**DATA5300 PROJECT – 1 Report**

on

# “ANALYSING DEPARTURE DELAYS ON UA CARRIER OF NYCFLIGHTS DATASET”

*Submitted*

*in the partial fulfilment of the requirements for*

*the award of the degree of*

**Master of Science**

in

**Data Science**

by

**Ms. Likhitha Veganti**

Under the guidance of

**Mr. J. McLean Sloughter**

**(Assistant Professor of Mathematics)**

**Seattle University**

901 12th Avenue, Seattle, WA, 98122.

# TABLE OF CONTENTS

# LIST OF FIGURES

# 1. PROJECT STATEMENT:

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

Suppose you work for United Airlines (carrier code UA). To improve both efficiency and customer satisfaction, you have been asked to study departure delays.

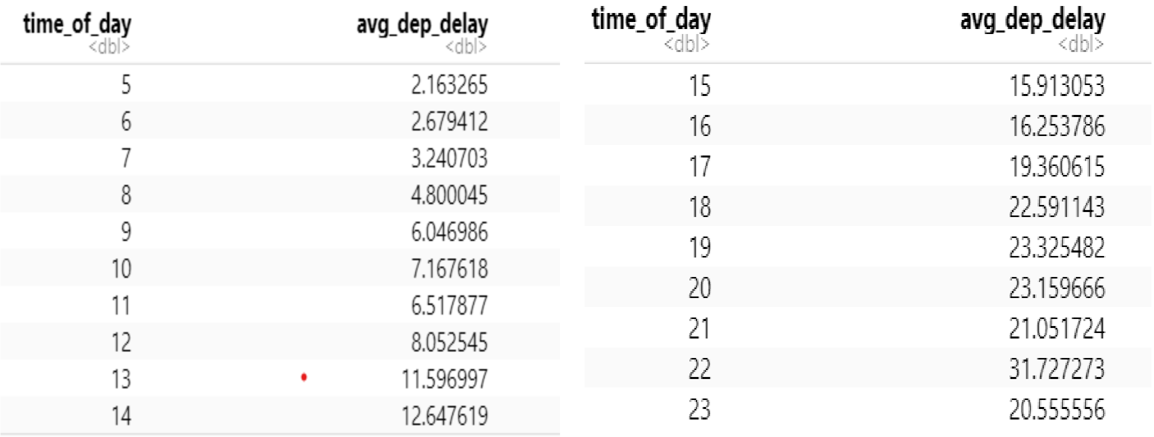
Your report should address the relationship between departure delays and each of the following:

1. Time of day
2. Time of year
3. Temperature
4. Wind speed
5. Precipitation
6. Visibility

For each of these factors, you can decide whether to use the original dep_delay variable, or the late and very_late variables you created for the first two homework assignments (for some questions, you may only be able to conduct a permutation test (based on what we have learned so far) if you use the late and very_late variables). If there are circumstances where you feel it is appropriate to create new variables based on the other factors being analyzed, you may do so, but be sure to clearly indicate that you have done that.

You should prepare a written report that includes appropriate graphs and tables, written for a non-technical audience (something you could give to someone in management to illustrate your results without getting into details about coding and such), along with an appendix that includes the code you used, clearly labeled to explain what part of the main report each section of code corresponds to.

## 2. DEPARTURE DELAY VS TIME OF DAY



| time_of_day<br><dbl> | avg_dep_delay<br><dbl> |
|---|---|
| 5 | 2.163265 |
| 6 | 2.679412 |
| 7 | 3.240703 |
| 8 | 4.800045 |
| 9 | 6.046986 |
| 10 | 7.167618 |
| 11 | 6.517877 |
| 12 | 8.052545 |
| 13 | 11.596997 |
| 14 | 12.647619 |
| 15 | 15.913053 |
| 16 | 16.253786 |
| 17 | 19.360615 |
| 18 | 22.591143 |
| 19 | 23.325482 |
| 20 | 23.159666 |
| 21 | 21.051724 |
| 22 | 31.727273 |
| 23 | 20.555556 |

Fig 2.1 Table having average departure delays at different times of day.

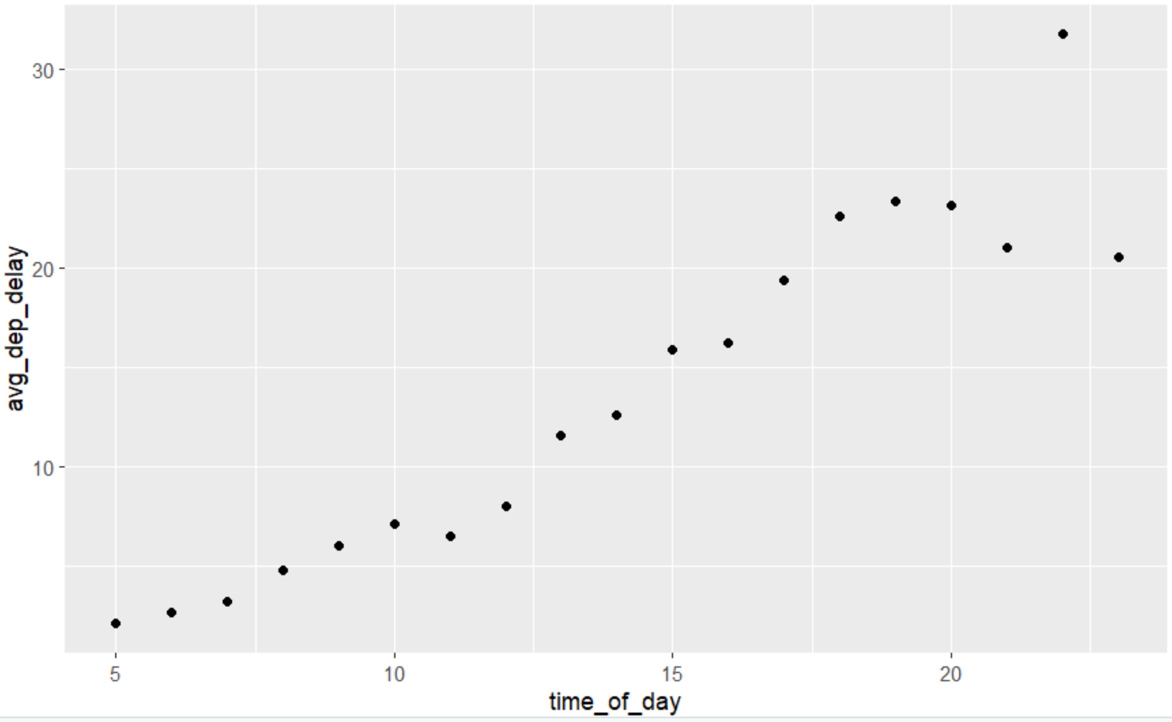


Fig 2.2 Graph between time of day and average departure delays.

From Fig 2.1 and 2.2, we can conclude that the average departure delays increase with increase in time of day. But there might be few abnormal values in the dep_delays that can cause the average departure delays to be in increasing order.

In order to ensure the above conclusion to be correct, we can perform the permutation test and see if there is evidence for any difference in departure delays at different time of a day. A permutation test is performed between morning (5 - 16) and evening (17 - 23) hours of day by making following hypothesis.

‘Null Hypothesis H0: Mean of departure delays in evening (17 – 23) hours of day = Mean of departure delays in morning (5 – 16) hours of day’.

‘Alternate Hypothesis H1: Mean of departure delays in evening (17 – 23) hours of day > Mean of departure delays in morning (5 – 16) hours of day’.

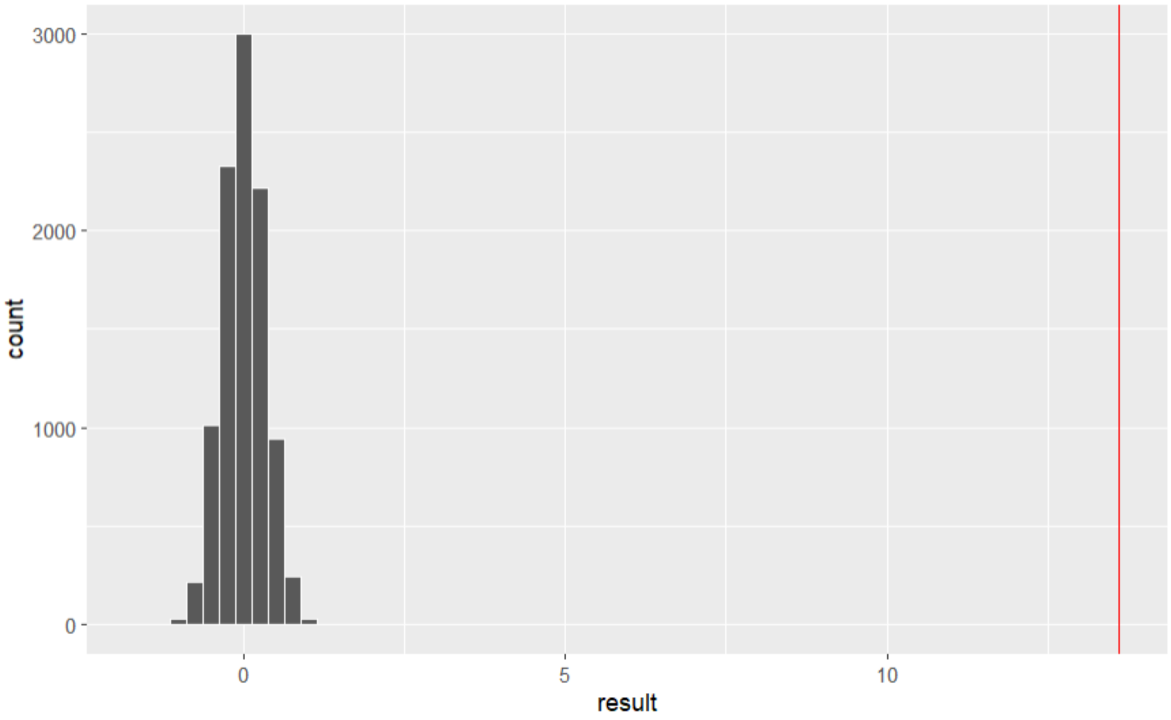
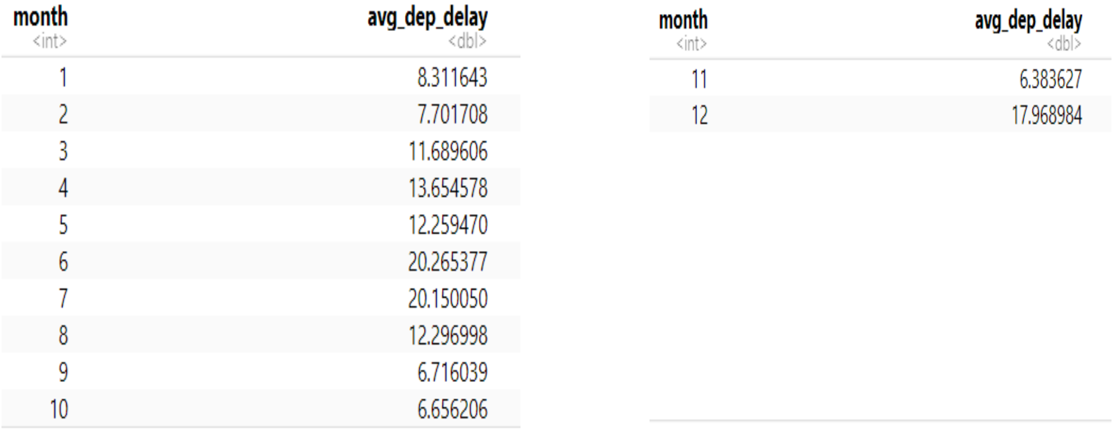


Fig 2.3 Graph of simulated and observed difference of morning and evening.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between morning and evening times of day. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in the evening hours of day > mean of departure delays in the morning hours of day.

Hence, we can conclude the relationship between departure delay and time of day is positive correlation i.e., the departure delay increases with increase in time of day (with an exception of 0-4 hours as we do not have any data related to those timings).

## 3. Departure Delay Vs Time of year



| month | avg_dep_delay |
|---|---|
| <int> | <dbl> |
| 1 | 8.311643 |
| 2 | 7.701708 |
| 3 | 11.689606 |
| 4 | 13.654578 |
| 5 | 12.259470 |
| 6 | 20.265377 |
| 7 | 20.150050 |
| 8 | 12.296998 |
| 9 | 6.716039 |
| 10 | 6.656206 |
| 11 | 6.383627 |
| 12 | 17.968984 |

Fig 3.1 Table having average departure delays at different times of year.

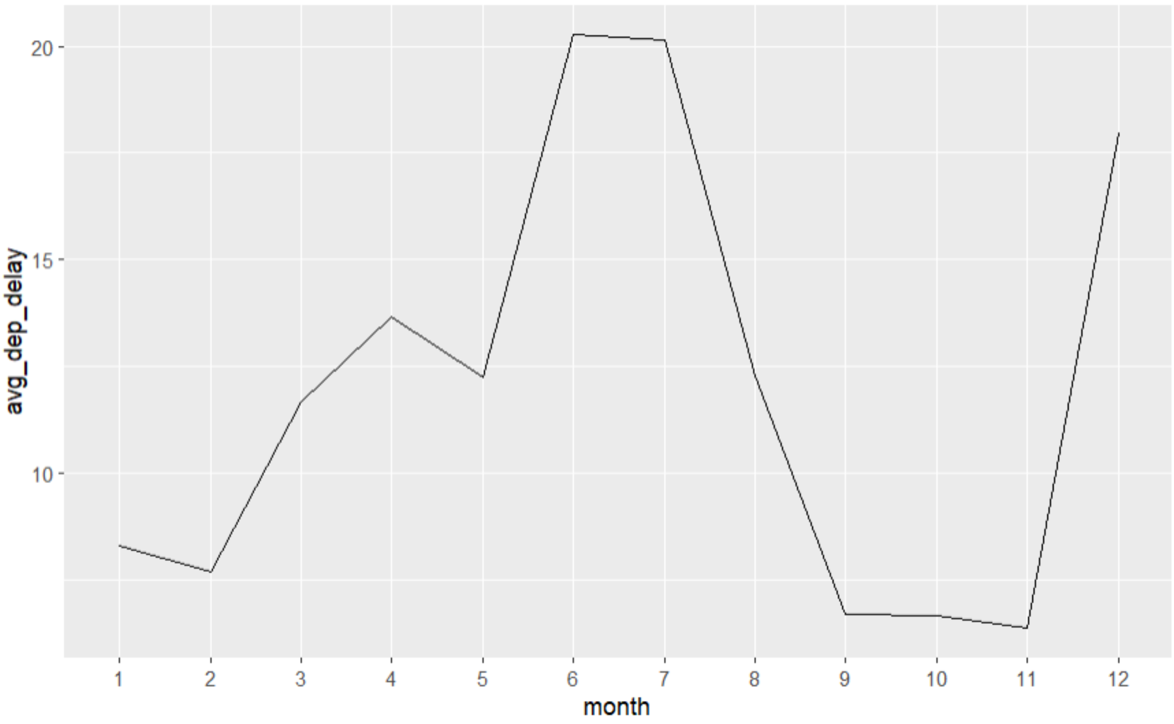


Fig 3.2 Graph of average departure delays and different times(months) of year.

From Fig 3.1 and 3.2, we can say that the average departure delay changes as the month changes. And, to further understand the pattern, we can perform some perform some permutation tests for different seasons. The seasons considered in the performed permutation tests are Spring (March – May), Summer (June – August), Autumn (September – November) and Winter (December – February).

### 3.1 Summer Vs Spring

Null Hypothesis H0: Mean of departure delays in Summer = Mean of departure delays in Spring.

Alternate Hypothesis H1: Mean of departure delays in Summer > Mean of departure delays in Spring.

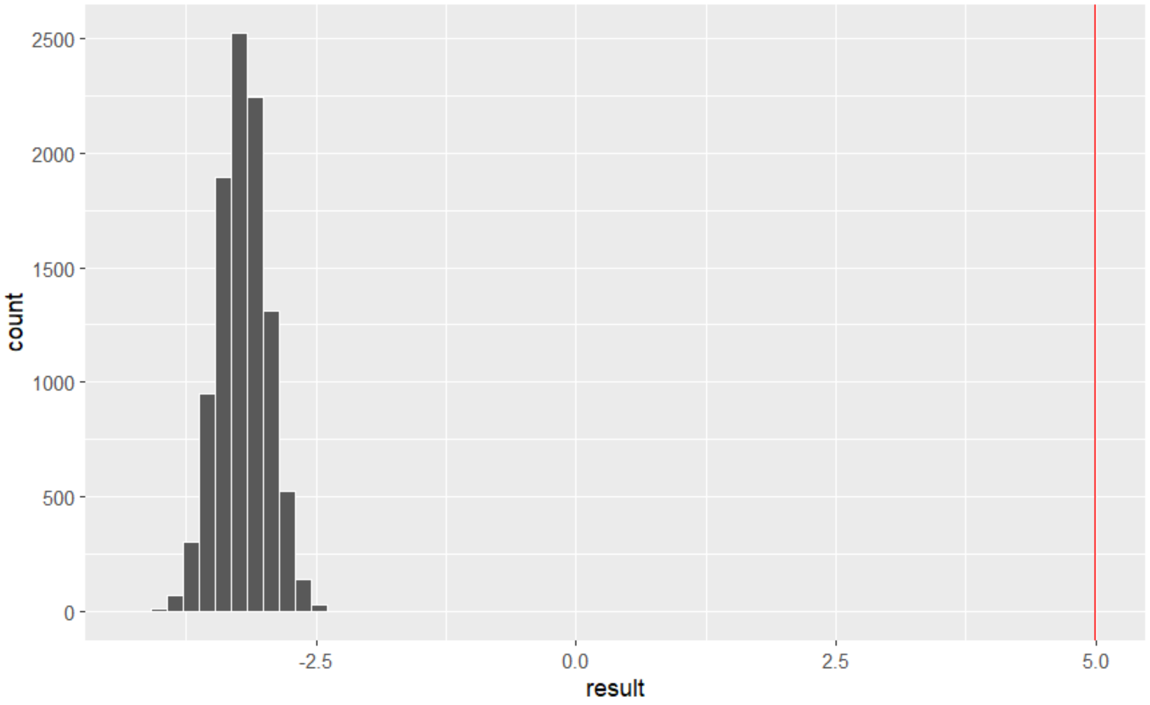


Fig 3.3 Graph of differences in Summer and Spring Seasons.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between summer and spring seasons. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in Summer > mean of departure delays in the Spring.

### 3.2 Summer Vs Autumn

Null Hypothesis H0: Mean of departure delays in Summer = Mean of departure delays in Autumn.

Alternate Hypothesis H1: Mean of departure delays in Summer > Mean of departure delays in Autumn.

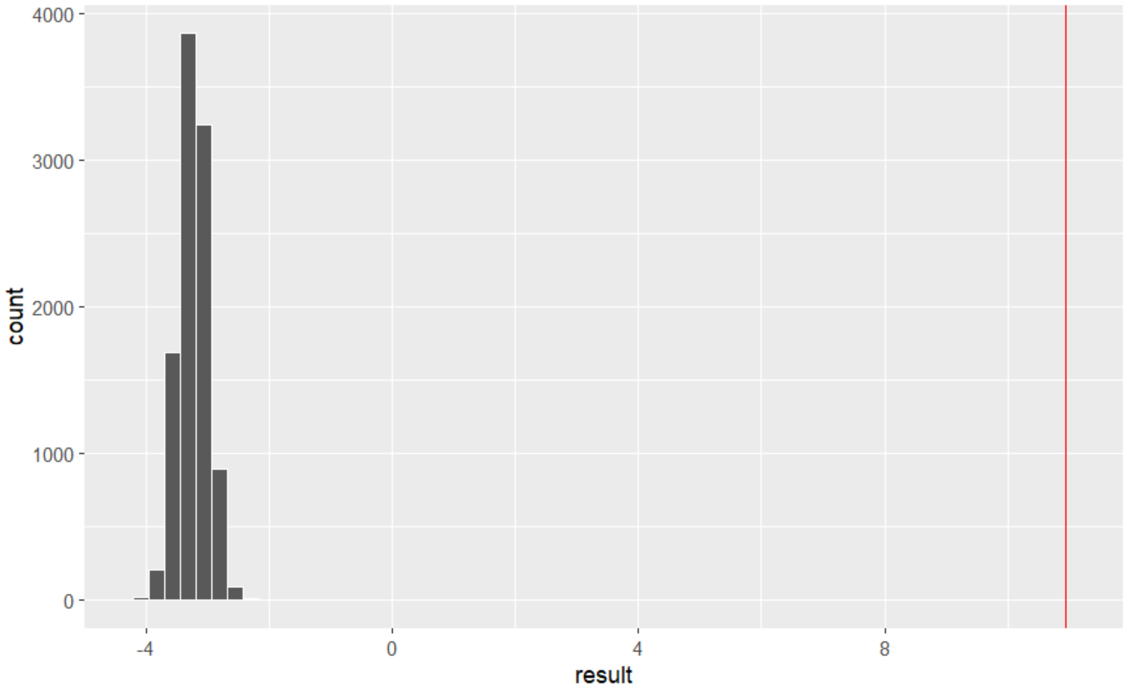


Fig 3.4 Graph of differences in Summer and Autumn Seasons.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between summer and autumn seasons. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in Summer > mean of departure delays in the Autumn.

### 3.3 Summer Vs Winter

Null Hypothesis H0: Mean of departure delays in Summer = Mean of departure delays in Winter.

Alternate Hypothesis H1: Mean of departure delays in Summer > Mean of departure delays in Winter.

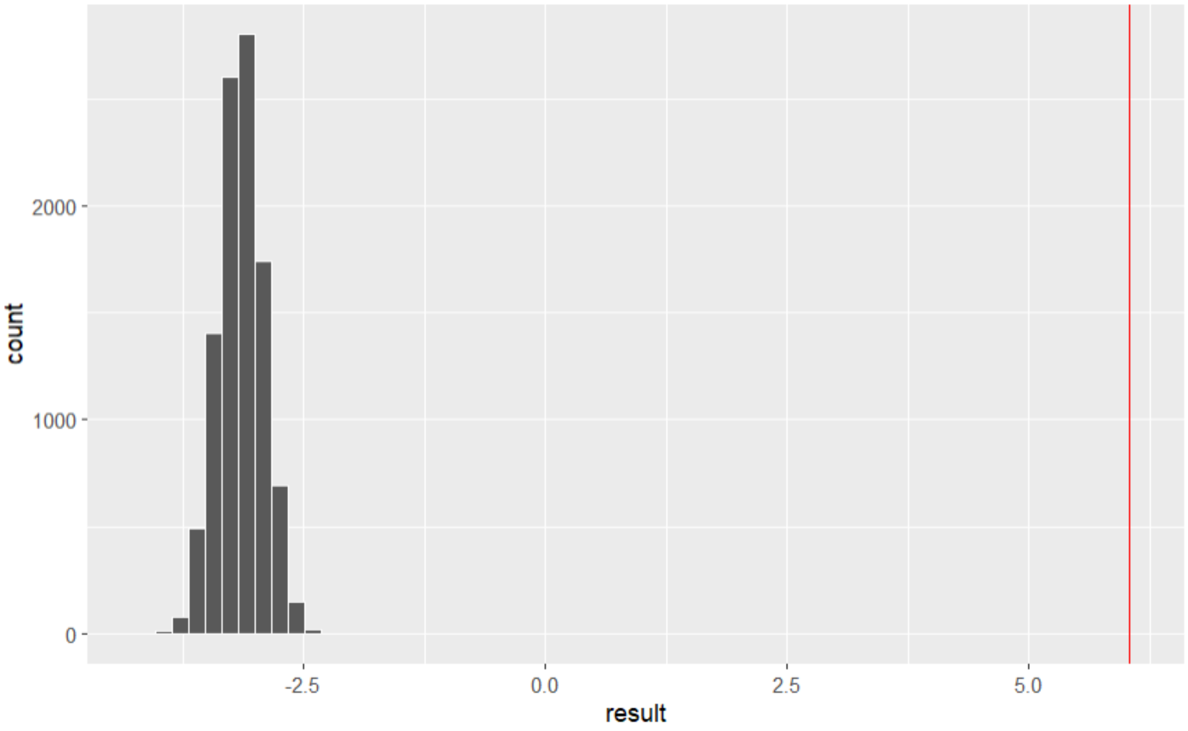


Fig 3.5 Graph of differences in Summer and Winter Seasons.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between summer and winter seasons. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in Summer > mean of departure delays in the Winter.

### 3.4 Spring Vs Autumn

Null Hypothesis H0: Mean of departure delays in Spring = Mean of departure delays in Autumn.

Alternate Hypothesis H1: Mean of departure delays in Spring > Mean of departure delays in Autumn.

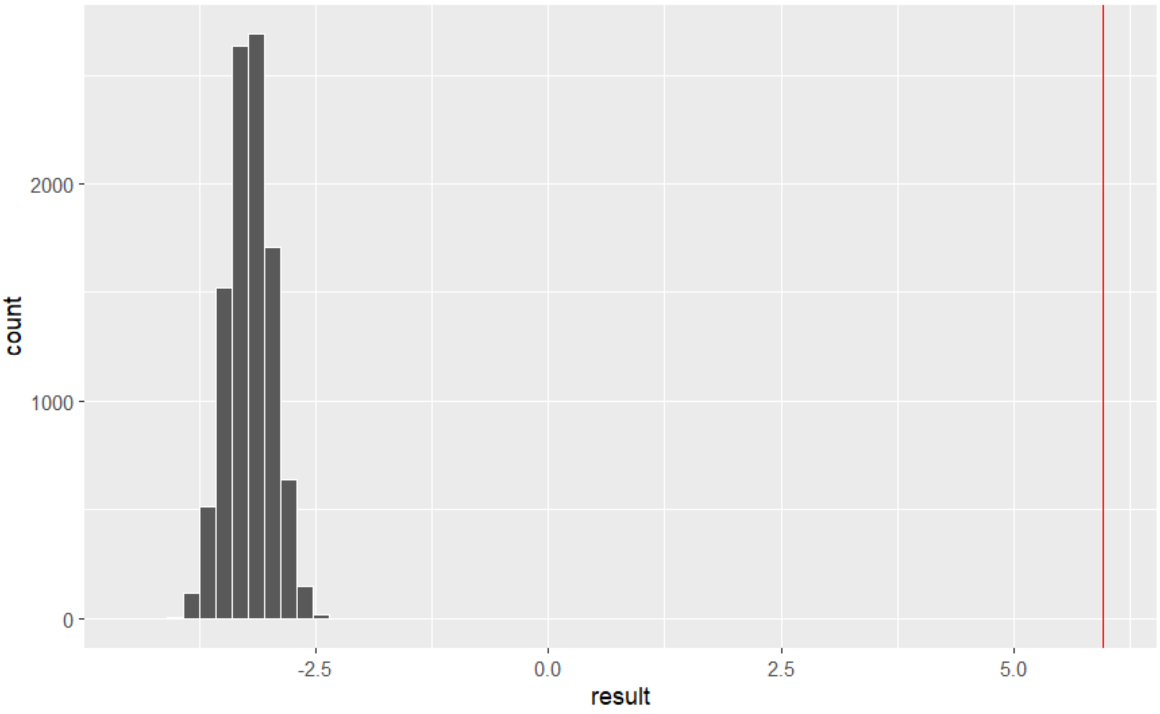


Fig 3.6 Graph of differences in Spring and Autumn Seasons.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between spring and autumn seasons. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in Spring > mean of departure delays in the Autumn.

### 3.5 Spring Vs Winter

Null Hypothesis H0: Mean of departure delays in Spring = Mean of departure delays in Winter.

Alternate Hypothesis H1: Mean of departure delays in Spring > Mean of departure delays in Winter.

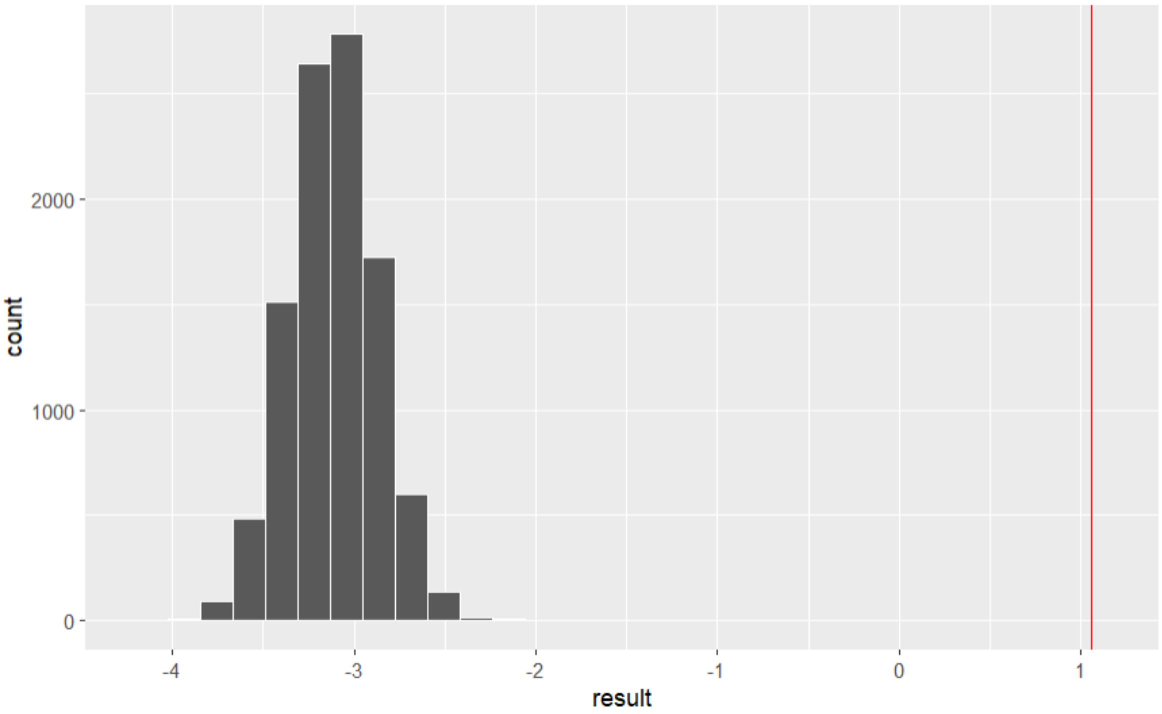


Fig 3.7 Graph of differences in Spring and Winter Seasons.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between spring and winter seasons. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in Spring > mean of departure delays in the Winter.

### 3.6 Winter Vs Autumn

Null Hypothesis H0: Mean of departure delays in Winter = Mean of departure delays in Autumn.

Alternate Hypothesis H1: Mean of departure delays in Winter > Mean of departure delays in Autumn.

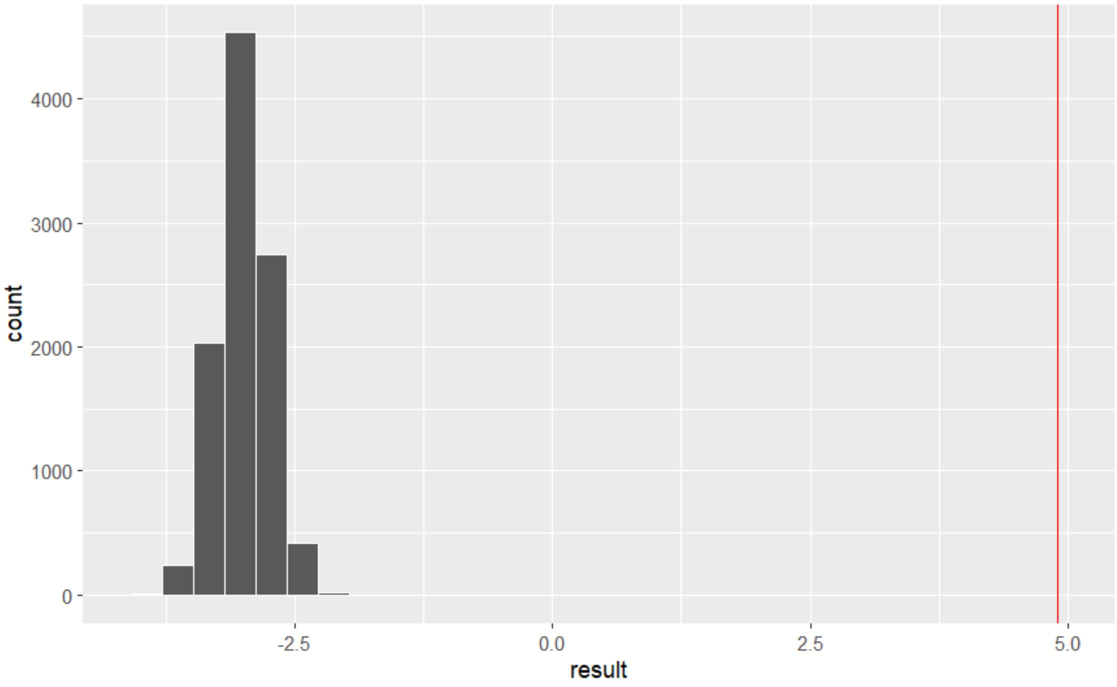


Fig 3.8 Graph of differences in Winter and Autumn Seasons.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between winter and autumn seasons. So, we can support our alternate hypothesis H1 i.e., mean of departure delays in Winter > mean of departure delays in the Autumn.

From the above permutation tests, we can conclude that Summer has more departure delays than any other seasons. Summer is a holiday season. It may be one of the reasons for high rate of departure delays as more people travels at this period.

The order of departure delays from the above permutation tests are:

Summer > Spring > Winter > Autumn.

And from Fig 3.2, we can see December also has more average departure delay value. It maybe because of Christmas month, more people tend to travel resulting in high values of delays in the departure of flights.

## 4. DEPARTURE DELAY VS TEMPERATURE

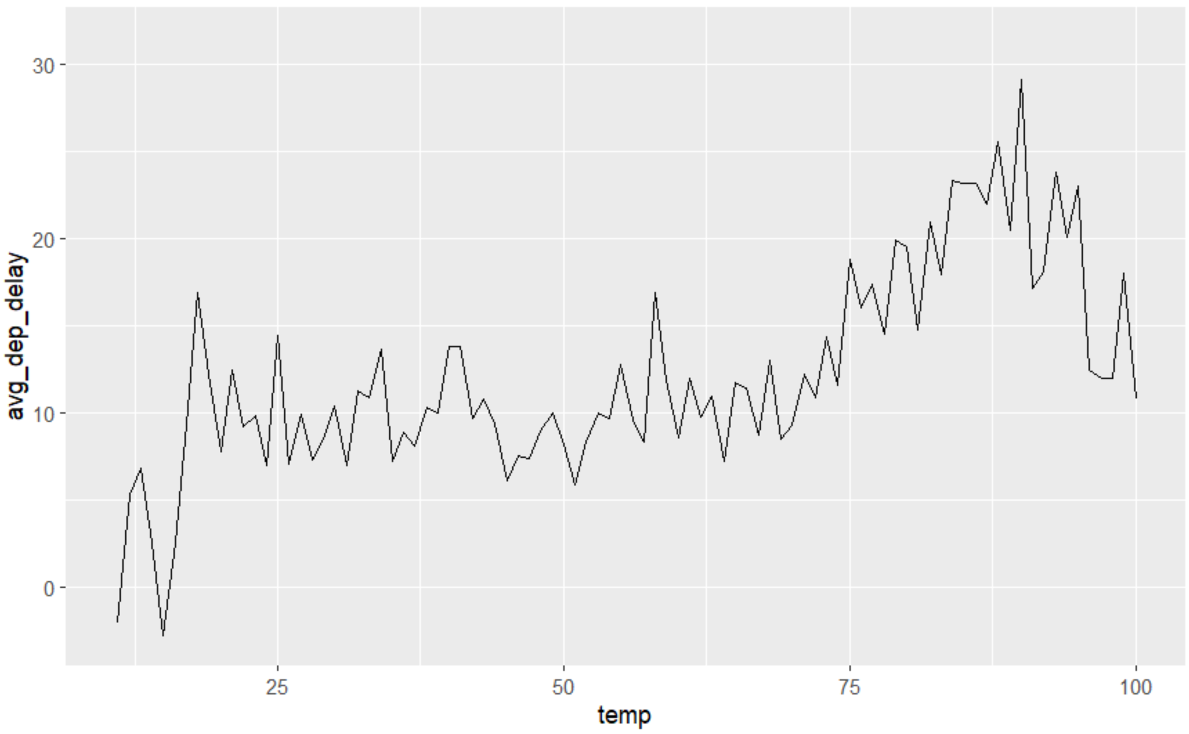


Fig 4.1 Graph between temperature and average departure delays.

From Fig 4.1, we can see the temperatures are approximately ranging from 11 to 100 plotted against the average departure delays at that temperature. We even have negative dep_delay values for very less temperature values which means possibility of no delays at very less temperature. This maybe because people may not like to travel much at very low temperature and more dense nature of cold air.

We can also observe increase in departure delays with increase in temperature and more dep_delays after the temperature of 75. To further confirm the results, we can also perform permutation test by taking a threshold of 50 degrees temperature to verify the results.

Null Hypothesis H0: Mean of dep_delays at (temp > 50) = Mean of dep_delays at (temp <= 50)

Alternate Hypothesis H1: Mean of dep_delays at (temp > 50) > Mean of dep_delays at (temp <= 50)

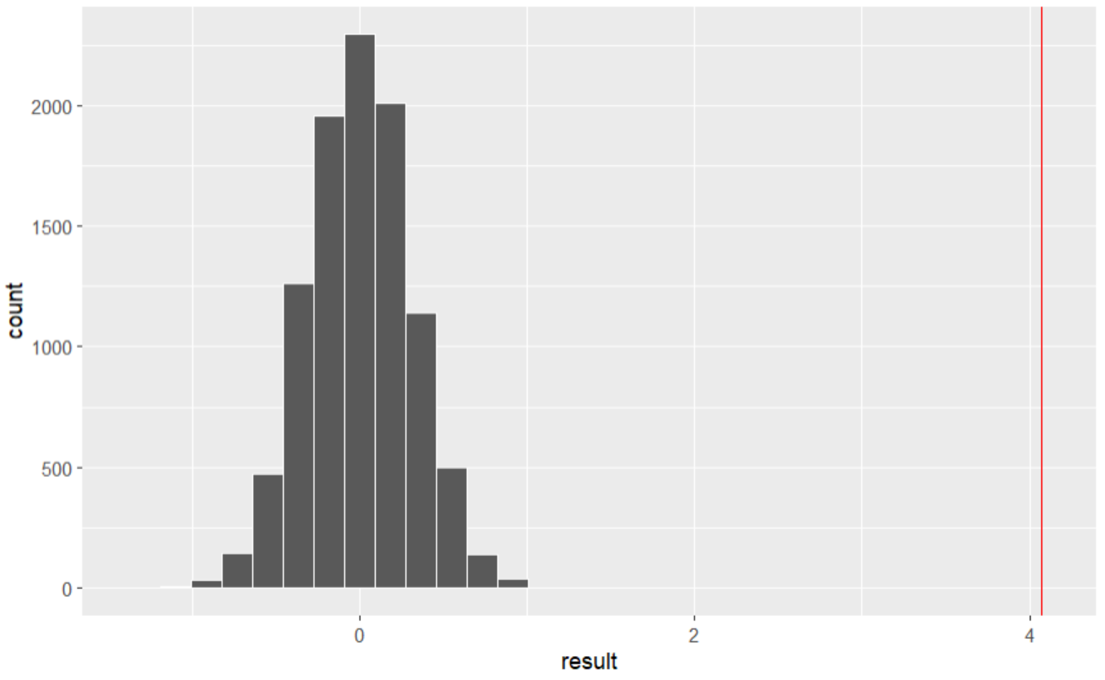


Fig 4.2 Graph of differences for temp <= 50 and >50.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between temp>50 and temp<=50. So, we can support our alternate hypothesis H1 i.e., mean of departure delays with (temp>50) > mean of departure delays with (temp<=50).

Hence, we can conclude that departure delays increase with increase in temperature. Hot air is less dense than cold air resulting in the flight requirement of more engine power to generate equal amount of thirst and lift as they would in cold climes. This maybe one of the reasons for increase in average departure delays at higher temperatures.

## 5. DEPARTURE DELAY VS WIND SPEED

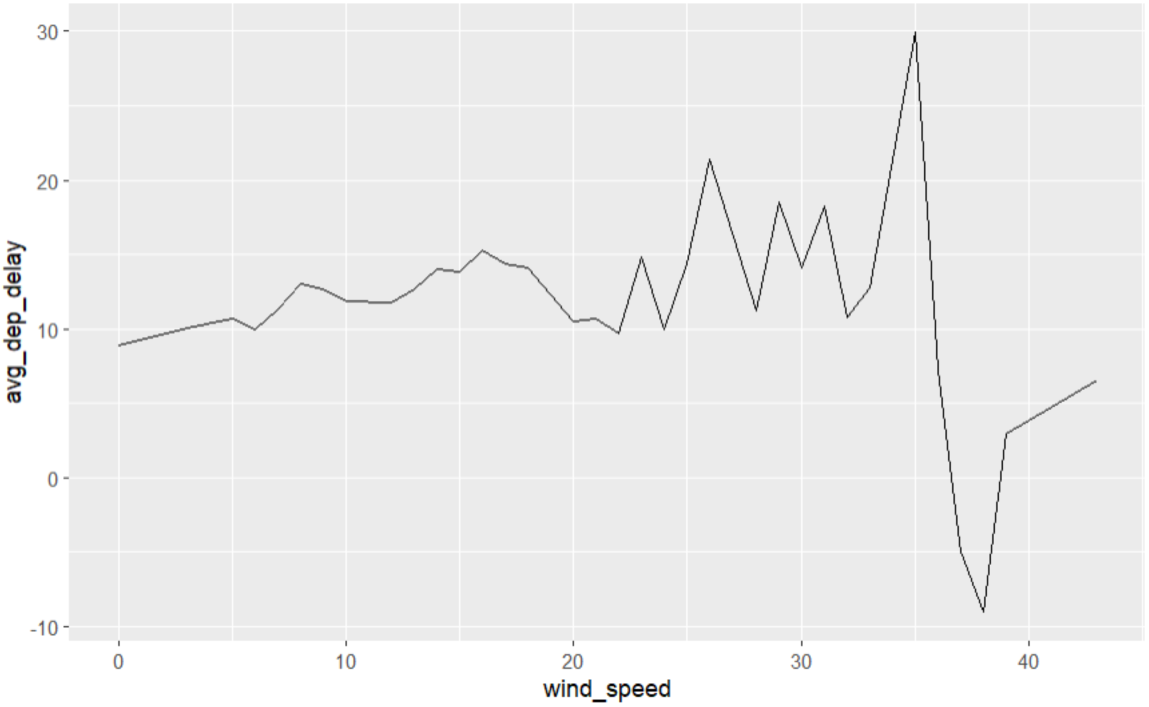


Fig 5.1 Graph between wind speed and average departure delays.

From Fig 5.1, we can see there is no much difference in average departure delays with change in wind speed except for few abnormal values. We can conduct permutation test to get a good idea about the relationship by taking a threshold value of 20.

Null Hypothesis H0: Mean of dep_delay with (wind speed > 20) = Mean of dep_delay with (wind speed <= 20)

Alternate Hypothesis H1: Mean of dep_delay with (wind speed > 20) > Mean of dep_delay with (wind speed > 20)

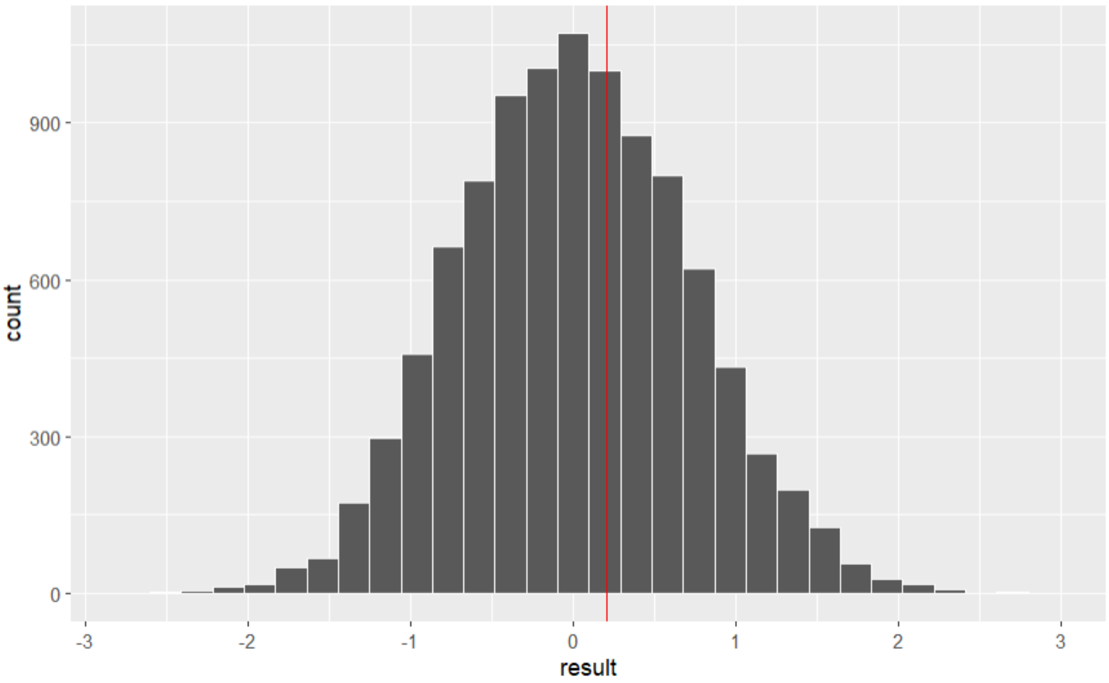


Fig 5.2 Graph of differences for wind speed <= 20 and >20.

From the permutation test the resultant p-value is 0.3863. As it is greater than 0.05, there is no evidence for significant difference in departure delays. So, we can support our null hypothesis H0 i.e., mean of departure delays with (wind speed > 20) = mean of departure delays with (wind speed <= 20).

Hence, there is no evidence to prove that wind speed influences the departure delay. Generally, the effect may be significant if the direction of wind is specified in the or against the direction of flight. Here, we cannot find any proper evidence to show the relationship between the departure delay and wind speed. So, we can conclude the departure delay does not depend on the wind speed.

# 6. DEPARTURE DELAY VS PRECIPITATION

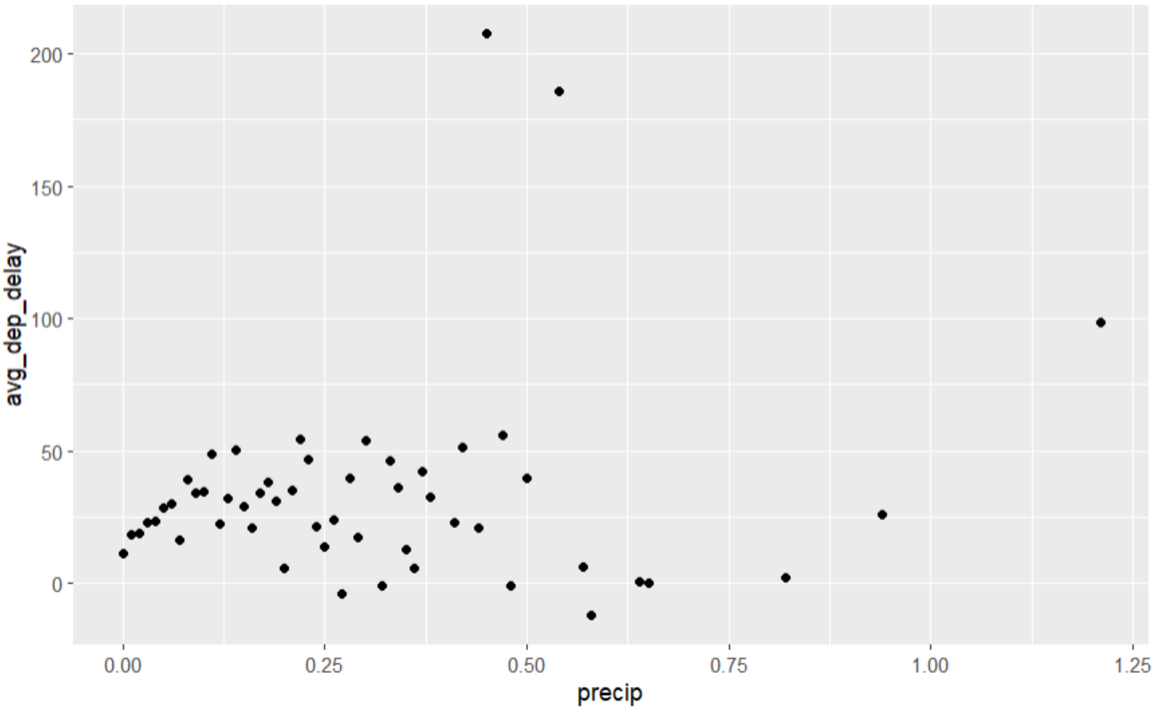


Fig 6.1 Graph between precipitation and average departure delays.

The Fig 6.1 does not give much information about the relationship between average departure delay ad precipitation. And we can also see there are only few precipitation values that are greater than 0.6. So, let us conduct a permutation test to further understand the effect of precipitation on departure delay.

Null Hypothesis H0: Mean of dep_delay with (precip > 0.6) = Mean of dep_delay with (precip <= 0.6)

Alternate Hypothesis H1: Mean of dep_delay with (precip > 0.6) > Mean of dep_delay with (precip <= 0.6)

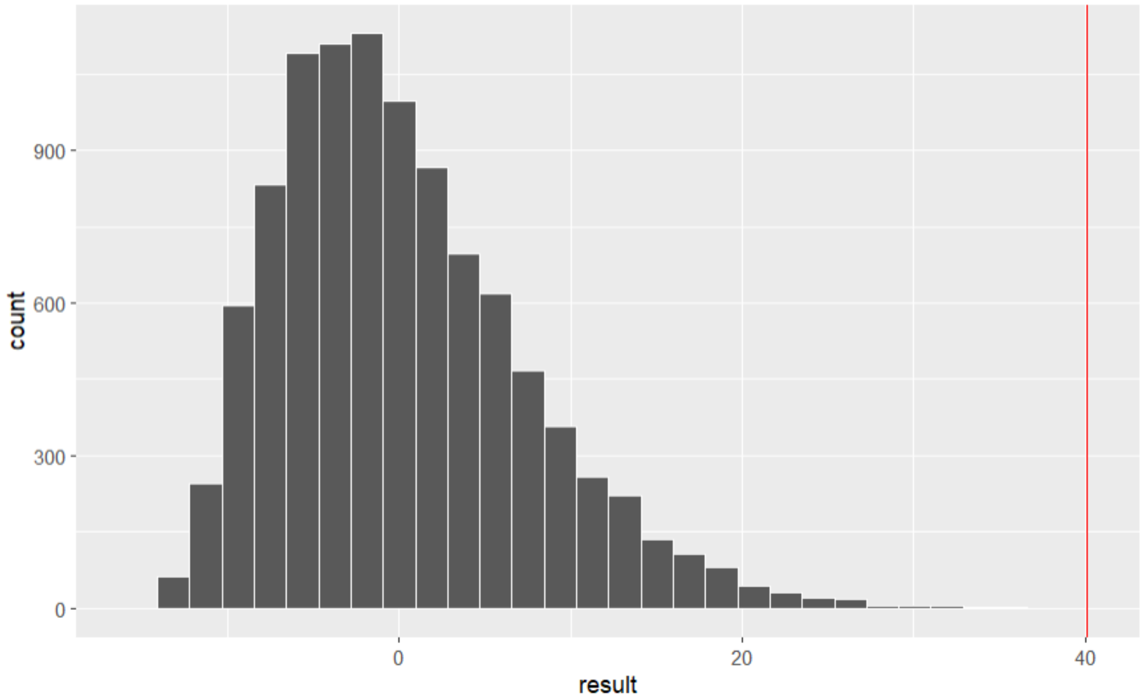
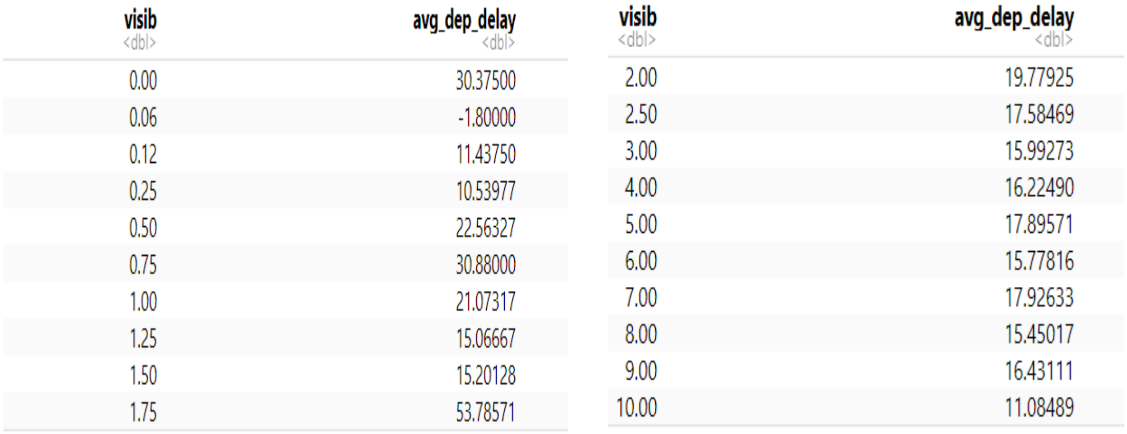


Fig 5.2 Graph of differences for precip <= 0.6 and > 0.6.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between precip > 0.6 and precip <= 0.6. So, we can support our alternate hypothesis H1 i.e., mean of departure delays with (precip > 0.6) > mean of departure delays with (precip <= 0.6).

So, we can conclude higher values of precipitation can delay the departure time of flights.

# 7. DEPARTURE DELAY VS VISIBILITY



| visib | avg_dep_delay |
|---|---|
| <dbl> | <dbl> |
| 0.00 | 30.37500 |
| 0.06 | -1.80000 |
| 0.12 | 11.43750 |
| 0.25 | 10.53977 |
| 0.50 | 22.56327 |
| 0.75 | 30.88000 |
| 1.00 | 21.07317 |
| 1.25 | 15.06667 |
| 1.50 | 15.20128 |
| 1.75 | 53.78571 |
| 2.00 | 19.77925 |
| 2.50 | 17.58469 |
| 3.00 | 15.99273 |
| 4.00 | 16.22490 |
| 5.00 | 17.89571 |
| 6.00 | 15.77816 |
| 7.00 | 17.92633 |
| 8.00 | 15.45017 |
| 9.00 | 16.43111 |
| 10.00 | 11.08489 |

Fig 7.1 Table having average departure delays at different visibilities.

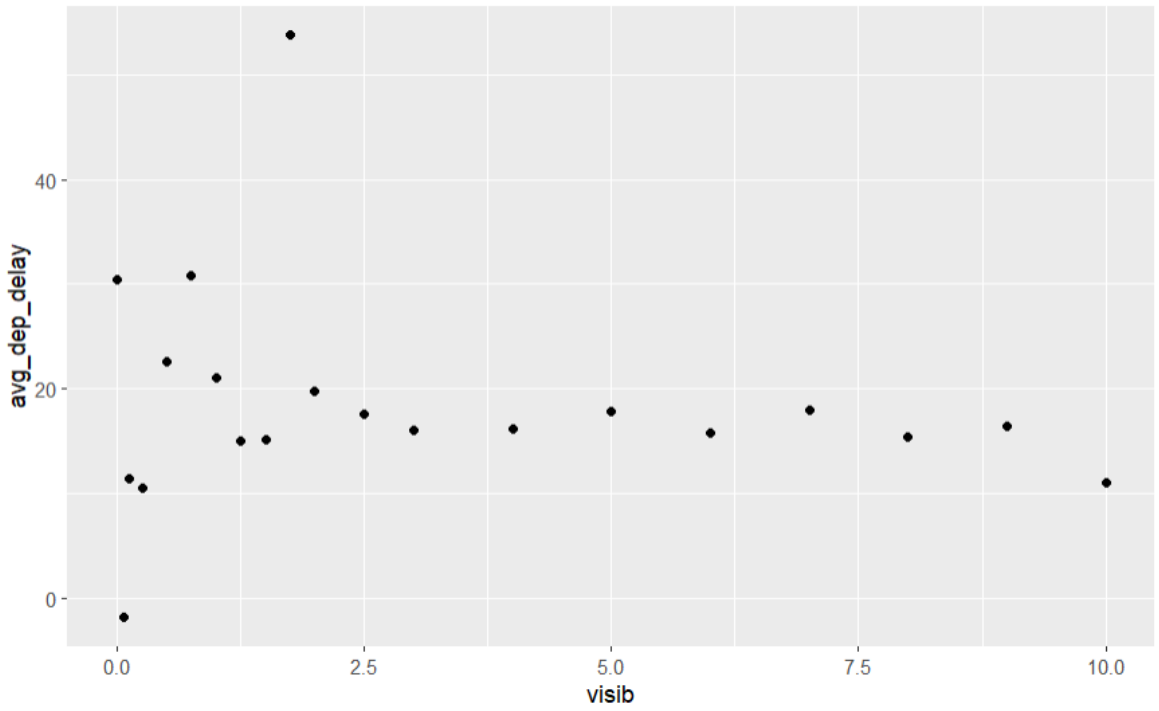


Fig 7.2 Graph between visibilities and average departure delay.

The Fig 7.2 does not give much information about the relationship between average departure delay ad visibility. And we can also see departure delays are more for the visibility values < 2.5. So, let us conduct a permutation test to further understand the effect of visibility on departure delay.

Null Hypothesis H0: Mean of dep_delay with (visib <= 2.5) = Mean of dep_delay with (visib > 2.5)

Alternate Hypothesis H1: Mean of dep_delay with (visib <= 2.5) > Mean of dep_delay with (visib > 2.5)

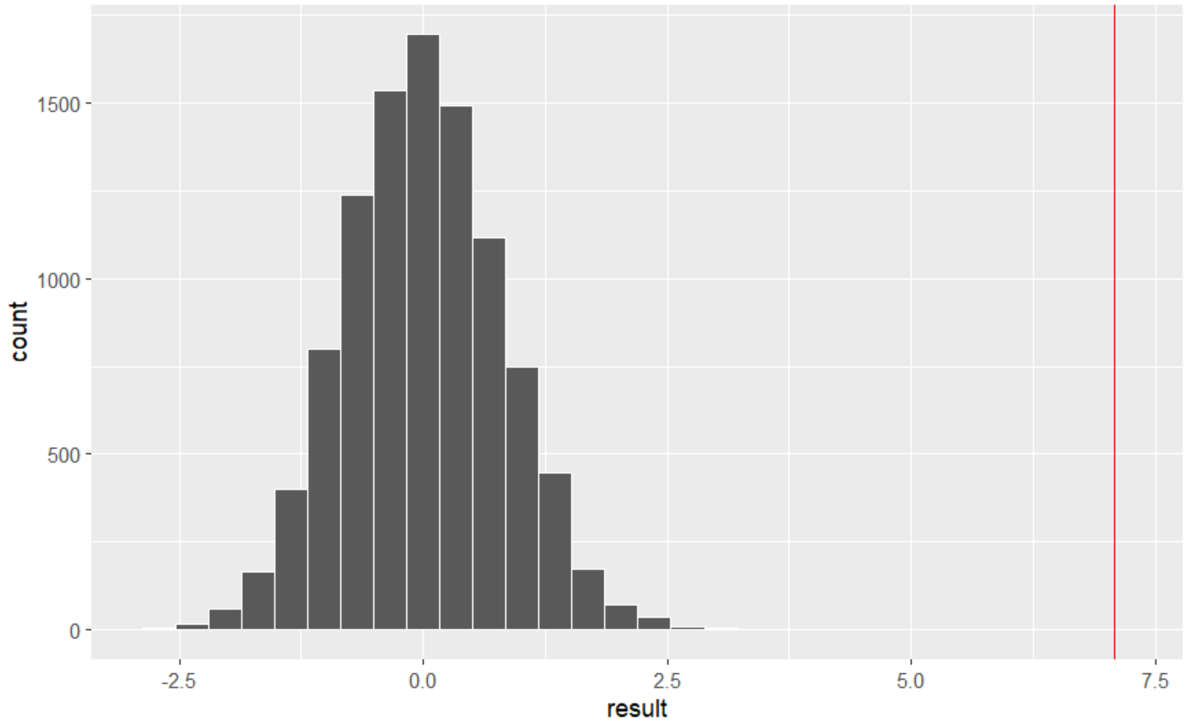


Fig 7.3 Graph of differences for visibilities <= 2.5 and > 2.5.

From the permutation test the resultant p-value is 1e-4. As it is less than 0.05, there is evidence for significant difference in departure delays between visib <= 2.5 and visib > 2.5. So, we can support our alternate hypothesis H1 i.e., mean of departure delays with (visib <= 2.5) > mean of departure delays with (visib > 2.5).

So, we can conclude less values of visibility can delay the departure time of flights.

## 8. CONCLUSION

The project deals with finding relationship between departure delay and some other variables. Different variables influence the departure delay in the following ways:

The departure delay increases with increase in the time of day and the delays in the seasons Summer > Spring > Winter > Autumn. This maybe because of holidays in summer season. Also, departure delays in the month of December are high (maybe due to Christmas month). Coming to the temperature, the departure delay increases with increase in temperature. This may be due to the density variations in cold and hot climates. Generally, hot air is less dense compared to cold air and the flight requires more engine power to generate equal thrust as of cold air. This maybe a reason for increase in departure delay.

And we could not find much relationship between wind speed and departure delays. The flight speed may increase if the wind is in the same direction and decrease if the wind is against the flight direction. So, perfect relationship could not be found between wind speed and departure delays. And for precipitation, departure delays are higher for high values of precipitation. So, we can say it is a positive correlation. Coming to visibility, departure delays are high for less visibilities.

## 9. APPENDIX

### # Import the required libraries

```
library(ggplot2)

library(tidyverse)

library(nycflights13)
```

### # Creating a new dataset named 'myflights' in order to filter the carrier as 'UA'.

```
myflights <- flights %>%

  filter(carrier == 'UA') %>%

  mutate(

    late = dep_delay>0,

    very_late = dep_delay>30

  )

glimpse(myflights)
```

### # Appending both the data sets ‘myflights’ and ‘weather’ into a new data set, ‘data’ using inner join.

```
data <- myflights %>% inner_join(weather, by = c('origin', 'year', 'month', 'day', 'hour', 'time_hour'))

glimpse(data)
```

## 1. Departure Delay Vs Time of day

**# Creating a temporary dataset named 'data1' that stores average departure delay group by hour.**

```
data1 <- data %>% group_by(time_of_day = hour) %>% summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE))

data1
```

**# Plotting a scatter plot between time and average departure delay**

```
ggplot(data1, aes(y = avg_dep_delay, x = time_of_day))+

  geom_point()
```

**# Permutation test between morning and evening times of day**

```
observed <- mean(data$dep_delay[data$hour > 16 ], na.rm = TRUE) - mean(data$dep_delay[data$hour <= 16], na.rm = TRUE)

observed

N <- 10^4-1

set.seed(2000)

sample.size = nrow(data)

#group.1.size = the number of observations in the first group

day = data %>% filter(hour > 16)

group.1.size = nrow(day)

#create a blank vector to store the simulation results

result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N
```

```
for(i in 1:N)

{

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

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

}

#plot a histogram of the simulated differences

#add a vertical line at the observed difference

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

geom_histogram(bins = 60, color = 'white') +

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

#Calculate the p-value

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


cat("The difference is significant.\nThe departure delays at evening times of a day are greater than departure delays at morning times of a day.")
```

## 2. Departure delay Vs Time of Year

**# Creating a temporary dataset named 'data1' that stores average departure delay group by month.**

```
data1 <- data %>% group_by(month) %>% summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE))

data1
```

**# Plotting a line plot between month and average departure delay**

```
ggplot(data1, aes(x = month, y = avg_dep_delay))+

geom_line() +

scale_x_discrete(limits = c("1","2","3","4","5","6","7","8","9","10","11","12"))
```

**# Permutation tests**

**a) Summer - Spring**

```
observed <-  mean(data$dep_delay[data$month >= 6 & data$month <= 8], na.rm = TRUE) -
mean(data$dep_delay[data$month >=3 & data$month <=5], na.rm = TRUE)

observed

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)

#sample.size = the number of observations in our sample

summer_spring = data %>% filter(month >=3 & month <=8)

sample.size = nrow(summer_spring)

#group.1.size = the number of observations in the first group

summer = data %>% filter(month >= 6 & month <= 8)

group.1.size = nrow(summer)

#create a blank vector to store the simulation results

result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N
```

```
for(i in 1:N)

{

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

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

}

#plot a histogram of the simulated differences

#add a vertical line at the observed difference

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

geom_histogram(bins = 60, color = 'white') +

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

#Calculate the p-value

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


cat("The difference is significance. The departure delays in Summer are greater than Spring.")
```

**b) Summer – Autumn**

```
observed <-  mean(data$dep_delay[data$month >= 6 & data$month <= 8], na.rm = TRUE) - mean(data$dep_delay[data$month >=9 & data$month <=11], na.rm = TRUE)

observed

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)

#sample.size = the number of observations in our sample
```

```
summer_autumn = data %>% filter(month >= 6 & month <=11)
sample.size = nrow(summer_autumn)
#group.1.size = the number of observations in the first group
summer = data %>% filter(month >= 6 & month <= 8)
group.1.size = nrow(summer)
#create a blank vector to store the simulation results
result <- numeric(N)
#use a for loop to cycle through values of i ranging from 1 to N
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result[i] = mean(data$dep_delay[index], na.rm = TRUE) - mean(data$dep_delay[-index], na.rm = TRUE)
}
#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram(bins = 60, color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)


cat(‘The difference is significant. The departure delays in summer are greater than departure delays in autumn.)
```

**c) Summer – Winter**

```
observed <-  mean(data$dep_delay[data$month >= 6 & data$month <= 8], na.rm = TRUE) - mean(data$dep_delay[data$month >11 | data$month <=2], na.rm = TRUE)

observed

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)

#sample.size = the number of observations in our sample

summer_winter = data %>% filter((month >= 6 & month <=8) | (month >11 | month <=2))

sample.size = nrow(summer_winter)

#group.1.size = the number of observations in the first group

summer = data %>% filter(month >= 6 & month <= 8)

group.1.size = nrow(summer)

#create a blank vector to store the simulation results

result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N

for(i in 1:N)

{

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

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

}

#plot a histogram of the simulated differences

#add a vertical line at the observed difference

ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram(bins = 60, color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)


cat('The difference is significant. The departure delays at Summer are greater than winter.')
```

**d) Spring – Autumn**

```
observed <- mean(data$dep_delay[data$month >=3 & data$month <=5], na.rm = TRUE) - mean(data$dep_delay[data$month >= 9 & data$month <= 11], na.rm = TRUE)
observed
# Spring - Autumn
#N = number of simulations we will use
N <- 10^4-1
set.seed(2000)
#sample.size = the number of observations in our sample
spring_autumn = data %>% filter((month >=9 & month <=11) | (month >=3 & month<=5))
sample.size = nrow(spring_autumn)
#group.1.size = the number of observations in the first group
spring = data %>% filter(month >= 3 & month <= 5)
group.1.size = nrow(spring)
#create a blank vector to store the simulation results
result <- numeric(N)
#use a for loop to cycle through values of i ranging from 1 to N
for(i in 1:N)
```

```
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result[i] = mean(data$dep_delay[index], na.rm = TRUE) - mean(data$dep_delay[-index], na.rm = TRUE)
}
#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram(bins = 60, color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)


cat('The difference is significant. The departure delays in spring are greater than autumn.')
```

**e) Spring – Winter**

```
observed <- mean(data$dep_delay[data$month >=3 & data$month <=5], na.rm = TRUE) - mean(data$dep_delay[data$month >11 | data$month <= 2], na.rm = TRUE)
observed
# Spring - Winter
#N = number of simulations we will use
N <- 10^4-1
set.seed(2000)
#sample.size = the number of observations in our sample
spring_winter = data %>% filter((month > 11 | month <= 2) | (month >=3 & month<=5))
```

```
sample.size = nrow(spring_winter)

#group.1.size = the number of observations in the first group

spring = data %>% filter(month >= 3 & month <= 5)

group.1.size = nrow(spring)

#create a blank vector to store the simulation results

result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N

for(i in 1:N)

{

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

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

}

#plot a histogram of the simulated differences

#add a vertical line at the observed difference

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

geom_histogram(color = 'white') +

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

#Calculate the p-value

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


cat('The difference is significant. The departure delays in spring are greater than winter.')
```

**f) Winter – Autumn**

```
observed <- mean(data$dep_delay[data$month >11 | data$month <= 2], na.rm = TRUE) - mean(data$dep_delay[data$month >=9 & data$month <=11], na.rm = TRUE)

observed

# Winter - Autumn

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)

#sample.size = the number of observations in our sample

winter_autumn = data %>% filter((month > 11 | month <= 2) | (month >=9 & month<=11))

sample.size = nrow(winter_autumn)

#group.1.size = the number of observations in the first group

winter = data %>% filter(month > 11 | month <= 2)

group.1.size = nrow(winter)

#create a blank vector to store the simulation results

result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N

for(i in 1:N)

{

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

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

}

#plot a histogram of the simulated differences

#add a vertical line at the observed difference
```

```
ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram(color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)
cat('The difference is significant. The departure delays in winter are greater than autumn.')
```

## 3. Departure Delay Vs Temperature

**# Create a temporary dataset,'data1' that contains average departure delays grouped by temp.**

```
data1 <- data %>% mutate(temp = round(temp))
data1 <- data1 %>% group_by(temp) %>% summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE))
data1
```

**# Plotting a lineplot between temperature and average departure delays.**

```
ggplot(data1, aes(x = temp, y = avg_dep_delay))+
  geom_line()
```

**# Permutation test**

```
observed <- mean(data$dep_delay[data$temp > 50 ], na.rm = TRUE) - mean(data$dep_delay[data$temp <= 50], na.rm = TRUE)
observed
#N = number of simulations we will use
N <- 10^4-1
set.seed(2000)
#sample.size = the number of observations in our sample
```

```
sample.size = nrow(data)
#group.1.size = the number of observations in the first group
temp1 = data %>% filter(temp > 50)
group.1.size = nrow(temp1)
#create a blank vector to store the simulation results
result <- numeric(N)
#use a for loop to cycle through values of i ranging from 1 to N
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result[i] = mean(data$dep_delay[index], na.rm = TRUE) - mean(data$dep_delay[-index], na.rm = TRUE)
}
#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram( bins = 30, color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)
cat('The difference is significance. The departure delays with temp>50 is greater than temp<=50.')
```

## 4. Departure Delay Vs Wind Speed

**# Create a temporary dataset, data1 that contains the average departure delay group by wind speed.**

```
data1 <- data %>% group_by(wind_speed = round(wind_speed)) %>%
summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE))

data1
```

**# Create a lineplot for windspeed and average departure delay**

```
ggplot(data1, aes(x = wind_speed, y = avg_dep_delay))+

  geom_line()
```

**# Permutation test**

```
observed <- mean(data$dep_delay[data$wind_speed > 20 ], na.rm = TRUE) -
mean(data$dep_delay[data$wind_speed <= 20], na.rm = TRUE)

observed

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)

#sample.size = the number of observations in our sample

sample.size = nrow(data)

#group.1.size = the number of observations in the first group

ws = data %>% filter(wind_speed > 20)

group.1.size = nrow(ws)

#create a blank vector to store the simulation results

result <- numeric(N)
```

```
#use a for loop to cycle through values of i ranging from 1 to N
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result[i] = mean(data$dep_delay[index], na.rm = TRUE) - mean(data$dep_delay[-index], na.rm = TRUE)
}
#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram( bins = 30, color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)
cat('There is no significant difference. The departure delay for wind speed > 20 is equal to wind speed <=20.')
```

## 5. Departure Delay Vs Precipitation

**# Create a temporary dataset, data1 that contains the average departure delay group by precip.**

```
data1 <- data %>% group_by(precip) %>% summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE))
data1
```

**# Plotting a scatter plot between precipitation and average departure delay.**

```
ggplot(data1, aes(x = precip, y = avg_dep_delay))+
  geom_point()

# Precipitation test

observed <- mean(data$dep_delay[data$precip > 0.6 ], na.rm = TRUE) - mean(data$dep_delay[data$precip <= 0.6], na.rm = TRUE)

observed

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)

#sample.size = the number of observations in our sample

sample.size = nrow(data)

#group.1.size = the number of observations in the first group

pre = data %>% filter(precip > 0.6)

group.1.size = nrow(pre)

#create a blank vector to store the simulation results

result <- numeric(N)

#use a for loop to cycle through values of i ranging from 1 to N

for(i in 1:N)

{

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

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

}

#plot a histogram of the simulated differences

#add a vertical line at the observed difference

ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram( bins = 30, color = 'white') +

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

#Calculate the p-value

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

cat("The difference is significant. The departure delays with precip > 0.6 are greater than departure delays with precip <= 0.6.")
```

## 6. Departure Delay Vs Visibility

**# Create a temporary dataset, data1 that contains the average departure delay group by visib.**

```
data1 <- data %>% group_by(visib) %>% summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE))

data1
```

**# Plotting a scatter plot between average departure delay and visibility.**

```
ggplot(data1, aes(x = visib, y = avg_dep_delay))+
  geom_point()
```

**# Permutation test**

```
observed <- mean(data$dep_delay[data$visib <= 2.5 ], na.rm = TRUE) - mean(data$dep_delay[data$visib > 2.5], na.rm = TRUE)

observed

#N = number of simulations we will use

N <- 10^4-1

set.seed(2000)
```

```
#sample.size = the number of observations in our sample
sample.size = nrow(data)
#group.1.size = the number of observations in the first group
vis = data %>% filter(visib <= 2.5)
group.1.size = nrow(vis)
#create a blank vector to store the simulation results
result <- numeric(N)
#use a for loop to cycle through values of i ranging from 1 to N
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result[i] = mean(data$dep_delay[index], na.rm = TRUE) - mean(data$dep_delay[-index], na.rm = TRUE)
}
#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result), mapping = aes(x=result)) +
geom_histogram( bins = 30, color = 'white') +
geom_vline(xintercept = observed, color = "red")
#Calculate the p-value
(sum(result >= observed) + 1) / (N + 1)


cat("The difference is significant. The departure delays with visibility <= 2.5 are greater than visibility > 2.5")
```