Project Topic: United Airlines Advanced Analysis Based On Gain Per Flight

TABLE OF CONTENT

Title	Page No
Executive Summary	3
Introduction	4
First Point to address	5
Second point to address	9
Third point to address	15
Fourth Point to address	19
Appendix	21

EXECUTIVE SUMMARY

In this report, we focus on the United Airlines data, which is part of flights dataset. Flights dataset is available in nycflights13 package. We will 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. We make use of t-test as the dataset satisfies the condition that the dataset should be large to use a t-test.

The report addresses 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?

The report utilizes confidence intervals and hypothesis tests, alongside appropriate exploratory data analysis, to address the above questions

INTRODUCTION

The goals of the analysis is to determine:

1. Whether average gain differ for flights that departed late versus those that did not? What about for flights that departed more than 30 minutes late?

   Here we make use of t-test and data analysis to find solution to above question

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

   Here we make use of confidence intervals, data analysis, mean(), groupby() and other methods to find solution to above question

3. Whether 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?

   Here we make use of t-test and data analysis to find solution to above question

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

   Here we make use of t-test and data analysis to find solution to above question

In this report, we make use of nycflights13 package. We focus on the United Airlines data. We make use of flights and airports datasets. We keep all the variables in the two datasets. Additionally we add new variables which are: ‘net_gain’, ‘late’, ‘very_late’, ‘duration_in_hours’, ‘gain_per_hour’ , and ’ shorter_flight’

OUTCOME OF EACH POINT

First we fetch dataset corresponding to only United Airlines from ‘flights’ dataset and name it ‘UAflights’. We remove all rows containing any missing values. We create a column called ‘net_gain’ by subtracting ‘arr_delay’ from ‘dep_delay’. We create column called ‘late’ where ‘TR’ corresponds to ‘dep_delay>0’ and ‘FA’ correspond to ‘dep_delay<0’. We create column ‘very_late’ where ‘TR’ corresponds to ‘dep_delay>30’ and ‘FA’ corresponds to ‘dep_delay<30’. As our dataset is large, we can use t test for entire project.

FIRST POINT TO ADDRESS: 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?

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

->Test

Here the Null and Alternative hypothesis are:

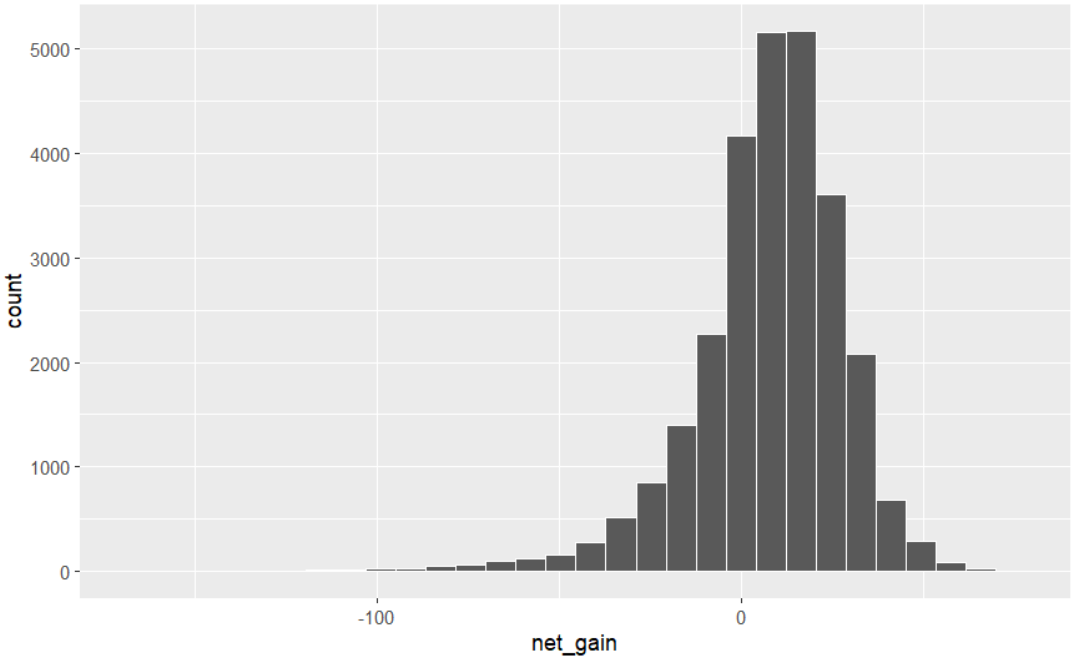
NULL: meanGain_departedLate=meanGain_departedNotLate

ALTERNATIVE: meanGain_departedLate!=meanGain_departedNotLate

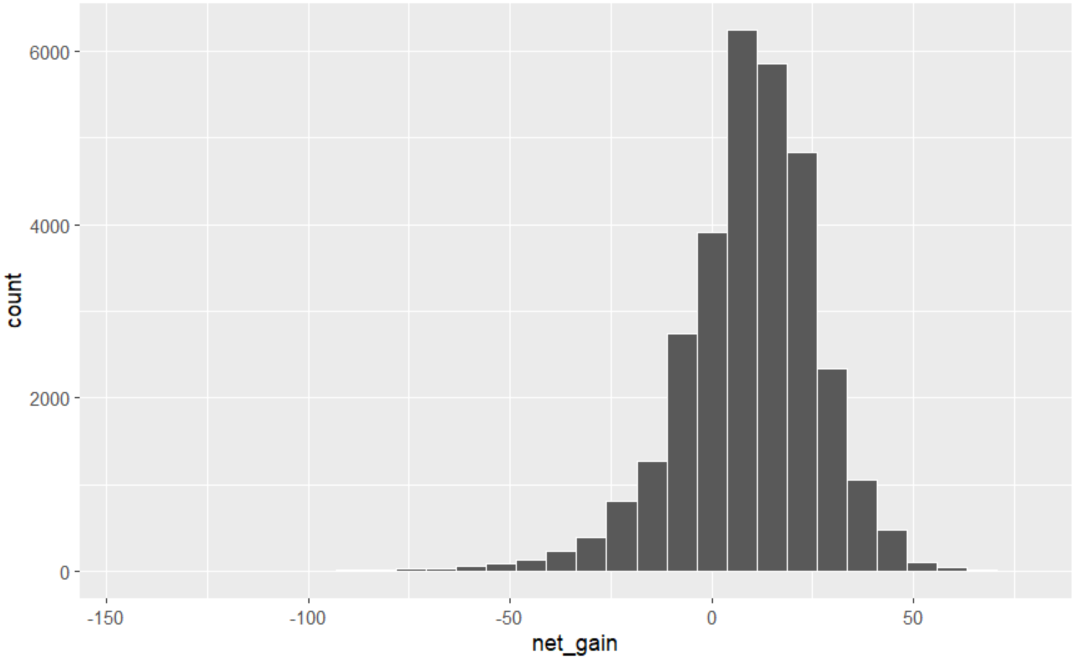
We do two sided t-test for net_gain~late. We get p value as: p-value < 2.2e-16. As p value is < than 0.05, so reject null hypothesis. There is evidence for mean gain of departed late flights different from mean gain of not departed late flights.

->Data Analysis

Histogram for ‘net_gain’ of departed late flights



Histogram for ‘net_gain’ of not departed late flights



Based on shape of above two graphs, average gain may differ for flights that departed late versus those that did not

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

->Test

Here the Null and Alternative hypothesis are:

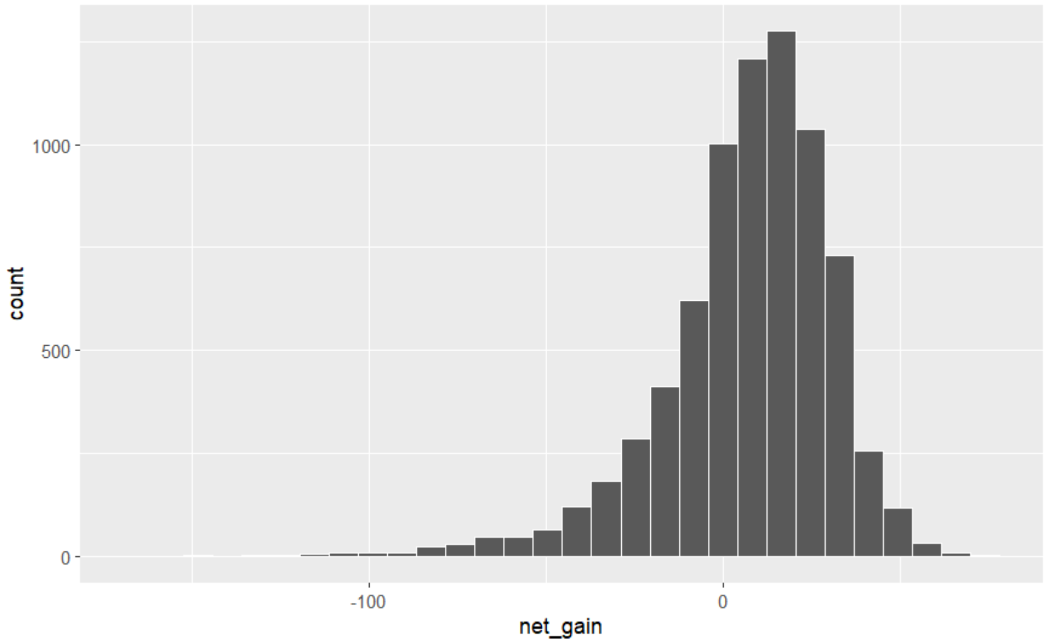
NULL:
meanGain_departedMoreThan30MinsLate=meanGain_didNotdepartedMoreThan30MinsLate

ALTERNATIVE:
meanGain_departedMoreThan30MinsLate!=meanGain_didNotdepartedMoreThan30MinsLate

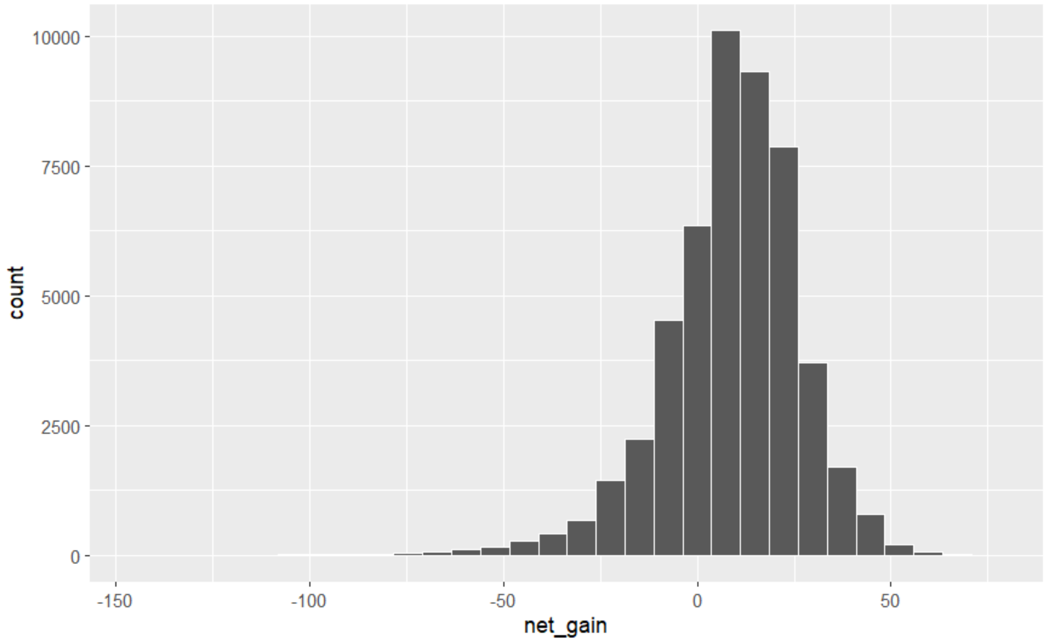
We do two sided t-test for net_gain~very_late. We get p value as: p-value = 3.215e-10
As p value is < than 0.05, so reject null hypothesis. There is evidence for mean gain of flights departed more than 30 minutes late different from mean gain of flights that did not depart more than 30 minutes late

->Data Analysis

Histogram for net_gain of departed very_late flights



Histogram for net_gain of not departed very_late flights



Based on shape of above two graphs, average gain may differ for flights that departed more than 30 minutes late versus those that did not

SECOND POINT TO ADDRESS: 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.

We do a inner join of UAflights and airports based on "dest" = "faa" columns and call the result as UAflights_JoinAirport. We make use of UAflights_JoinAirport for SECOND POINT TO ADDRESS

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

For UAflights_JoinAirport data frame we use group_by(), count, arrange(), head() to retrieve five most common destination airports for United Airlines flights from New York City

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

1)George Bush Intercontinental

2)Chicago Ohare Intl

3)San Francisco Intl

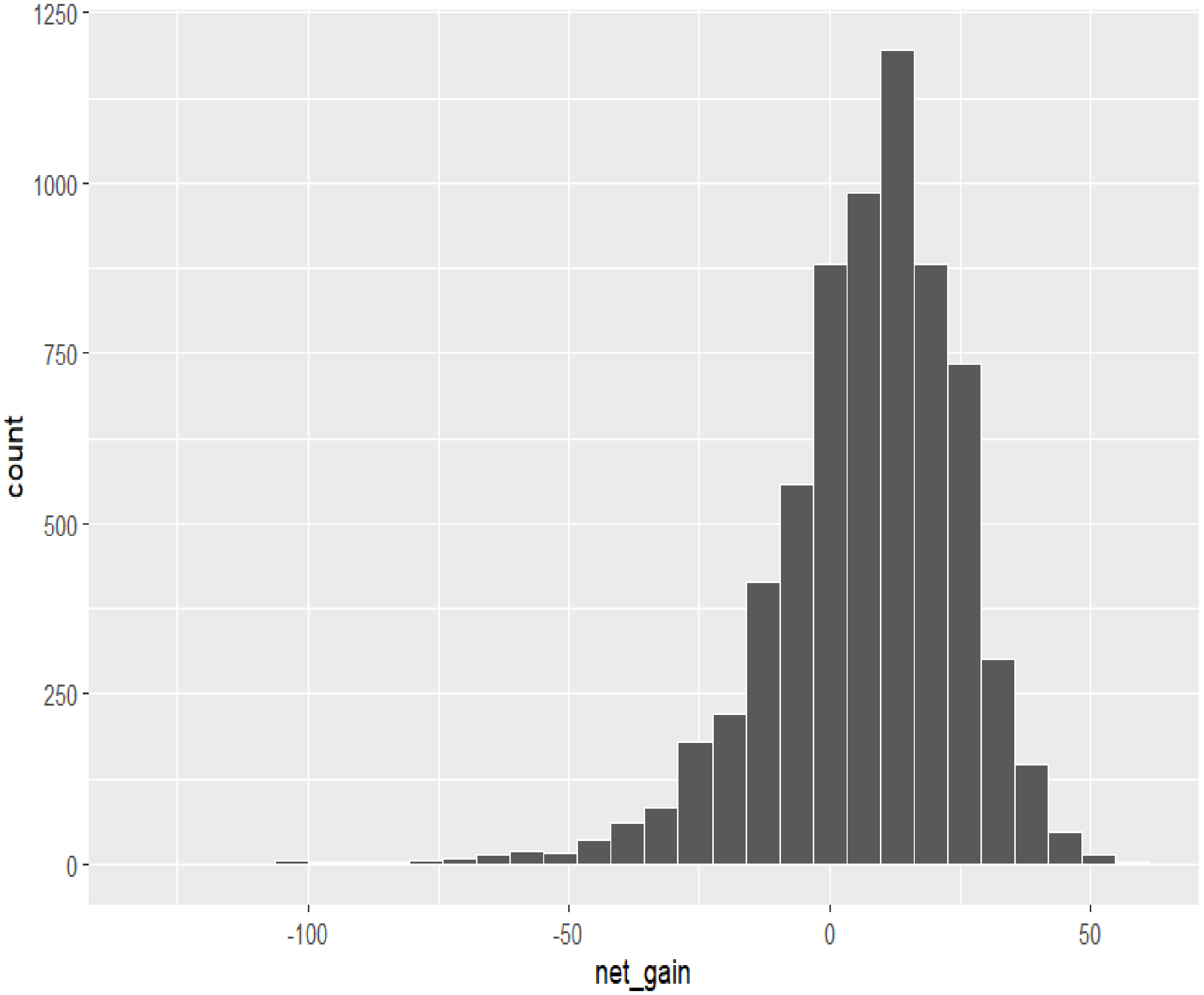
4)Los Angeles Intl

5)Denver Intl

Second Part: Describe the distribution and the average gain for each of these five airports

1)George Bush Intercontinental

-> We do histogram of net_gain of flights to George Bush Intercontinental airport



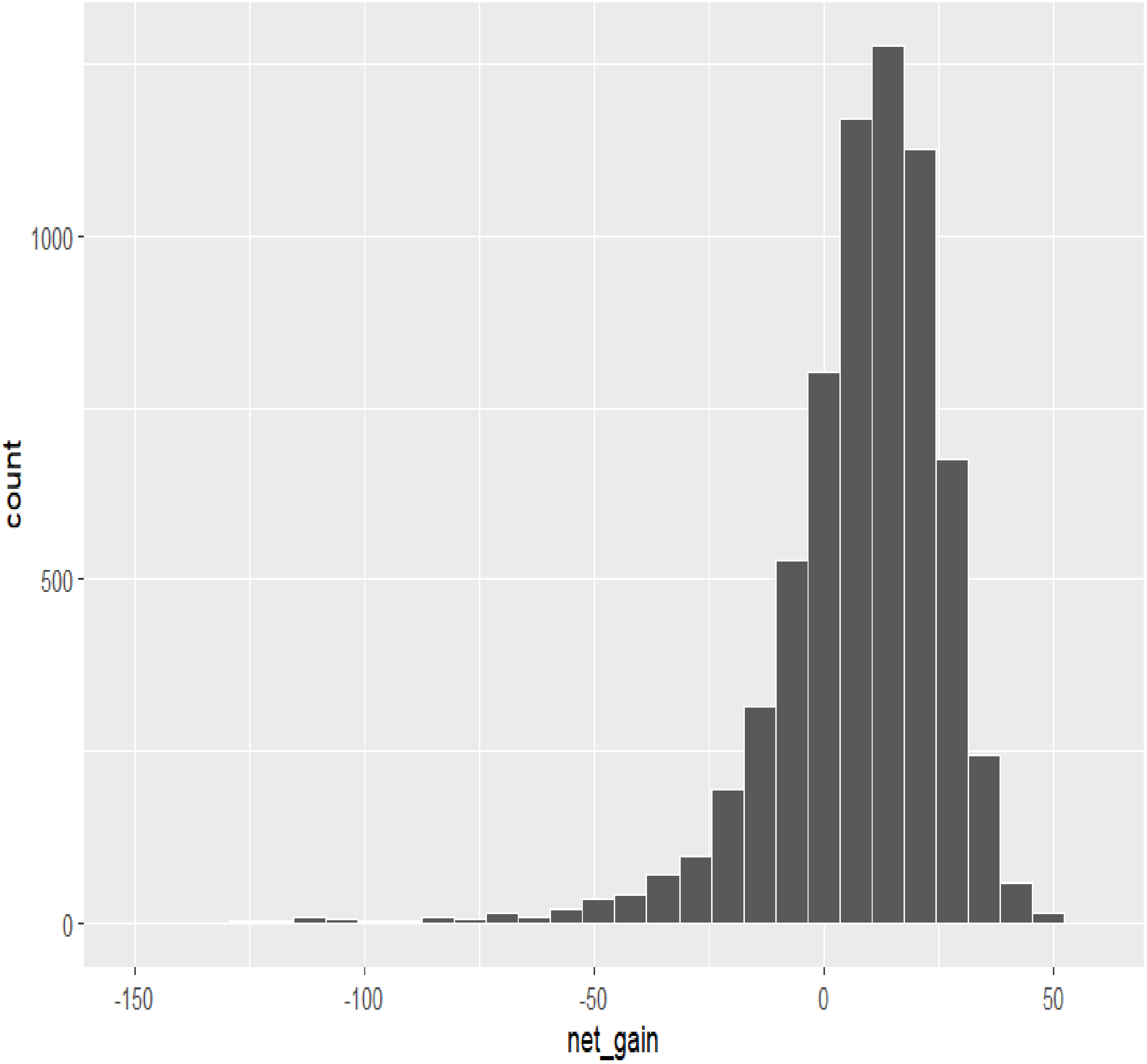
The distribution is left skewed

->Mean of net_gain for flights going to George Bush Intercontinental is 6.861755. The average net_gain is least among list of average net_gain of top five most common destination airports for United Airlines flights from New York City

->We find confidence interval for net_gain for flights going to George Bush Intercontinental. With 95% confidence, the mean net_gain of flights to George Bush Intercontinental airport is between 6.423820 and 7.299691

2)Chicago Ohare Intl

-> We do histogram of net_gain of flights to Chicago Ohare Intl airport



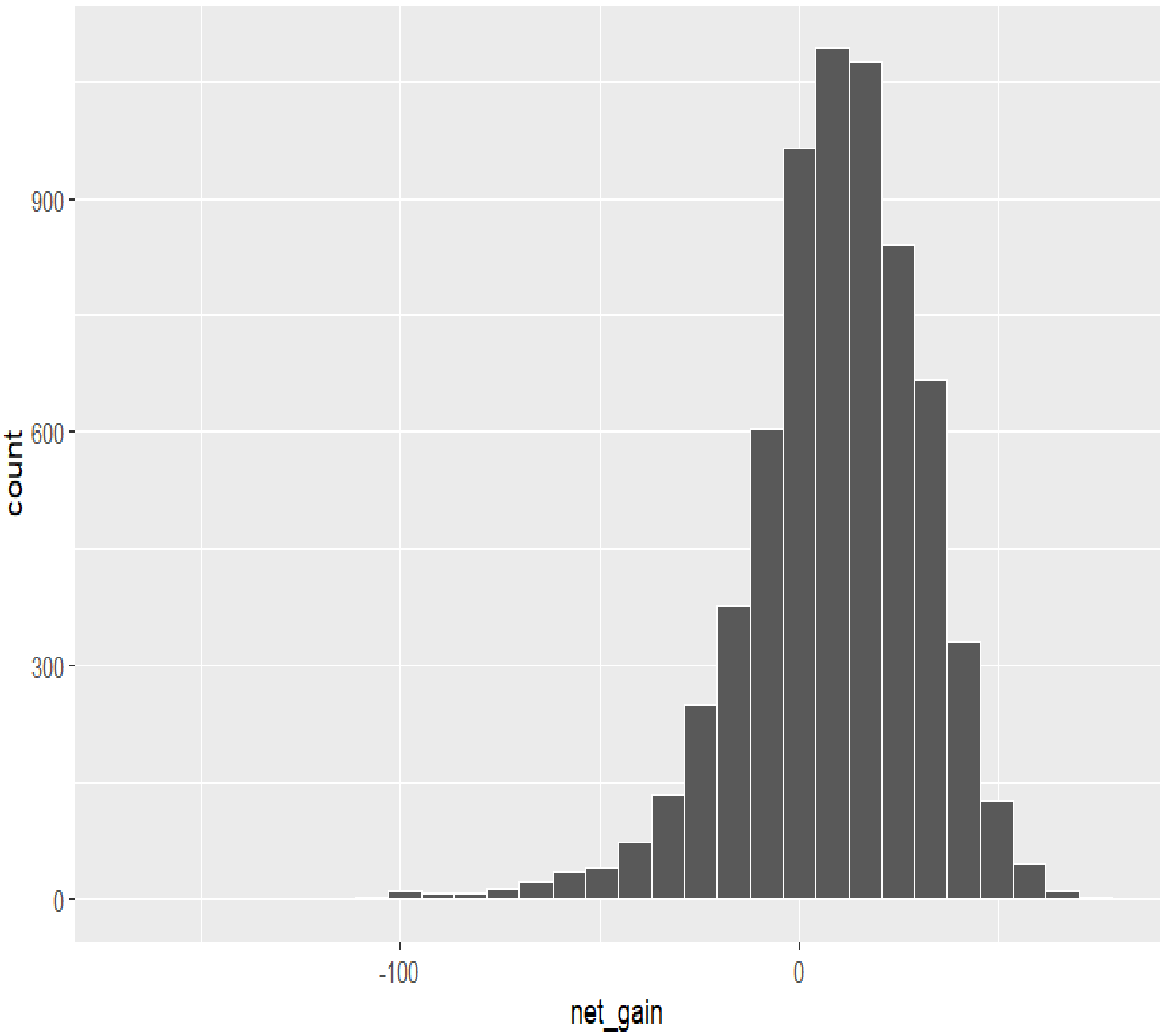
The distribution is left skewed

->Mean of net_gain for flights going to Chicago Ohare Intl is 7.777432. The average net_gain is third highest among list of average net_gain of top five most common destination airports for United Airlines flights from New York City

->We find confidence interval for net_gain for flights going to Chicago Ohare Intl. With 95% confidence, the mean net_gain of flights to Chicago Ohare Intl airport is between 7.320135 and 8.234729

3)San Francisco Intl

-> We do histogram of net_gain of flights to San Francisco Intl



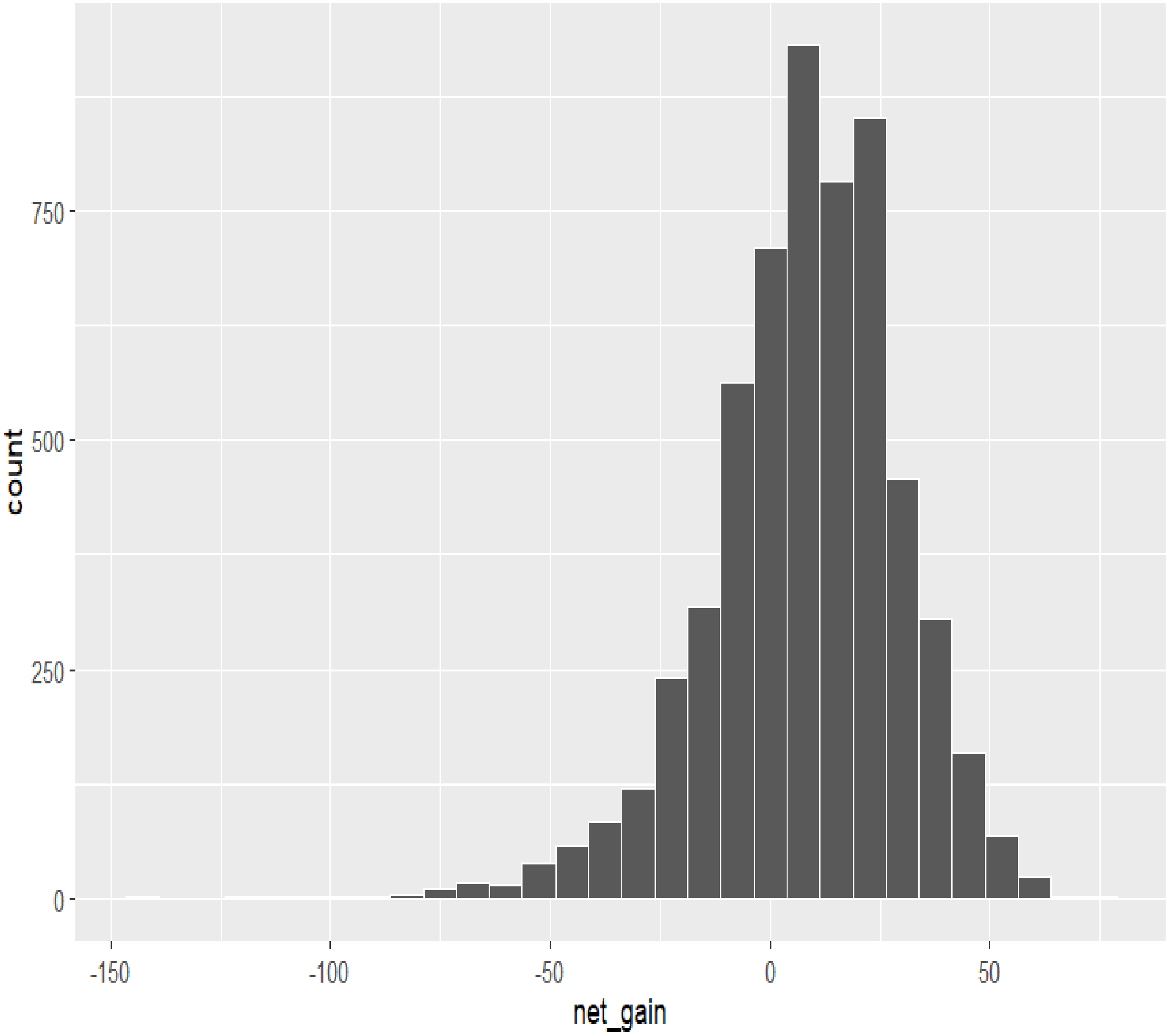
The distribution is left skewed

->Mean of net_gain for flights going to San Francisco Intl is 8.695006. The average net_gain is highest among list of average net_gain of top five most common destination airports for United Airlines flights from New York City

->We find confidence interval for net_gain for flights going to San Francisco Intl. With 95% confidence, the mean net_gain of flights to San Francisco Intl airport is between 8.159475 and 9.230536

4)Los Angeles Intl

-> We do histogram of net_gain of flights to Los Angeles Intl



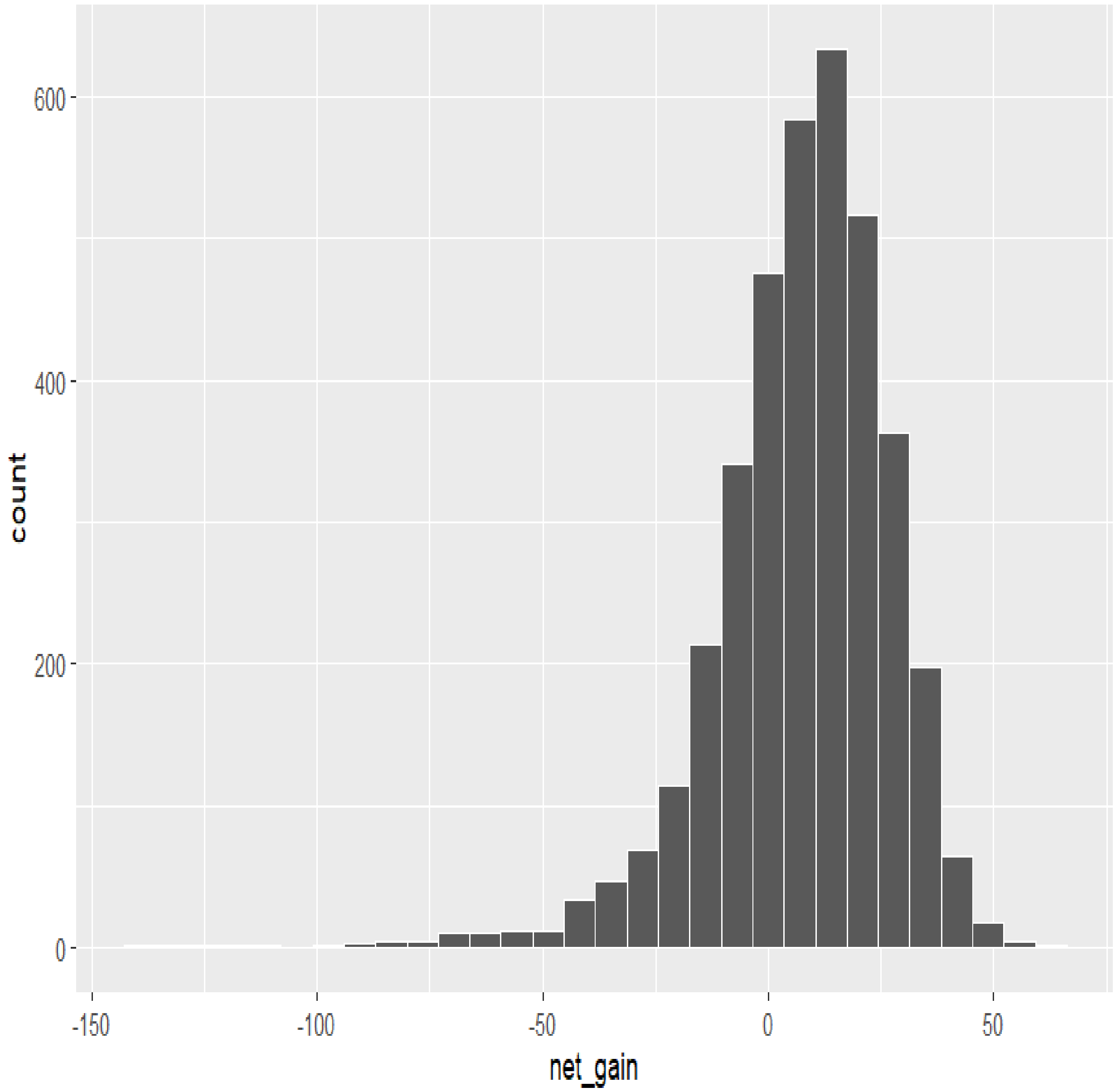
The distribution is left skewed

->Mean of net_gain for flights going to Los Angeles Intl is 7.825303
. The average net_gain is second highest among list of average net_gain of top five most common destination airports for United Airlines flights from New York City

->We find confidence interval for net_gain for flights going to Los Angeles Intl. With 95% confidence, the mean net_gain of flights to Los Angeles Intl airport is between 7.259681 and 8.390925

5)Denver Intl

-> We do histogram of net_gain of flights to Denver Intl



The distribution is left skewed

->Mean of net_gain for flights going to Denver Intl is 7.302382. The average net_gain is fourth highest among list of average net_gain of top five most common destination airports for United Airlines flights from New York City

->We find confidence interval for net_gain for flights going to Denver Intl. With 95% confidence, the mean net_gain of flights to Denver Intl airport is between 6.659348 and 7.945415

THIRD POINT TO ADDRESS: 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?

We make use of UAflights data frame for THIRD POINT TO ADDRESS

First part: 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?

We add a column 'duration_in_hours' by dividing 'air_time' column by 60 to get duration in hours of each flight. We add a column 'gain_per_hour' by dividing columns 'net_gain' by 'duration_in_hours'.

->Test

Here the Null and Alternative hypothesis are:

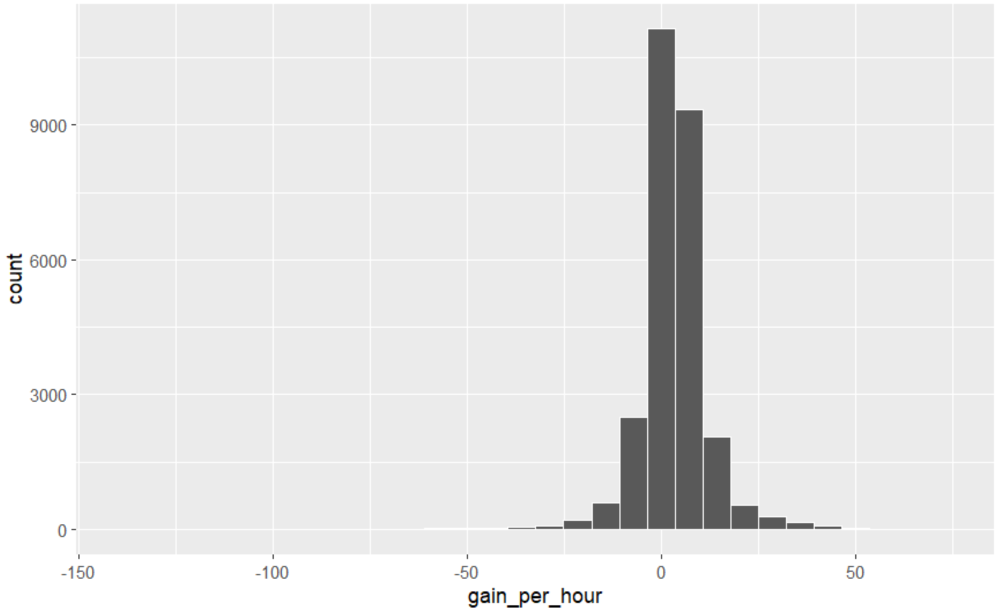
NULL: meanGainPerHour_late=meanGainPerHour_notLate

ALTERNATIVE: meanGainPerHour_late!=meanGainPerHour_notLate

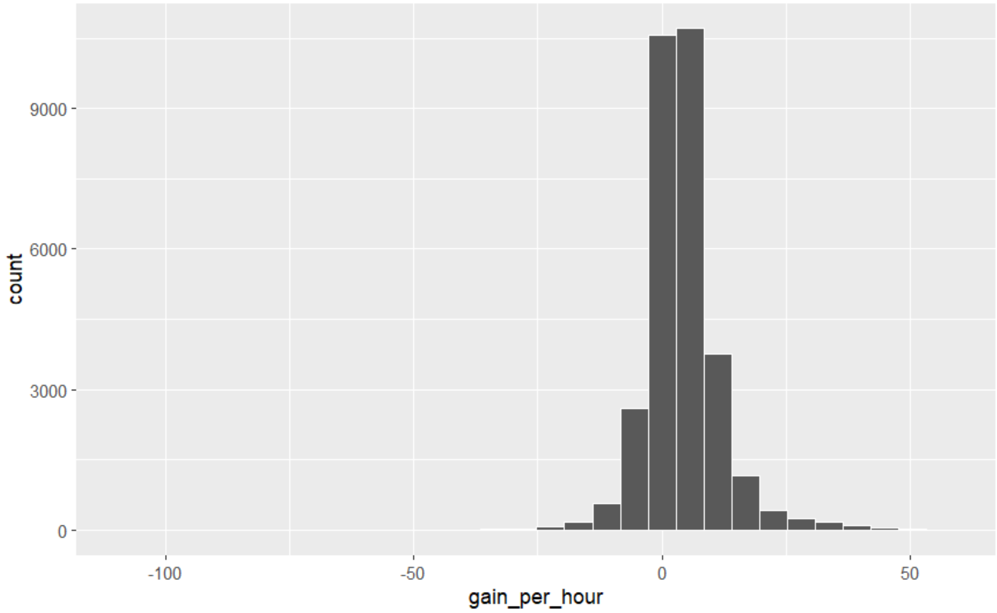
We do two sided t-test for gain_per_hour~late. We get p value as: p-value < 2.2e-16. As p value<0.05, we reject null hypothesis. We have evidence that mean Gain per hour for flights departed late differs from mean gain per hour for flights did not depart late

->Data Analysis

Histogram for gain_per_hour for departed late fights



Histogram for gain_per_hour for not departed late fights



Based on shape of above graphs,the average gain per hour may differ for flights that departed late versus those that did not

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

->Test

Here the Null and Alternative hypothesis are:

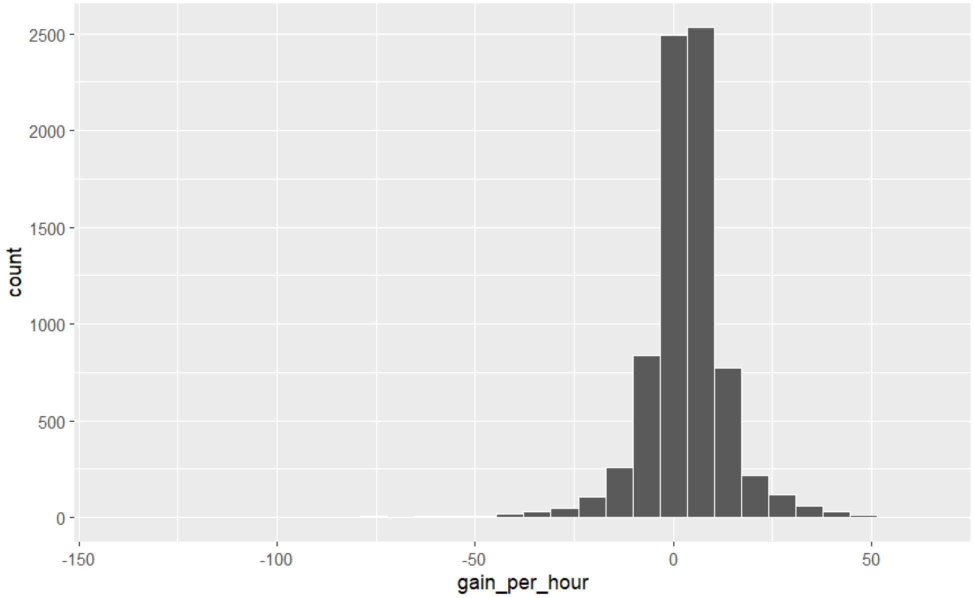
NULL: meanGainPerHour_moreThan30late=meanGainPerHour_notMoreThan30Late

ALTERNATIVE: meanGainPerHour_moreThan30late!=meanGainPerHour_notMoreThan30Late

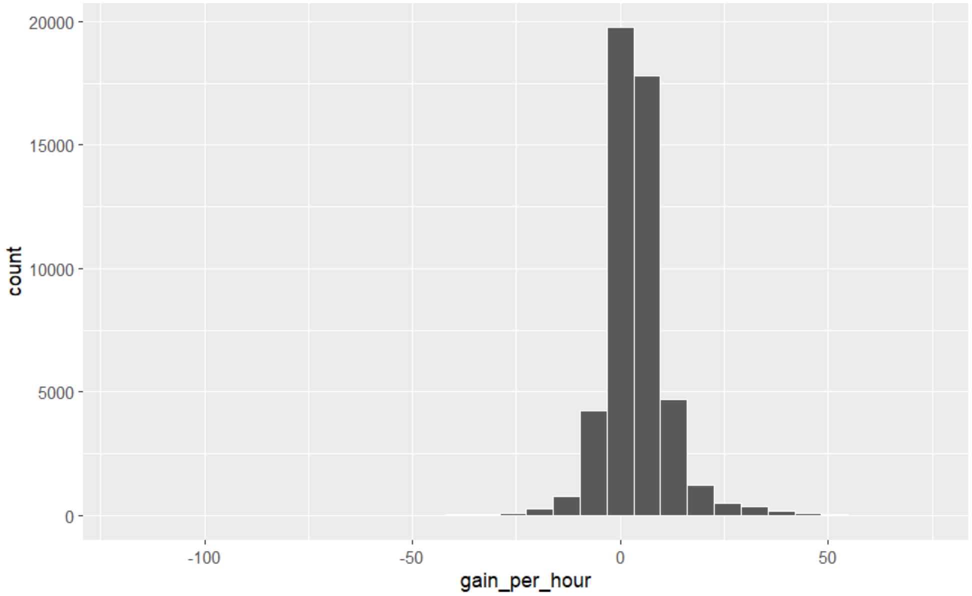
We do two sided t-test for gain_per_hour~very_late. We get p value as: p-value = 1.372e-06. As p value<0.05, we reject null hypothesis. We have evidence that mean gain per hour for flights departed more than 30 mins late is different from mean gain per hour for flights departed not more than 30 mins late

->Data analysis

Histogram for gain_per_hour for flights departed very late



Histogram for gain_per_hour for flights not departed very late



Based on shape of above graphs,the average gain per hour may differ for flights that departed more than 30 minutes late versus those that did not

FOURTH POINT TO ADDRESS: Does the average gain per hour differ for longer flights versus shorter flights?

We make use of UAflights data frame for FOURTH POINT TO ADDRESS

We first filter out UAflights for consisting rows only for " (duration_in_hours<=3) | (duration_in_hours>=6) ", meaning keeping only flights whose duration is short or long, thus removing medium duration flights. We then add a column called 'shorter_flight' in which if 'duration_in_hours<=3' is labeled as 'TR' and 'duration_in_hours>3' is labeled as 'FA'

->Test

Here the Null and Alternative hypothesis are:

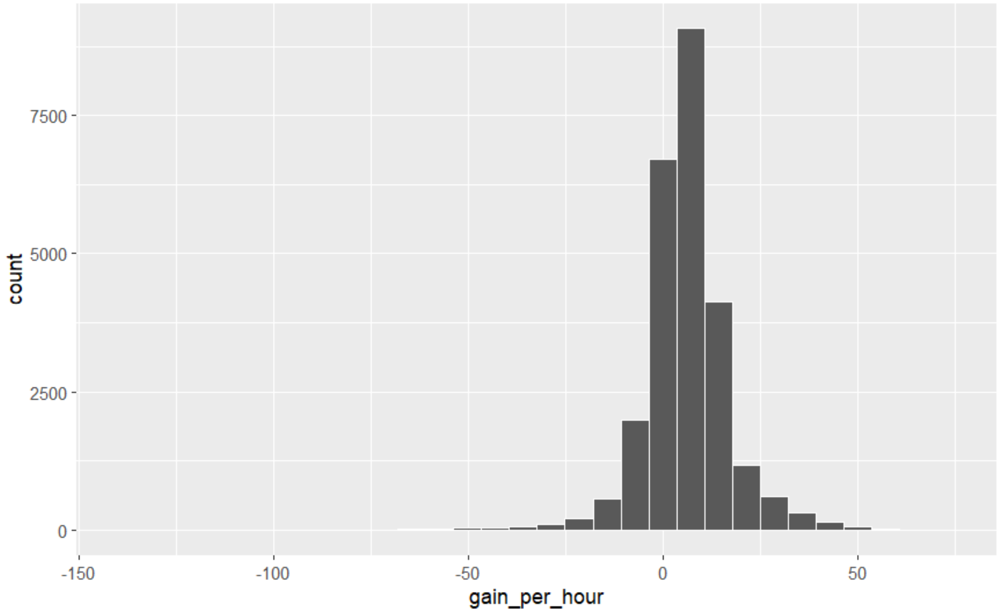
NULL: average gain per hour for longer flights=average gain per hour for shorter flights

ALTERNATIVE: average gain per hour for longer flights!=average gain per hour for shorter flights

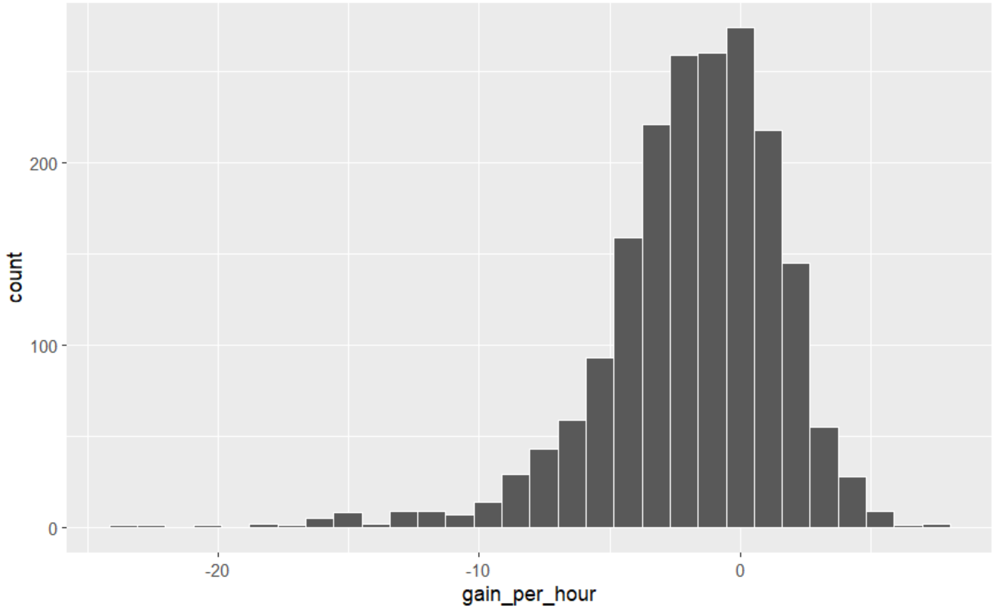
We do two sided t-test for gain_per_hour~shorter_flight. We get p value as: p-value < 2.2e-16. As p value<0.05, we reject null. We have evidence for average gain per hour for longer flights different from average gain per hour for shorter flights

->Data Analysis

We do histogram for gain_per_hour for shorter duration flights



We do histogram for gain_per_hour for longer duration flights



Based on shape of above graphs, average gain per hour may differ for longer flights versus shorter flights

APPENDIX

FIRST POINT TO ADDRESS: 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?

```
library(nycflights13)
library(tidyverse)

UAflights<-flights%>%
  filter(carrier=="UA")

UAflights<-na.omit(UAflights)

UAflights <- UAflights %>%
  mutate(net_gain = dep_delay-arr_delay)

UAflights <- UAflights %>%
  mutate(
    late=ifelse(dep_delay>0, "TR", "FA")
  )

UAflights <- UAflights %>%
  mutate(
    very_late=ifelse(dep_delay>30, "TR", "FA")
  )

t.test(net_gain~late,data=UAflights, alternative = "two.sided")
```

```
latef<-UAflights %>%
  filter(late=="TR")
ggplot(data = latef, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

notLate<-UAflights %>%
  filter(late=="FA")
ggplot(data = notLate, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

t.test(net_gain~very_late,data=UAflights, alternative = "two.sided")

latef<-UAflights %>%
  filter(very_late=="TR")
ggplot(data = latef, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

notLate<-UAflights %>%
  filter(very_late=="FA")
ggplot(data = notLate, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")
```

SECOND POINT TO ADDRESS: 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.

```
UAflights_JoinAirport <- UAflights %>%
  inner_join(airports, by = c("dest" = "faa"))


UAflights_JoinAirport<-na.omit(UAflights_JoinAirport)


mostCommon<-UAflights_JoinAirport%>%
  group_by(name) %>%
  summarize(count = n())


topr<-mostCommon %>%
  arrange(desc(count))


top5<-head(topr,5)
top5


top5$name


george<-UAflights_JoinAirport %>%
  filter(name=="George Bush Intercontinental")
ggplot(data = george, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")


mean(george$net_gain)


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

```
chicago<-UAflights_JoinAirport %>%
  filter(name=="Chicago Ohare Intl")
ggplot(data = chicago, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

mean(chicago$net_gain)

t.test(chicago$net_gain)$conf

san<-UAflights_JoinAirport %>%
  filter(name=="San Francisco Intl")
ggplot(data = san, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

mean(san$net_gain)

t.test(san$net_gain)$conf

los<-UAflights_JoinAirport %>%
  filter(name=="Los Angeles Intl")
ggplot(data = los, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

mean(los$net_gain)

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

```
den<-UAflights_JoinAirport %>%
  filter(name=="Denver Intl")
ggplot(data = den, mapping = aes(x = net_gain)) +
  geom_histogram(color = "white")

mean(den$net_gain)

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

THIRD POINT TO ADDRESS: 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?

```
UAflights <- UAflights %>%
  mutate(duration_in_hours = air_time/60)

UAflights <- UAflights %>%
  mutate(gain_per_hour = net_gain/duration_in_hours)

t.test(gain_per_hour~late,data=UAflights, alternative = "two.sided")

latef<-UAflights %>%
  filter(late=="TR")
ggplot(data = latef, mapping = aes(x = gain_per_hour)) +
  geom_histogram(color = "white")

latef<-UAflights %>%
  filter(late=="FA")
ggplot(data = latef, mapping = aes(x = gain_per_hour)) +
  geom_histogram(color = "white")

t.test(gain_per_hour~very_late,data=UAflights, alternative = "two.sided")

latef<-UAflights %>%
  filter(very_late=="TR")
ggplot(data = latef, mapping = aes(x = gain_per_hour)) +
  geom_histogram(color = "white")
```

```
latef<-UAflights %>%
  filter(very_late=="FA")
ggplot(data = latef, mapping = aes(x = gain_per_hour)) +
  geom_histogram(color = "white")
```

FOURTH POINT TO ADDRESS: Does the average gain per hour differ for longer flights versus shorter flights?

```
max(UAflights$duration_in_hours)
min(UAflights$duration_in_hours)

short_long<-UAflights %>%
  filter((duration_in_hours<=3) | (duration_in_hours>=6))

short_long <- short_long %>%
  mutate(
    shorter_flight=ifelse(duration_in_hours<=3, "TR", "FA")
  )

t.test(gain_per_hour~shorter_flight,data=short_long, alternative = "two.sided")

latef<-short_long %>%
  filter(shorter_flight=="TR")
ggplot(data = latef, mapping = aes(x = gain_per_hour)) +
  geom_histogram(color = "white")

latef<-short_long %>%
  filter(shorter_flight=="FA")
ggplot(data = latef, mapping = aes(x = gain_per_hour)) +
  geom_histogram(color = "white")
```