SEATTLEU

Department of Data Science

DATA 5300 - Applied Stat Infer & Exp Des

United Airlines: NYC Flight Delays

Author:
Akanksha Sharma

Date : 2 December 2022

Table of Contents

List of Figures

List of Tables

1 Introduction

The report will attempt to analyse the flight gains for United Airlines aircraft departing from New York City in 2013. The goal is to see how much quicker the flight ended up being than planned.The study focuses on various characteristics of the flights such as duration of the flight, distance, air time and destination airports.

2 About Dataset

Our analysis will be carried out by leveraging the nycflights13 dataset. This dataset contains the departure timings for all flights departing from New York City's three airports - La Guardia (LGA), John F. Kennedy (JFK), and Newark Liberty International Airport (EWR) in 2013. For the scope of this project, we will be focusing on the data related to United Airlines.

Summary Statistics about the dataset:

Total Number of records for the UA flights : 58,665

Mean gain per flight : -8.54 minutes

In this case study we will be using 6 derived variables based on the following conditions.

Derived Variable	Condition
Gain per flight	Arrival Delay - Departure Delay
Late	True if delay is greater than 0 else False
Very Late	True if delay is greater than 30 else False
Relative Gain	Gain / air time
Flight Short Distance	True if distance is less than 1800 else False

Table 1: Gain, Late,Very Late,Relative Gain,Flight Short : Derived Columns

Before creating derived columns, we found out that there are missing values in the departure delay, arrival delay and air time.

Column Name	Count of Missing value
Arrival Delay	883
Departure Delay	686
Air Time	883

Table 2: Missing Value Count

To handle the missing values we will impute the missing values with the mean of departure delays, arrival delays and air time.

3 Analysis of UA : Gain per flight

Let's examine the distribution of gain per flight for the United Airlines We can observe that the gain per flight is symmetrical, follows a normal distribution, and is centered at -8.548 minutes. This indicates that the arrival delay is often less than the departure delay.

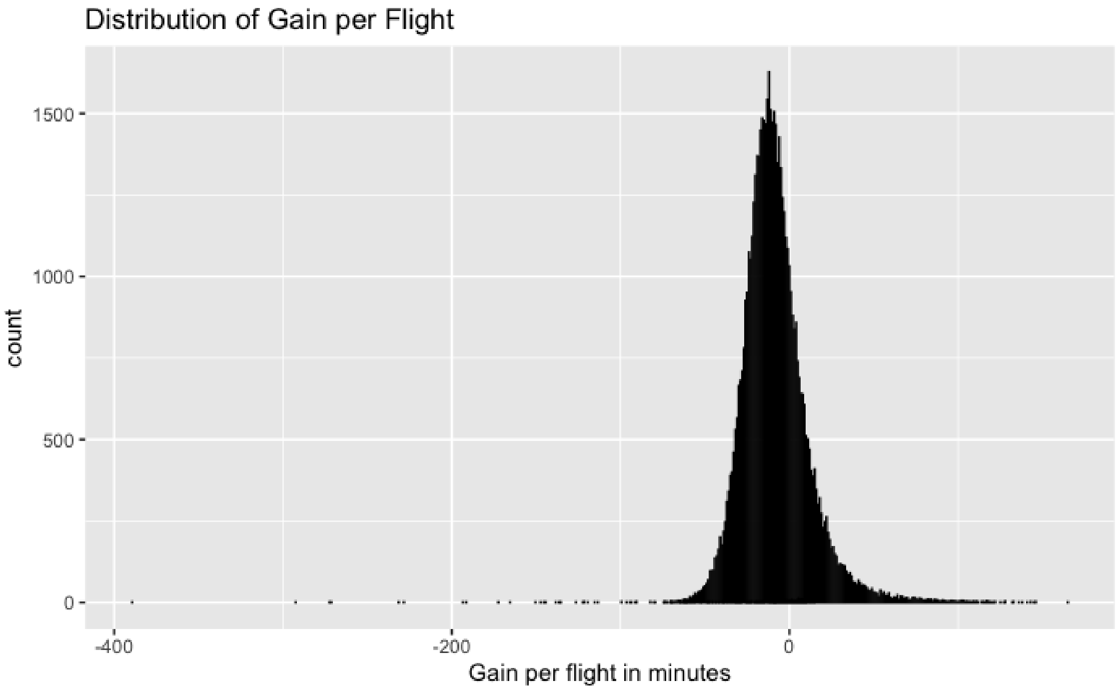


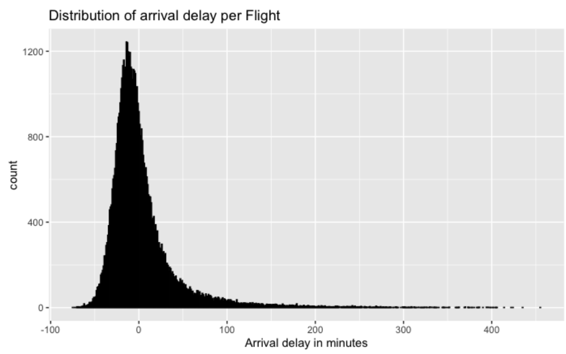
Figure 1: Histogram of Gain in minutes

Since gain depends on arrival and departure delays, it is important to examine how each flight's arrival and departure delays are distributed.

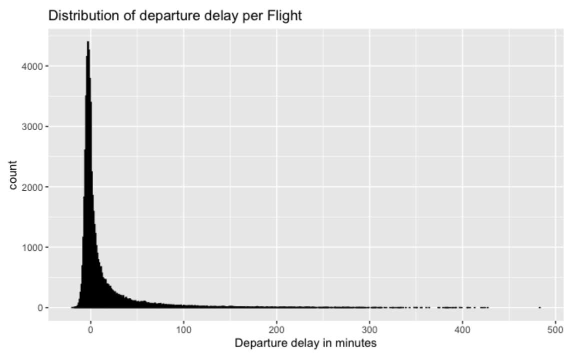
Arrival and Departure delay

We can see that both the arrival and departure delay follows a normal distribution. Both of the arrival delay and departure delay have outliers. But the spread of departure delay is less compare to the arrival delay.

It's interesting to note that the arrival delay is centered at -6.000 minutes whereas departure delay is centered at 0 minutes which means that most of the flights are arriving before the scheduled arrival time. It is safe to say that even though flights are leaving later than scheduled, they are still arriving on time. The major source of flight getting delay is the delay propagation, in which late arrival of an incoming flight leads to late departure and subsequently late arrival of the subsequent outgoing flight. This means that the flight is losing time after each flight trip. We must address this as it may cause more delays.



(a) Histogram of Arrival Delay



(b) Histogram of departure delay

Figure 2: Histogram of Arrival and Departure Delay of UA flights

Let's analyse gain based on late and very late derived variables.

Late	Mean Gain	Median Gain	Standard Deviation Gain	Minimum Gain	Maximum Gain
False	-9.236472	-11	17.28238	-73.000	143
True	-7.791394	-10	21.35567	-389.442	165

Table 3: Late Gain Statistics

Very Late	Mean Gain	Median Gain	Standard Deviation Gain	Minimum Gain	Maximum Gain
False	-8.676587	-10	17.97710	-73.000	143
True	-7.686703	-11	26.74247	-389.442	165

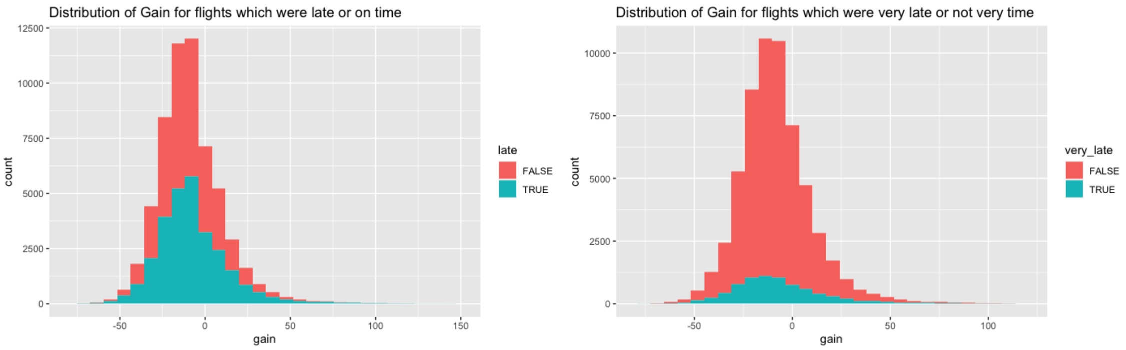
Table 4: Very Late Gain Statistics

Based on the above statistical summary (Table 3) for the flights which were late or on time have almost the same median gain. But it's interesting to note that the spread of gain for the flights which were late is more compared to the flights which were on time.

We can see that there are few outliers in our dataset where the maximum gain and minimum gain is more than 2 hours.

From Table 4, we can conclude that the median gain for the flights which were delayed by more than 30 minutes is almost as same as flights which were not delayed by 30 minutes. There's a huge difference between the spread for the flights which were very late vs which were late by less than 30 minutes.

Let's see the distribution of the gain for the flights which were late or very late.



(a) Late: Histogram of Gain per flight

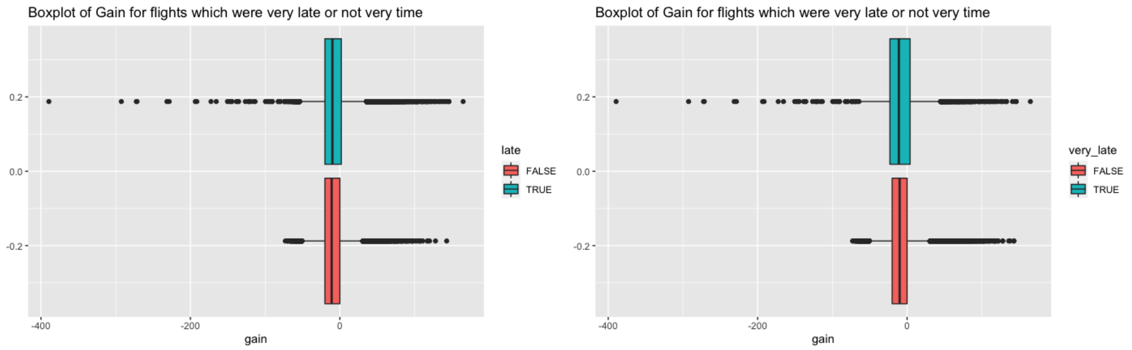
(b) Very Late : Histogram of Gain per flight

Figure 3: Histogram of Gain of flights

Based on the above graph we can say that the gain for each late and very late category is following a normal distribution and they are symmetric in nature.

In Figure 3.a, The distribution of flights which were late or on time are overlapping. Hence, we can say that both of them are similar and there's not much difference between gain of flights which were late or on time.
It goes same for the Figure 3.b where in the distribution of the gain for the flights which were delayed by 30 minutes or less than that are following normal distribution.

We can also see the box plot for both of the very late and late and see how it differs.



(a) Late: Histogram of Gain per flight

(b) Very Late : Histogram of Gain per flight

Figure 4: Histogram of Gain of flights

Even the boxplot for both the Late and Very late is same. Hence, It means that gain for each of these flights is same.

NOTE : We will be using hypothesis testing and confidence interval to determine whether results are significant or not.

Let's do the hypothesis testing for late variable.

H0 : Average gain for late and flight on time is same

average(gain for late) = average(gain for flight on time)

Ha : Average gain for late and flights on time is different

average(gain for late) != average(gain for flight on time)

Statistic	Value
p-value	2.2e-16
95 % Confidence Interval	-1.761377 -1.128780

Table 5: Late : Hypothesis testing p-value and Confidence Interval

We can see that p-value is very small. It means that we can reject the null hypothesis. Which means that there's a evidence that the average gain for late and flights which were on time is different. We can also see that the 95% confidence interval doesn't include 0 in the range which aligns with the p-value.

Let's do the hypothesis test for very late

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same

average(gain for very late flights) = average(gain for flight where delays is less than 30 minutes)

Ha : Average gain for very late and flight which were having delay less than 30 minutes is different

average(gain for very late flights) != average(gain for flight where delays is less than 30 minutes)

Statistic	Value
p-value	0.001773
95 % Confidence Interval	-1.6104501 -0.3693194

Table 6: Late : Hypothesis testing p-value and Confidence Interval

At 5 % significance level, the p-value is less than 0.05 hence it means that we can safely reject our hypothesis. Which means that the average gain of very late and flight which were not late is different and there's a evidence that the average difference between the gains can be different.We can also see that the 95% confidence interval doesnt include 0 in the range which aligns with the p-value.

Note: It's sometimes useful to recognize how much of a difference actually matters in the real world. The average difference between the delays reveals that it is off by a small amount. In the actual world, that figure is not important. It indicates that even if we have evidence that there is a difference between the average gains in the statistical hypothesis testing, we can fairly assume that the average difference is the same.

As mentioned earlier, we do have outliers in the dataset. Outliers are always a concern and should be inspected: do they represent an extreme value from the population or a recording mistake? Conduct the test with and without the outlier to determine if the outlier is influential. If the conclusion changes, then you should report both outcomes: it is not acceptable to report the results without the outlier unless there is a clear identifiable reason why that observation does not belong in the sample.

Let's identify the outliers in the dataset and remove them and then again do the hypothesis testing and see if there's any change in the results.

We can identify outliers based on this condition Outliers = Observations with z-scores greater than 3 or less than -3

Number of outliers in the dataset based on the condition : 735 records

Hypothesis for the flights gain based on the Late variable:

H0 : Average gain for late and flight on time is same

average(gain for late) = average(gain for flight on time)

Ha : Average gain for late and flights on time is different

average(gain for late) != average(gain for flight on time)

The p-value is smaller than 0.05 which means that we can reject our Null hypothesis. Hence, we can conclude that there's a evidence that there's difference between the gain of flights which were late or on time.

Statistic without Outlier	Value
p-value	1.215e-08
95 % Confidence Interval	-1.0781134 -0.5262771

Table 7: Late Without Outliers: Hypothesis testing p-value and Confidence Interval

Hypothesis for the very late derived variable :

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same

average(gain for very late flights) = average(gain for flight where delays is less than 30 minutes)

Ha : Average gain for very late and flight which were having delay less than 30 minutes is different

average(gain for very late flights) != average(gain for flight where delays is less than 30 minutes)

Interestingly for the very late variable the p-value is greater than 0.05 which means that there's a evidence that there's no difference between the average gain for the flights which were very late or were not very late.

Let's try to do the bootstrap t test to compare the means of relative gain for the late and very

Statistic without Outlier	Value
p-value	0.8648
95 % Confidence Interval	-0.5197887 0.4366856

Table 8: Very Late Without Outliers: Hypothesis testing p-value and Confidence Interval

late flights.

We will use the same hypothesis as earlier for the late and very late variable but this time with the bootstrapped sample.

H0 : Average gain for late and flight on time is same

average(gain for late) = average(gain for flight on time)

Ha : Average gain for late and flights on time is different

average(gain for late) != average(gain for flight on time)

The p-value for the bootstrapped hypothesis testing yields the p-value as 2e-05 which is smaller than 0.05. It means that we can reject our null hypothesis.There's a evidence that the average gain for the flights on time is different for the flights which were late vs on time. This is the same result which we got from with the t.test.

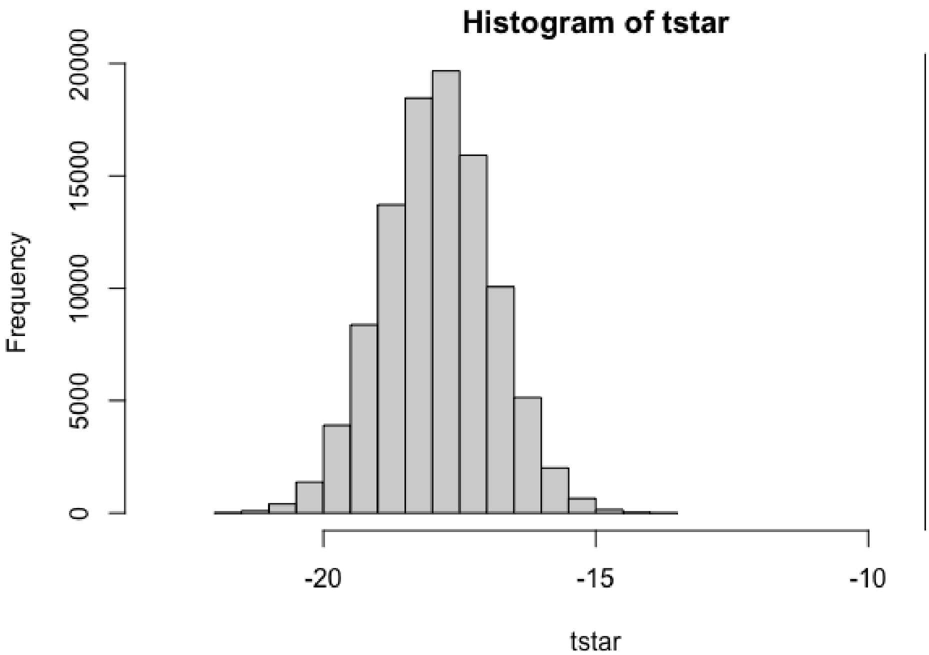


Figure 5: Bootstrap : Hypothesis Testing

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same

average(gain for very late flights) = average(gain for flight where delays is less than 30 minutes)

Ha : Average gain for very late and flight which were having delay less than 30 minutes is different

average(gain for very late flights) != average(gain for flight where delays is less than 30 minutes)

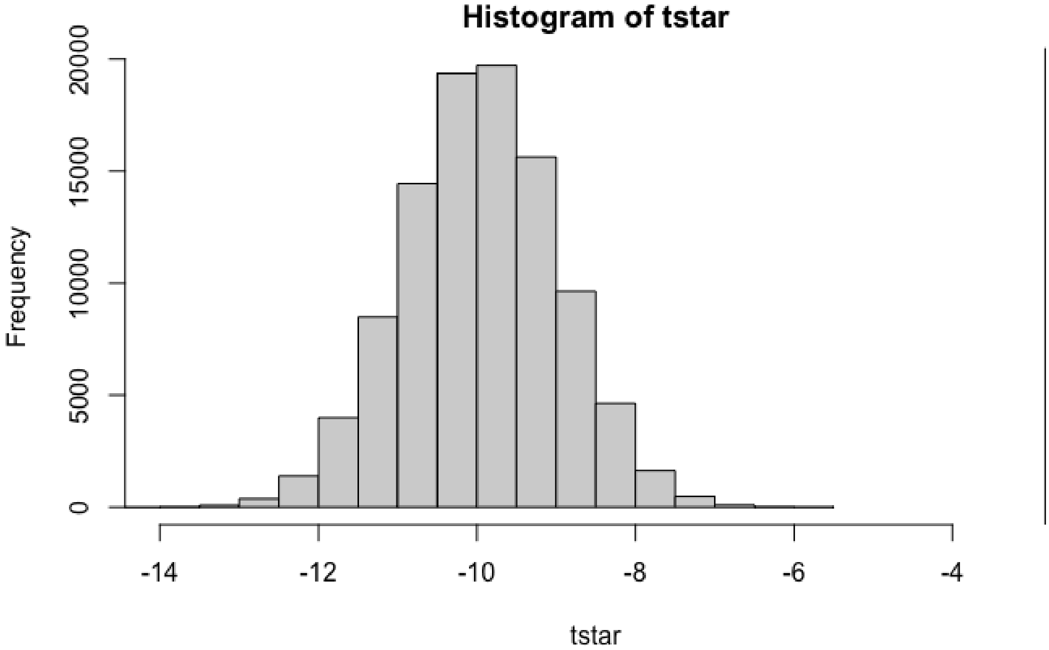


Figure 6: Bootstrap : Hypothesis Testing

Even for the very late variable the p-value is 0.0018 which is less than the 0.05. Hence, we can reject our null hypothesis and we have a evidence that the mean gain for very late and the flights which were having delay less than 30 minutes is different.

4 Average gain for the airports

We know that average gain is heavily dependent on the destination airport. Let's take a look at the most popular destination airports for flights departing from New York City in 2013.

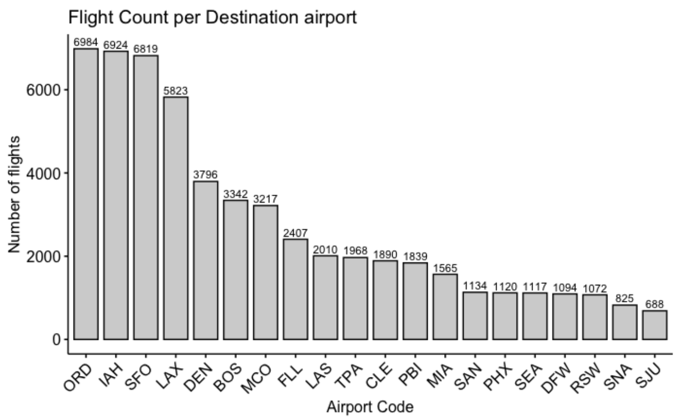


Figure 7: Departure delay based on hour

According to the graph above, these five airports are the most frequently visited by UA planes.:

Let's try to find the average gain , median gain, standard deviation for each of the top 5 destined

Airport Code	Count
ORD O'Hare International Airport	6984
IAH George Bush Intercontinental Airport	6924
SFO San Francisco International Airport	6819
LAX Los Angeles International Airport	5823
DEN Denver International Airport	3796

Table 9: Top 5 Destination Airports

airports.Based on the table 10, we can see that the median gain or mean gain for all the top 5 airports is negative which means that most of the time departure delay is more than the arrival delay.We can also see that standard deviation for the SFO airport is very high.

Destination Airport	Mean gain	Median gain	Standard Deviation gain
DEN	-7.437359	-9	20.67609
IAH	-6.946199	-9	18.63452
LAX	-7.841434	-9	21.86556
ORD	-7.936062	-10	19.44381
SFO	-8.873921	-10	23.21225

Table 10: Top 5 Destination Airports Gain Statistic

late	Destination Airport	Mean gain	Median gain	StandardDeviation gain	MinGain	MaxGain
FALSE	DEN	-8.771821	-10	17.65013	-56	94
TRUE	DEN	-6.036606	-8.548062	23.35886	-271.44199	136
FALSE	IAH	-7.40963	-9	17.89738	-59	117
TRUE	IAH	-6.421396	-8.548062	19.42498	-228.44199	128
FALSE	LAX	-8.495085	-9	20.1015	-73	118
TRUE	LAX	-7.136694	-9	23.60374	-90.44199	145
FALSE	ORD	-9.024221	-11	16.63262	-57	128
TRUE	ORD	-6.557586	-8.548062	22.43362	-272.44199	146
FALSE	SFO	-9.264197	-10	20.23446	-70	143
TRUE	SFO	-8.43072	-10	26.18268	-389.44199	165

Table 11: Top 5 Destination Airports Gain Statistic based on Late

very_late	Destination Airport	Mean gain	Median gain	StandardDeviation gain	MinGain	MaxGain
FALSE	DEN	-7.648294	-9	18.60662	-66	136
TRUE	DEN	-5.98707	-11	31.39922	-271.44199	133
FALSE	IAH	-7.11204	-9	17.68014	-59	117
TRUE	IAH	-5.626554	-9	24.93649	-228.44199	128
FALSE	LAX	-8.174897	-9	20.72673	-73	118
TRUE	LAX	-5.2239	-9	29.19492	-90.44199	145
FALSE	ORD	-8.285262	-10	17.53602	-57	138
TRUE	ORD	-5.940252	-10	27.87973	-272.44199	146
FALSE	SFO	-9.197185	-10	21.02795	-70	143
TRUE	SFO	-6.692259	-10	34.44927	-389.44199	165

Table 12: Top 5 Destination Airports Gain Statistic based on Very Late

ORD O'Hare International Airport :

Let's find the distribution of the gain per flight for the ORD airport. We can see that the gain per flight for the ORD airport is following a normal distribution and centered around -7.437359.

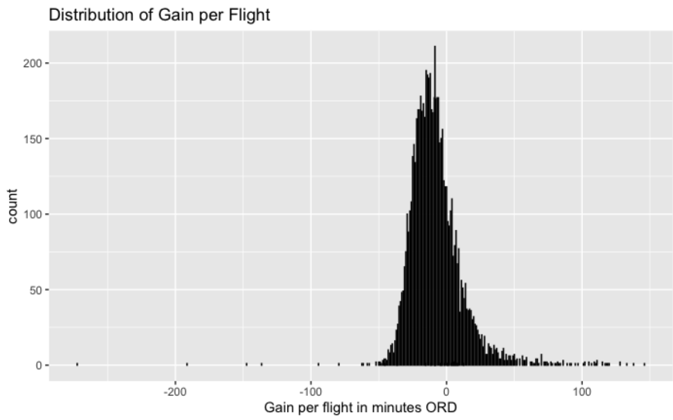


Figure 8: ORD : Distribution of Gain per flight

In figure 7.A, we can see that the distribution of gain for the flights which were on time or delayed is following a normal distribution. Both of the distribution are overlapping with each other.

In figure 7.B, we tried to plot the distribution based on the flights which were delayed for more than 30 minutes vs the flights which were not delayed for more than 30 minutes. Even the gain for both of these is following a normal distribution and overlapping with each other.

Let's try to plot the boxplot for both of the variables.

In figure 8.A, The boxplots look similar for the gain for the flights which are on time or delayed.But there's a difference between the boxplots of the flight based on very late variable. The flights which were not delayed by 30 minutes have less gain.

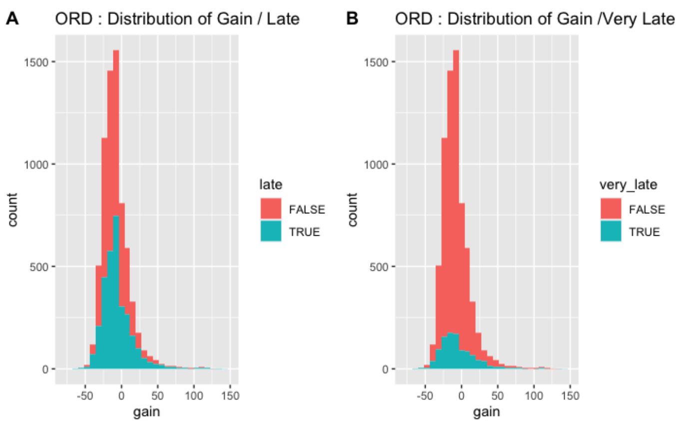


Figure 9: ORD : Distribution of Gain per flight based on Late and Very late

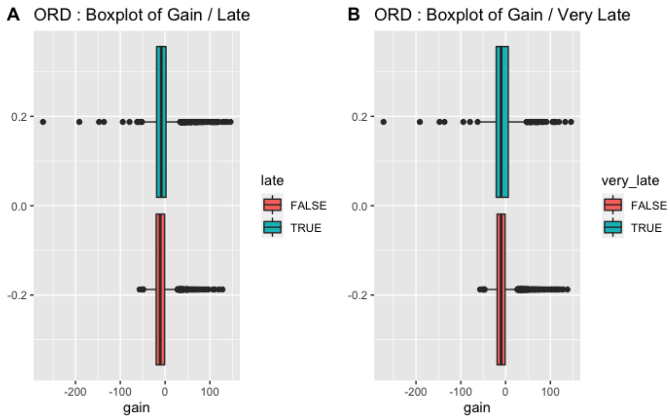


Figure 10: ORD : Boxplot of Gain per flight based on Late and Very Late

Based on the graphs, here is the conclusion for the remaining airports:

1. Distribution of the gains for the IAH, SFO, LAX and DEN : We can see that the gain per flight is following a normal distribution and symmetric in nature.

It's interesting question to answer that why the gains for all the airports is following a normal distribution. One reason is the central limit theorem: When an outcome is produced by many independent effects that act additively, the result will be normally distributed. Hence, we see that the gains are following a normal distribution.

2. Even we if try to see the distribution of gains for the IAH, SFO,LAX and DEN based on Late and very late variable they do follow a normal distribution and identical in nature and overlapping. Hence, it's difficult to segregate it.

2. The boxplots for the late and very late variable for the gains is also identical.

4. We can see the five point summary for the flights which were late and very late in table 11 and 12 for each of the airport.

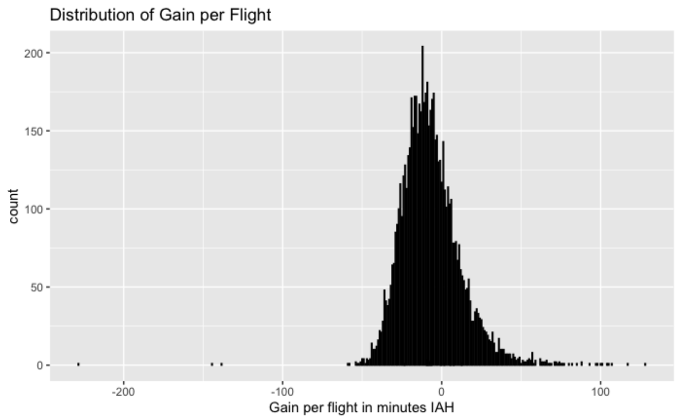


Figure 11: IAH : Distribution of Gain per flight

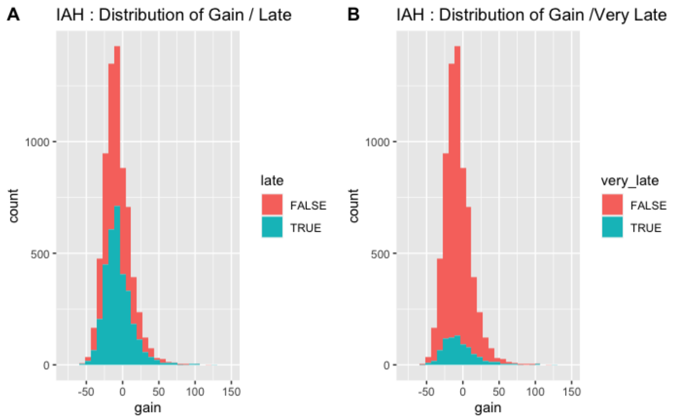


Figure 12: IAH : Distribution of Gain per flight based on Late

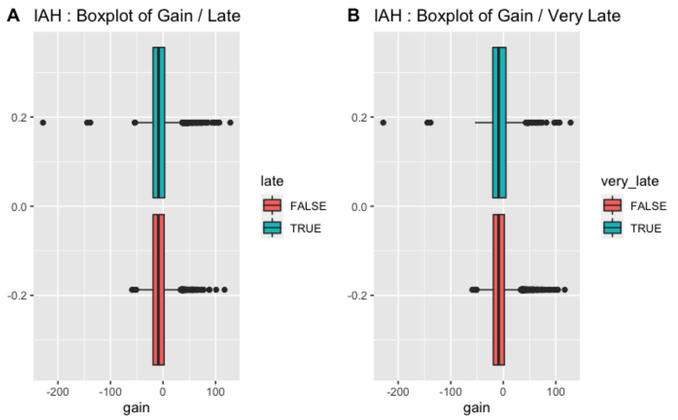


Figure 13: IAH : Boxplot of Gain per flight based on Very Late

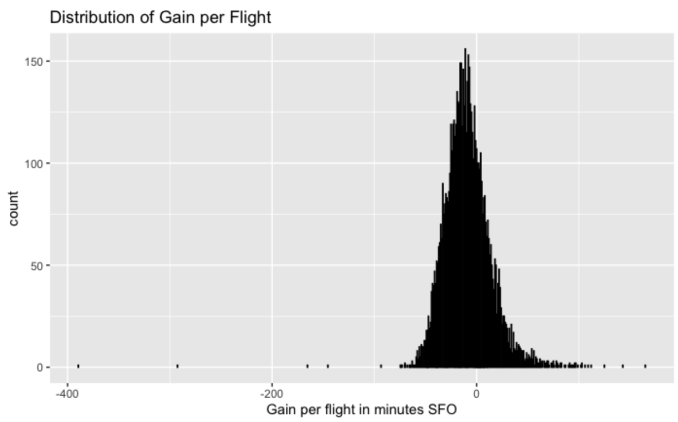


Figure 14: SFO : Distribution of Gain per flight

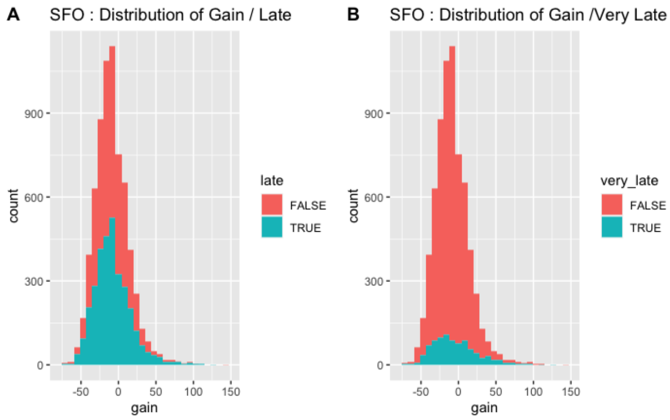


Figure 15: SFO : Distribution of Gain per flight based on Late

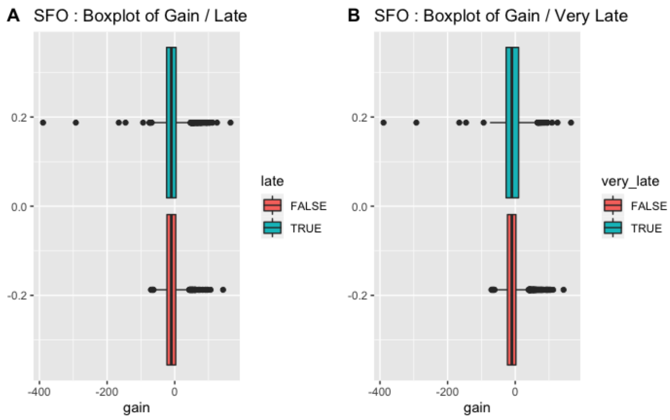


Figure 16: SFO : Boxplot of Gain per flight based on Very Late

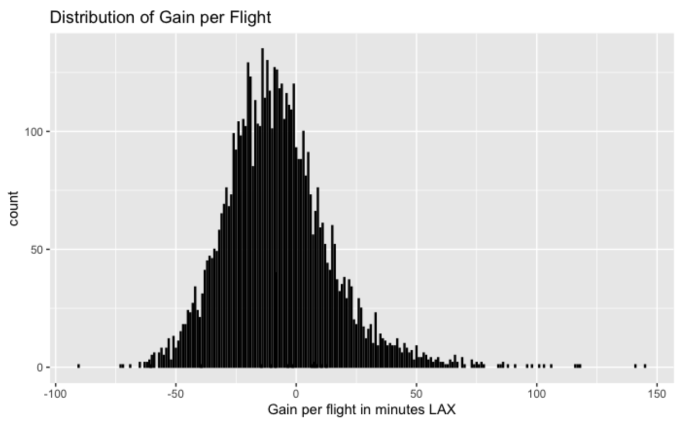


Figure 17: LAX : Distribution of Gain per flight

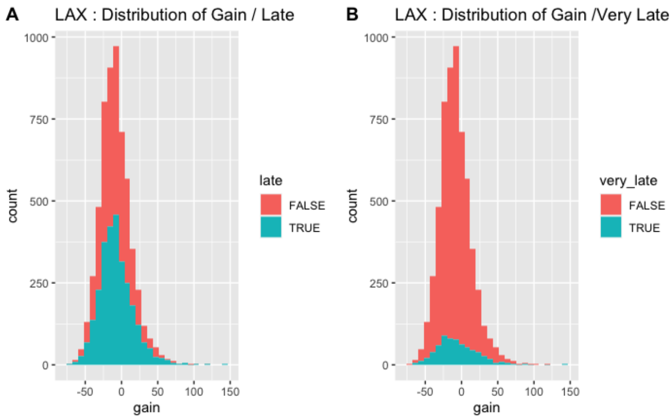


Figure 18: LAX : Distribution of Gain per flight based on Late

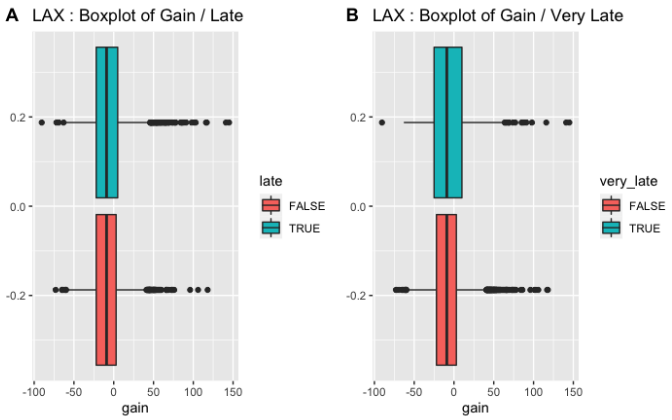


Figure 19: LAX : Boxplot of Gain per flight based on Very Late

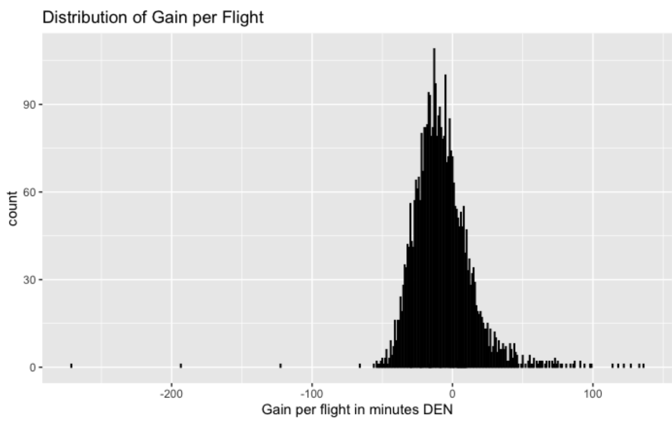


Figure 20: DEN : Distribution of Gain per flight

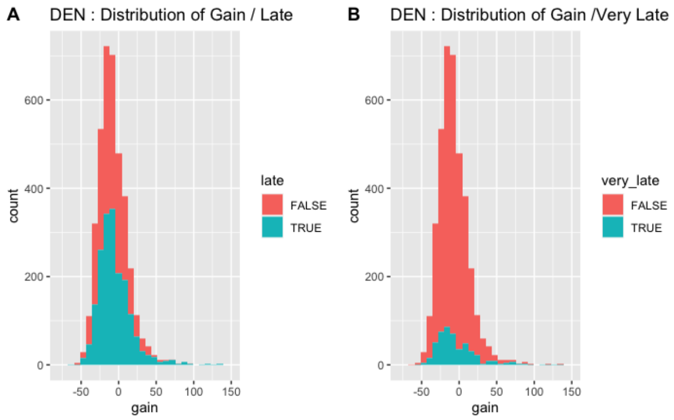


Figure 21: DEN : Distribution of Gain per flight based on Late

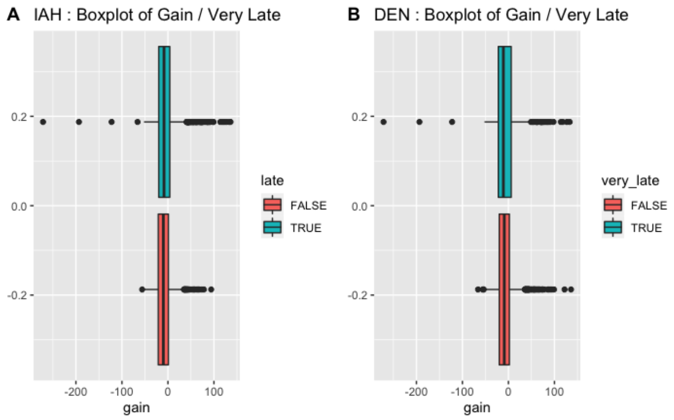


Figure 22: DEN : Boxplot of Gain per flight based on Very Late

Let's do the hypothesis testing for ORD, IAH, SFO, LAX and DEN airport. We will conduct the test with and without the outlier to determine if the outlier is influential. If the conclusion changes, then you will report both outcomes: it is not acceptable to report the results without the outlier unless there is a clear identifiable reason why that observation does not belong in the sample.

Hypothesis testing based on the the flights which are late or on time.

H0 : Average gain for late and flight on time is same

average(gain for late) = average(gain for flight on time)

Ha : Average gain for late and flights on time is not same

average(gain for late) != average(gain for flight on time)

Airport	p-value(outlier)	Confidence Interval(outlier)	p-value(no outliers)	Confidence Interval(no outliers)
ORD	3.572E-07	-3.415404 -1.517867	0.0001379	-2.219516 - 0.71259
IAH	0.02844	-1.8721676 -0.1043001	0.02637	-1.679892 -0.1048095
SFO	0.1454	-1.9555410 0.2885874	0.6636	-0.7400808 1.1622203
LAX	0.01854	-2.4889607 -0.2278215	0.4768	-1.3888319 0.6491234
DEN	5.11E-05	-4.05756 -1.41287	0.006759	-2.6206117 -0.4204751

Table 13: Hypothesis results for gain based on Late variable

The p-value for the LAX airport changes if we remove the outlier from the dataset. If we are considering outliers, then we have a evidence to reject the null hypothesis because p-value is less than 0.05. But if we remove outliers from the dataset then the p-value is greater than 0.05 which is 0.4768. We can conclude that without outliers we have a evidence that the average gain for the flights which are on time or delayed is same. Hence, we can accept the null hypothesis. It means that in this case the outliers play a major role.

We can deduce the same thing using confidence interval for the LAX airport. With outliers the 95 % confidence interval doesn't include 0 but without outliers we can see that the 0 lies in the range.

It's interesting to note that the p-value for the SFO airport is more than 0.05 for the dataset which include/exclude the outliers. That is, we have proof that the null hypothesis is correct. As a result, there is a chance that the average gain for the SFO airport is the same for flights that were on time or late. The confidence interval in this situation comprises zero, which corresponds to the p-value conclusion.

We can observe that the p-value for ORD, IAH, and DEN airports is less than 0.05 for data with or without outliers. It means that we can reject our Null hypothesis.As a result, we may conclude that there is a difference in the gain of planes that were late or on time for the ORD, IAH, and DEN airports.For these three airports, the confidence interval doesn't contain 0 and based on the CI we can also conclude that there might be a difference between the average gain for flights which were late or on time.

Let's do the hypothesis test for very late variable for ORD, IAH ,SFO ,LAX and DEN airport with and without outliers.

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same

average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues)

Ha : Average gain for very late and flight which were having delay less than 30 minutes is different

average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

Airport	p-value(outlier)	Confidence Interval(outlier)	p-value(no outliers)	Confidence Interval(no outliers)
ORD	0.008821	-4.0988779 -0.5911423	0.05975	-2.48130035 0.04998776
IAH	0.1086	-3.3005792 0.3296066	0.1565	-2.4466507 0.3940638
SFO	0.036	-4.8458829 -0.1639685	0.189	-2.8380858 0.5611672
LAX	0.01217	-5.2559574 -0.6460371	0.7529	-2.221192 1.607126
DEN	0.2577	-4.541608 1.219161	0.6096	-1.383834 2.356828

Table 14: Hypothesis results for gain based on Very Late variable

Even for the very late variable we will do hypothesis testing for the late and very late variable.

It's interesting to note that for the very late variable the p-value for the ORD , SFO, LAX changes if we are doing hypothesis testing with outlier or without outliers.

The average gain per flight with outliers for ORD, SFO, and LAX airports is less than 0.05 at the 5% significance level. It indicates that we can confidently reject our null hypothesis and have evidence that the average gain per flight differs for flights that are delayed by more than 30 minutes vs flights that are delayed by less than 30 minutes.

However, removing the outlier from the sample reveals that the p-value is more than 0.05. As a result, we have evidence that the null hypothesis can be true. As a result, we have proof showing the average gain per flight is the same whether the flight was delayed by 30 minutes or not.

The p-value with or without outliers for the IAH and DEN is larger than 0.05 at the 5% significance level. This suggests we have evidence that the null hypothesis is possible. As a result, we have evidence that the average gain per flight is the same for planes that were delayed by 30 minutes or less.

5 Analysis of Relative gain to the duration of the flight

Let's analyse the relative gain to the duration of the flight.

The average gain per flight follows a normal distribution. We can see that average gain mean is centered at 0.

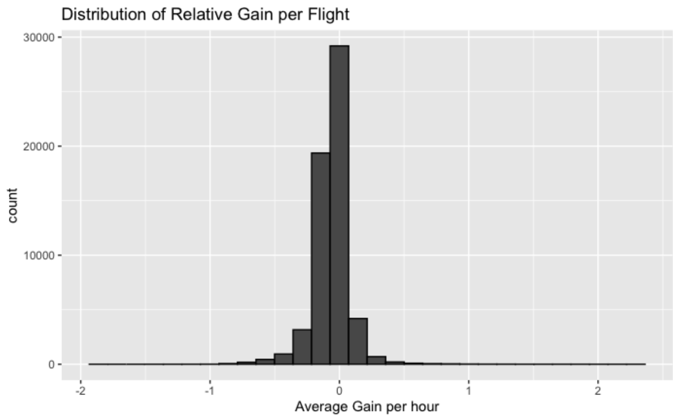


Figure 23: Distribution of Relative gain to the duration of the flight

We also attempted to determine whether there is a difference in relative flight gain depending on the late and extremely late variables. We can observe that they both have a normal distribution and overlap with one another.

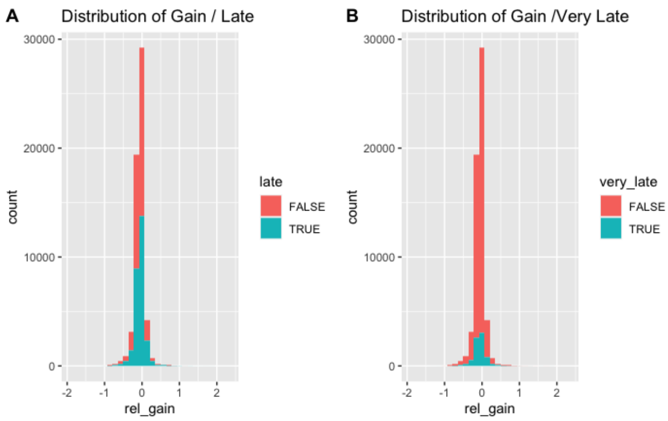


Figure 24: Distribution of Relative Gain per flight based on late and very late variable

Even the five point summary for both the very late and late variables is same. Let's do hypothesis testing and see if there's any difference.

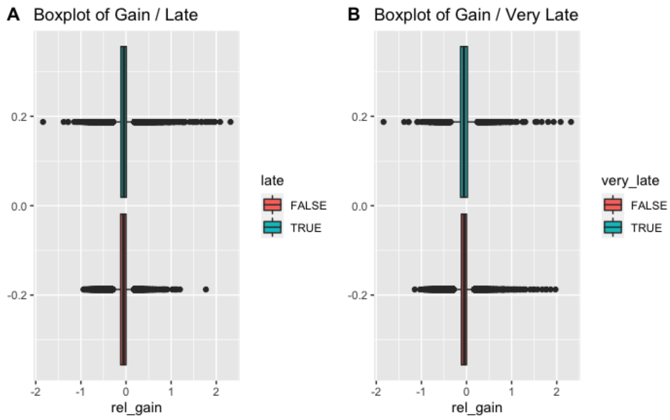


Figure 25: Boxplot of Relative Gain per flight based on Late and very late variable

Late	Mean Gain	Median Gain	Standard Deviation Gain	Minimum Gain	Maximum Gain
False	-0.06631539	-0.05527638	0.1289658	-0.9411765	1.770492
True	-0.05373506	-0.04761905	0.1541525	-1.8388002	2.315789

Table 15: Late Relative Gain Statistics

We can observe from the summary statistic table for the late variable that the data has a pretty small standard deviation.

The mean,median value for the relative gains for flights that were on time or late is nearly identical and only differs by a small margin.

However, the minimum and maximum gain for planes that were on time or late differs significantly.

Late	Mean Gain	Median Gain	Standard Deviation Gain	Minimum Gain	Maximum Gain
False	-0.06114270	-0.05116279	0.1333962	-1.147059	1.972222
True	-0.05482439	-0.05691057	0.1878199	-1.838800	2.315789

Table 16: Very Late Relative Gain Statistics

Even the very late variable has a pretty small standard deviation.

The mean,median,min and maximum value for the relative gains for flights that were delayed more than 30 minutes or less is nearly identical and only differs by a small margin.

Hypothesis testing based on the the flights which are late or on time for the relative gain

H0 : Mean of relative gain per hour for late and flight on time is same

Mean(relative gain per hour for late) = average(relative gain per hour for flight on time)

Ha : Mean of relative gain per hour for late and flight on time is different

Mean(relative gain per hour for late) != mean(relative gain per hour for flight on time)

Statistic	Value
p-value	2.2e-16
95 % Confidence Interval	-0.01489259 -0.01026808

Table 17: Late : Hypothesis testing p-value and Confidence Interval for Relative Gain

We can see that the p-value is very small and less than 0.05. Hence, we can reject the null hypothesis and there's evidence that the alternate hypothesis can be true. Which means that there's a evidence that the mean of relative gain per hour for the flight which are late and or on time have a difference.

H0 : Mean of relative gain per hour for very late and flight which were having delay less than 30 minutes is same

Mean(relative gain per hour for very late flights) = Mean(relative gain per hour for flight where delays is less than 30 minutes)

Ha : Mean of relative gain per hour for very late and flight which were having delay less than 30 minutes is different

Mean(relative gain per hour for very late flights) != Mean(relative gain per hour for flight where delays is less than 30 minutes)

Statistic	Value
p-value	0.004646
95 % Confidence Interval	-0.010692684 -0.001943938

Table 18: Late : Hypothesis testing p-value and Confidence Interval for Relative Gain

Even for the very late variable the p-value is less than 0.05. It means that we can reject the null hypothesis.We have a evidence that there might be a difference between the relative gain for the flights which were very late or having departure delay less than 30 minutes. Even we can conclude the same thing by seeing the confidence interval for the difference between means of the flights because it doesn't contain 0.

Let's do the bootstrap hypothesis testing for the late and very late variable based on the relative gain.

We will use the same hypothesis as earlier for the late and very late variable but this time with the bootstrapped sample.

Hypothesis testing for relative gain for the late variable: **H0 : Average gain for late and flight on time is same**

average(gain for late) = average(gain for flight on time)

Ha : Average gain for late and flights on time is different

average(gain for late) != average(gain for flight on time)

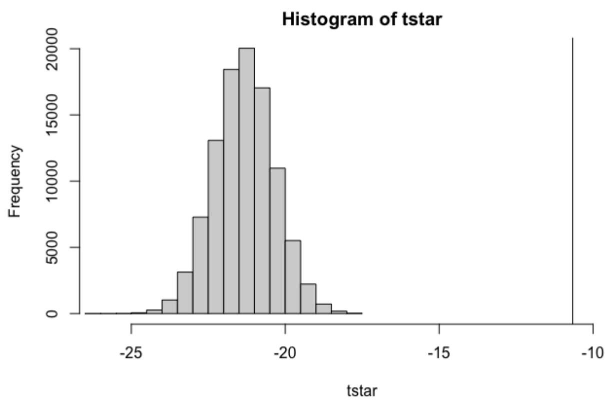


Figure 26: Bootstrap Late: Hypothesis Testing for Relative gain

The p-value for the bootstrapped hypothesis testing is 2e-05. It is very small. At 5 % significance level we can reject the null hypothesis. We can safely say that we have a evidence that the average relative gain for the flights which were delayed or on time is different.

Let's do the bootstrapped hypothesis testing based on the very late variable and see how the results is changing.

H0 : Mean of relative gain per hour for very late and flight which were having delay less than 30 minutes is same

Mean(relative gain per hour for very late flights) = Mean(relative gain per hour for flight where delays is less than 30 minutes)

Ha : Mean of relative gain per hour for very late and flight which were having delay less than 30 minutes is different

Mean(relative gain per hour for very late flights) != Mean(relative gain per hour for flight where delays is less than 30 minutes)

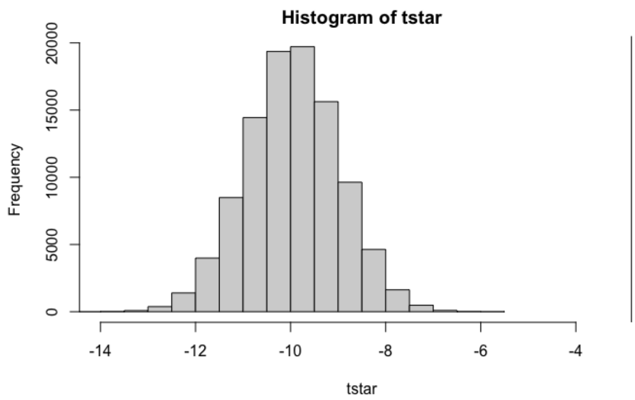


Figure 27: Bootstrap Very Late: Hypothesis Testing for Relative gain

Even in this case, the p-value is 2e-05 which is very small. We can reject our null hypothesis at 5% significance level. The result didn't change for the bootstrapped hypothesis testing for both late and very late variable.

6 Gain per hour:Longer flights versus Shorter flights

We use the distance variable to determine which flights are longer or shorter. If the flight distance is more than 1800 than we are saying that the flight is longer else shorter. We can observe that

Parameter	Count
Longer flights	38814
Shorter flights	19851

Table 19: Duration : Shorter / Longer flight Count

19851 flights are going over lesser distances and there are 38814 flights which are departing from New York for longer distance.

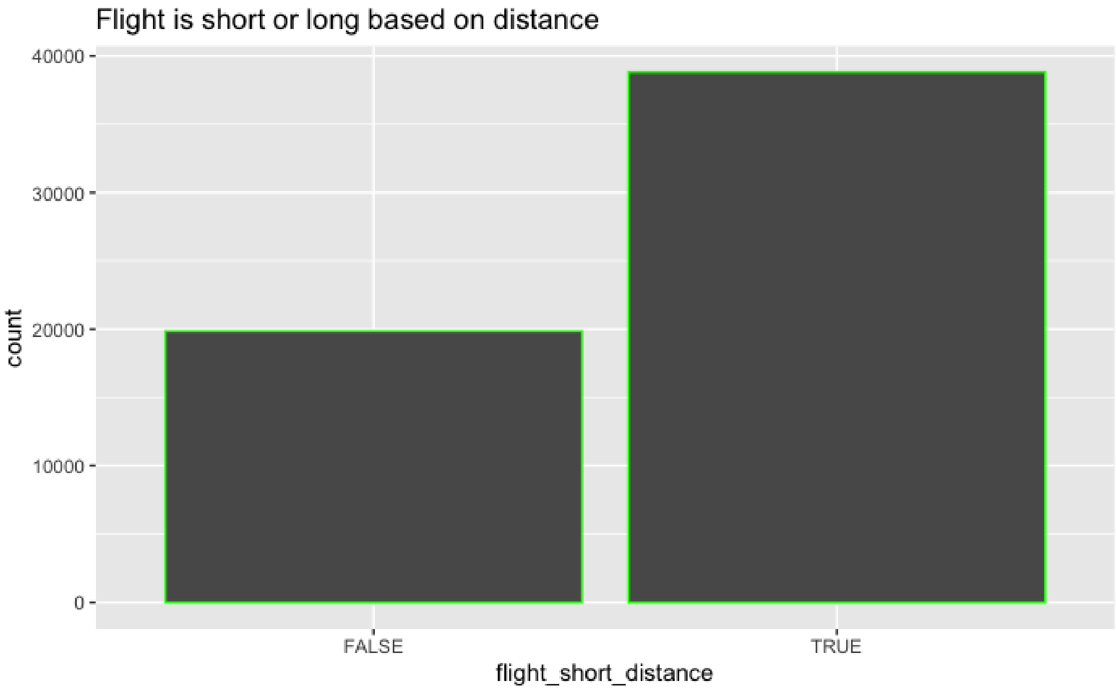


Figure 28: Flight: Longer or Short based on Distance

Let's see the distribution of gain per flight based on the shorter and longer flight distance.

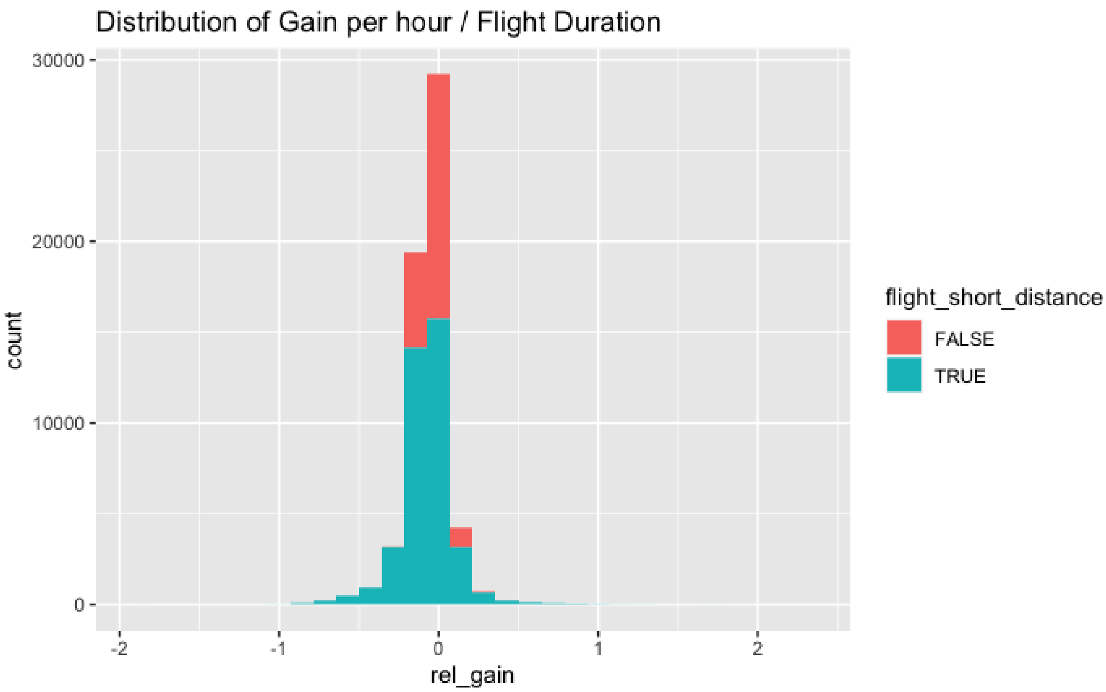


Figure 29: Flight: Longer or Short based on Distance

We can observe that the relative gain for planes going a shorter distance is the same as for flights traveling a longer distance. Both of them have a normal distribution of gain. Let's see the 5 point summary based on the boxplot for the short and longer flights.

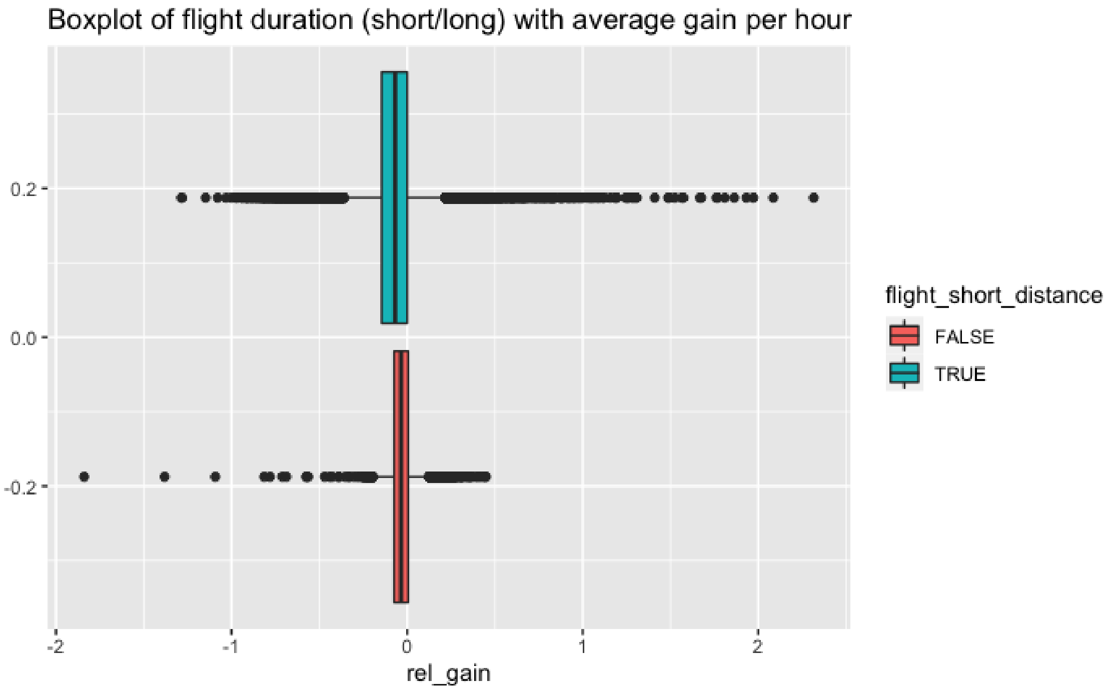


Figure 30: Flight: Longer or Short based on Distance

We can see that the relative gain for the longer duration have wide range of values compare to the flight which are short.

Based on the summary statistic table we can see that the relative mean gain is different for both the flights which travelled short/longer distance. The standard deviation is more comparatively more for the flights which travelled distance less than 1800.

Shortest Flight	Mean Gain	Median Gain	Standard Deviation Gain	Minimum Gain	Maximum Gain
False	-0.03214139	-0.03438395	0.0708344	-1.83880	0.4496855
True	-0.07473518	-0.06818182	0.1647795	-1.28637	2.315789

Table 20: Distance: Relative Gain Statistics

Let's do the hypothesis test and see if there's any difference between the average gain per flight for the longer and shorter duration flights.
Hypothesis testing based on the the relative gain for the flights which travleled longer or shorter distance
H0 : Mean of relative gain per hour for flights which travelled longer or shorter distance is same
Mean(relative gain per hour for longer flights) = average(relative gain per hour for shorter flights)
Ha : Mean of relative gain per hour for flights which travelled longer or shorter distance is different
Mean(relative gain per hour for longer flights) != average(relative gain per hour for shorter flights)

Statistic	Value
p-value	2.2e-16
95 % Confidence Interval	0.04068109 0.04450648

Table 21: Shorter/Longer : Hypothesis testing p-value and Confidence Interval for Relative Gain

We can see that the p-value is less than 0.05 which means that we can safely reject the null hypothesis. It means that we have a evidence that there's a difference between the average relative gain per hour for the flights which are shorter and longer. The same can be concluded from the 95% confidence interval.

Let's do the same hypothesis testing but without outliers. There's no impact on the p-value even without the outliers. AS the p-value is less than 2.2e-16 which is very small at 5% significance level. Hence, we can reject the null hypothesis and conclude that we have evidence that the relative gain are different for the flights which travelled for small distance vs large distance.

Statistic without outliers	Value
p-value	2.2e-16
95 % Confidence Interval	0.04409890 0.04754034

Table 22: Shorter/Longer Without outliers : Hypothesis testing p-value and Confidence Interval for Relative Gain

Let's run the bootstrap t test to determine whether there's a difference in the p-value for the relative gain for aircraft that traveled the smallest distance vs planes that traveled the longest distance.

Hypothesis testing for relative gain based on the distance: **H0 : Mean of relative gain per hour for flights which travelled longer or shorter distance is same**
Mean(relative gain per hour for longer flights) = average(relative gain per hour for shorter flights)
Ha : Mean of relative gain per hour for flights which travelled longer or shorter distance is different
Mean(relative gain per hour for longer flights) != average(relative gain per hour for shorter flights)

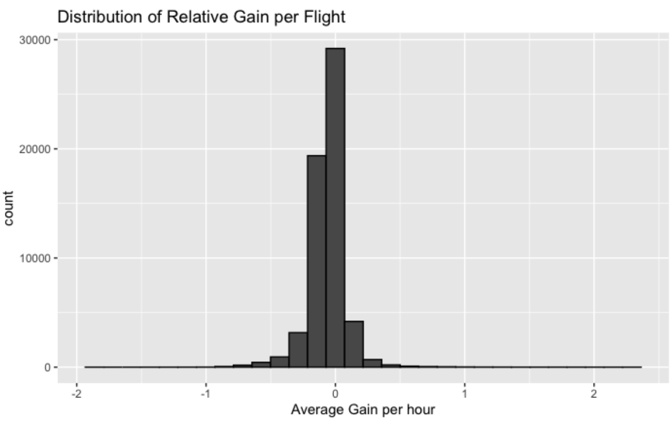


Figure 31: Bootstrap Distance: Hypothesis Testing for Relative gain

The p-value for the bootstrapped hypothesis testing is 2. At 5 % significance level we can accept the null hypothesis because the p-value is greater than 0.05. We can say that we have a evidence that the average of relative gain for the flights which travelled for shorter distance vs the flights which travelled for longer distance is same. Note: We didn't get the same result for the bootstrap test and t.test.

7 Conclusions

We conducted exploratory analysis, hypothesis testing, and bootstrapped hypothesis testing in our report.In several situations, we obtained data that revealed a considerable disparity between average gains and relative gains. But it's critical to recognize whether that distinction truly matters in the real world. It might be beneficial to recognize how big of a difference matters in the real world.

Milestone2

Akanksha Sharma

2022-11-19

Import librarires

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ purrr   0.3.4
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(nycflights13)
library(ggpubr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
##
## The following object is masked from 'package:ggpubr':
##
##     get_legend
```

Let's try to understand more about the data:

Filter out the data based on the United Airlines carrier

```
UA_flight = flights %>%
  filter(carrier == 'UA')
```

We are going to use UA_flight data for further analysis of this project.
How many rows are there for the United Airlines ?

```
print(paste('Size of dataset for the United Airlines', nrow(UA_flight)))
```

```
## [1] "Size of dataset for the United Airlines 58665"
```

What are the type of variables?

```
glimpse(UA_flight)
```

```
## Rows: 58,665
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 517, 533, 554, 558, 558, 559, 607, 611, 623, 628, 643, …
## $ sched_dep_time <int> 515, 529, 558, 600, 600, 600, 607, 600, 627, 630, 646, …
## $ dep_delay      <dbl> 2, 4, -4, -2, -2, -1, 0, 11, -4, -2, -3, 8, 1, 1, -4, -…
## $ arr_time       <int> 830, 850, 740, 924, 923, 854, 858, 945, 933, 1016, 922,…
## $ sched_arr_time <int> 819, 830, 728, 917, 937, 902, 915, 931, 932, 947, 940, …
## $ arr_delay      <dbl> 11, 20, 12, 7, -14, -8, -17, 14, 1, 29, -18, -9, -6, -7…
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "…
## $ flight         <int> 1545, 1714, 1696, 194, 1124, 1187, 1077, 303, 496, 1665…
## $ tailnum        <chr> "N14228", "N24211", "N39463", "N29129", "N53441", "N765…
## $ origin         <chr> "EWR", "LGA", "EWR", "JFK", "EWR", "EWR", "EWR", "JFK",…
## $ dest           <chr> "IAH", "IAH", "ORD", "LAX", "SFO", "LAS", "MIA", "SFO",…
## $ air_time       <dbl> 227, 227, 150, 345, 361, 337, 157, 366, 229, 366, 146, …
## $ distance       <dbl> 1400, 1416, 719, 2475, 2565, 2227, 1085, 2586, 1416, 24…
## $ hour           <dbl> 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7…
## $ minute         <dbl> 15, 29, 58, 0, 0, 0, 7, 0, 27, 30, 46, 36, 45, 45, 0, 0…
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
```

```
cat('Number of flights for which the departure delay is missing' , sum(is.na(UA_flight$d
ep_delay)),'\n')
```

```
## Number of flights for which the departure delay is missing 686
```

```
cat('Percentage of missing data for departure delays for the UA carrier' ,sum((is.na(UA_
flight$dep_delay))/nrow(UA_flight))*100,'\n')
```

```
## Percentage of missing data for departure delays for the UA carrier 1.169351
```

```
perct <- c(sum(is.na(UA_flight$dep_delay)),sum((is.na(UA_flight$dep_delay))/nrow(UA_flig
ht))*100)
perct
```

```
## [1] 686.000000   1.169351
```

```
tab <- matrix(c(sum(is.na(UA_flight$dep_delay)),sum((is.na(UA_flight$dep_delay))/nrow(UA
_flight))*100), ncol=2, byrow=TRUE)
colnames(tab) <- c('Null values in dataset','Percentage of null values')

kable(tab) %>%
  kable_styling()
```

Null values in dataset	Percentage of null values

Null values in dataset	Percentage of null values
686	1.169351

```
cat('Number of flights for which the arrival delay is missing' , sum(is.na(UA_flight$arr
_delay)),'\n')
```

```
## Number of flights for which the arrival delay is missing 883
```

```
cat('Percentage of missing data for arrival delay for the UA carrier' ,sum((is.na(UA_fli
ght$arr_delay))/nrow(UA_flight))*100,'\n')
```

```
## Percentage of missing data for arrival delay for the UA carrier 1.505156
```

```
perct <- c(sum(is.na(UA_flight$arr_delay)),sum((is.na(UA_flight$arr_delay))/nrow(UA_flig
ht))*100)
perct
```

```
## [1] 883.000000   1.505156
```

```
tab <- matrix(c(sum(is.na(UA_flight$arr_delay)),sum((is.na(UA_flight$arr_delay))/nrow(UA
_flight))*100), ncol=2, byrow=TRUE)
colnames(tab) <- c('Null values in dataset','Percentage of null values')

kable(tab) %>%
  kable_styling()
```

Null values in dataset	Percentage of null values
883	1.505156

```
cat('Number of flights for which the air  time is missing' , sum(is.na(UA_flight$air_tim
e)),'\n')
```

```
## Number of flights for which the air  time is missing 883
```

```
cat('Percentage of missing data for air time for the UA carrier' ,sum((is.na(UA_flight$a
ir_time))/nrow(UA_flight))*100,'\n')
```

```
## Percentage of missing data for air time for the UA carrier 1.505156
```

```
perct <- c(sum(is.na(UA_flight$air_time)),sum((is.na(UA_flight$dep_delay))/nrow(UA_fligh
t))*100)
perct
```

```
## [1] 883.000000   1.169351
```

```
tab <- matrix(c(sum(is.na(UA_flight$air_time)),sum((is.na(UA_flight$air_time))/nrow(UA_f
light))*100), ncol=2, byrow=TRUE)
colnames(tab) <- c('Null values in dataset','Percentage of null values')

kable(tab) %>%
  kable_styling()
```

Null values in dataset	Percentage of null values
883	1.505156

```
cat('Number of flights for which the distance is missing' , sum(is.na(UA_flight$distanc
e)),'\n')
```

```
## Number of flights for which the distance is missing 0
```

```
cat('Percentage of missing data for distance for the UA carrier' ,sum((is.na(UA_flight$d
istance))/nrow(UA_flight))*100,'\n')
```

```
## Percentage of missing data for distance for the UA carrier 0
```

```
perct <- c(sum(is.na(UA_flight$distance)),sum((is.na(UA_flight$distance))/nrow(UA_fligh
t))*100)
perct
```

```
## [1] 0 0
```

```
tab <- matrix(c(sum(is.na(UA_flight$distance)),sum((is.na(UA_flight$distance))/nrow(UA_f
light))*100), ncol=2, byrow=TRUE)
colnames(tab) <- c('Null values in dataset','Percentage of null values')

kable(tab) %>%
  kable_styling()
```

Null values in dataset	Percentage of null values
0	0

```
# Impute missing values with mean in departure delay column
UA_flight$dep_delay <- with(UA_flight, impute(dep_delay, mean))
UA_flight$arr_delay <- with(UA_flight, impute(arr_delay, mean))
UA_flight$air_time <-  with(UA_flight, impute(air_time, mean))
```

Add Late, Very_late and gain variable in the dataset

```
#Add late and Very Late columns in the dataset
UA_flight <- UA_flight %>%
  mutate(late = case_when(dep_delay > 0 ~ TRUE,
                          dep_delay <=0 ~ FALSE ),
         very_late = case_when(dep_delay > 30 ~ TRUE,
                          dep_delay <= 30 ~ FALSE ),
         gain = arr_delay - dep_delay)
glimpse(UA_flight)
```

```
## Rows: 58,665
## Columns: 22
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 517, 533, 554, 558, 558, 559, 607, 611, 623, 628, 643, …
## $ sched_dep_time <int> 515, 529, 558, 600, 600, 600, 607, 600, 627, 630, 646, …
## $ dep_delay      <dbl> 2, 4, -4, -2, -2, -1, 0, 11, -4, -2, -3, 8, 1, 1, -4, -…
## $ arr_time       <int> 830, 850, 740, 924, 923, 854, 858, 945, 933, 1016, 922,…
## $ sched_arr_time <int> 819, 830, 728, 917, 937, 902, 915, 931, 932, 947, 940, …
## $ arr_delay      <dbl> 11, 20, 12, 7, -14, -8, -17, 14, 1, 29, -18, -9, -6, -7…
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "…
## $ flight         <int> 1545, 1714, 1696, 194, 1124, 1187, 1077, 303, 496, 1665…
## $ tailnum        <chr> "N14228", "N24211", "N39463", "N29129", "N53441", "N765…
## $ origin         <chr> "EWR", "LGA", "EWR", "JFK", "EWR", "EWR", "EWR", "JFK",…
## $ dest           <chr> "IAH", "IAH", "ORD", "LAX", "SFO", "LAS", "MIA", "SFO",…
## $ air_time       <dbl> 227, 227, 150, 345, 361, 337, 157, 366, 229, 366, 146, …
## $ distance       <dbl> 1400, 1416, 719, 2475, 2565, 2227, 1085, 2586, 1416, 24…
## $ hour           <dbl> 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7…
## $ minute         <dbl> 15, 29, 58, 0, 0, 0, 7, 0, 27, 30, 46, 36, 45, 45, 0, 0…
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
## $ late           <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FA…
## $ very_late      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ gain           <dbl> 9, 16, 16, 9, -12, -7, -17, 3, 5, 31, -15, -17, -7, -8,…
```

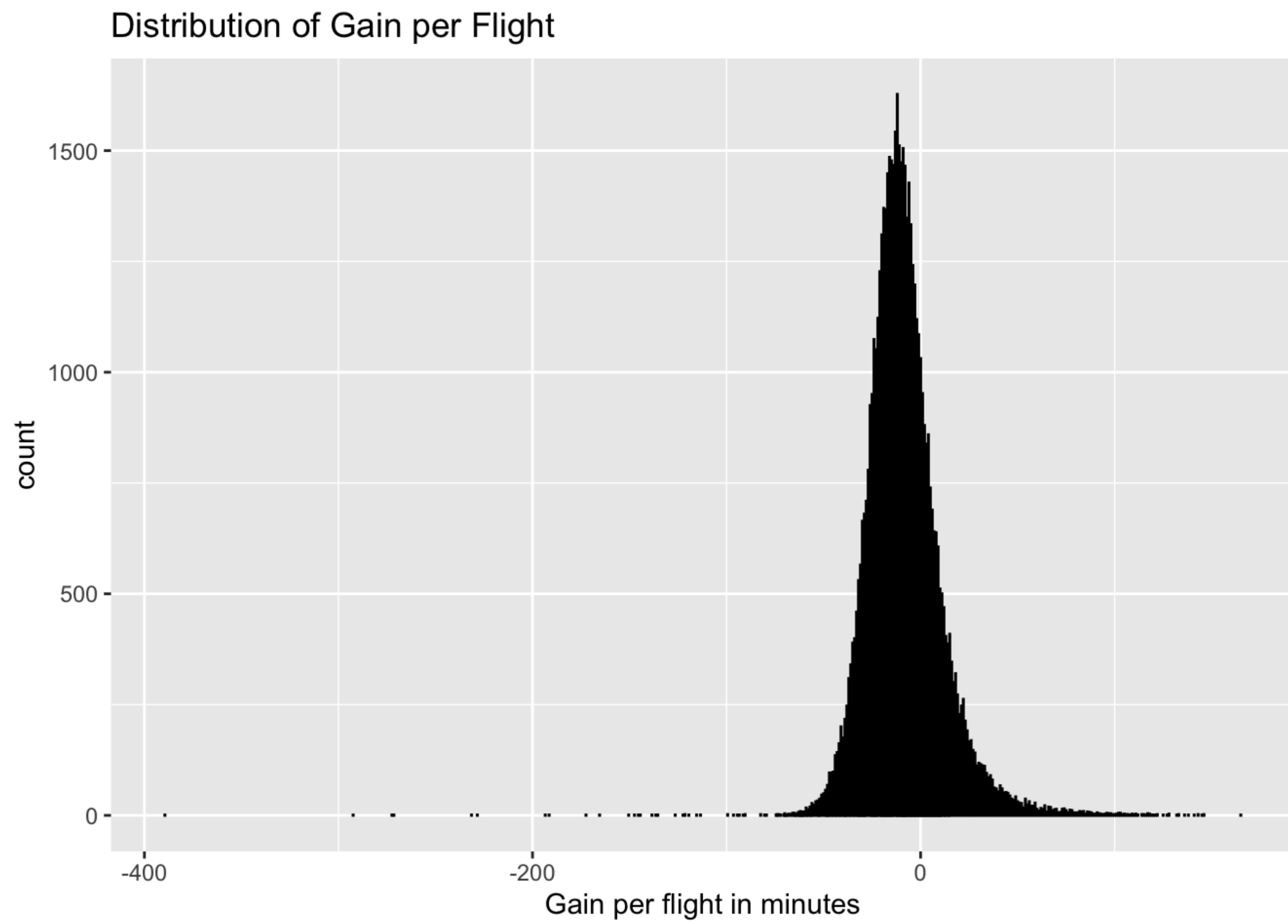
```
UA_flight[UA_flight$arr_delay > UA_flight$dep_delay,]
```

```
## # A tibble: 14,812 × 22
##     year month   day dep_time sched_de…¹ dep_d…² arr_t…³ sched…⁴ arr_d…⁵ carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      554        558      -4     740     728      12 UA
##  4  2013     1     1      558        600      -2     924     917       7 UA
##  5  2013     1     1      611        600      11     945     931      14 UA
##  6  2013     1     1      623        627      -4     933     932       1 UA
##  7  2013     1     1      628        630      -2    1016     947      29 UA
##  8  2013     1     1      709        700       9     852     832      20 UA
##  9  2013     1     1      715        713       2     911     850      21 UA
## 10  2013     1     1      727        730      -3     959     952       7 UA
## # … with 14,802 more rows, 12 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, late <lgl>, very_late <lgl>, gain <dbl>,
## #   and abbreviated variable names ¹sched_dep_time, ²dep_delay, ³arr_time,
## #   ⁴sched_arr_time, ⁵arr_delay
```

#Let's analyse the gain per flight for the UA carrier flight

```
#Create a bar plot
ggplot(data = UA_flight , aes(x= gain ))+
  geom_bar(color = 'black') +
  labs(x = "Gain per flight in minutes", title = "Distribution of Gain per Flight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
summary(UA_flight$gain)
```

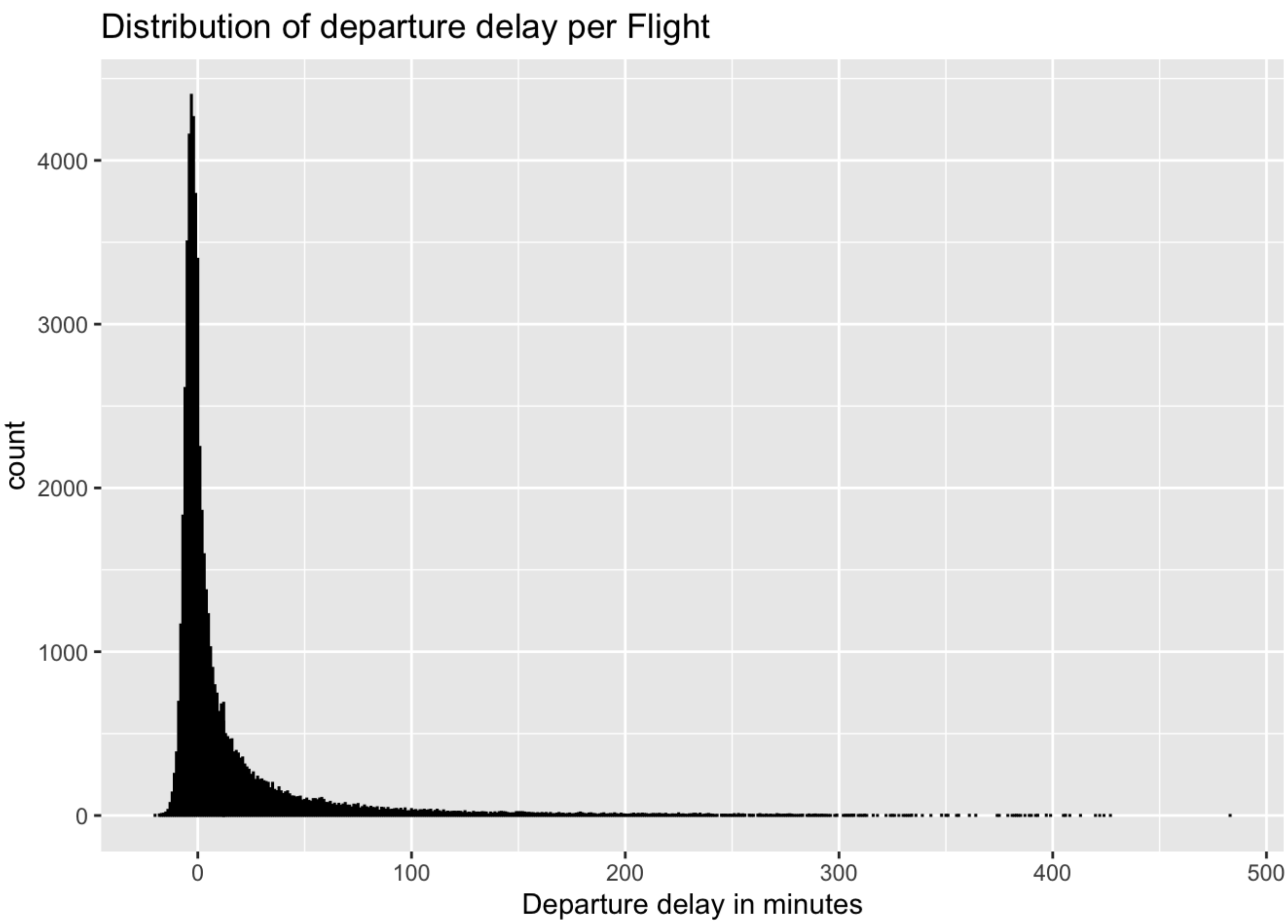
```
##
## Imputed Values:
##
## UA_flight$gain
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      883        0       94    0.531   -14.38     16.4  -56.042  -10.442
##      .25      .50      .75      .90      .95
##   -8.548   -8.548   -8.548   -2.642    5.558
##
## lowest : -389.441989 -292.441989 -272.441989 -271.441989 -231.441989
## highest:    8.558011    9.558011   10.558011   12.558011   14.558011
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -389.442  -20.000  -10.000   -8.548    1.000  165.000
```

The mean gain per flight is -8.54 which means that the most of the time the flights were delayed by 8 minutes.

```
#Create a bar plot
ggplot(data = UA_flight , aes(x= dep_delay ))+
  geom_bar(color = 'black') +
  labs(x = "Departure delay in minutes", title = "Distribution of departure delay per Fl
ight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



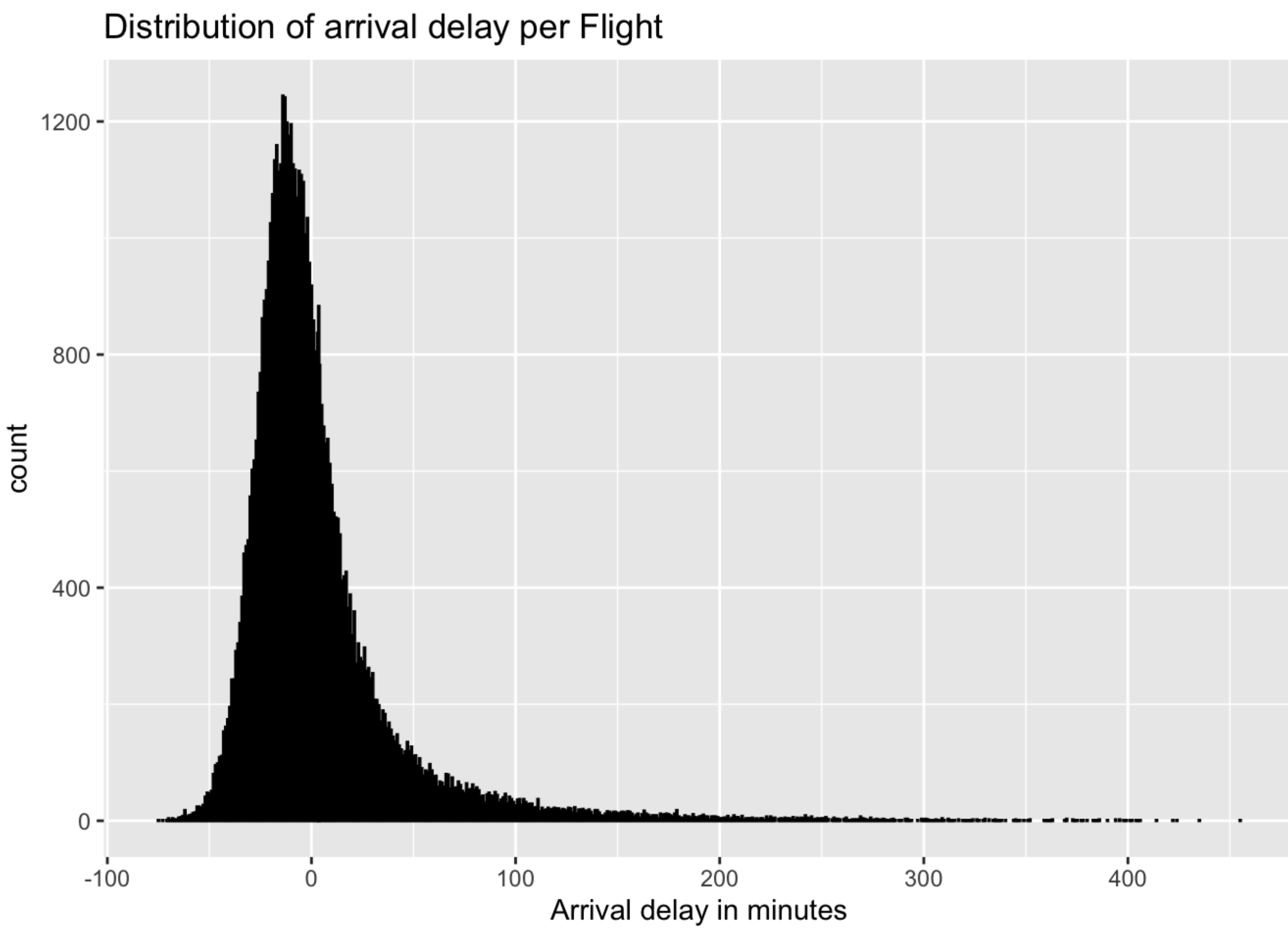

```
summary(UA_flight$dep_delay)
```

```
##
##  686 values imputed to 12.10607
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -20.00   -4.00    0.00   12.11   12.00  483.00
```

```r
#Create a bar plot
ggplot(data = UA_flight , aes(x= arr_delay ))+
  geom_bar(color = 'black') +
  labs(x = "Arrival delay in minutes", title = "Distribution of arrival delay per Fligh
t")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```r
(summary(UA_flight$arr_delay))
```

```
##
##  883 values imputed to 3.558011
```

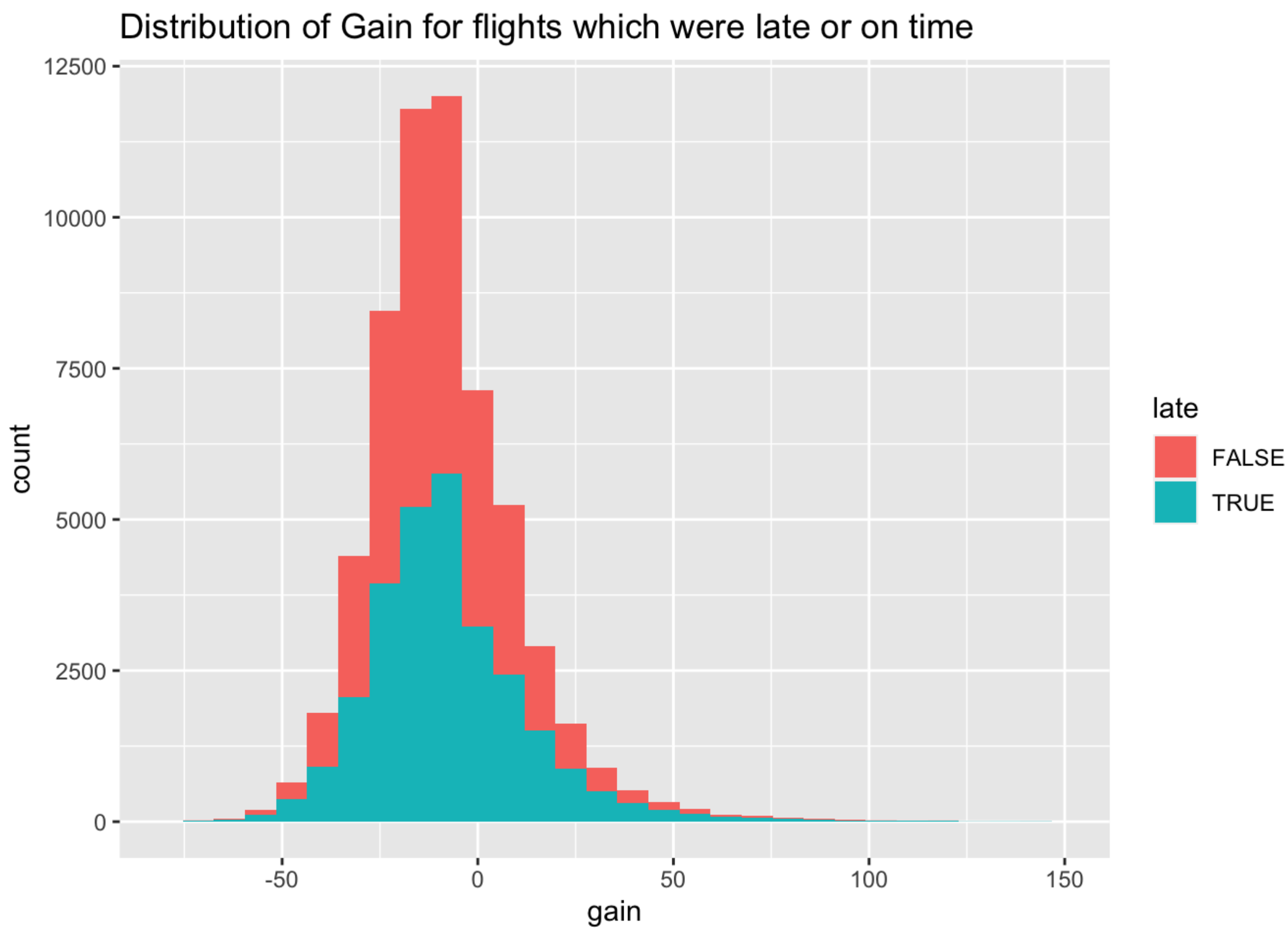
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -75.000 -18.000  -6.000   3.558  11.000 455.000
```

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?

```
ggplot(UA_flight,aes(gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'Distribution of Gain for flights which were late or on time')+
  xlim(-80,150)
```

```
## Warning: Removed 36 rows containing non-finite values (stat_bin).
```

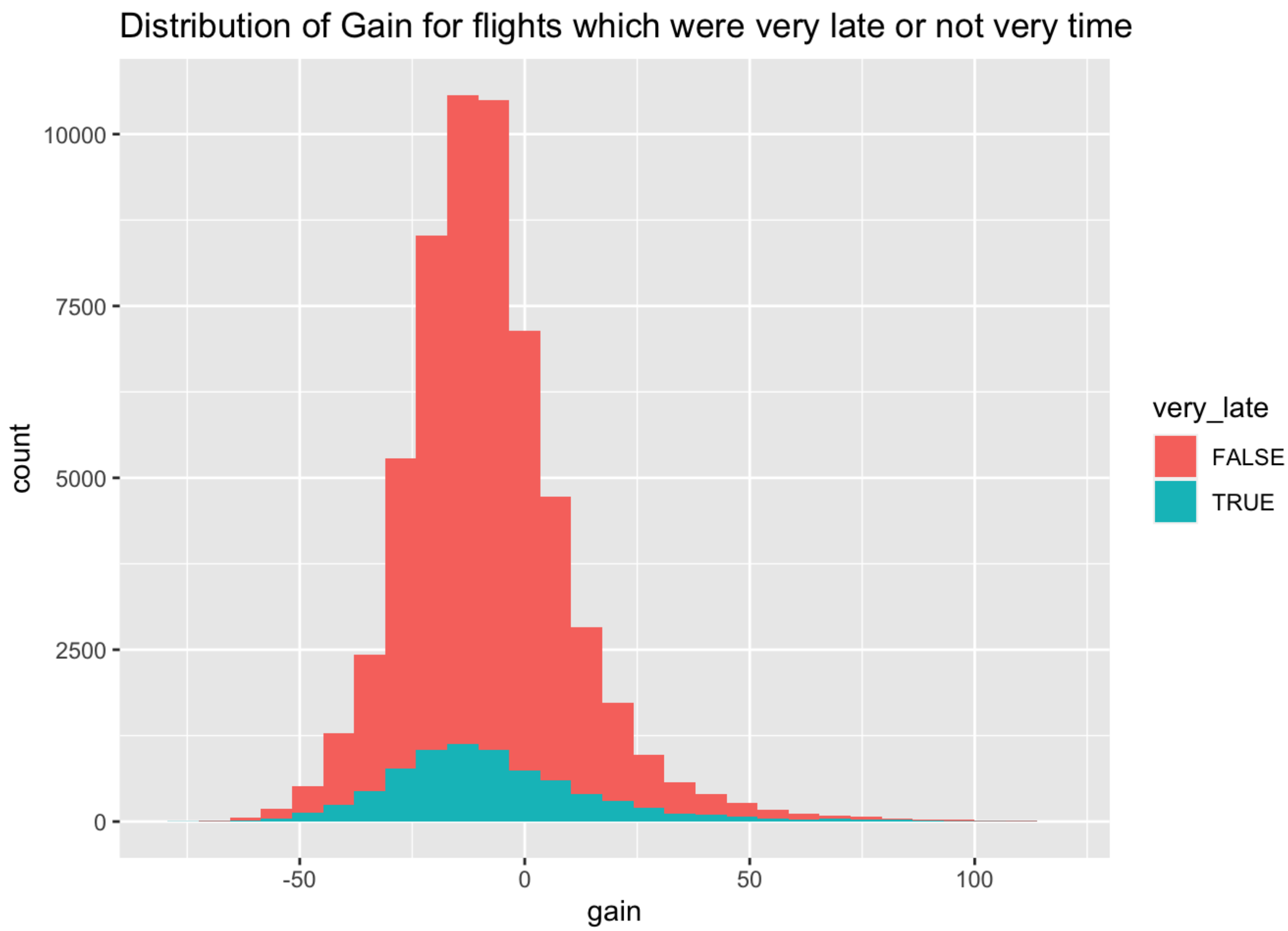
```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
ggplot(UA_flight,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_histogram(bins = 30)+
  xlim(-80,120)+

  labs(title = 'Distribution of Gain for flights which were very late or not very time')
```

```
## Warning: Removed 55 rows containing non-finite values (stat_bin).
```

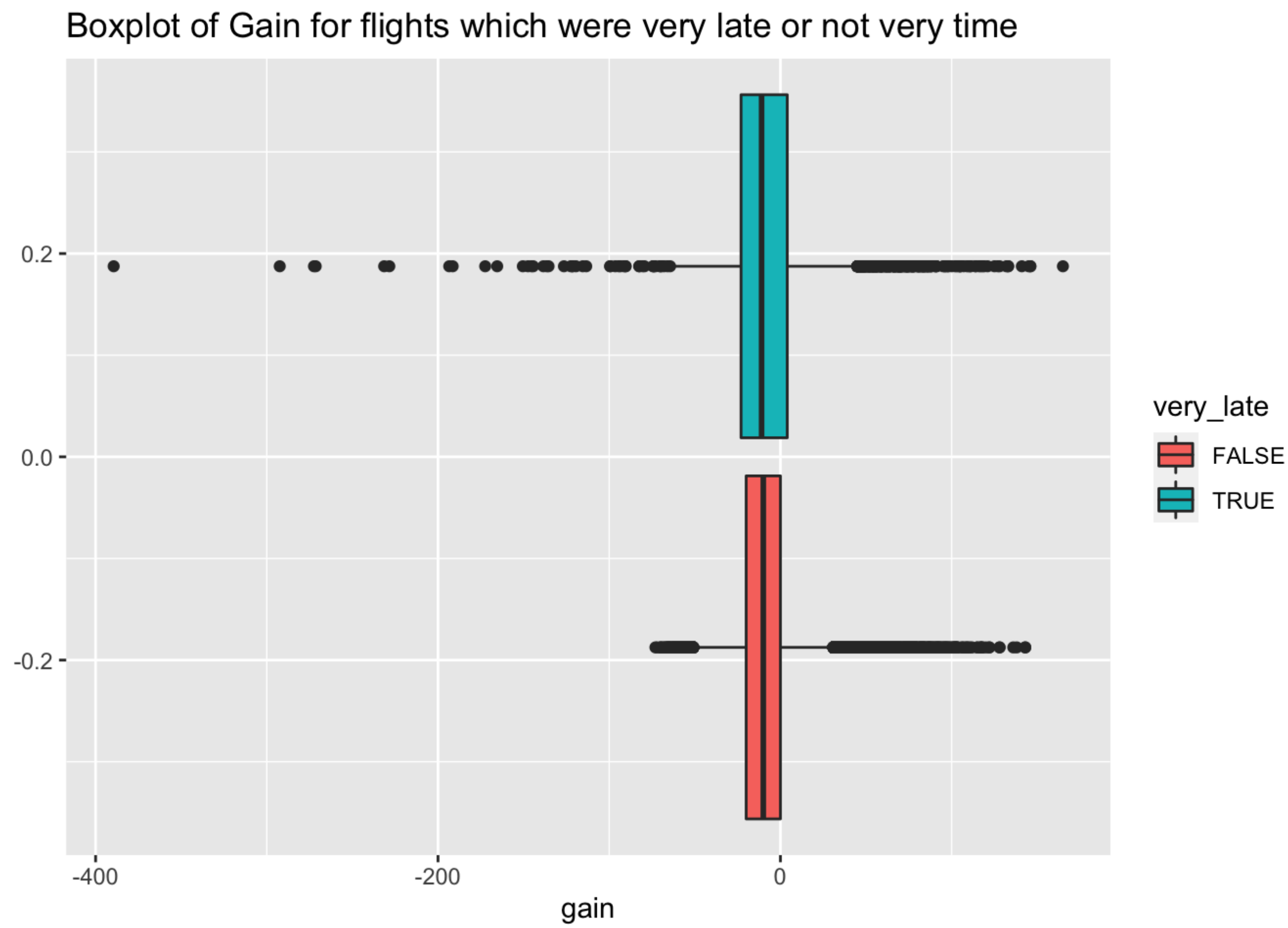
```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
ggplot(UA_flight,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'Boxplot of Gain for flights which were very late or not very time')
```

```
## Warning: Ignoring unknown parameters: bins
```

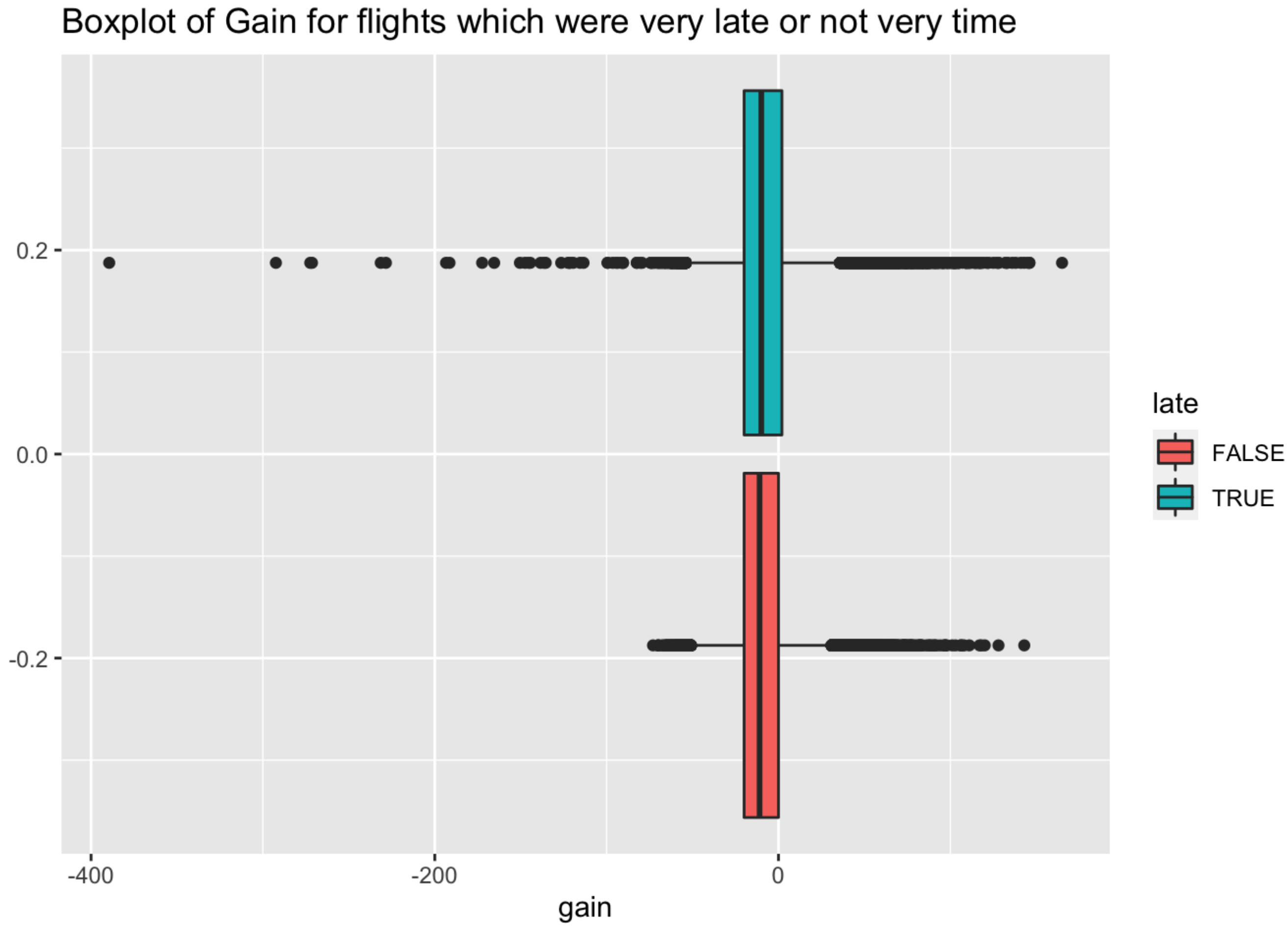
```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
ggplot(UA_flight,aes(gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'Boxplot of Gain for flights which were very late or not very time')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
glimpse(UA_flight)
```

```
## Rows: 58,665
## Columns: 22
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 517, 533, 554, 558, 558, 559, 607, 611, 623, 628, 643, …
## $ sched_dep_time <int> 515, 529, 558, 600, 600, 600, 607, 600, 627, 630, 646, …
## $ dep_delay      <dbl> 2, 4, -4, -2, -2, -1, 0, 11, -4, -2, -3, 8, 1, 1, -4, -…
## $ arr_time       <int> 830, 850, 740, 924, 923, 854, 858, 945, 933, 1016, 922,…
## $ sched_arr_time <int> 819, 830, 728, 917, 937, 902, 915, 931, 932, 947, 940, …
## $ arr_delay      <dbl> 11, 20, 12, 7, -14, -8, -17, 14, 1, 29, -18, -9, -6, -7…
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "…
## $ flight         <int> 1545, 1714, 1696, 194, 1124, 1187, 1077, 303, 496, 1665…
## $ tailnum        <chr> "N14228", "N24211", "N39463", "N29129", "N53441", "N765…
## $ origin         <chr> "EWR", "LGA", "EWR", "JFK", "EWR", "EWR", "EWR", "JFK",…
## $ dest           <chr> "IAH", "IAH", "ORD", "LAX", "SFO", "LAS", "MIA", "SFO",…
## $ air_time       <dbl> 227, 227, 150, 345, 361, 337, 157, 366, 229, 366, 146, …
## $ distance       <dbl> 1400, 1416, 719, 2475, 2565, 2227, 1085, 2586, 1416, 24…
## $ hour           <dbl> 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7…
## $ minute         <dbl> 15, 29, 58, 0, 0, 0, 7, 0, 27, 30, 46, 36, 45, 45, 0, 0…
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
## $ late           <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FA…
## $ very_late      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ gain           <dbl> 9, 16, 16, 9, -12, -7, -17, 3, 5, 31, -15, -17, -7, -8,…
```

```
UA_flight %>%
  group_by(late) %>%
  dplyr::summarize(Mean_gain = mean(gain),
                   Median_gain = median(gain),
                   StandardDeviation_gain = sd(gain),
                   MinGain =min(gain),
                   MaxGain = max(gain)
                   )
```

```
## # A tibble: 2 × 6
##   late  Mean_gain Median_gain StandardDeviation_gain MinGain MaxGain
##   <lgl>     <dbl>       <dbl>                  <dbl>   <dbl>   <dbl>
## 1 FALSE     -9.24         -11                   17.3     -73     143
## 2 TRUE      -7.79         -10                   21.4   -389.     165
```

```
UA_flight %>%
  group_by(very_late) %>%
  dplyr::summarize(Mean_gain = mean(gain),
                   Median_gain = median(gain),
                   StandardDeviation_gain = sd(gain),
                   MinGain =min(gain),
                   MaxGain = max(gain)
                   )
```

```
## # A tibble: 2 × 6
##   very_late Mean_gain Median_gain StandardDeviation_gain MinGain MaxGain
##   <lgl>         <dbl>       <dbl>                  <dbl>   <dbl>   <dbl>
## 1 FALSE         -8.68         -10                   18.0     -73     143
## 2 TRUE          -7.69         -11                   26.7    -389.    165
```

Hypothesis Testing for Late variable

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=UA_flight, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -8.9547, df = 53794, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.761377 -1.128780
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.236472           -7.791394
```

Hypothesis testing for Very Late variable

```
ggplot(UA_flight)
```

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=UA_flight, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -3.1268, df = 8671.8, p-value = 0.001773
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.6104501 -0.3693194
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.676587           -7.686703
```

```
z_scores <- as.data.frame(sapply(UA_flight$gain, function(df) (abs(df-(-8.548062))/(19.3
4348))),colnames = c('score'))
colnames(z_scores) <- c('score')
without_outlier <- subset(UA_flight, (z_scores$score < 3) & (z_scores$score > -3))
```

Hypothesis Testing for Late variable Without Outlier

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=without_outlier, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -5.6985, df = 55715, p-value = 1.215e-08
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.0781134 -0.5262771
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.753196           -8.951000
```

Hypothesis testing for Very Late variable

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

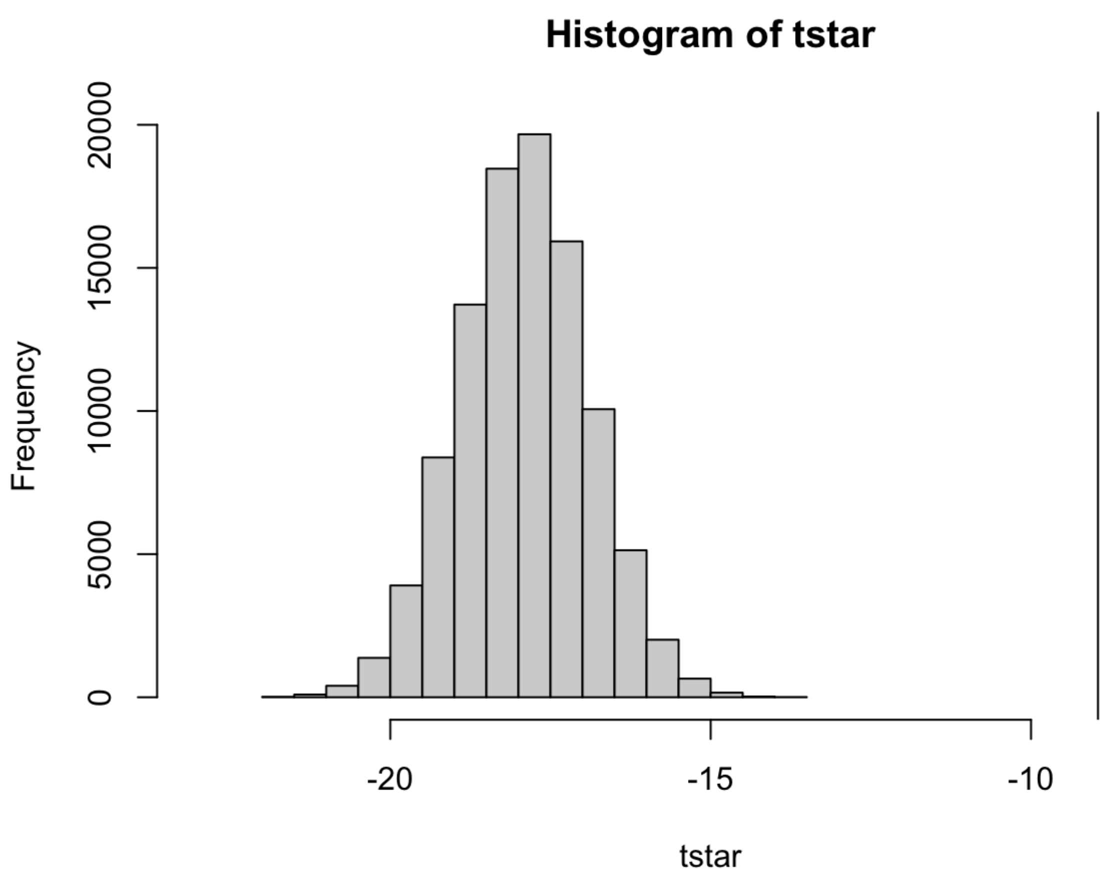
```
t.test(gain~very_late,data=without_outlier, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -0.17031, df = 8846, p-value = 0.8648
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -0.5197887  0.4366856
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.378605           -9.337053
```

Bootstrap t test to see if there’s difference between the means for the flights which were late or on time.

```
UA_flight_late <-subset(UA_flight,gain,subset = late ==TRUE,drop=T)
UA_flight_notlate <- subset(UA_flight,gain,subset = late ==FALSE,drop=T)
```

```
tstat <- function(x , y , mu)
{
  (mean(y) - mean(x) - mu)/sqrt(var(y)/length(y) + var(x)/length(x))

}
observed <- tstat(UA_flight_late,UA_flight_notlate,0)
thetahat <- mean(UA_flight_late) - mean(UA_flight_notlate)
n1 <- length(UA_flight_late)
n2 <- length(UA_flight_notlate)

N <- 10^5-1
tstar <- numeric(N)
set.seed(5)
for (i in 1:N)
{
  boot1 <- sample(UA_flight_late,n1,replace = TRUE)
  boot2 <- sample(UA_flight_notlate,n2,replace = TRUE)
  tstar[i] <- tstat(boot1,boot2,thetahat)
}
hist(tstar, xlim = c(-23,-7))
abline(v=observed)
```

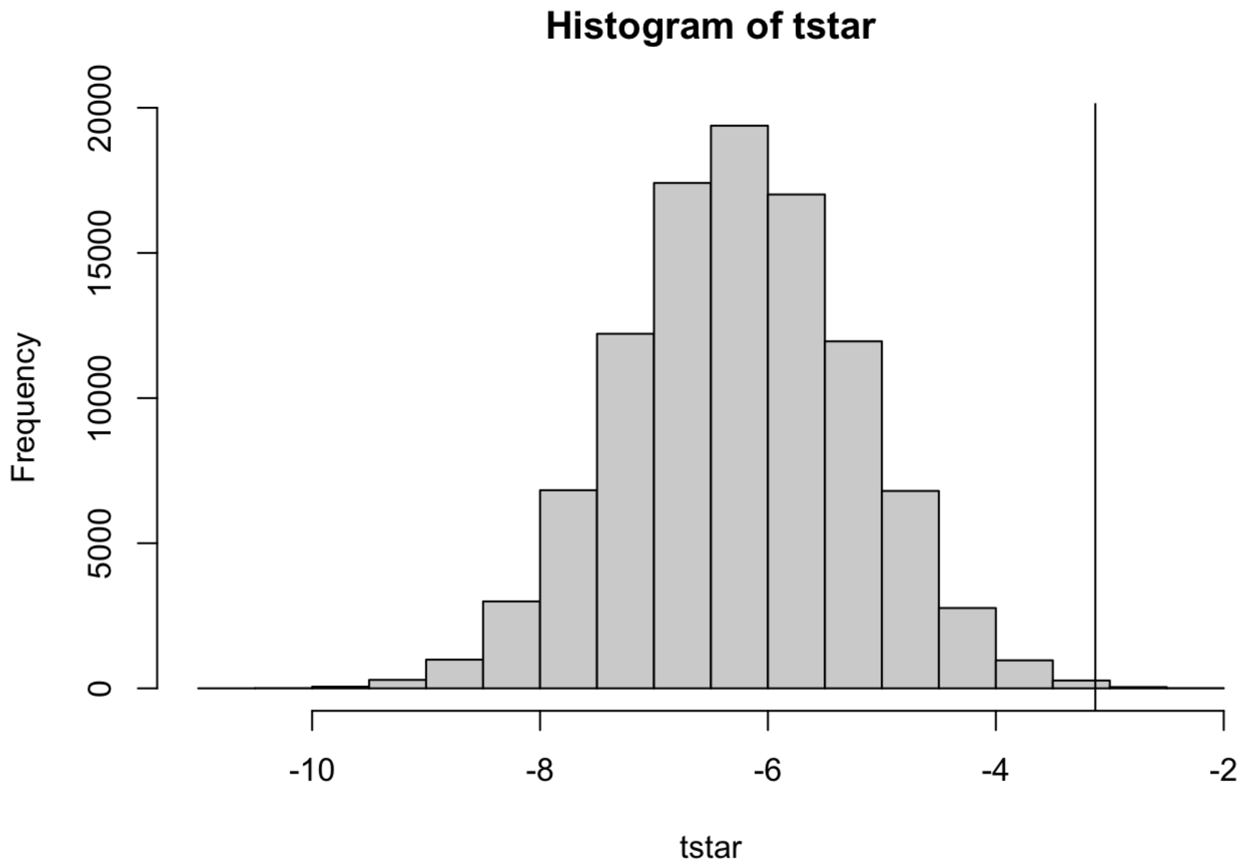



```
cat('The p-value is :',2*(sum(tstar >= observed)+1)/(N+1))
```

```
## The p-value is : 2e-05
```

```
UA_flight_verylate <-subset(UA_flight,gain,subset = very_late ==TRUE,drop=T)
UA_flight_notverylate <- subset(UA_flight,gain,subset = very_late ==FALSE,drop=T)
```

```
tstat <- function(x , y , mu)
{
  (mean(y) - mean(x) - mu)/sqrt(var(y)/length(y) + var(x)/length(x))

}
observed <- tstat(UA_flight_verylate,UA_flight_notverylate,0)
thetahat <- mean(UA_flight_verylate) - mean(UA_flight_notverylate)
n1 <- length(UA_flight_verylate)
n2 <- length(UA_flight_notverylate)

N <- 10^5-1
tstar <- numeric(N)
set.seed(5)
for (i in 1:N)
{
  boot1 <- sample(UA_flight_verylate,n1,replace = TRUE)
  boot2 <- sample(UA_flight_notverylate,n2,replace = TRUE)
  tstar[i] <- tstat(boot1,boot2,thetahat)
}
hist(tstar)
abline(v=observed)
```




```
cat('The p-value is :',2*(sum(tstar >= observed)+1)/(N+1))
```

```
## The p-value is : 0.0018
```

Let’s do the bootstrap t-test for the very_late variable.

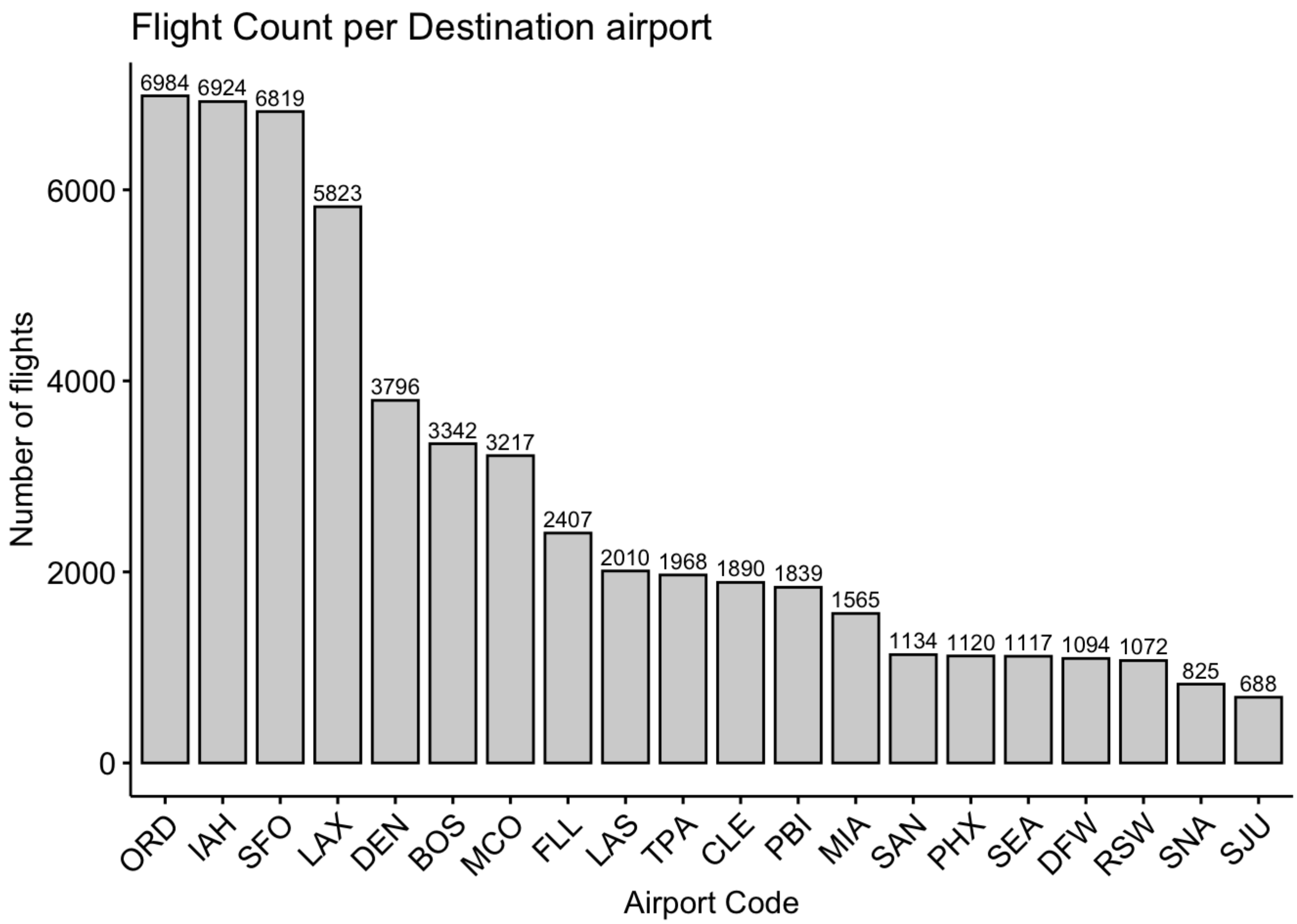
What are the five most common destination airports for United Airlines flights from New York City? Describe the distribution and the average gain for each of these five airports.

```
airport_freq = as.data.frame(table(UA_flight$dest))
```

```
airport_freq
```

```
##    Var1 Freq
## 1   ANC    8
## 2   ATL  103
## 3   AUS  670
## 4   BDL    8
## 5   BOS 3342
## 6   BQN  297
## 7   BZN   36
## 8   CHS    1
## 9   CLE 1890
## 10  CLT    2
## 11  DCA    2
## 12  DEN 3796
## 13  DFW 1094
## 14  DTW    1
## 15  EGE  110
## 16  FLL 2407
## 17  HDN   15
## 18  HNL  365
## 19  IAD    1
## 20  IAH 6924
## 21  IND    3
## 22  JAC   23
## 23  LAS 2010
## 24  LAX 5823
## 25  MCO 3217
## 26  MIA 1565
## 27  MSP    2
## 28  MSY  269
## 29  MTJ   15
## 30  OMA    2
## 31  ORD 6984
## 32  PBI 1839
## 33  PDX  571
## 34  PHX 1120
## 35  PIT    2
## 36  RDU    1
## 37  RSW 1072
## 38  SAN 1134
## 39  SAT  330
## 40  SDF    3
## 41  SEA 1117
## 42  SFO 6819
## 43  SJU  688
## 44  SNA  825
## 45  STL    2
## 46  STT  189
## 47  TPA 1968
```

```r
ggbarplot(airport_freq, x = "Var1", y = "Freq",
         fill = "lightgray", width = 0.8,
         xlab = "Airport Code", ylab = "Number of flights",
         label = TRUE, lab.pos = "out", lab.col = "black",lab.size = 3,
         sort.val = "desc", # Sort in descending order
         top = 20,          # select top 20 most citated genes
         x.text.angle = 45 , # x axis text rotation angle
         title = "Flight Count per Destination airport"
         )
```




```r
UA_flight %>%
  filter(dest %in% c('ORD','IAH','SFO','LAX','DEN')) %>%
  group_by(dest) %>%
  dplyr::summarize(Mean_gain = mean(gain),
                   Median_gain = median(gain),
                   StandardDeviation_gain = sd(gain),
                   MinGain =min(gain),
                   MaxGain = max(gain)
                   )
```

```
## # A tibble: 5 × 6
##   dest  Mean_gain Median_gain StandardDeviation_gain MinGain MaxGain
##   <chr>     <dbl>       <dbl>                  <dbl>   <dbl>   <dbl>
## 1 DEN       -7.44          -9                   20.7  -271.      136
## 2 IAH       -6.95          -9                   18.6  -228.      128
## 3 LAX       -7.84          -9                   21.9   -90.4     145
## 4 ORD       -7.94         -10                   19.4  -272.      146
## 5 SFO       -8.87         -10                   23.2  -389.      165
```

```
UA_flight %>%
  filter(dest %in% c('ORD','IAH','SFO','LAX','DEN')) %>%
  group_by(late,dest) %>%
  dplyr::summarize(Mean_gain_ = mean(gain),
                   Median_gain_ = median(gain),
                   StandardDeviation_gain_ = sd(gain),
                   MinGain_ =min(gain),
                   MaxGain_ = max(gain)
                   )
```

```
## `summarise()` has grouped output by 'late'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 10 × 7
## # Groups:   late [2]
##    late  dest  Mean_gain_ Median_gain_ StandardDeviation_gain_ MinGain_ MaxGain_
##    <lgl> <chr>      <dbl>        <dbl>                   <dbl>    <dbl>    <dbl>
##  1 FALSE DEN        -8.77       -10                       17.7    -56         94
##  2 FALSE IAH        -7.41        -9                       17.9    -59        117
##  3 FALSE LAX        -8.50        -9                       20.1    -73        118
##  4 FALSE ORD        -9.02       -11                       16.6    -57        128
##  5 FALSE SFO        -9.26       -10                       20.2    -70        143
##  6 TRUE  DEN        -6.04        -8.55                    23.4   -271.       136
##  7 TRUE  IAH        -6.42        -8.55                    19.4   -228.       128
##  8 TRUE  LAX        -7.14        -9                       23.6    -90.4      145
##  9 TRUE  ORD        -6.56        -8.55                    22.4   -272.       146
## 10 TRUE  SFO        -8.43       -10                       26.2   -389.       165
```

```
UA_flight %>%
  filter(dest %in% c('ORD','IAH','SFO','LAX','DEN')) %>%
  group_by(very_late,dest) %>%
  dplyr::summarize(Mean_gain = mean(gain),
                   Median_gain = median(gain),
                   StandardDeviation_gain = sd(gain),
                   MinGain =min(gain),
                   MaxGain = max(gain)
                   )
```

```
## `summarise()` has grouped output by 'very_late'. You can override using the
## `.groups` argument.
```

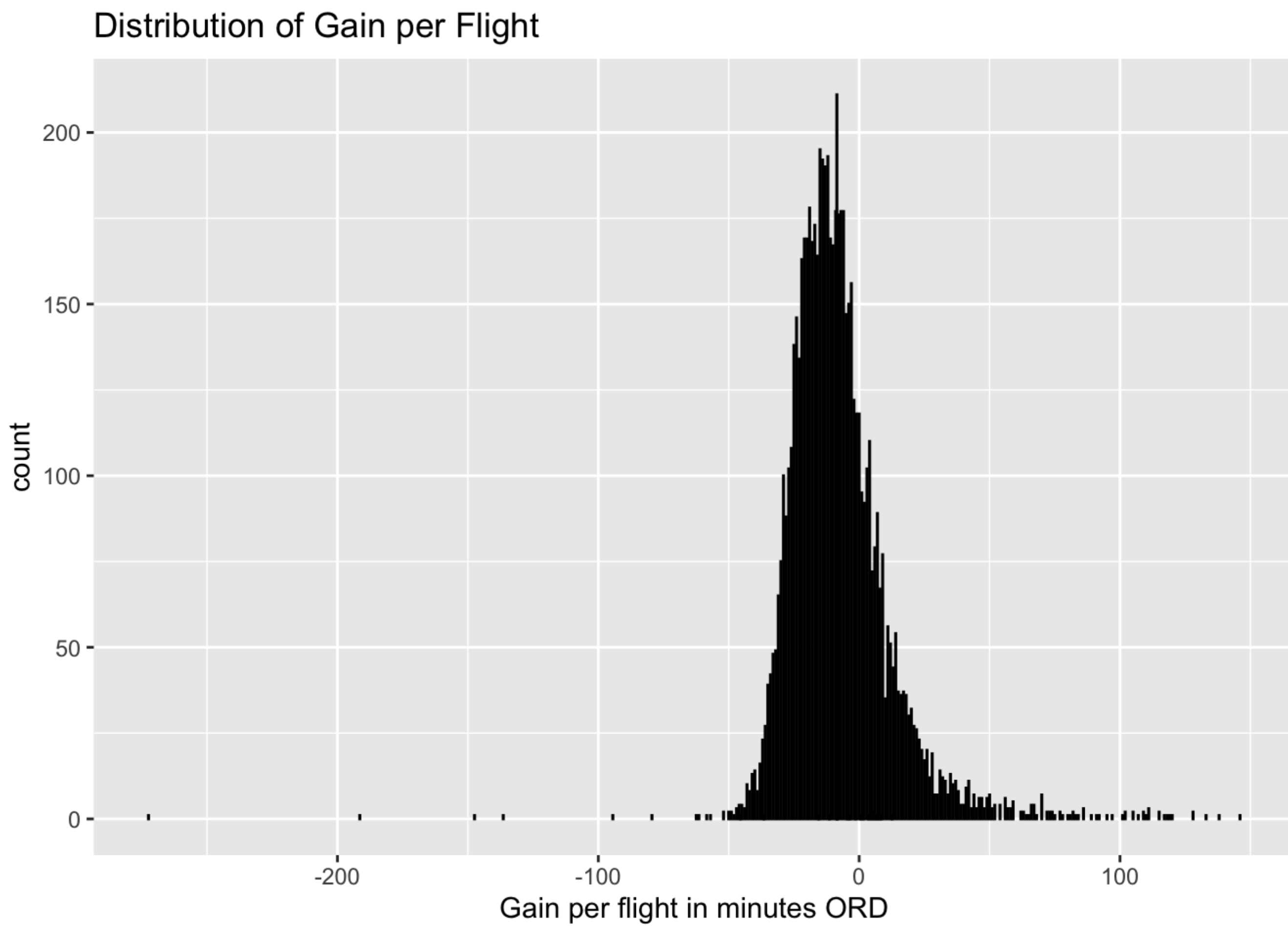
```
## # A tibble: 10 × 7
## # Groups:   very_late [2]
##    very_late dest  Mean_gain Median_gain StandardDeviation_gain MinGain MaxGain
##    <lgl>     <chr>     <dbl>       <dbl>                  <dbl>   <dbl>   <dbl>
##  1 FALSE     DEN       -7.65          -9                   18.6   -66       136
##  2 FALSE     IAH       -7.11          -9                   17.7   -59       117
##  3 FALSE     LAX       -8.17          -9                   20.7   -73       118
##  4 FALSE     ORD       -8.29         -10                   17.5   -57       138
##  5 FALSE     SFO       -9.20         -10                   21.0   -70       143
##  6 TRUE      DEN       -5.99         -11                   31.4  -271.      133
##  7 TRUE      IAH       -5.63          -9                   24.9  -228.      128
##  8 TRUE      LAX       -5.22          -9                   29.2   -90.4     145
##  9 TRUE      ORD       -5.94         -10                   27.9  -272.      146
## 10 TRUE      SFO       -6.69         -10                   34.4  -389.      165
```

```
UA_flight_ORD <- UA_flight %>%
  filter(dest == 'ORD')
```

analysis for ORD

```
#Create a bar plot
ggplot(data = UA_flight_ORD , aes(x= gain ))+
  geom_bar(color = 'black') +
  labs(x = "Gain per flight in minutes ORD", title = "Distribution of Gain per Flight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



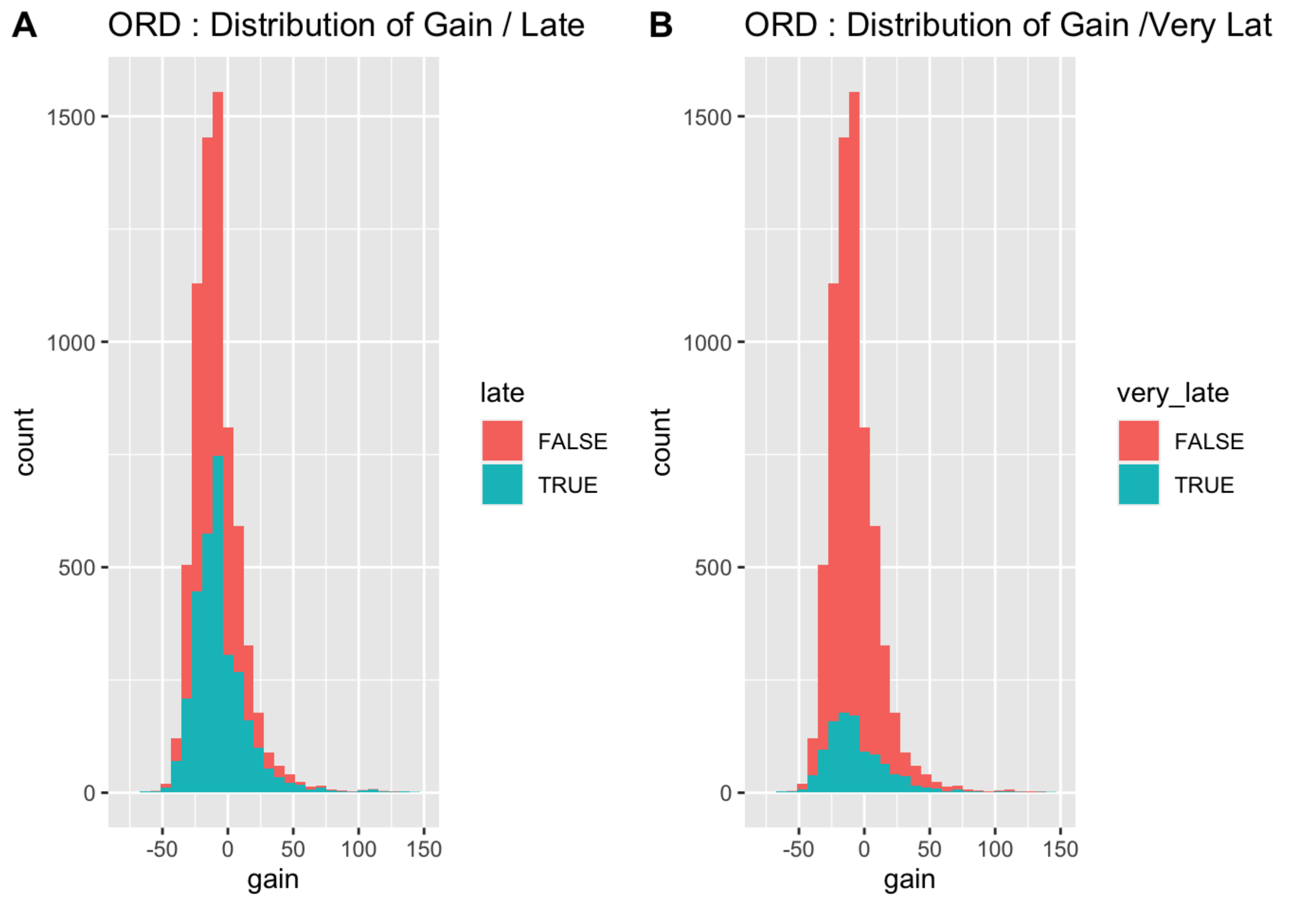

```
h1 = ggplot(UA_flight_ORD,aes(gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'ORD : Distribution of Gain / Late')+
  xlim(-80,150)
h2 = ggplot(UA_flight_ORD,aes(gain,fill = very_late))+
  geom_histogram(bins = 30)+
  labs(title = 'ORD : Distribution of Gain /Very Late')+
  xlim(-80,150)
plot_grid(h1, h2, labels="AUTO")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
h1 = ggplot(UA_flight_ORD,aes(gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'ORD : Boxplot of Gain / Late')
```

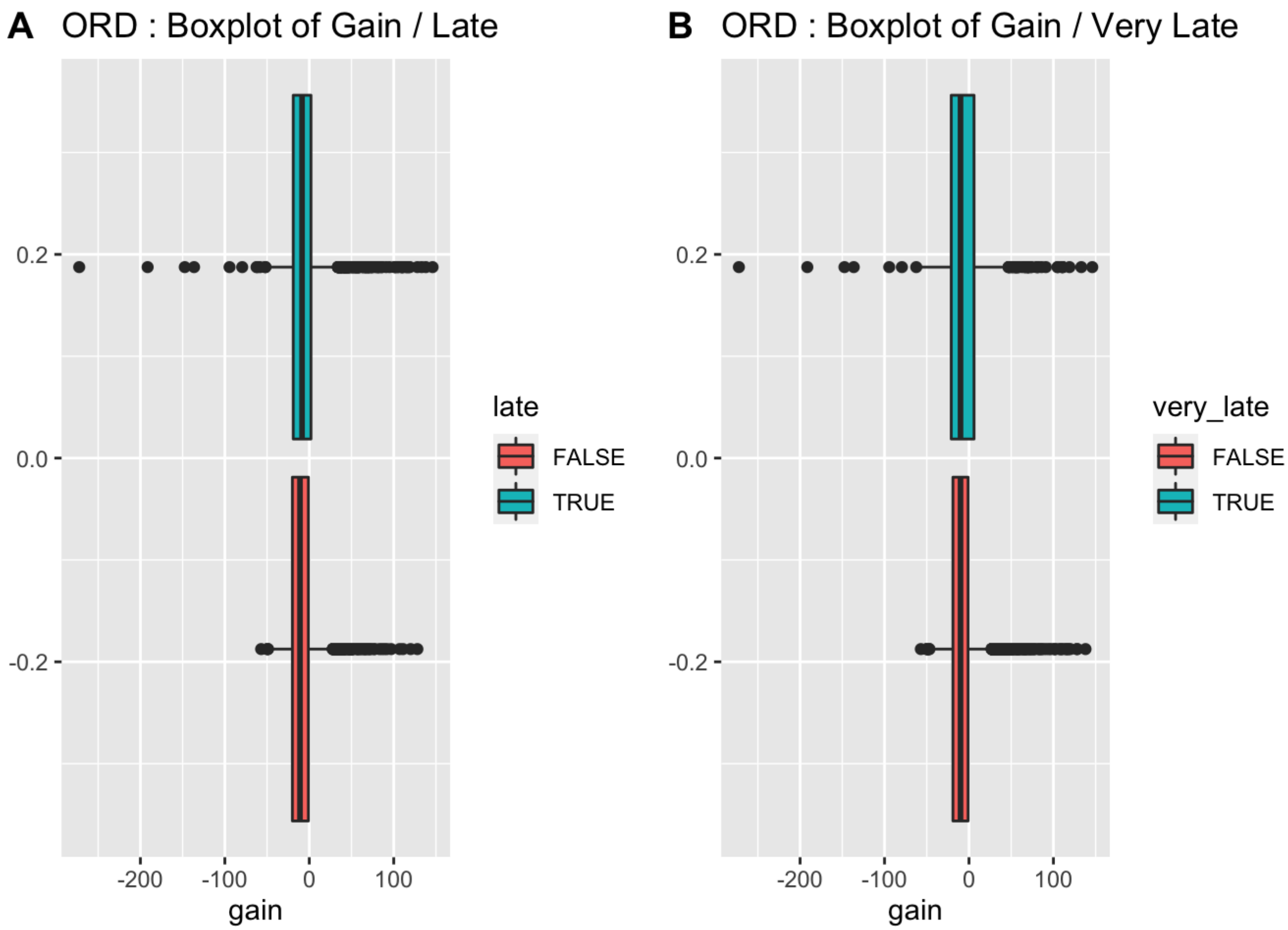
```
## Warning: Ignoring unknown parameters: bins
```

```
h2 = ggplot(UA_flight_ORD,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'ORD : Boxplot of Gain / Very Late')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



ORD Hypothesis Testing for Late variable

H0 : Average gain for late and flight on time is same for ORD destination average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same for ORD destination average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=UA_flight_ORD, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -5.0967, df = 5513.5, p-value = 3.572e-07
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -3.415404 -1.517867
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.024221           -6.557586
```

ORD Hypothesis testing for Very Late variable

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=UA_flight_ORD, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -2.6232, df = 1186.8, p-value = 0.008821
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -4.0988779 -0.5911423
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.285262           -5.940252
```

```
without_outlier_ORD <- without_outlier %>%
  filter(dest =='ORD')
```

Hypothesis Testing for Late variable Without Outlier

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=without_outlier_ORD, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -3.8144, df = 6056, p-value = 0.0001379
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.219516 -0.712596
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.732369           -8.266313
```

Hypothesis testing for Very Late variable Without Outlier

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=without_outlier_ORD, alternative = "two.sided")
```

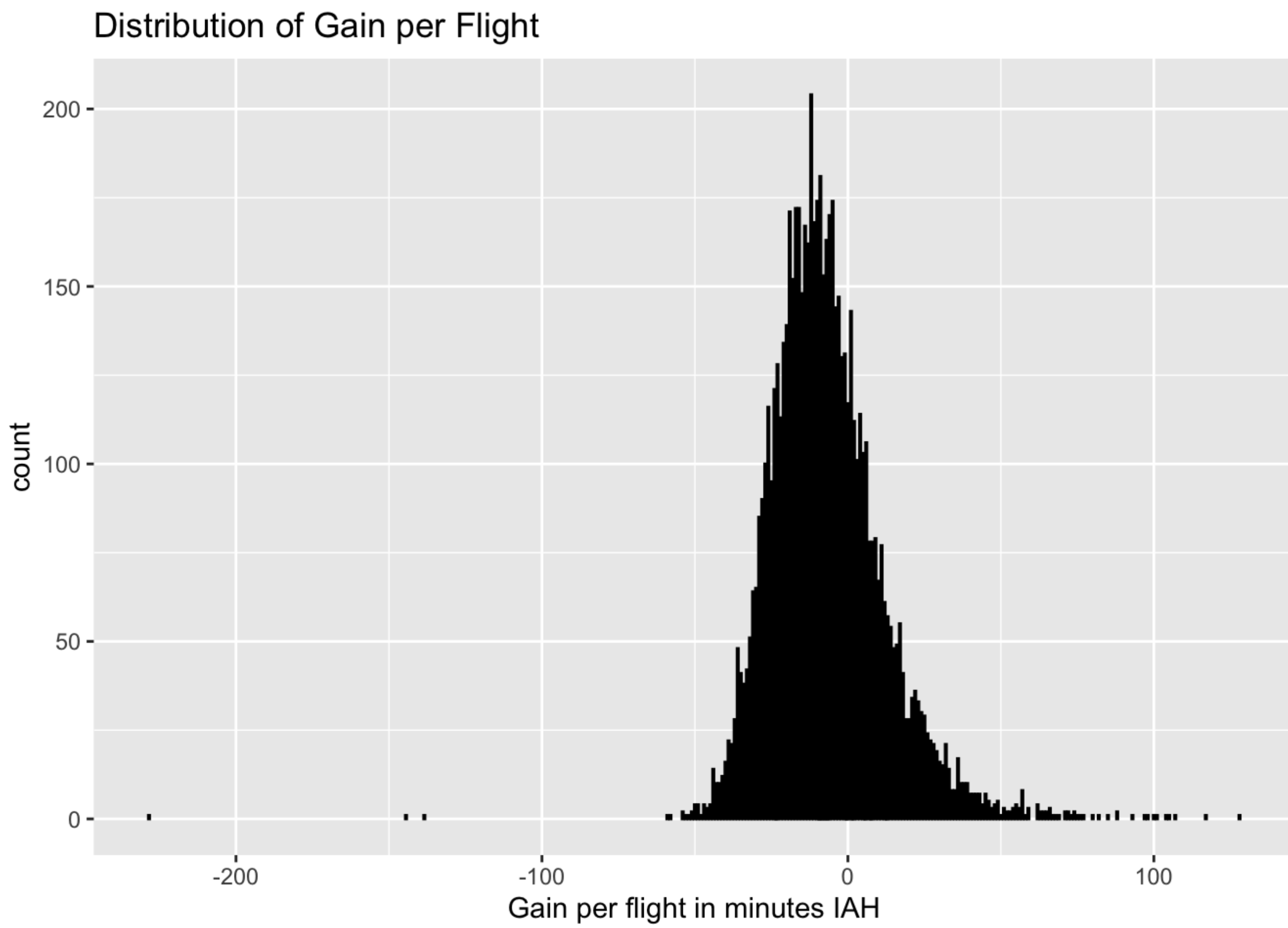
```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -1.8845, df = 1202, p-value = 0.05975
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.48130035  0.04998776
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.267971           -8.052315
```

Analysis for IAH Airport

```
UA_flight_IAH <- UA_flight %>%
  filter(dest == 'IAH')
```

```
#Create a bar plot
ggplot(data = UA_flight_IAH , aes(x= gain ))+
  geom_bar(color = 'black') +
  labs(x = "Gain per flight in minutes IAH", title = "Distribution of Gain per Flight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



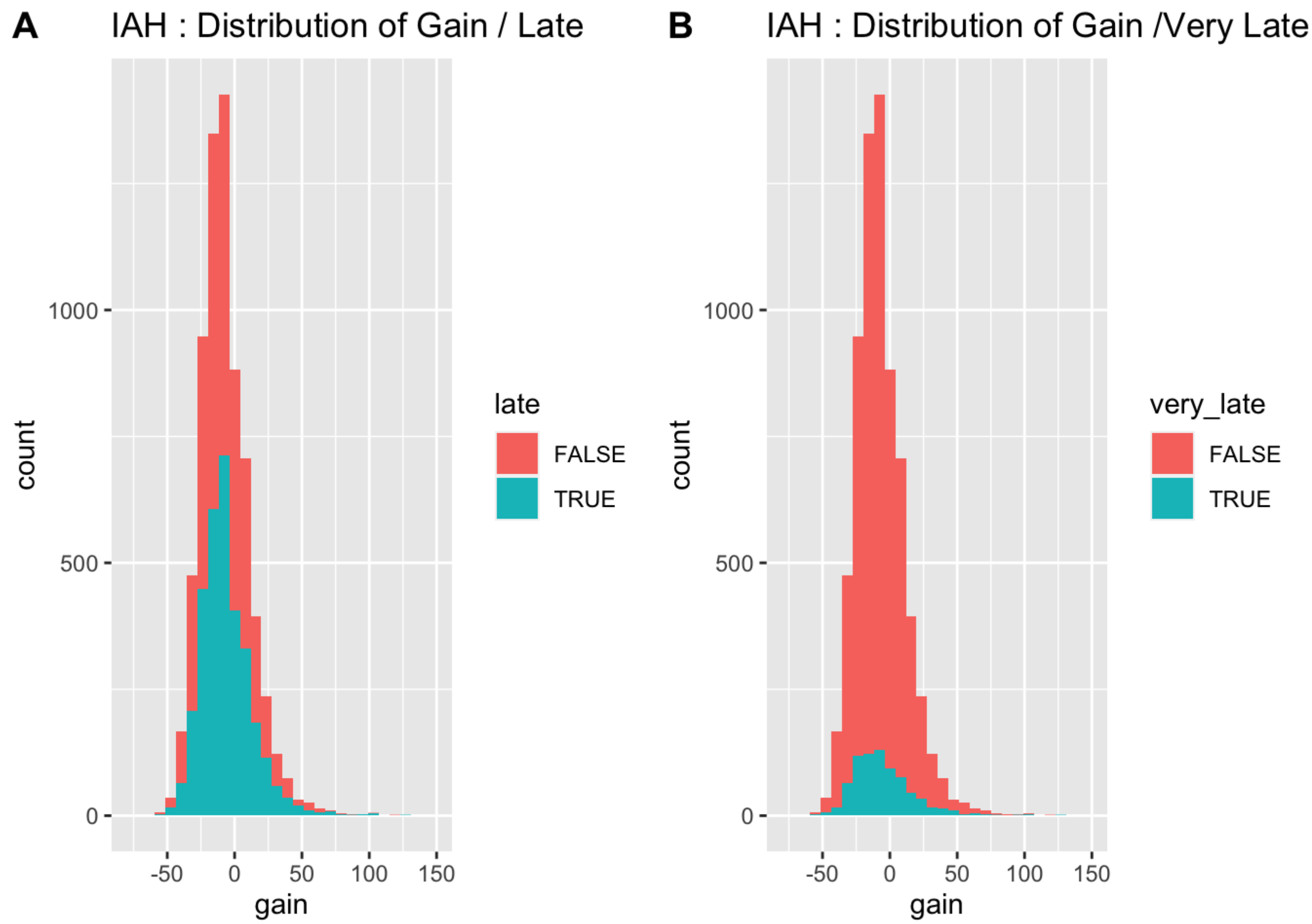

```
h1 = ggplot(UA_flight_IAH,aes(gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'IAH : Distribution of Gain / Late')+
  xlim(-80,150)
h2 = ggplot(UA_flight_IAH,aes(gain,fill = very_late))+
  geom_histogram(bins = 30)+
  labs(title = 'IAH : Distribution of Gain /Very Late')+
  xlim(-80,150)
plot_grid(h1, h2, labels="AUTO")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
h1 = ggplot(UA_flight_IAH,aes(gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'IAH : Boxplot of Gain / Late')
```

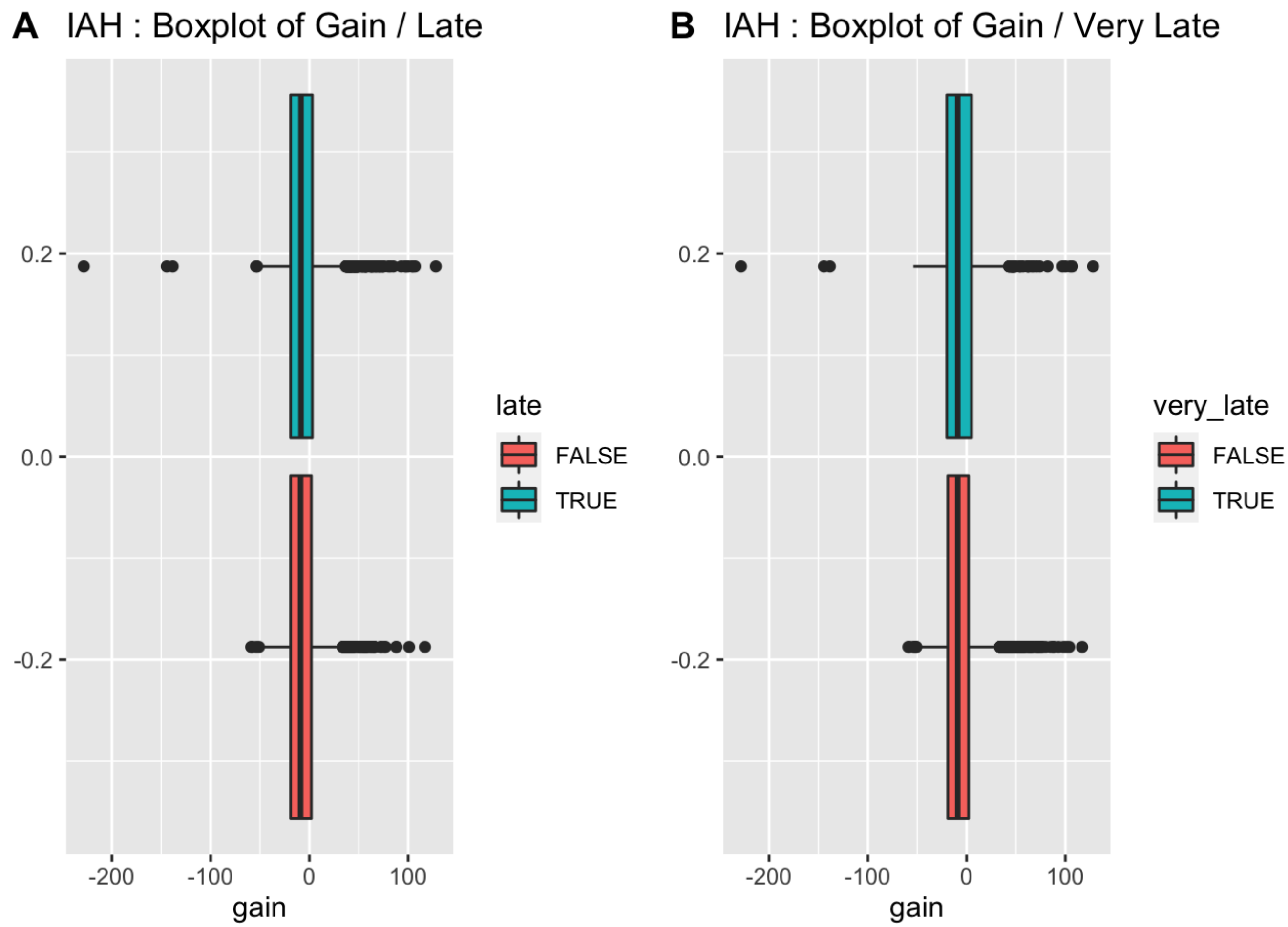
```
## Warning: Ignoring unknown parameters: bins
```

```
h2 = ggplot(UA_flight_IAH,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'IAH : Boxplot of Gain / Very Late')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



IAH Hypothesis Testing for Late variable

H0 : Average gain for late and flight on time is same for IAH destination average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same for IAH destination average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=UA_flight_IAH, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -2.1916, df = 6641.2, p-value = 0.02844
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.8721676 -0.1043001
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -7.409630           -6.421396
```

IAH Hypothesis testing for Very Late variable

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=UA_flight_IAH, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -1.6063, df = 872.18, p-value = 0.1086
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -3.3005792  0.3296066
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -7.112040           -5.626554
```

```
without_outlier_IAH <- without_outlier %>%
  filter(dest =='IAH')
```

Hypothesis Testing for Late variable Without Outlier

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=without_outlier_IAH, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -2.2212, df = 6723.5, p-value = 0.02637
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.6798920 -0.1048095
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.079113           -7.186763
```

Hypothesis testing for Very Late variable Without Outlier

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=without_outlier_IAH, alternative = "two.sided")
```

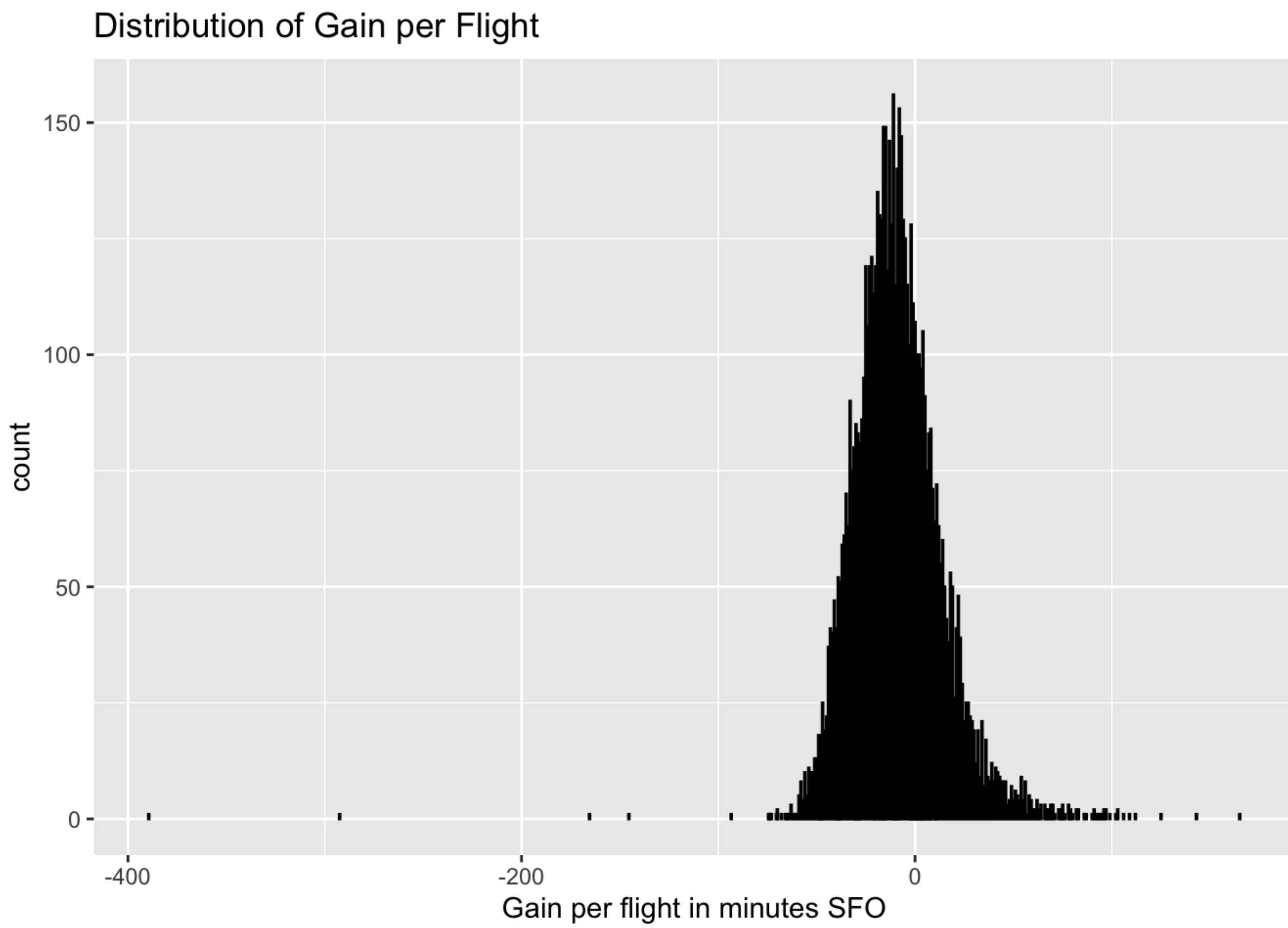
```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -1.4181, df = 894.15, p-value = 0.1565
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.4466507  0.3940638
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -7.774304           -6.748011
```

Analysis for IAH Airport

```
UA_flight_SFO <- UA_flight %>%
  filter(dest == 'SFO')
```

```
#Create a bar plot
ggplot(data = UA_flight_SFO , aes(x= gain ))+
  geom_bar(color = 'black') +
  labs(x = "Gain per flight in minutes SFO", title = "Distribution of Gain per Flight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



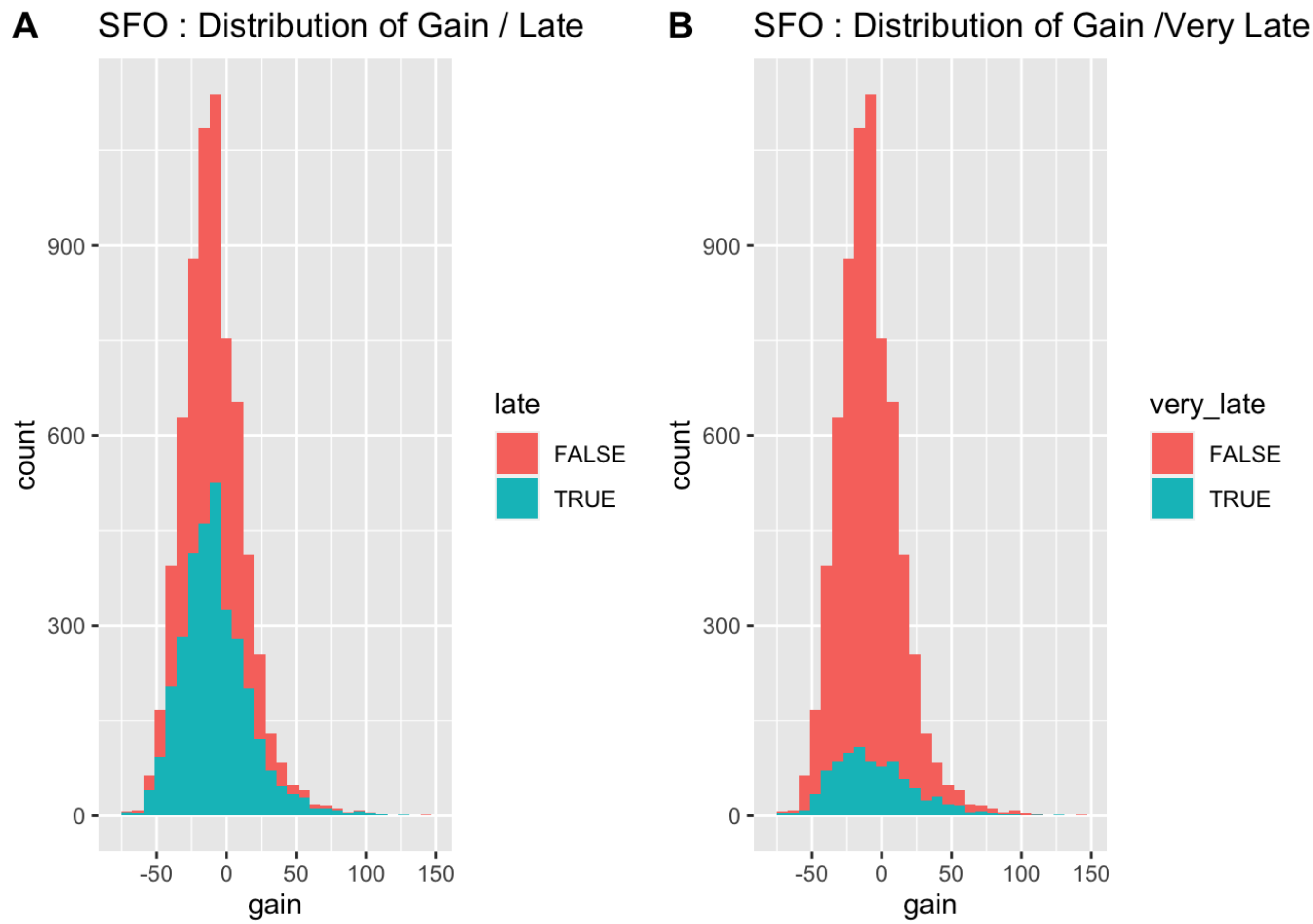

```
h1 = ggplot(UA_flight_SFO,aes(gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'SFO : Distribution of Gain / Late')+
  xlim(-80,150)
h2 = ggplot(UA_flight_SFO,aes(gain,fill = very_late))+
  geom_histogram(bins = 30)+
  labs(title = 'SFO : Distribution of Gain /Very Late')+
  xlim(-80,150)
plot_grid(h1, h2, labels="AUTO")
```

```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
h1 = ggplot(UA_flight_SFO,aes(gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'SFO : Boxplot of Gain / Late')
```

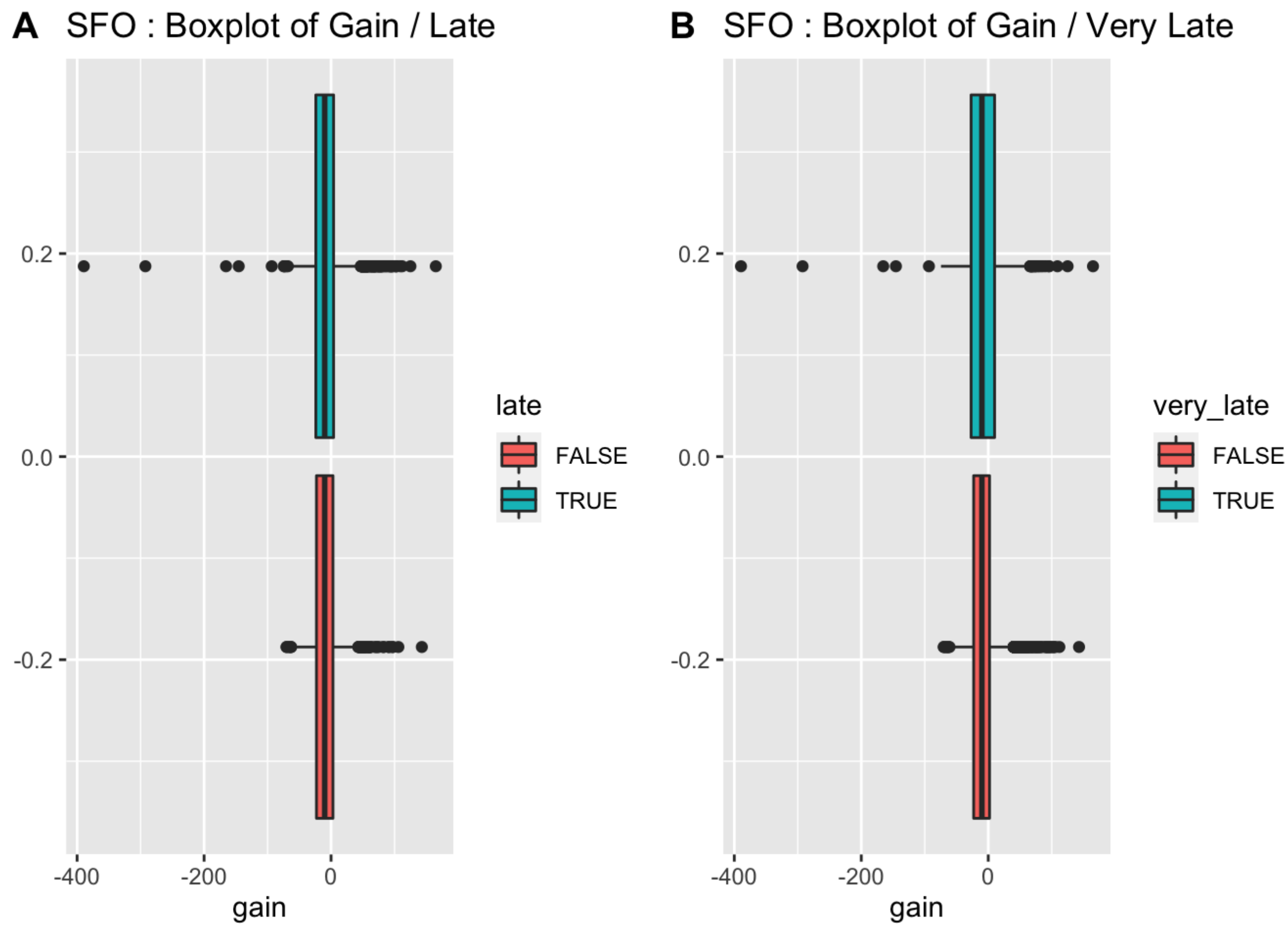
```
## Warning: Ignoring unknown parameters: bins
```

```
h2 = ggplot(UA_flight_SFO,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'SFO : Boxplot of Gain / Very Late')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



SFO Hypothesis Testing for Late variable

H0 : Average gain for late and flight on time is same for IAH destination average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same for IAH destination average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=UA_flight_SFO, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -1.4562, df = 5976.7, p-value = 0.1454
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.9555410  0.2885874
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.264197           -8.430720
```

SFO Hypothesis testing for Very Late variable

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=UA_flight_SFO, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -2.0998, df = 978.29, p-value = 0.036
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -4.8458829 -0.1639685
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.197185           -6.692259
```

```
without_outlier_SFO <- without_outlier %>%
  filter(dest =='SFO')
```

Hypothesis Testing for Late variable Without Outlier

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=without_outlier_SFO, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = 0.43502, df = 6315.5, p-value = 0.6636
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -0.7400808  1.1622203
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.944578          -10.155647
```

Hypothesis testing for Very Late variable Without Outlier

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=without_outlier_SFO, alternative = "two.sided")
```

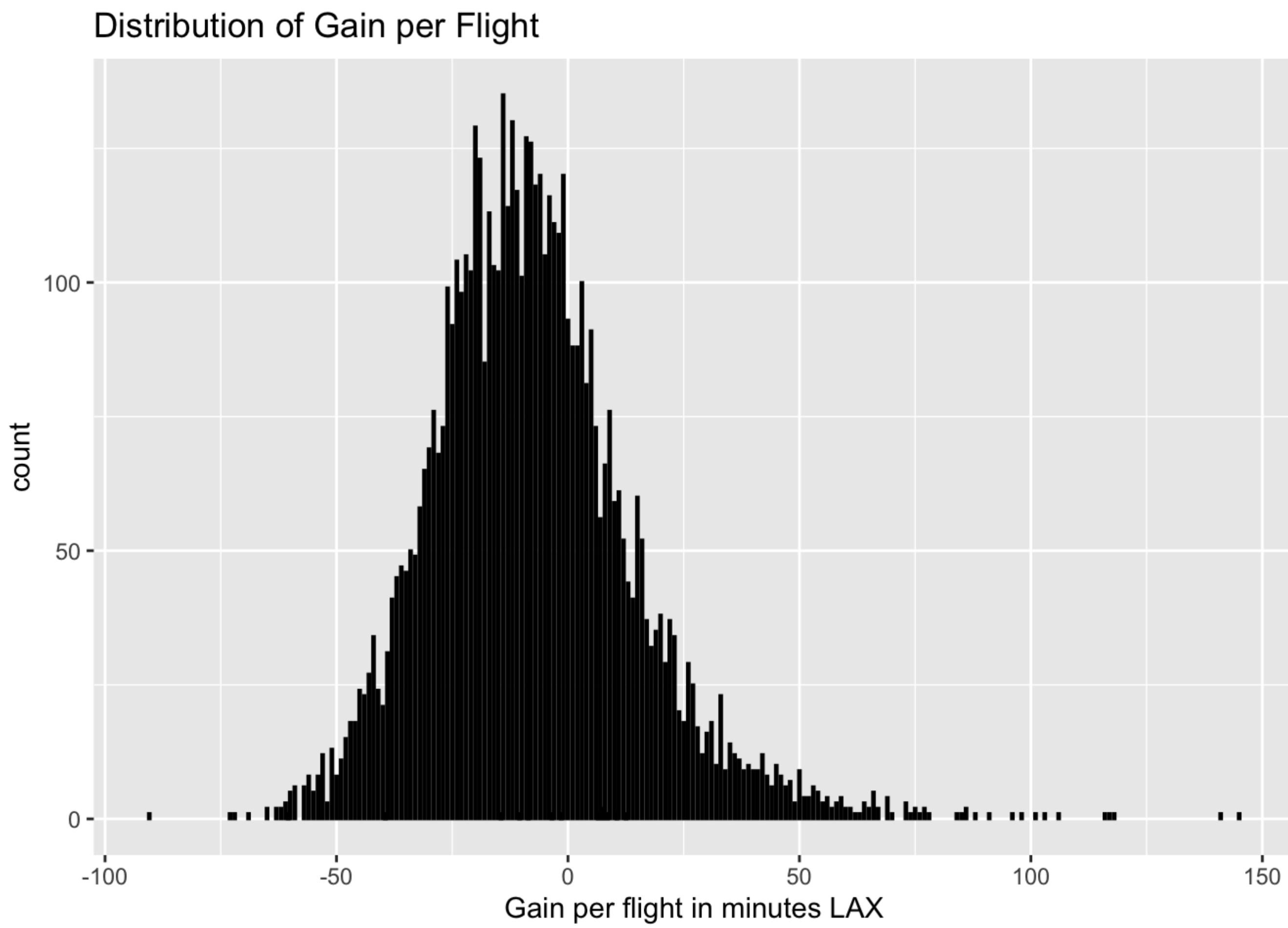
```
## 
##  Welch Two Sample t-test
## 
## data:  gain by very_late
## t = -1.3145, df = 979.95, p-value = 0.189
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.8380858  0.5611672
## sample estimates:
## mean in group FALSE  mean in group TRUE 
##          -10.183259           -9.044799
```

Analysis for LAX Airport

```
UA_flight_LAX <- UA_flight %>%
  filter(dest == 'LAX')
```

```
#Create a bar plot
ggplot(data = UA_flight_LAX , aes(x= gain ))+
  geom_bar(color = 'black') +
  labs(x = "Gain per flight in minutes LAX", title = "Distribution of Gain per Flight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



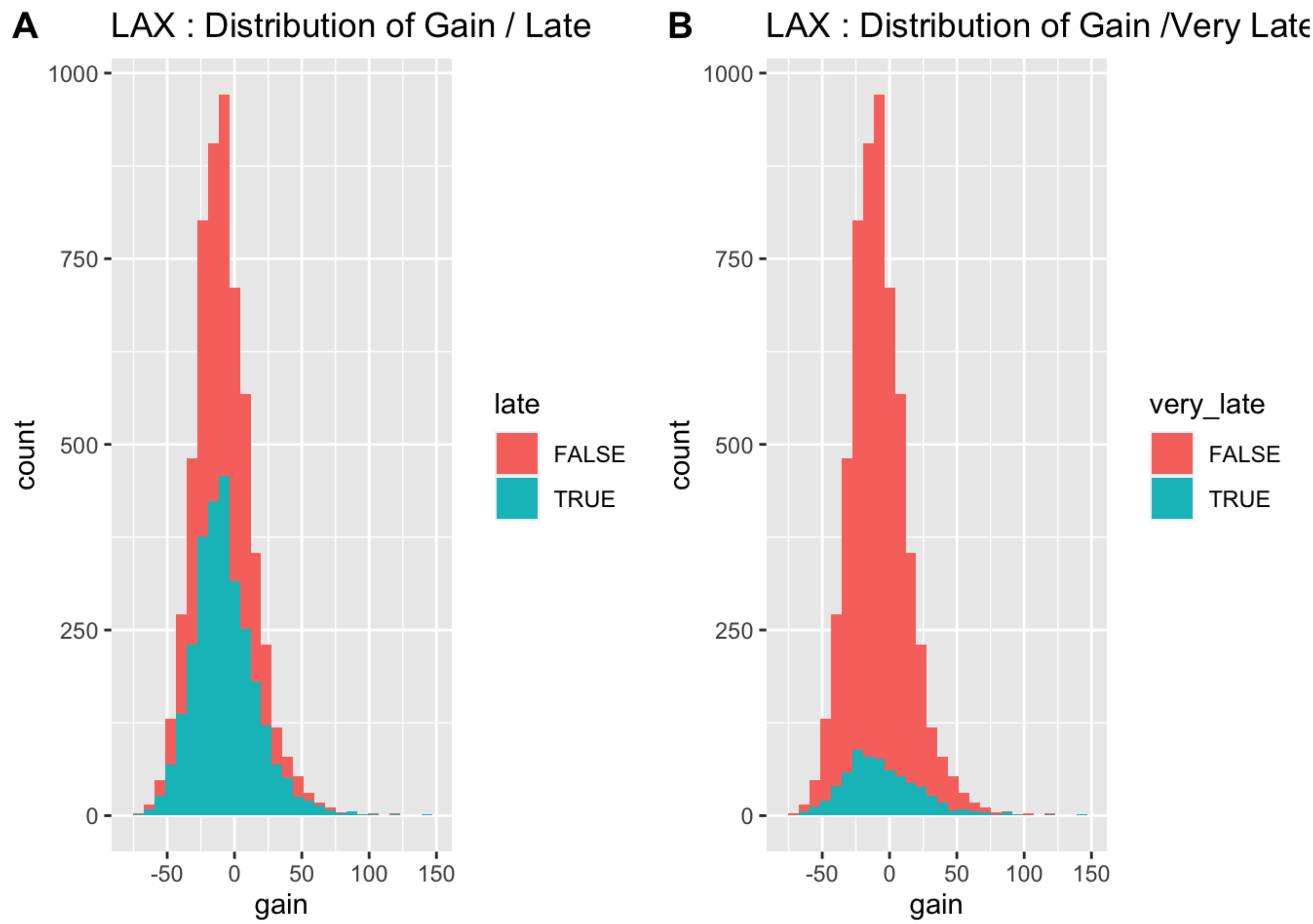

```
h1 = ggplot(UA_flight_LAX,aes(gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'LAX : Distribution of Gain / Late')+
  xlim(-80,150)
h2 = ggplot(UA_flight_LAX,aes(gain,fill = very_late))+
  geom_histogram(bins = 30)+
  labs(title = 'LAX : Distribution of Gain /Very Late')+
  xlim(-80,150)
plot_grid(h1, h2, labels="AUTO")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
h1 = ggplot(UA_flight_LAX,aes(gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'LAX : Boxplot of Gain / Late')
```

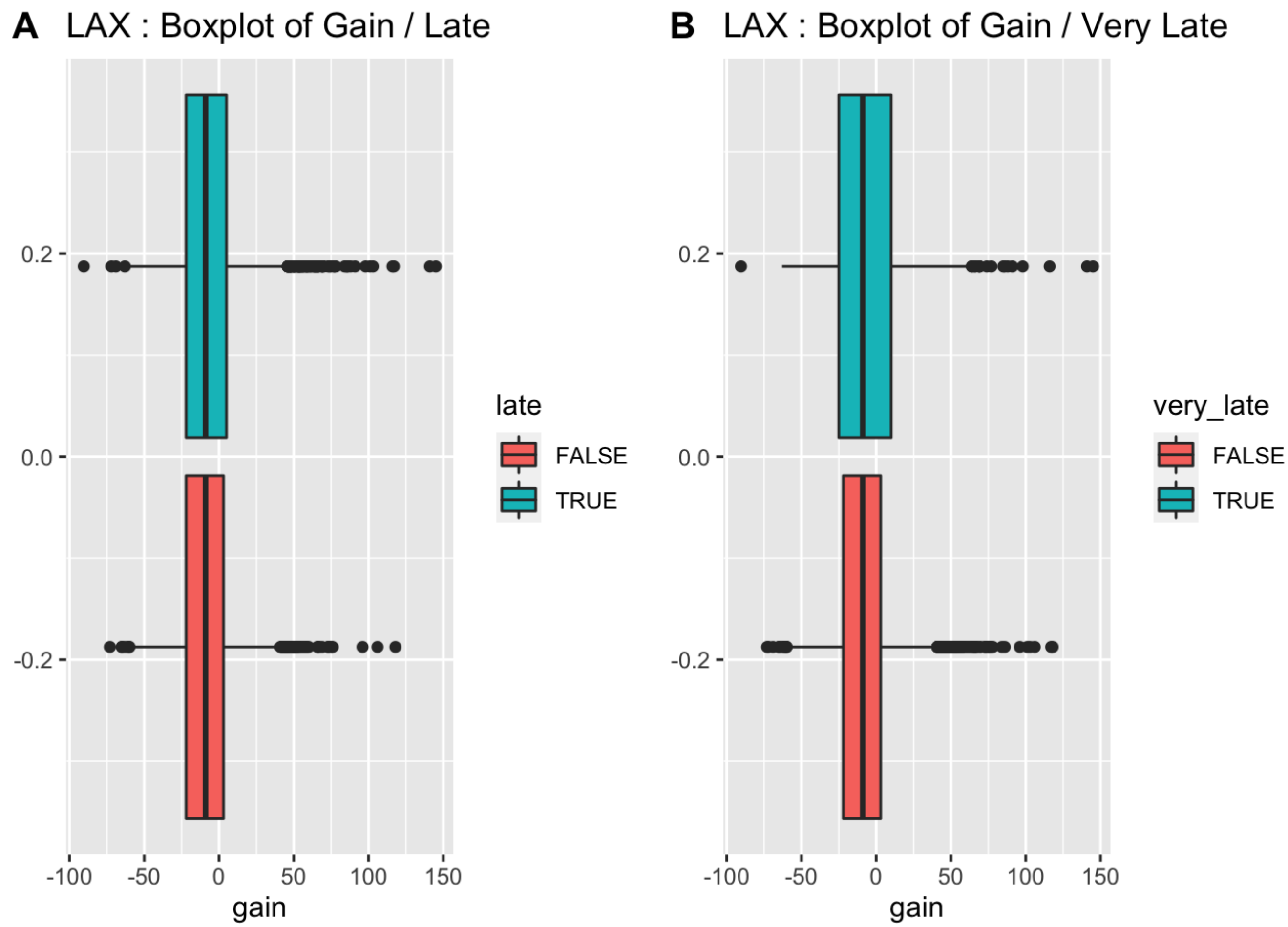
```
## Warning: Ignoring unknown parameters: bins
```

```
h2 = ggplot(UA_flight_LAX,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'LAX : Boxplot of Gain / Very Late')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



LAX Hypothesis Testing for Late variable

H0 : Average gain for late and flight on time is same for LAX destination average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same for LAX destination average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=UA_flight_LAX, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -2.3554, df = 5520.1, p-value = 0.01854
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.4889607 -0.2278215
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.495085           -7.136694
```

LAX Hypothesis testing for Very Late variable

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=UA_flight_LAX, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -2.5134, df = 743.69, p-value = 0.01217
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -5.2559574 -0.6460371
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.174897           -5.223900
```

```
without_outlier_LAX <- without_outlier %>%
  filter(dest =='LAX')
```

Hypothesis Testing for Late variable Without Outlier

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=without_outlier_LAX, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -0.71156, df = 5556.8, p-value = 0.4768
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.3888319  0.6491234
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.193529           -8.823675
```

Hypothesis testing for Very Late variable Without Outlier

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=without_outlier_LAX, alternative = "two.sided")
```

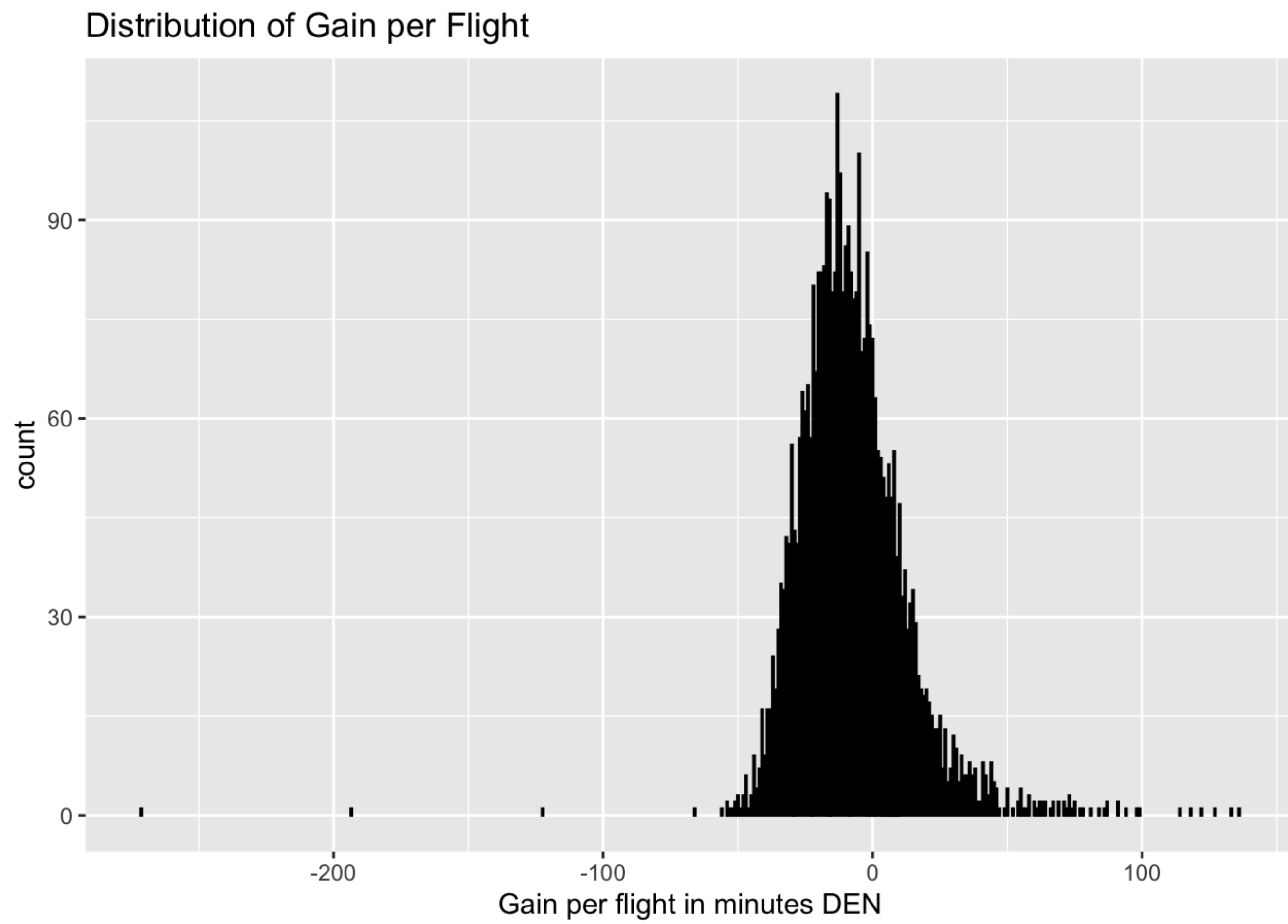
```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -0.3149, df = 731.82, p-value = 0.7529
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.221192  1.607126
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.050479           -8.743446
```

Analysis for DEN Airport

```
UA_flight_DEN <- UA_flight %>%
  filter(dest == 'DEN')
```

```
#Create a bar plot
ggplot(data = UA_flight_DEN , aes(x= gain ))+
  geom_bar(color = 'black') +
  labs(x = "Gain per flight in minutes DEN", title = "Distribution of Gain per Flight")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



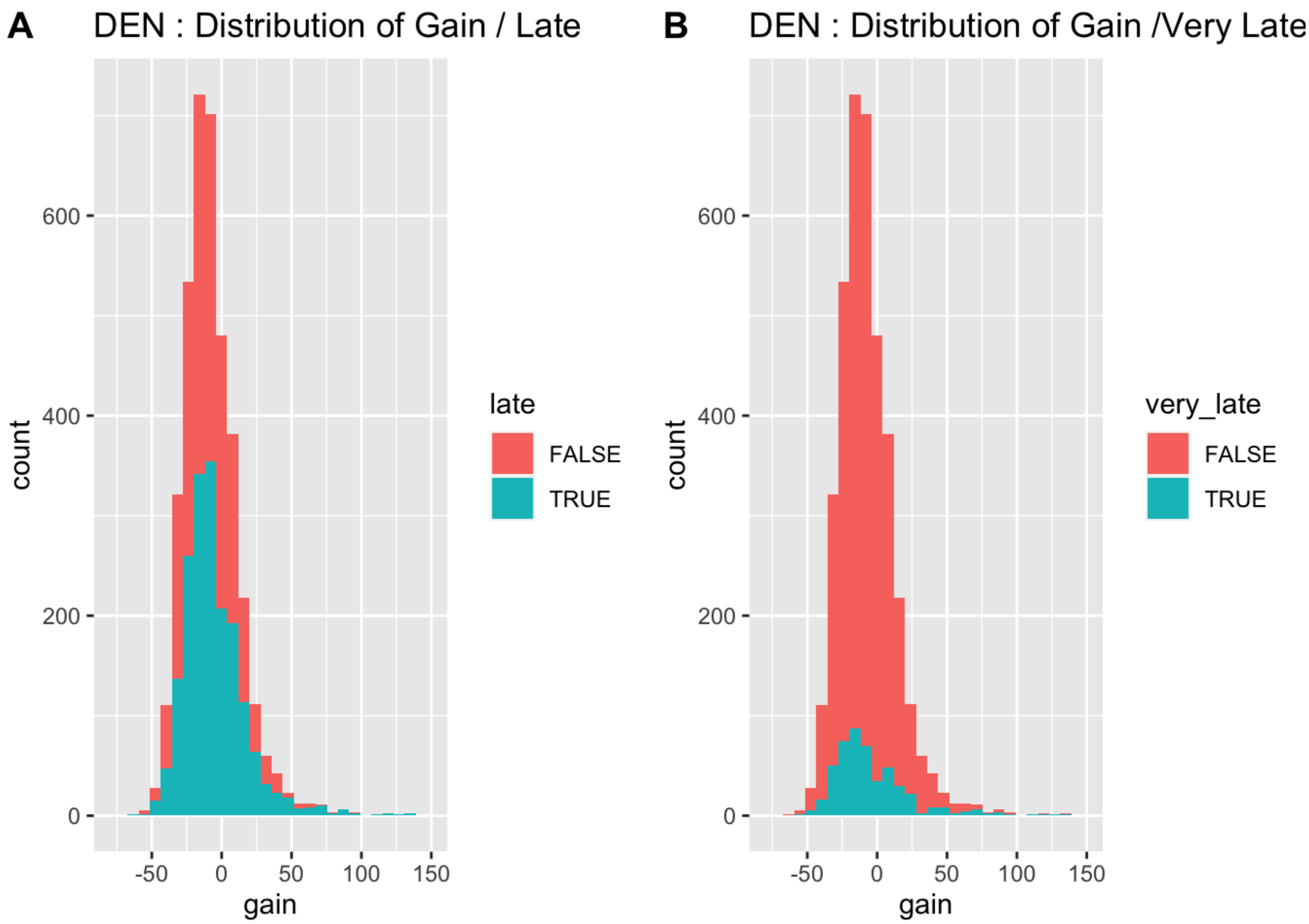

```
h1 = ggplot(UA_flight_DEN,aes(gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'DEN : Distribution of Gain / Late')+
  xlim(-80,150)
h2 = ggplot(UA_flight_DEN,aes(gain,fill = very_late))+
  geom_histogram(bins = 30)+
  labs(title = 'DEN : Distribution of Gain /Very Late')+
  xlim(-80,150)
plot_grid(h1, h2, labels="AUTO")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
h1 = ggplot(UA_flight_DEN,aes(gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'IAH : Boxplot of Gain / Late')
```

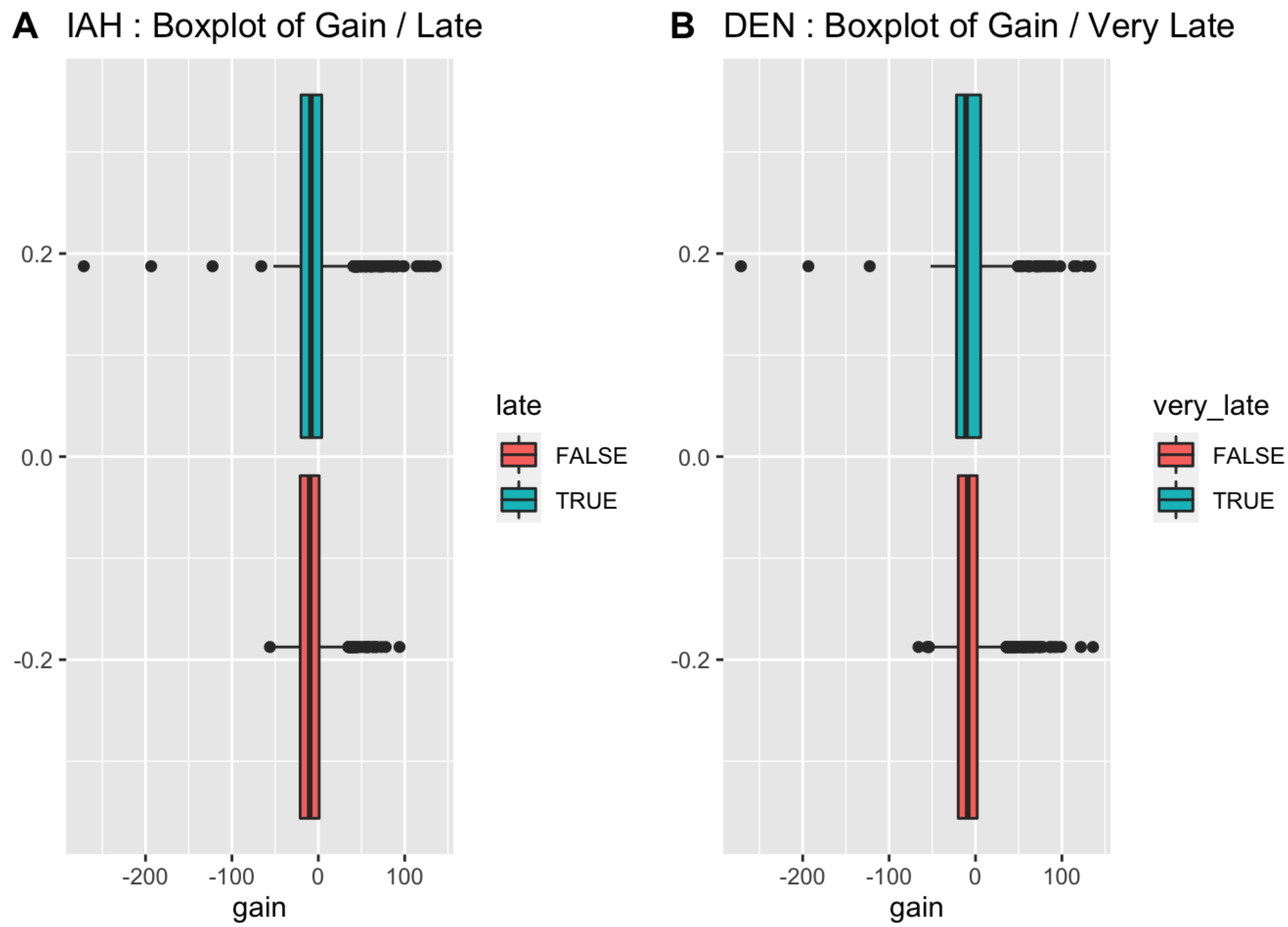
```
## Warning: Ignoring unknown parameters: bins
```

```
h2 = ggplot(UA_flight_DEN,aes(gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'DEN : Boxplot of Gain / Very Late')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```



IAH Hypothesis Testing for Late variable

H0 : Average gain for late and flight on time is same for DEN destination average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same for DEN destination average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=UA_flight_DEN, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -4.0555, df = 3442.1, p-value = 5.113e-05
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -4.05756 -1.41287
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.771821           -6.036606
```

IAH Hypothesis testing for Very Late variable

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=UA_flight_DEN, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = -1.133, df = 531.19, p-value = 0.2577
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -4.541608  1.219161
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -7.648294           -5.987070
```

```
without_outlier_DEN <- without_outlier %>%
  filter(dest =='DEN')
```

Hypothesis Testing for Late variable Without Outlier

H0 : Average gain for late and flight on time is same average(gain for late) = average(gain for flight on time) Ha : Average gain for late and flights on time is not same average(gain for late) != average(gain for flight on time)

```
t.test(gain~late,data=without_outlier_DEN, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by late
## t = -2.71, df = 3677, p-value = 0.006759
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -2.6206117 -0.4204751
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -9.264847           -7.744304
```

Hypothesis testing for Very Late variable Without Outlier

H0 : Average gain for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Average gain for very late and flight which were having delay less than 30 minutes is different average(gain for very late flights) != average(gain for flight where delays is less than 30 mintues)

```
t.test(gain~very_late,data=without_outlier_DEN, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by very_late
## t = 0.51092, df = 555.62, p-value = 0.6096
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -1.383834  2.356828
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           -8.470706           -8.957203
```

#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?

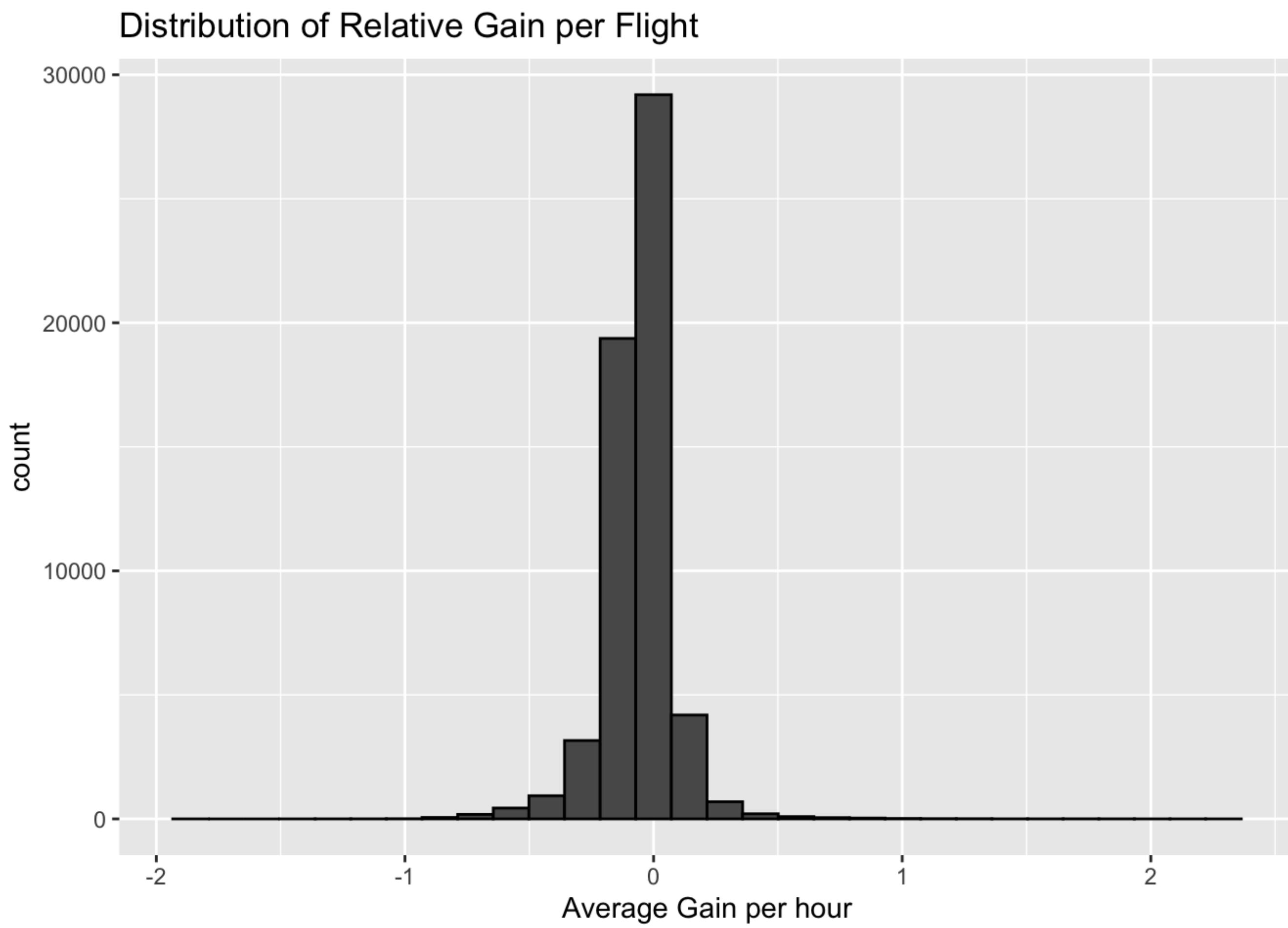
```
UA_flight <- UA_flight %>%
  mutate(rel_gain = UA_flight$gain/UA_flight$air_time)
glimpse(UA_flight)
```

```
## Rows: 58,665
## Columns: 23
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 517, 533, 554, 558, 558, 559, 607, 611, 623, 628, 643, …
## $ sched_dep_time <int> 515, 529, 558, 600, 600, 600, 607, 600, 627, 630, 646, …
## $ dep_delay      <dbl> 2, 4, -4, -2, -2, -1, 0, 11, -4, -2, -3, 8, 1, 1, -4, -…
## $ arr_time       <int> 830, 850, 740, 924, 923, 854, 858, 945, 933, 1016, 922,…
## $ sched_arr_time <int> 819, 830, 728, 917, 937, 902, 915, 931, 932, 947, 940, …
## $ arr_delay      <dbl> 11, 20, 12, 7, -14, -8, -17, 14, 1, 29, -18, -9, -6, -7…
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "…
## $ flight         <int> 1545, 1714, 1696, 194, 1124, 1187, 1077, 303, 496, 1665…
## $ tailnum        <chr> "N14228", "N24211", "N39463", "N29129", "N53441", "N765…
## $ origin         <chr> "EWR", "LGA", "EWR", "JFK", "EWR", "EWR", "EWR", "JFK",…
## $ dest           <chr> "IAH", "IAH", "ORD", "LAX", "SFO", "LAS", "MIA", "SFO",…
## $ air_time       <dbl> 227, 227, 150, 345, 361, 337, 157, 366, 229, 366, 146, …
## $ distance       <dbl> 1400, 1416, 719, 2475, 2565, 2227, 1085, 2586, 1416, 24…
## $ hour           <dbl> 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7…
## $ minute         <dbl> 15, 29, 58, 0, 0, 0, 7, 0, 27, 30, 46, 36, 45, 45, 0, 0…
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
## $ late           <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FA…
## $ very_late      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ gain           <dbl> 9, 16, 16, 9, -12, -7, -17, 3, 5, 31, -15, -17, -7, -8,…
## $ rel_gain       <dbl> 0.039647577, 0.070484581, 0.106666667, 0.026086957, -0.…
```

```
#Create a bar plot
ggplot(data = UA_flight , aes(x= rel_gain ))+
  geom_histogram(color = 'black') +
  labs(x = "Average Gain per hour", title = "Distribution of Relative Gain per Flight")
```

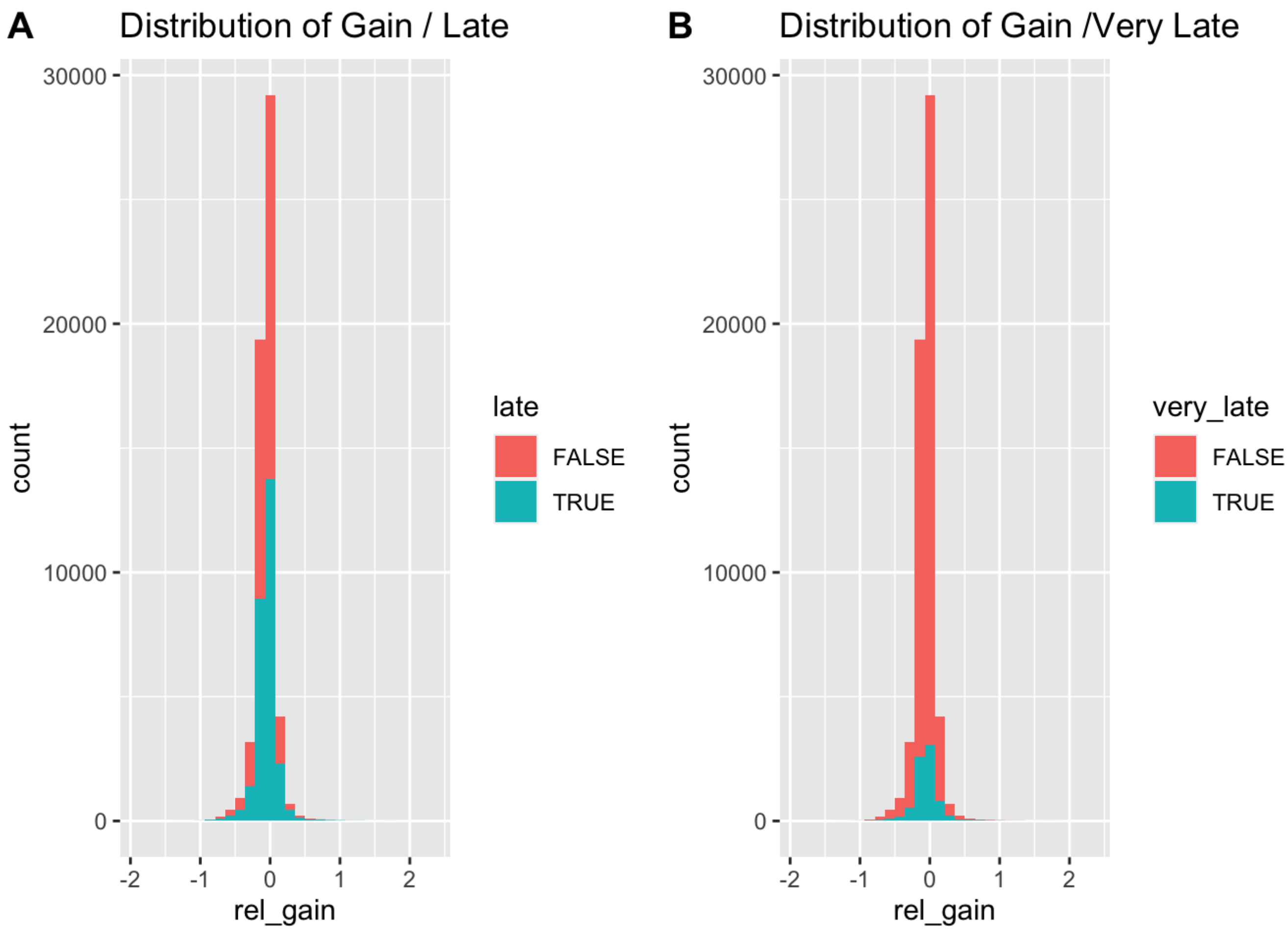
```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
h1 = ggplot(UA_flight,aes(rel_gain,fill = late))+
  geom_histogram(bins = 30)+
  labs(title = 'Distribution of Gain / Late')
h2 = ggplot(UA_flight,aes(rel_gain,fill = very_late))+
  geom_histogram(bins = 30)+
  labs(title = 'Distribution of Gain /Very Late')
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
h1 = ggplot(UA_flight,aes(rel_gain,fill = late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'Boxplot of Gain / Late')
```

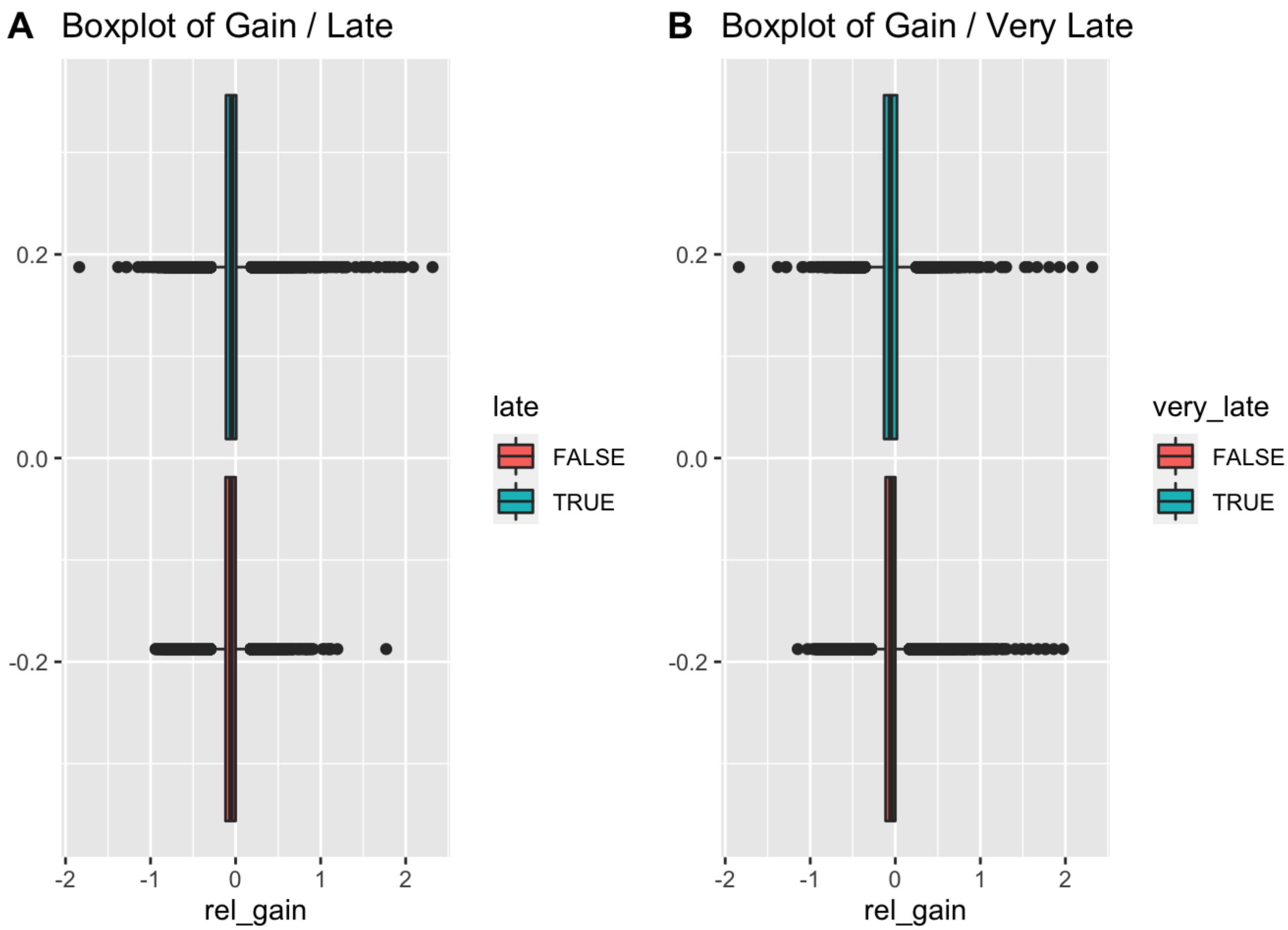
```
## Warning: Ignoring unknown parameters: bins
```

```
h2 = ggplot(UA_flight,aes(rel_gain,fill = very_late))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'Boxplot of Gain / Very Late')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
plot_grid(h1, h2, labels="AUTO")
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
UA_flight %>%
  group_by(late) %>%
  dplyr::summarize(Mean_gain_ = mean(rel_gain),
                   Median_gain_ = median(rel_gain),
                   StandardDeviation_gain_ = sd(rel_gain),
                   MinGain_ =min(rel_gain),
                   MaxGain_ = max(rel_gain)
                   )
```

```
## # A tibble: 2 × 6
##   late  Mean_gain_ Median_gain_ StandardDeviation_gain_ MinGain_ MaxGain_
##   <lgl>      <dbl>        <dbl>                   <dbl>    <dbl>    <dbl>
## 1 FALSE    -0.0663      -0.0553                   0.129   -0.941     1.77
## 2 TRUE     -0.0537      -0.0476                   0.154   -1.84      2.32
```

```
UA_flight %>%
  group_by(very_late) %>%
  dplyr::summarize(Mean_gain_ = mean(rel_gain),
                   Median_gain_ = median(rel_gain),
                   StandardDeviation_gain_ = sd(rel_gain),
                   MinGain_ =min(rel_gain),
                   MaxGain_ = max(rel_gain)
                   )
```

```
## # A tibble: 2 × 6
##   very_late Mean_gain_ Median_gain_ StandardDeviation_gain_ MinGain_ MaxGain_
##   <lgl>          <dbl>        <dbl>                   <dbl>    <dbl>    <dbl>
## 1 FALSE        -0.0611      -0.0512                   0.133    -1.15     1.97
## 2 TRUE         -0.0548      -0.0569                   0.188    -1.84     2.32
```

Hypothesis Testing for Late variable

H0 : Mean of average gain per hour for late and flight on time is same Mean(average gain per hour for late) = average(average gain per hour for flight on time) Ha : Average gain for late and flights on time is not same Mean(average gain per hour for late) != average(average gain per hour for flight on time)

```
t.test(rel_gain~late,data=UA_flight, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  rel_gain by late
## t = -10.664, df = 54692, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -0.01489259 -0.01026808
## sample estimates:
## mean in group FALSE  mean in group TRUE
##         -0.06631539         -0.05373506
```

Hypothesis testing for Very Late variable

H0 : Mean of average gain per hour for very late and flight which were having delay less than 30 minutes is same average(gain for very late flights) = average(gain for flight where delays is less than 30 mintues) Ha : Mean of average gain per hour for very late and flight which were having delay less than 30 minutes is different average(average gain per hour for very late flights) != average(verage gain per hour for flight where delays is less than 30 mintues)

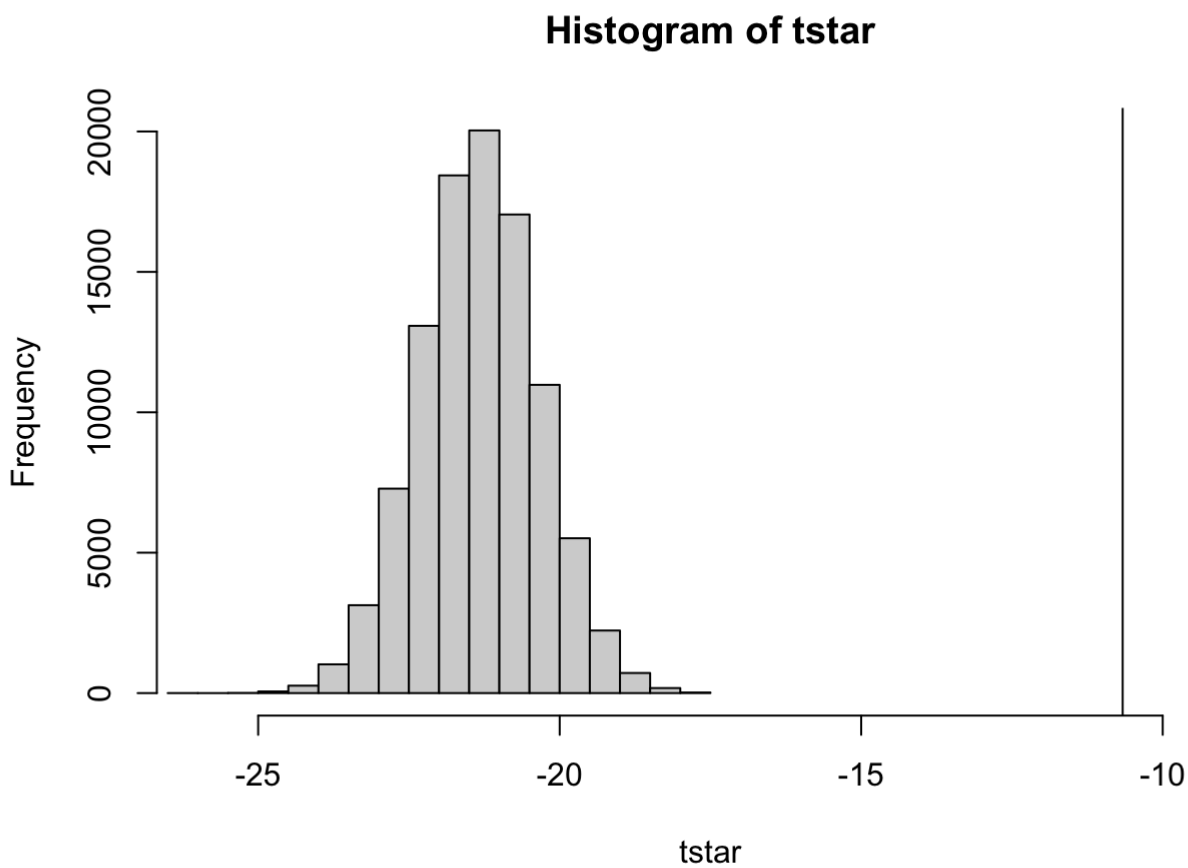
```
t.test(rel_gain~very_late,data=UA_flight, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  rel_gain by very_late
## t = -2.8313, df = 8798.2, p-value = 0.004646
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  -0.010692684 -0.001943938
## sample estimates:
## mean in group FALSE  mean in group TRUE
##         -0.06114270         -0.05482439
```

```
UA_flight_rellate <-subset(UA_flight,rel_gain,subset = late ==TRUE,drop=T)
UA_flight_relnotlate <- subset(UA_flight,rel_gain,subset = late ==FALSE,drop=T)

tstat <- function(x , y , mu)
{
  (mean(y) - mean(x) - mu)/sqrt(var(y)/length(y) + var(x)/length(x))

}
observed <- tstat(UA_flight_rellate,UA_flight_relnotlate,0)
observed
```

```
## [1] -10.66385
```

```
thetahat <- mean(UA_flight_rellate) - mean(UA_flight_relnotlate)
n1 <- length(UA_flight_rellate)
n2 <- length(UA_flight_relnotlate)

N <- 10^5-1
tstar <- numeric(N)
set.seed(5)
for (i in 1:N)
{
  boot1 <- sample(UA_flight_rellate,n1,replace = TRUE)
  boot2 <- sample(UA_flight_relnotlate,n2,replace = TRUE)
  tstar[i] <- tstat(boot1,boot2,thetahat)
}
hist(tstar,xlim = c(-26,-9))
abline(v=observed)
```



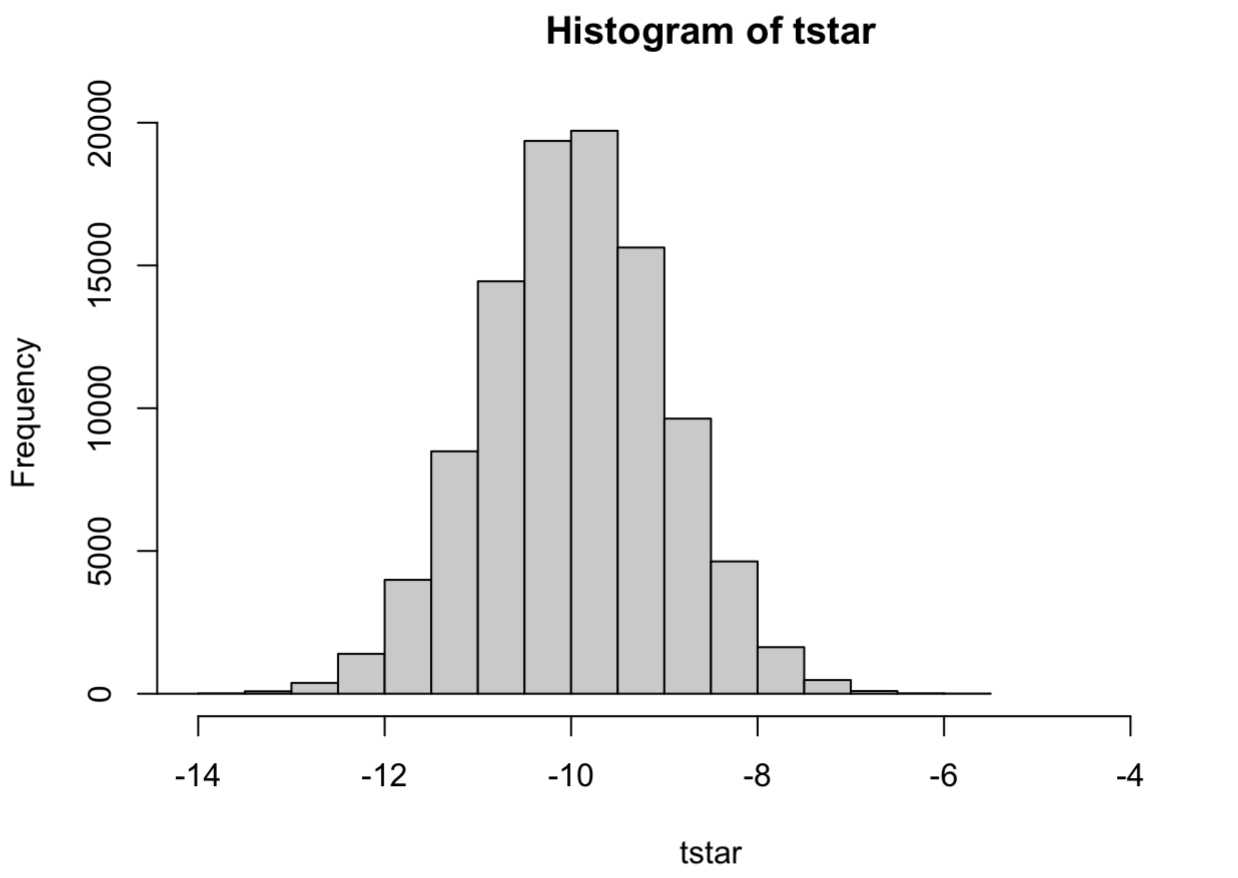

```
cat('The p-value is :',2*(sum(tstar >= observed)+1)/(N+1))
```

```
## The p-value is : 2e-05
```

```
UA_flight_relverylate <-subset(UA_flight,rel_gain,subset = very_late ==TRUE,drop=T)
UA_flight_relnotverylate <- subset(UA_flight,rel_gain,subset = very_late ==FALSE,drop=T)

tstat <- function(x , y , mu)
{
  (mean(y) - mean(x) - mu)/sqrt(var(y)/length(y) + var(x)/length(x))

}
observed <- tstat(UA_flight_relverylate,UA_flight_relnotverylate,0)
observed
```

```
## [1] -2.831346
```

```r
thetahat <- mean(UA_flight_relverylate) - mean(UA_flight_relnotverylate)
n1 <- length(UA_flight_rellate)
n2 <- length(UA_flight_relnotverylate)

N <- 10^5-1
tstar <- numeric(N)
set.seed(5)
for (i in 1:N)
{
  boot1 <- sample(UA_flight_relverylate,n1,replace = TRUE)
  boot2 <- sample(UA_flight_relnotverylate,n2,replace = TRUE)
  tstar[i] <- tstat(boot1,boot2,thetahat)
}
hist(tstar,xlim = c(-14,-3))
abline(v=observed)
```



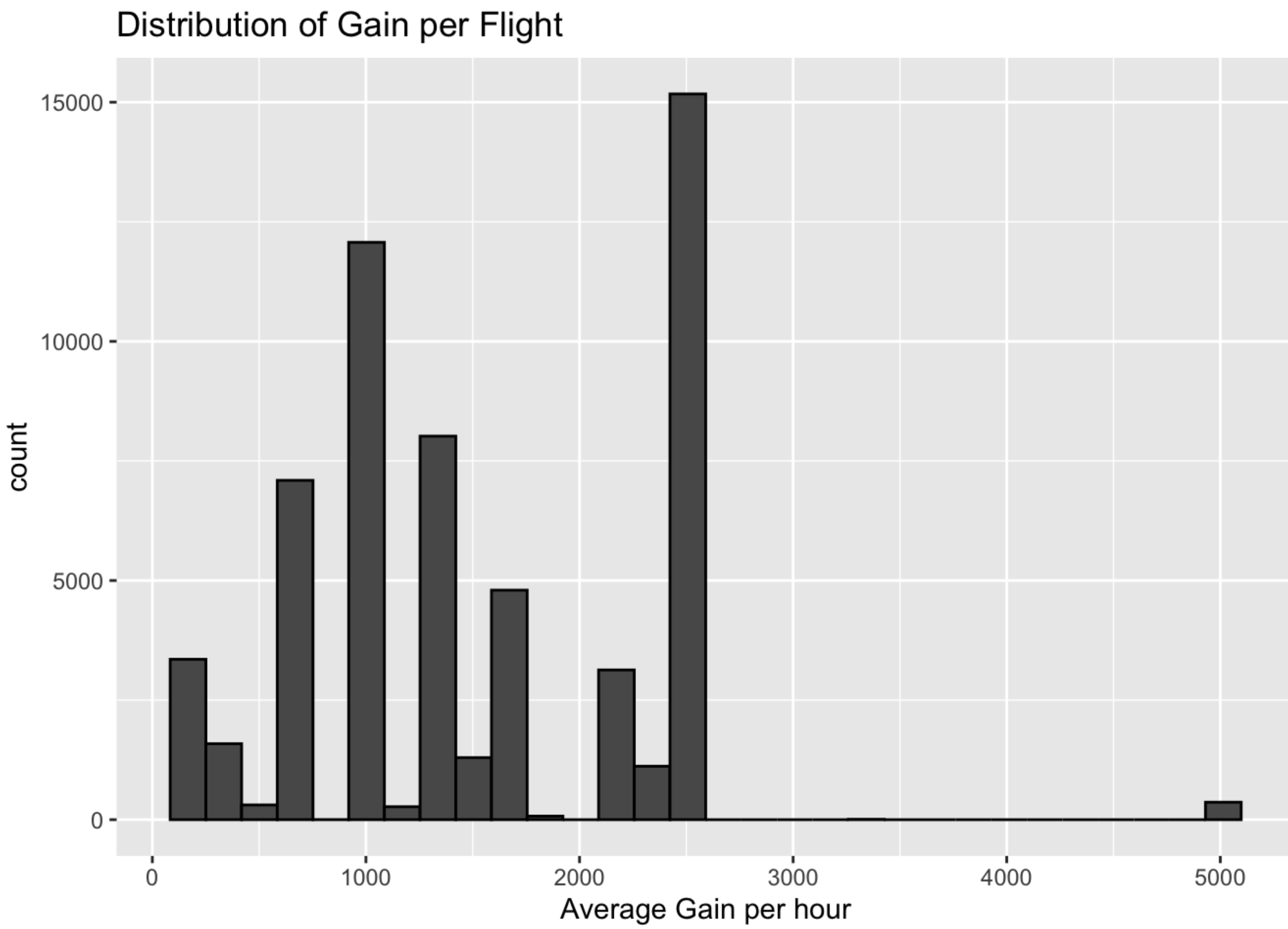

```r
cat('The p-value is :',2*(sum(tstar >= observed)+1)/(N+1))
```

```
## The p-value is : 2e-05
```

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

```
ggplot(data = UA_flight , aes(x= distance))+
  geom_histogram(color = 'black') +
  labs(x = "Average Gain per hour", title = "Distribution of Gain per Flight")
```

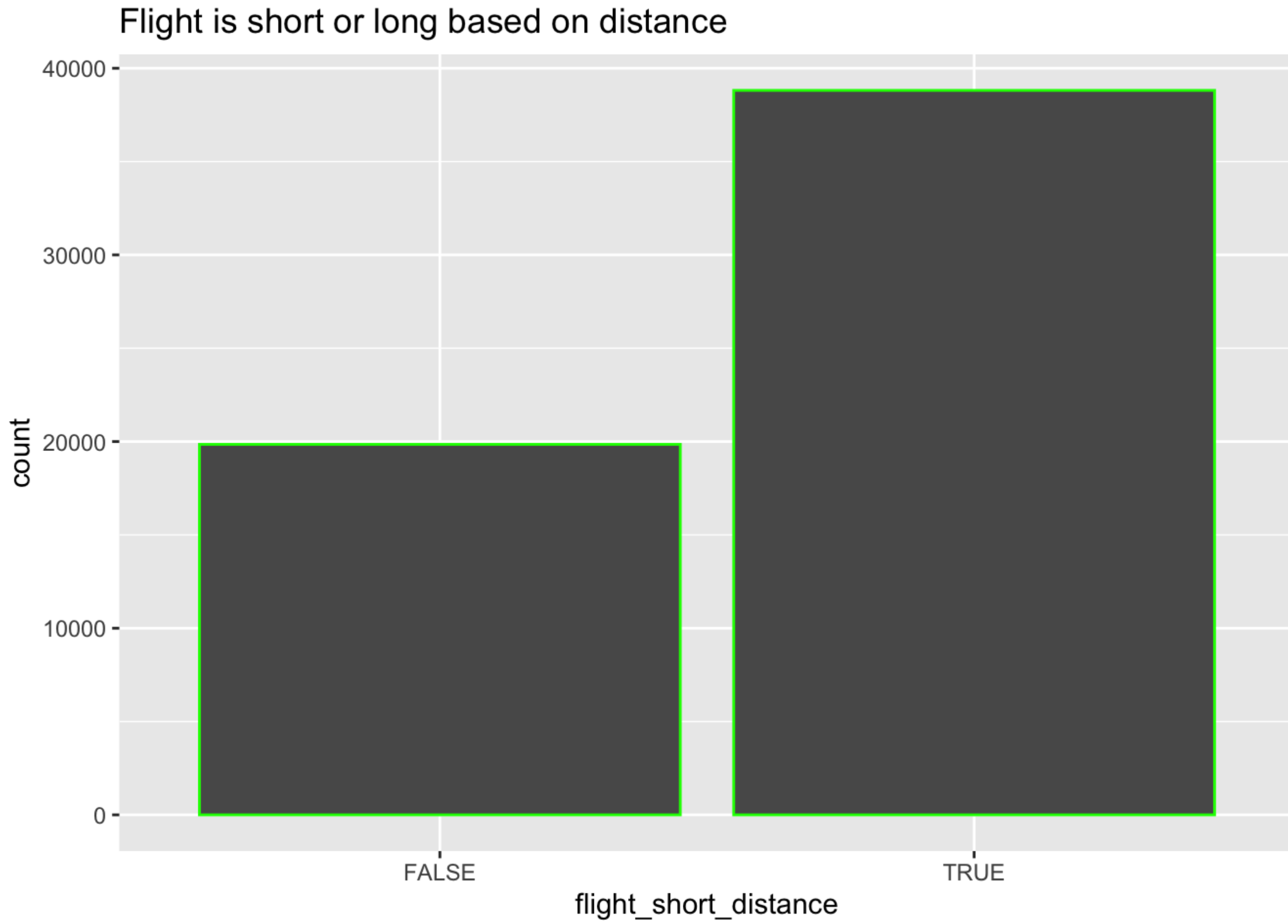
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
UA_flight <- UA_flight %>%
  mutate(flight_short_distance = case_when(distance < 1800 ~ TRUE,
                            distance >=1800 ~ FALSE ))
glimpse(UA_flight)
```

```
## Rows: 58,665
## Columns: 24
## $ year                <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, …
## $ month               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ day                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ dep_time            <int> 517, 533, 554, 558, 558, 559, 607, 611, 623, 628…
## $ sched_dep_time      <int> 515, 529, 558, 600, 600, 600, 607, 600, 627, 630…
## $ dep_delay           <dbl> 2, 4, -4, -2, -2, -1, 0, 11, -4, -2, -3, 8, 1, 1…
## $ arr_time            <int> 830, 850, 740, 924, 923, 854, 858, 945, 933, 101…
## $ sched_arr_time      <int> 819, 830, 728, 917, 937, 902, 915, 931, 932, 947…
## $ arr_delay           <dbl> 11, 20, 12, 7, -14, -8, -17, 14, 1, 29, -18, -9,…
## $ carrier             <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", …
## $ flight              <int> 1545, 1714, 1696, 194, 1124, 1187, 1077, 303, 49…
## $ tailnum             <chr> "N14228", "N24211", "N39463", "N29129", "N53441"…
## $ origin              <chr> "EWR", "LGA", "EWR", "JFK", "EWR", "EWR", "EWR",…
## $ dest                <chr> "IAH", "IAH", "ORD", "LAX", "SFO", "LAS", "MIA",…
## $ air_time            <dbl> 227, 227, 150, 345, 361, 337, 157, 366, 229, 366…
## $ distance            <dbl> 1400, 1416, 719, 2475, 2565, 2227, 1085, 2586, 1…
## $ hour                <dbl> 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, …
## $ minute              <dbl> 15, 29, 58, 0, 0, 0, 7, 0, 27, 30, 46, 36, 45, 4…
## $ time_hour           <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-…
## $ late                <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, T…
## $ very_late           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ gain                <dbl> 9, 16, 16, 9, -12, -7, -17, 3, 5, 31, -15, -17, …
## $ rel_gain            <dbl> 0.039647577, 0.070484581, 0.106666667, 0.0260869…
## $ flight_short_distance <lgl> TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, FAL…
```

```
ggplot(data = UA_flight , aes(x= flight_short_distance))+
  geom_bar(color = 'green') +
  ggtitle('Flight is short or long based on distance')
```

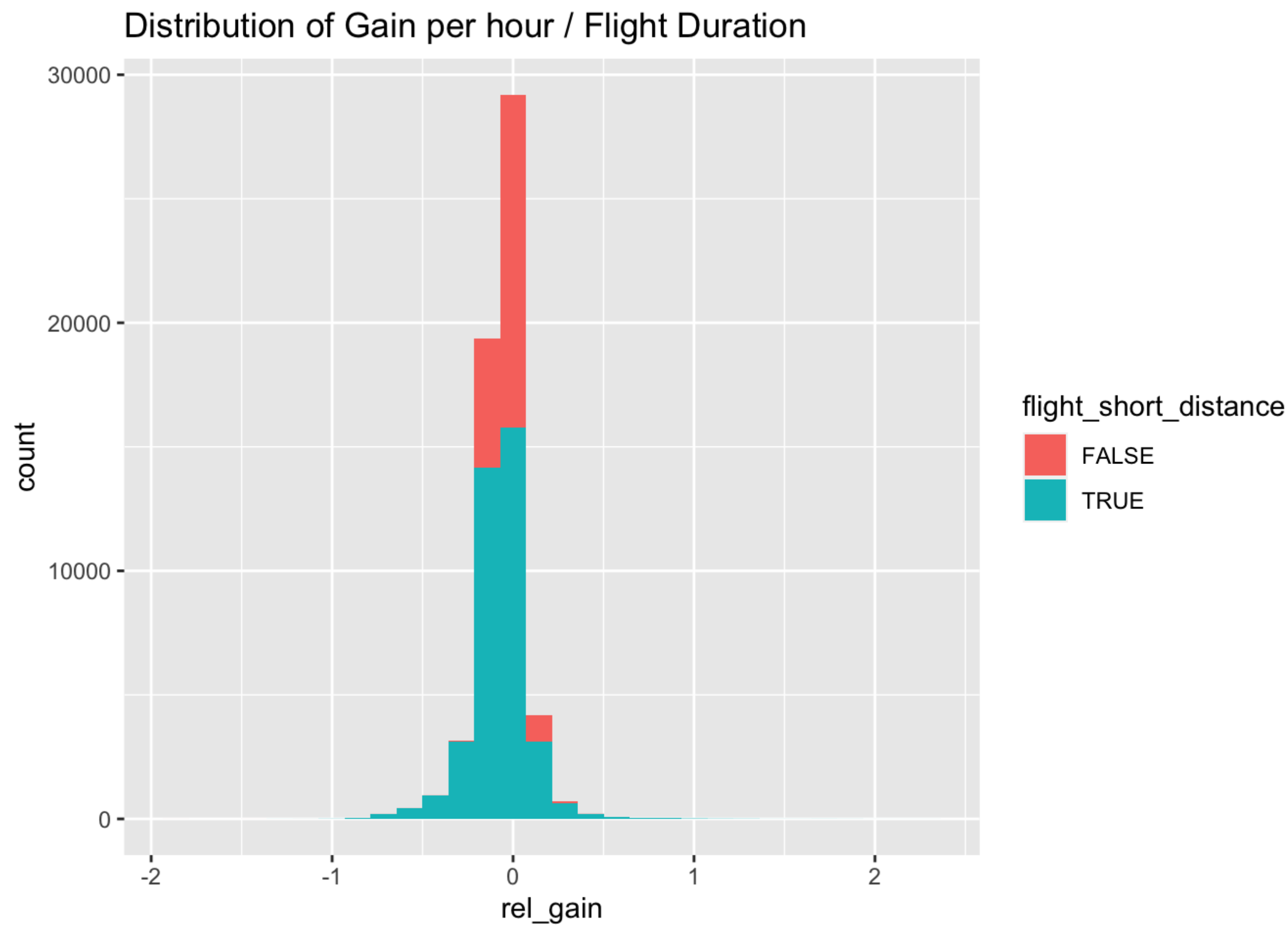



```
table(UA_flight$flight_short_distance)
```

```
##
## FALSE  TRUE
## 19851 38814
```

```
ggplot(UA_flight,aes(rel_gain,fill = flight_short_distance))+
  geom_histogram(bins = 30)+
  labs(title = 'Distribution of Gain per hour / Flight Duration')
```

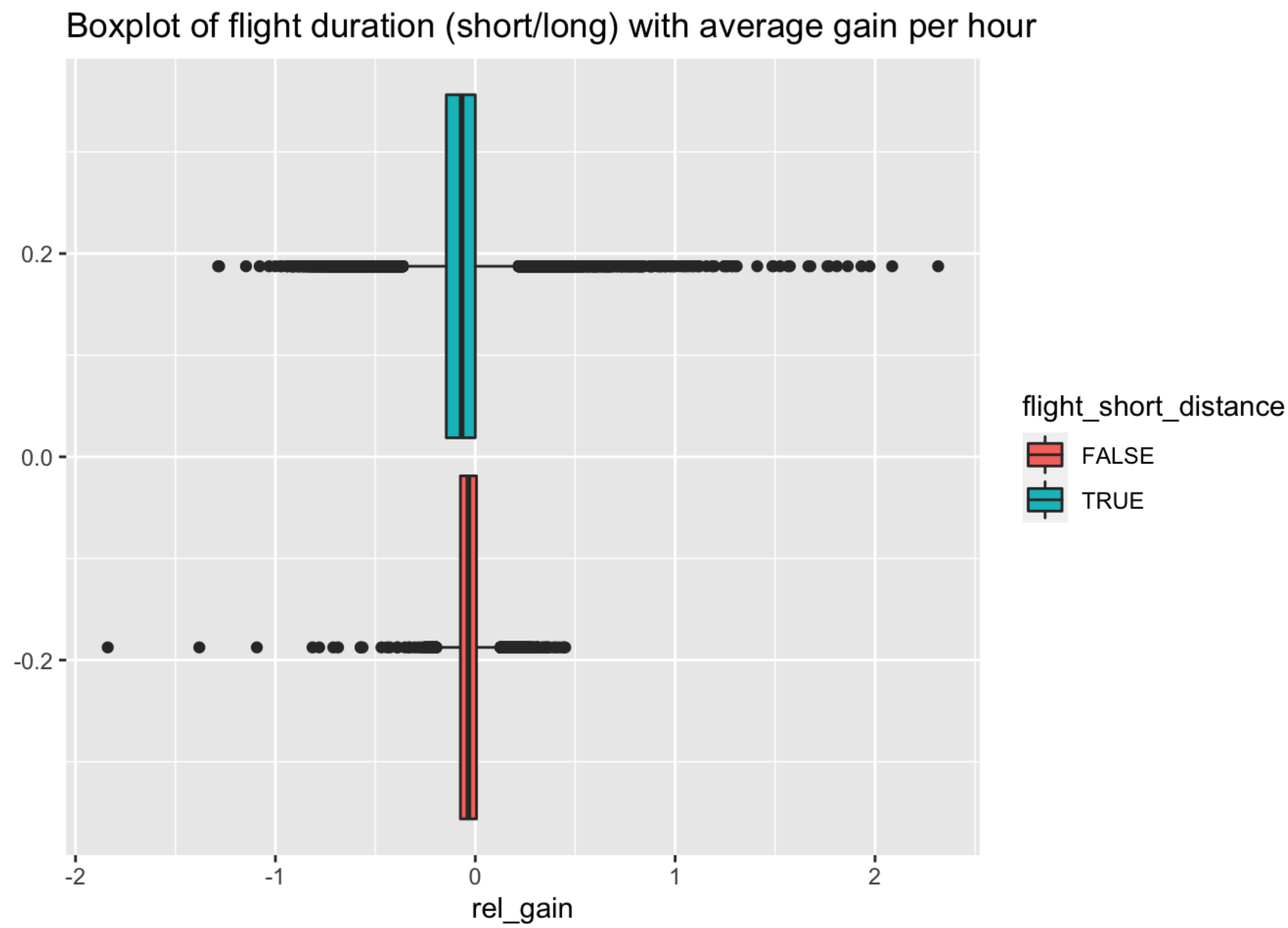
```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
ggplot(UA_flight,aes(rel_gain,fill = flight_short_distance))+
  scale_shape_discrete(name  ="Payer")+
  geom_boxplot(bins = 30)+
  labs(title = 'Boxplot of flight duration (short/long) with average gain per hour')
```

```
## Warning: Ignoring unknown parameters: bins
```

```
## Don't know how to automatically pick scale for object of type impute. Defaulting to c
ontinuous.
```




```
UA_flight %>%
  group_by(flight_short_distance) %>%
  dplyr::summarize(Mean_gain = mean(rel_gain),
                   Median_gain = median(rel_gain),
                   StandardDeviation_gain = sd(rel_gain),
                   MinGain =min(rel_gain),
                   MaxGain = max(rel_gain)
                   )
```

```
## # A tibble: 2 × 6
##   flight_short_distance Mean_gain Median_gain StandardDeviatio…¹ MinGain MaxGain
##   <lgl>                     <dbl>       <dbl>              <dbl>   <dbl>   <dbl>
## 1 FALSE                   -0.0321     -0.0344             0.0708   -1.84   0.450
## 2 TRUE                    -0.0747     -0.0682             0.165    -1.29   2.32
## # … with abbreviated variable name ¹StandardDeviation_gain
```

```
t.test(rel_gain~flight_short_distance,data=UA_flight, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  rel_gain by flight_short_distance
## t = 43.647, df = 57301, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  0.04068109 0.04450648
## sample estimates:
## mean in group FALSE  mean in group TRUE
##         -0.03214139         -0.07473518
```

```
glimpse(without_outlier)
```

```
## Rows: 57,930
## Columns: 22
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 517, 533, 554, 558, 558, 559, 607, 611, 623, 628, 643, …
## $ sched_dep_time <int> 515, 529, 558, 600, 600, 600, 607, 600, 627, 630, 646, …
## $ dep_delay      <dbl> 2, 4, -4, -2, -2, -1, 0, 11, -4, -2, -3, 8, 1, 1, -4, -…
## $ arr_time       <int> 830, 850, 740, 924, 923, 854, 858, 945, 933, 1016, 922,…
## $ sched_arr_time <int> 819, 830, 728, 917, 937, 902, 915, 931, 932, 947, 940, …
## $ arr_delay      <dbl> 11, 20, 12, 7, -14, -8, -17, 14, 1, 29, -18, -9, -6, -7…
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "…
## $ flight         <int> 1545, 1714, 1696, 194, 1124, 1187, 1077, 303, 496, 1665…
## $ tailnum        <chr> "N14228", "N24211", "N39463", "N29129", "N53441", "N765…
## $ origin         <chr> "EWR", "LGA", "EWR", "JFK", "EWR", "EWR", "EWR", "JFK",…
## $ dest           <chr> "IAH", "IAH", "ORD", "LAX", "SFO", "LAS", "MIA", "SFO",…
## $ air_time       <dbl> 227, 227, 150, 345, 361, 337, 157, 366, 229, 366, 146, …
## $ distance       <dbl> 1400, 1416, 719, 2475, 2565, 2227, 1085, 2586, 1416, 24…
## $ hour           <dbl> 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7…
## $ minute         <dbl> 15, 29, 58, 0, 0, 0, 7, 0, 27, 30, 46, 36, 45, 45, 0, 0…
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
## $ late           <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FA…
## $ very_late      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ gain           <dbl> 9, 16, 16, 9, -12, -7, -17, 3, 5, 31, -15, -17, -7, -8,…
```

```
without_outlier <- without_outlier %>%
  mutate(flight_short_distance = case_when(distance < 1800 ~ TRUE,
                        distance >=1800 ~ FALSE ),
         flight_short = case_when(air_time < 200 ~ TRUE,
                        air_time >=200 ~ FALSE ),
         rel_gain = without_outlier$gain/without_outlier$air_time
         )
```

```
t.test(rel_gain~flight_short_distance,data=without_outlier, alternative = "two.sided")
```

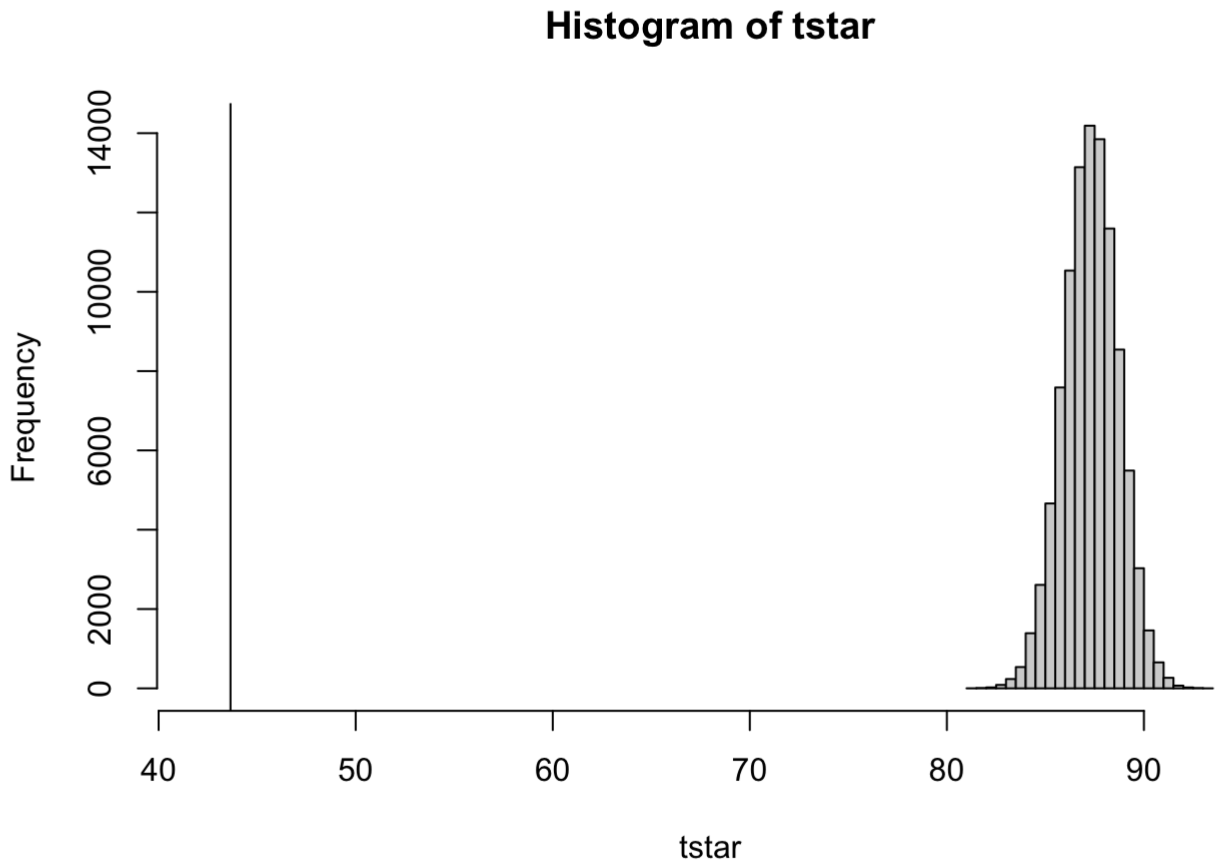
```
##
##  Welch Two Sample t-test
##
## data:  rel_gain by flight_short_distance
## t = 52.191, df = 55870, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE i
s not equal to 0
## 95 percent confidence interval:
##  0.04409890 0.04754034
## sample estimates:
## mean in group FALSE  mean in group TRUE
##         -0.03474258         -0.08056220
```

bootstrap t test for the distance and relative gain

```
UA_flight_short <-subset(UA_flight,rel_gain,subset = flight_short_distance ==TRUE,drop=T
)
UA_flight_notshort <- subset(UA_flight,rel_gain,subset = flight_short_distance ==FALSE,d
rop=T)

tstat <- function(x , y , mu)
{
  (mean(y) - mean(x) - mu)/sqrt(var(y)/length(y) + var(x)/length(x))

}
observed <- tstat(UA_flight_short,UA_flight_notshort,0)
observed
```

```
## [1] 43.64732
```

```
thetahat <- mean(UA_flight_short) - mean(UA_flight_notshort)
n1 <- length(UA_flight_short)
n2 <- length(UA_flight_notshort)

N <- 10^5-1
tstar <- numeric(N)
set.seed(5)
for (i in 1:N)
{
  boot1 <- sample(UA_flight_short,n1,replace = TRUE)
  boot2 <- sample(UA_flight_notshort,n2,replace = TRUE)
  tstar[i] <- tstat(boot1,boot2,thetahat)
}
hist(tstar,xlim = c(42,94))
abline(v=observed)
```




```
cat('The p-value is :',2*(sum(tstar >= observed)+1)/(N+1))
```

```
## The p-value is : 2
```