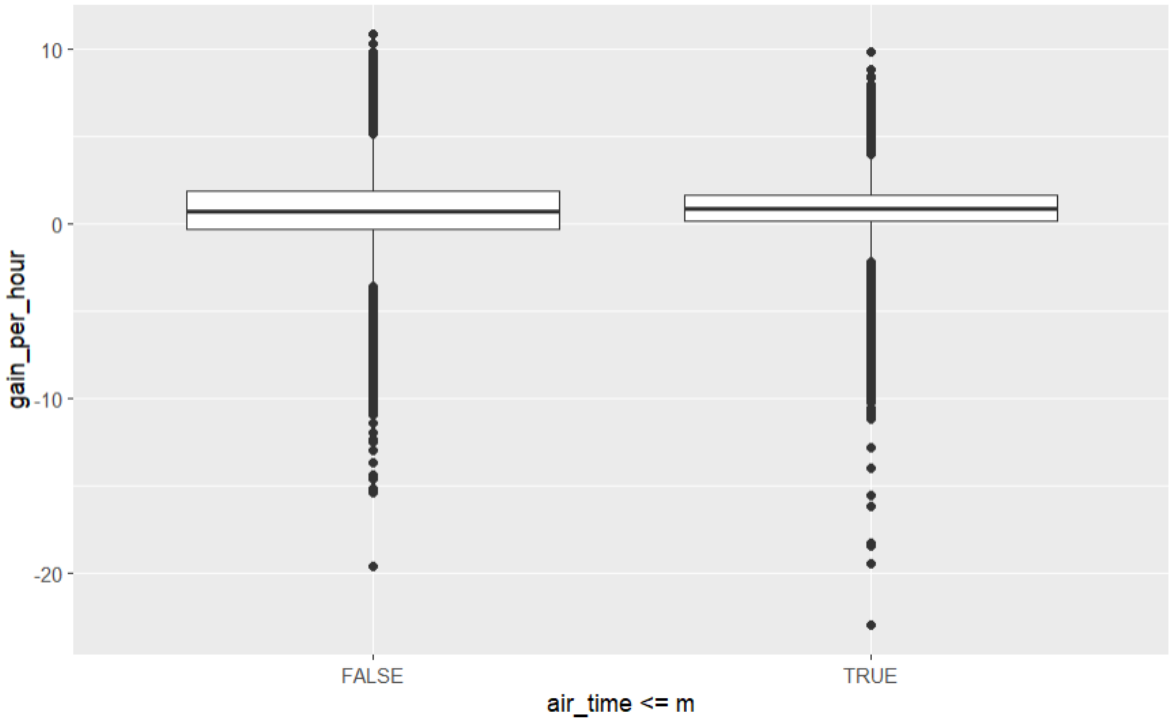

Fig 6.2 Boxplot of gain per hour for airtime <= mean m and > mean m.

We can conduct permutation test to get a good idea about the relationship by taking a threshold value of mean m.

Null Hypothesis H0: Mean of gain per hour with (airtime <= m) = Mean of gain per hour with (airtime > m)

Alternate Hypothesis H1: Mean of gain per hour with (airtime <= m) != Mean of gain per hour with (airtime > m)

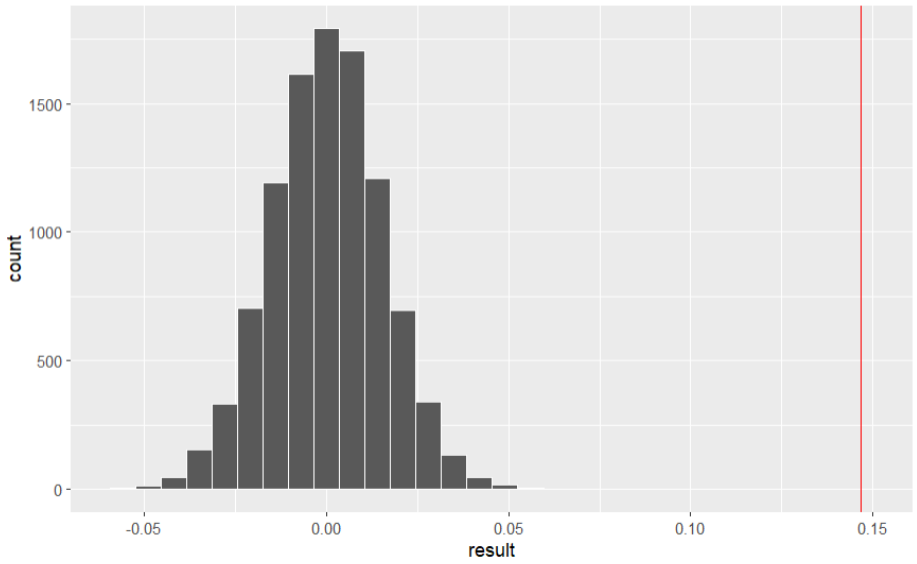

Fig 6.2 Graph for differences in Gain per hour for airtime<=m and airtime>m.

From the permutation test the resultant p-value is 1e-04. As it is less than 0.05, there is evidence for significant difference in gain per hour. So, we can support our alternate hypothesis H1 i.e., mean of gain per hour with (airtime <= m) != mean of gain per hour with (airtime > m)

```
        Welch Two Sample t-test

data:  gain_per_hour by air_time <= m
t = -9.263, df = 45829, 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:
 -0.1779203 -0.1157757
sample estimates:
mean in group FALSE  mean in group TRUE
          0.7053551           0.8522031
```

Fig 6.3 T-test for gain per hour for airtime <= m and airtime >m.

From the Fig 6.2 and 6.3, true difference in means between group FALSE and group TRUE is not equal to 0.

7. APPENDIX

Import the required libraries

```
library(ggplot2)

library(tidyverse)

library(nycflights13)

library(moderndive)
```

Creating a new dataset named 'myflights' in order to filter the carrier as 'UA' and add new columns late, very_late and gain.

```
myflights <- flights %>%
  filter(carrier == 'UA') %>%
  mutate(
    late = dep_delay>0,
    very_late = dep_delay>30,
    gain = dep_delay - arr_delay
  )

glimpse(myflights)
```

1

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

```
data1 <- myflights %>% group_by(late) %>% summarize(avg_gain = mean(gain, na.rm = TRUE))

data1
```

```
ggplot(data = myflights, aes(x = late, y = gain))+
  geom_boxplot()
observed <- mean(myflights$gain[myflights$late == FALSE ], na.rm = TRUE) -
mean(myflights$gain[myflights$late == TRUE ], na.rm = TRUE)
observed
N <- 10^4-1
sample.size = nrow(myflights)
notdelayed = myflights %>% filter(late == FALSE)
group.1.size = nrow(notdelayed)
result1 <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result1[i] = mean(myflights$gain[index],na.rm = TRUE) - mean(myflights$gain[-
index],na.rm = TRUE)
}
ggplot(data=tibble(result1), mapping = aes(x=result1)) +
  geom_histogram(color = 'white') +
  geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result1 >= observed) + 1) / (N + 1)
t.test(gain~late,data=myflights, alternative = 'greater')
```

b. What about for flights that departed more than 30 minutes late?

```
data1 <- myflights %>% group_by(very_late) %>% summarize(avg_gain = mean(gain, na.rm = TRUE))

data1

ggplot(data = myflights, aes(x = very_late, y = gain))+
  geom_boxplot()

observed <- mean(myflights$gain[myflights$very_late == FALSE ], na.rm = TRUE) - mean(myflights$gain[myflights$very_late == TRUE ], na.rm = TRUE)

observed

N <- 10^4-1

sample.size = nrow(myflights)

notdelayed30 = myflights %>% filter(very_late == FALSE)

group.1.size = nrow(notdelayed30)

result2 <- numeric(N)

for(i in 1:N)

{

  index = sample(sample.size, size=group.1.size, replace = FALSE)

  result2[i] = mean(myflights$gain[index],na.rm = TRUE) - mean(myflights$gain[-index],na.rm = TRUE)

}

ggplot(data=tibble(result2), mapping = aes(x=result2)) +
  geom_histogram(color = 'white') +
  geom_vline(xintercept = observed, color = "red")

#Calculate the p-value

(sum(result2 >= observed) + 1) / (N + 1)

t.test(gain~very_late,data=myflights, alternative = "greater")
```

2.

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

```
dest_airports <- myflights %>% group_by(dest) %>% count()

dest_airports <- arrange(dest_airports,desc(n))

glimpse(dest_airports)

cat("The five most destination airports from nyc are : IAH, ORD, SFO, LAX, DEN.")

most_common_data <- myflights %>% filter(dest %in% dest_airports[1:5,]$dest)

most_common_data %>% group_by(dest) %>% count()
```

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

```
ggplot(data = most_common_data, aes(x = gain, y = dest))+

  geom_boxplot()

data1 <- most_common_data %>% group_by(dest) %>% summarize(average_gain =
mean(gain, na.rm = TRUE))

data1

ggplot(data = data1, aes(x = dest, y = average_gain))+

  geom_point()

ggplot(data = most_common_data %>% filter(dest == 'ORD') , aes(x = gain))+

  geom_histogram(color = 'white', bins = 50)

cat('The distribution is normal distribution and looks left skewed.')

ggplot(data = most_common_data %>% filter(dest == 'IAH') , aes(x = gain))+

  geom_histogram(color = 'white', bins = 50)

cat('The distribution is normal and lightly left skewed.')

ggplot(data = most_common_data %>% filter(dest == 'SFO') , aes(x = gain))+
```

```
  geom_histogram(color = 'white', bins = 50)
cat('The distribution looks normal and lightly left-skewed.')
ggplot(data = most_common_data %>% filter(dest == 'LAX') , aes(x = gain))+
  geom_histogram(color = 'white', bins = 50)
cat('The distribution is normal and left skewed.')
ggplot(data = most_common_data %>% filter(dest == 'DEN') , aes(x = gain))+
  geom_histogram(color = 'white', bins = 50)
cat('The distribution is normal and left-skewed.')
```

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.

```
myflights <- myflights %>%
  mutate(
    gain_per_hour = gain/hour
  )
glimpse(myflights)
```

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

```
data1 <- myflights %>% group_by(late) %>% summarize(avg_gain = mean(gain_per_hour, na.rm = TRUE))
data1
ggplot(data = myflights, aes(x = late, y = gain_per_hour))+
  geom_boxplot()
```

```
observed <- mean(myflights$gain_per_hour[myflights$late == FALSE ], na.rm = TRUE) - mean(myflights$gain_per_hour[myflights$late == TRUE ], na.rm = TRUE)

observed

N <- 10^4-1

sample.size = nrow(myflights)

delayed = myflights %>% filter(late == TRUE)

group.1.size = nrow(delayed)

result1 <- numeric(N)

for(i in 1:N)

{

  index = sample(sample.size, size=group.1.size, replace = FALSE)

  result1[i] = mean(myflights$gain_per_hour[index],na.rm = TRUE) - mean(myflights$gain_per_hour[-index],na.rm = TRUE)

}

ggplot(data=tibble(result1), mapping = aes(x=result1)) +

  geom_histogram(color = 'white') +

  geom_vline(xintercept = observed, color = "red")

#Calculate the p-value

(sum(result1 >= observed) + 1) / (N + 1)

t.test(gain_per_hour~late,data=myflights, alternative = "greater")
```

What about for flights that departed more than 30 minutes late?

```
data1 <- myflights %>% group_by(very_late) %>% summarize(avg_gain = mean(gain_per_hour, na.rm = TRUE))

data1

ggplot(data = myflights, aes(x = very_late, y = gain_per_hour))+
```

```
  geom_boxplot()

observed <- mean(myflights$gain_per_hour[myflights$very_late == FALSE ], na.rm =
TRUE) - mean(myflights$gain_per_hour[myflights$very_late == TRUE ], na.rm = TRUE)

observed

N <- 10^4-1

sample.size = nrow(myflights)

delayed30 = myflights %>% filter(very_late == TRUE)

group.1.size = nrow(delayed30)

result1 <- numeric(N)

for(i in 1:N)

{

  index = sample(sample.size, size=group.1.size, replace = FALSE)

  result1[i] = mean(myflights$gain_per_hour[index],na.rm = TRUE) -
mean(myflights$gain_per_hour[-index],na.rm = TRUE)

}

ggplot(data=tibble(result1), mapping = aes(x=result1)) +

  geom_histogram(color = 'white') +

  geom_vline(xintercept = observed, color = "red")

#Calculate the p-value

(sum(result1 >= observed) + 1) / (N + 1)

t.test(gain_per_hour~very_late,data=myflights, alternative = "greater")
```

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

```
m = mean(myflights$air_time, na.rm = TRUE)

cat('Mean of air_time is',m)
```

```
data1 <- myflights %>% group_by(air_time <= m) %>% summarize(avg_gain =
mean(gain_per_hour, na.rm = TRUE))

data1

ggplot(data = myflights, aes(x = air_time <= m , y = gain_per_hour))+

  geom_boxplot()

observed <- mean(myflights$gain_per_hour[myflights$air_time <= m], na.rm = TRUE) -
mean(myflights$gain_per_hour[myflights$air_time > m], na.rm = TRUE)

observed

N <- 10^4-1

sample.size = nrow(myflights)

group1 <- myflights %>% filter(air_time<=m)

group.1.size = nrow(group1)

result <- numeric(N)


for(i in 1:N)

{

  index = sample(sample.size, size=group.1.size, replace = FALSE)


  result[i] = mean(myflights$gain_per_hour[index],na.rm = TRUE) -
mean(myflights$gain_per_hour[-index],na.rm = TRUE)

}


ggplot(data=tibble(result), mapping = aes(x=result)) +

  geom_histogram(color = 'white') +

  geom_vline(xintercept = observed, color = "red")
```

```
#Calculate the p-value

2*(sum(result >= observed) + 1) / (N + 1)

t.test(gain_per_hour~air_time<=m, data=myflights, alternative = "two.sided")
```