NAME: VIVEK REDDY KARRA

PROJECT REPORT

This project is about working and analyzing the departure delays for United Airlines flights where we study both efficiency and customer satisfaction. In this project we compare the relationship of departure delays with six various factors and perform exploratory data analysis and perform.

The six various factors that we are comparing with the relationship between departure delays are :

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

These factors are chosen because departure delays of flights are based on hours, seasons, temperature conditions, wind speed , whether it is raining or not and the visibility conditions.

In this project we first load two dataframes “flights” and “weather” from library “nycflights2013”. After loading the two datasets, we join these two dataframes with join() function and then filter according to the carrier on which you want to perform. After going through the data, I feel that based on the factors, new variables are to be constructed based on the situation.

The new variables that are being added to the dataframe are categorical variables. The variables that are added to the variables are :

1. time_of_day : In this variable the values are added based on the hours such as morning (5am to 12pm) , noon(12pm-5pm), evening(5pm-8pm) and night(8pm-5am).

2. time_of_year : In this variable the values are added based on months value and the values are categorised as Fall(months – 9,10,11), Winter (months – 12,1,2), Spring (months – 3,4,5), Summer(months - 6,7,8).

3. temperature : In this variable the values are categorized as Cold (Below 55 F), Mild (55 to 85 F) and Hot (Above 85 F).

4. wind_speeds : In this variable the values are categorized as Low (below 30mph) and High speeds(Above 30 mph).

5. precipitation : In this variables the values are categorized as Non-rain (0 in) and raining(more than 0 in) from precip variable.

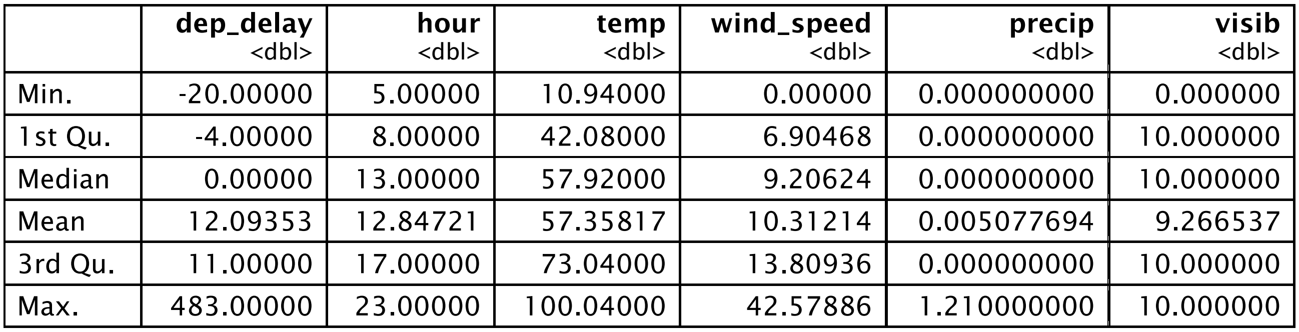
6. visibility : In this variable the values are categorized as “0-4m” and “5-10m” from the visib variable.

7. late and very_late : In some cases late and very_late variables are also constructed. One variable called "late" which is TRUE if the departure delay was greater than 0 and FALSE otherwise, and one variable called "very_late" which is TRUE if the departure delay was greater than 30 minutes and FALSE otherwise

EXPLORATORY DATA ANALYSIS

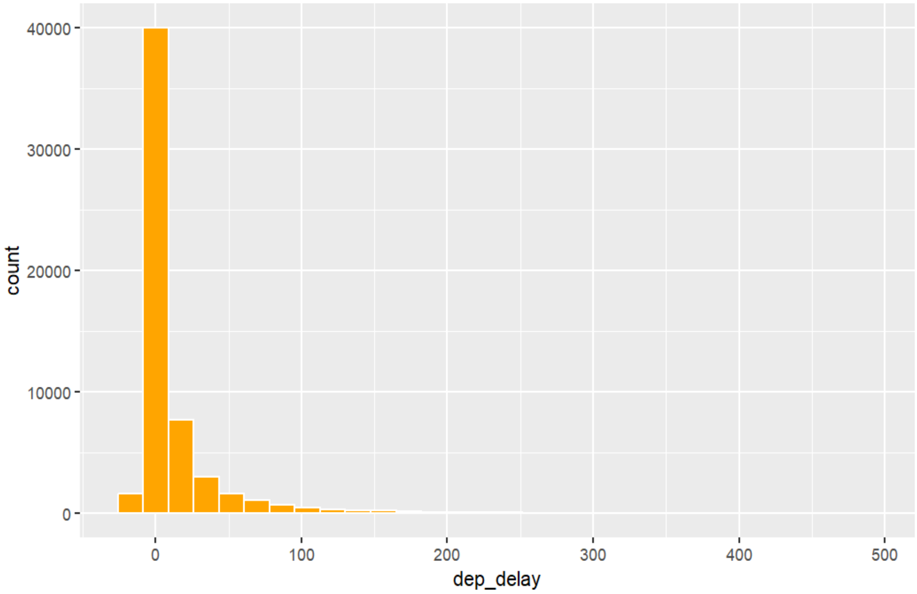
Exploratory Data Analysis (EDA) is one of the main step while performing data analysis on the data. It involves calculating statistical summaries such as mean, median, quantile etc and also data visualization such as plotting barplots, histograms, scatter plots , boxplots etc. which is used for comparing the relationship between two variables.

Summary Table of UA flights

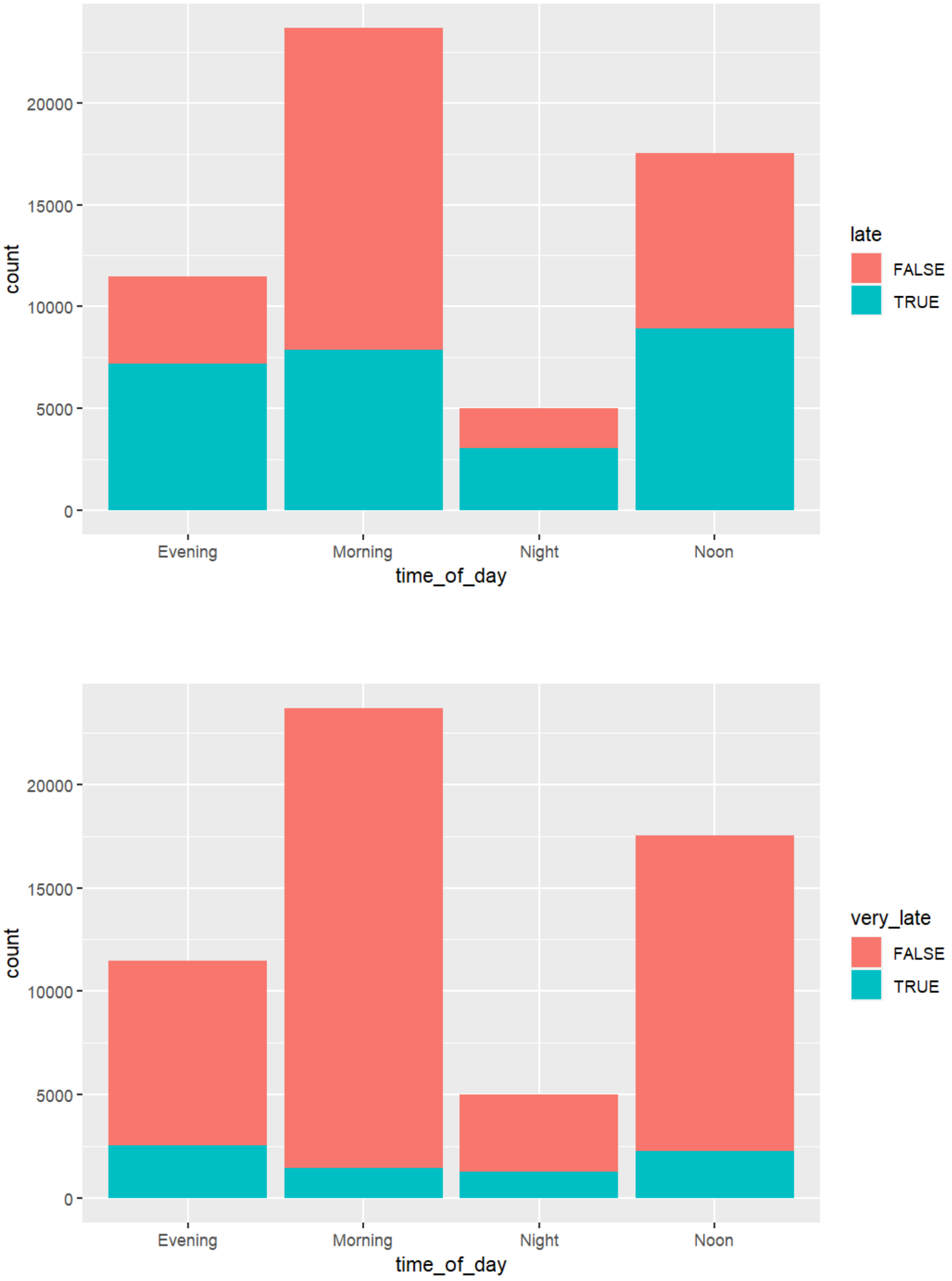


	dep_delay <dbl>	hour <dbl>	temp <dbl>	wind_speed <dbl>	precip <dbl>	visib <dbl>
Min.	-20.00000	5.00000	10.94000	0.00000	0.000000000	0.000000
1st Qu.	-4.00000	8.00000	42.08000	6.90468	0.000000000	10.000000
Median	0.00000	13.00000	57.92000	9.20624	0.000000000	10.000000
Mean	12.09353	12.84721	57.35817	10.31214	0.005077694	9.266537
3rd Qu.	11.00000	17.00000	73.04000	13.80936	0.000000000	10.000000
Max.	483.00000	23.00000	100.04000	42.57886	1.210000000	10.000000

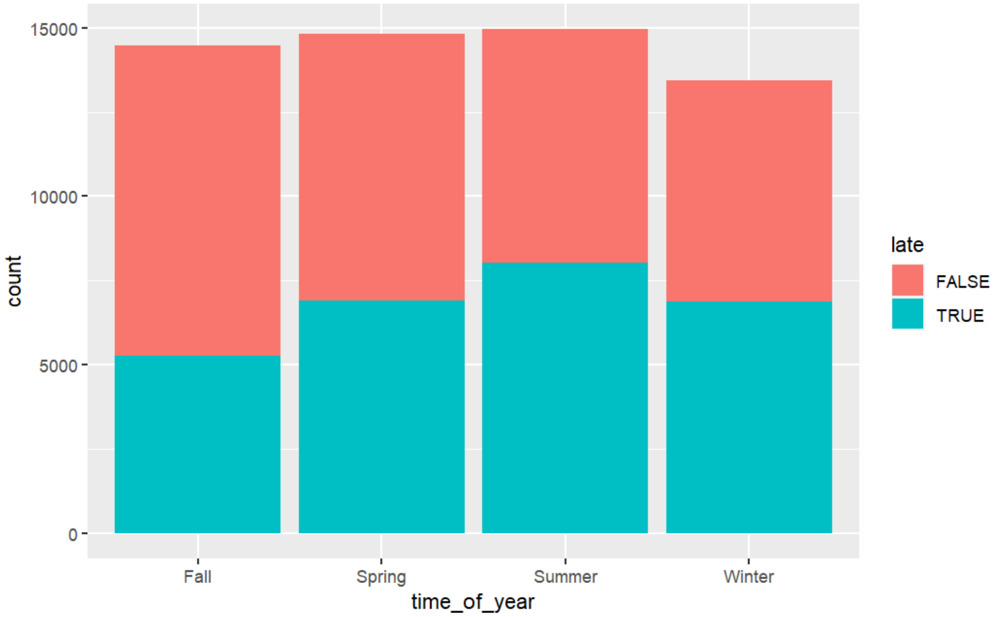
Histogram plot for departure delay of UA flights

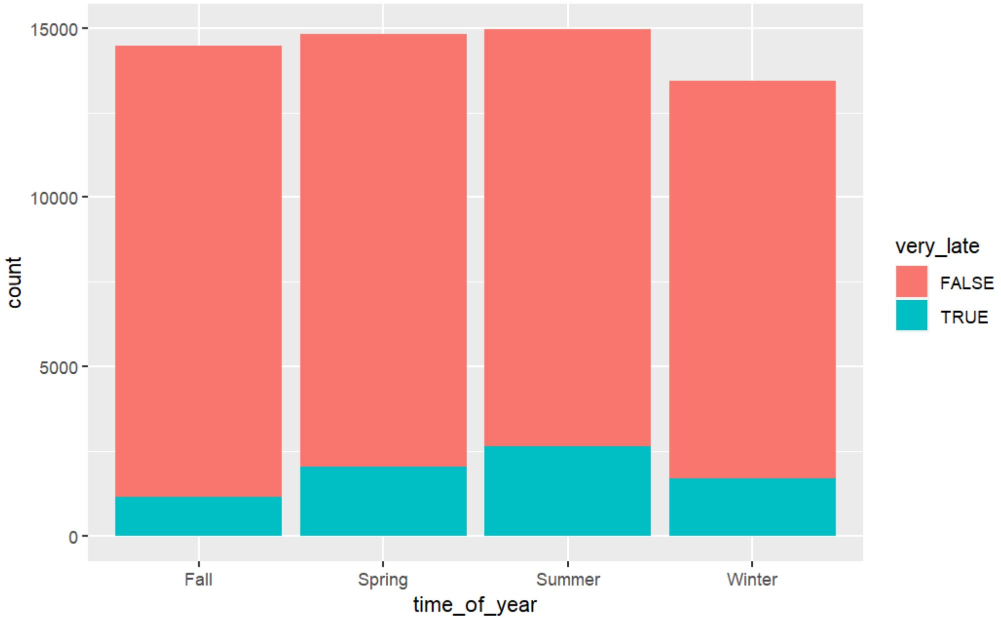


Relationship between departure delay and time_of_day(bar plot)

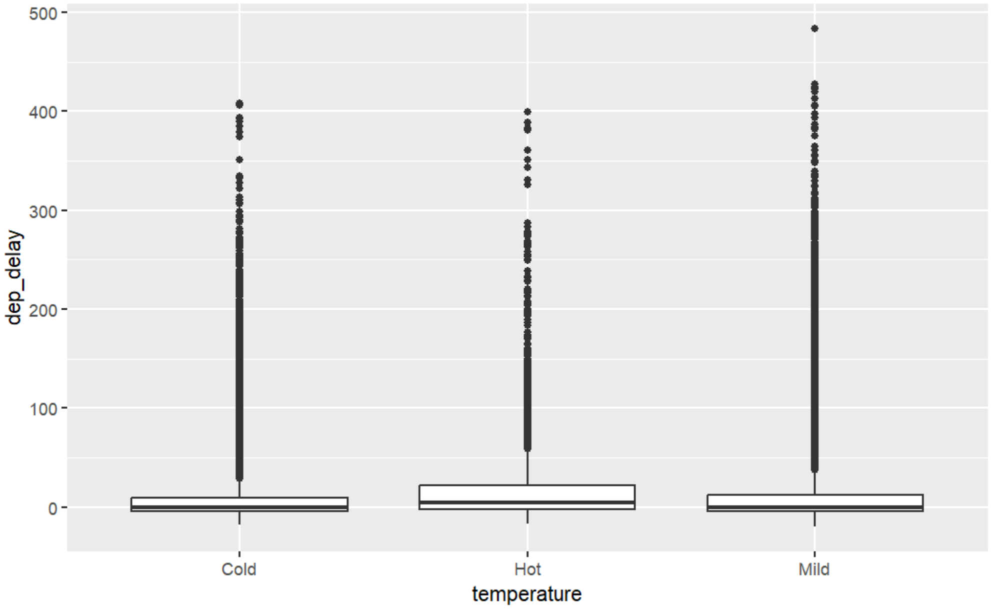


Relationship between departure delay and time_of_year

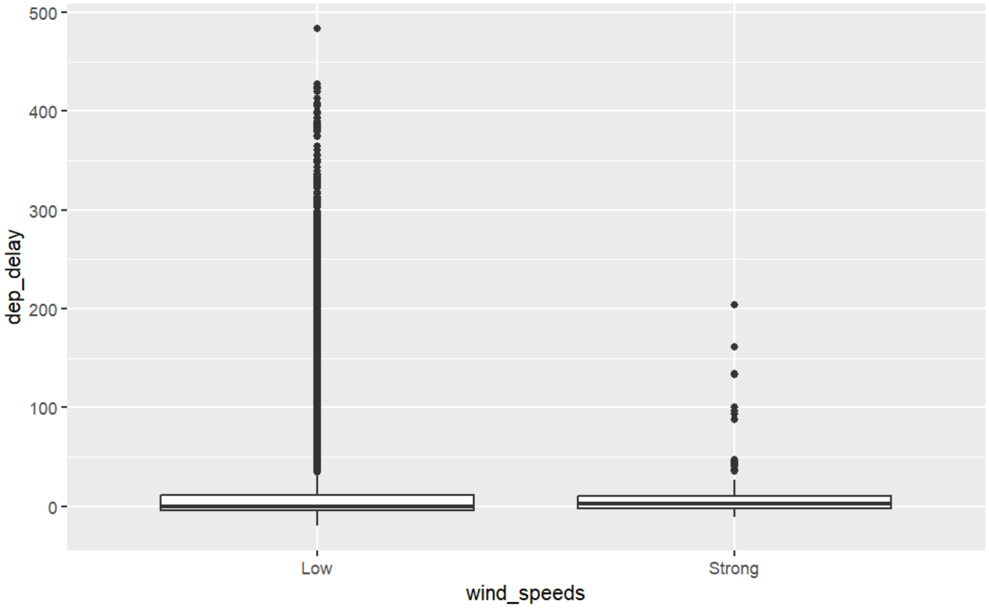




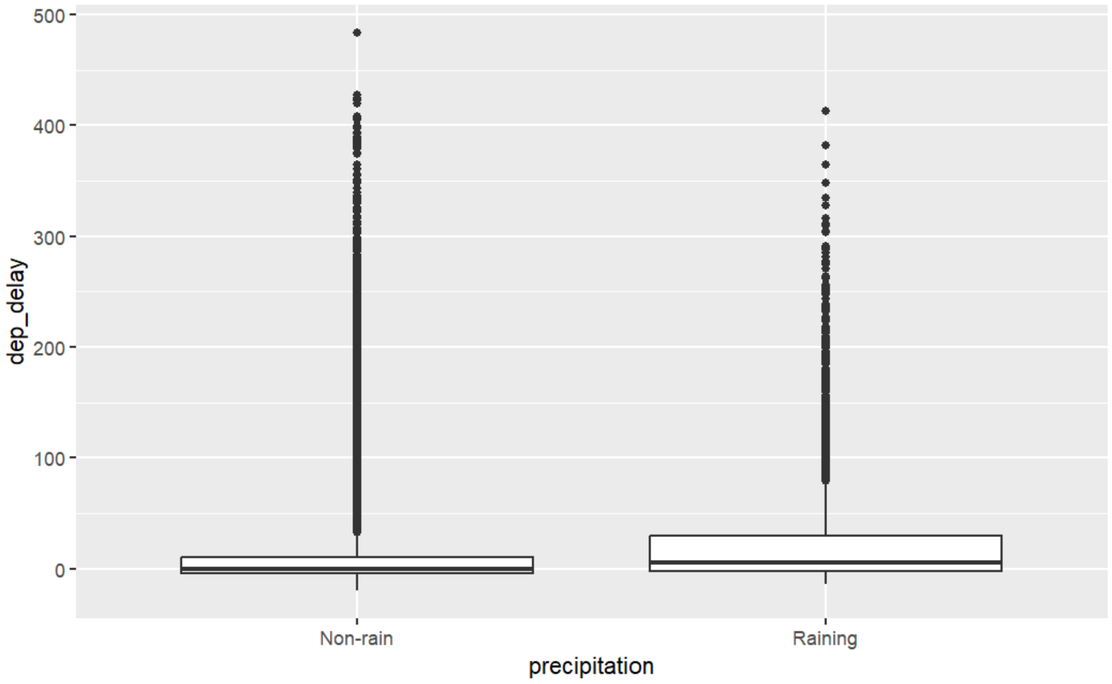
Relationship between departure delay and temperature



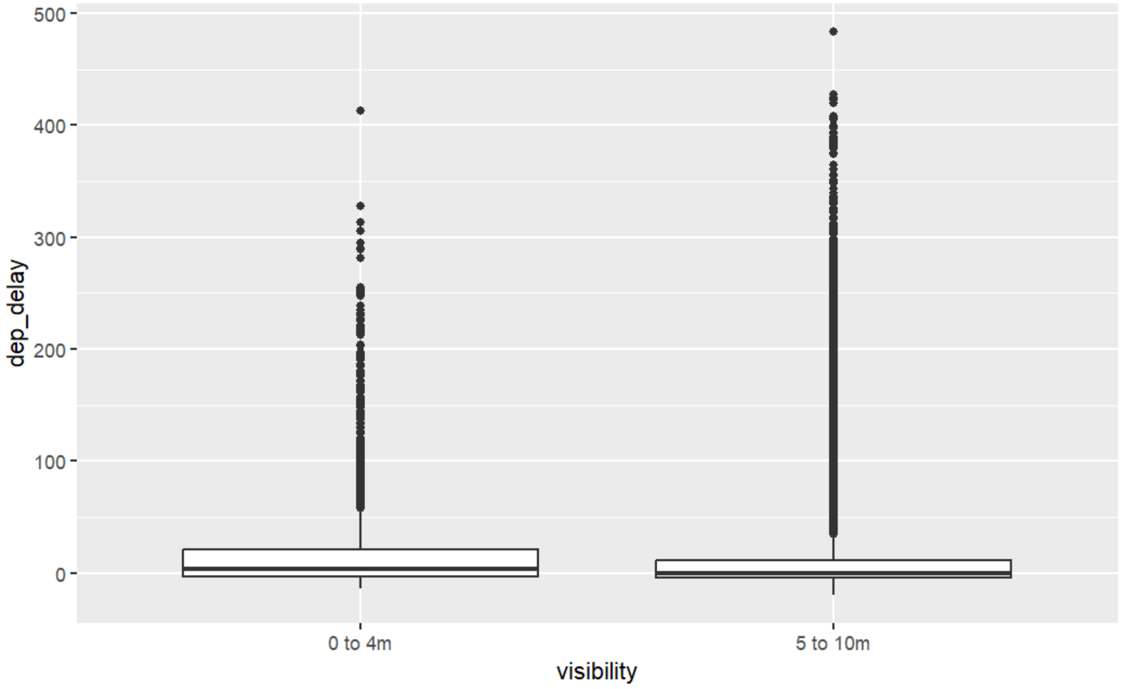
Relationship between departure delay and windspeed



Relationship between departure delay and precipitation



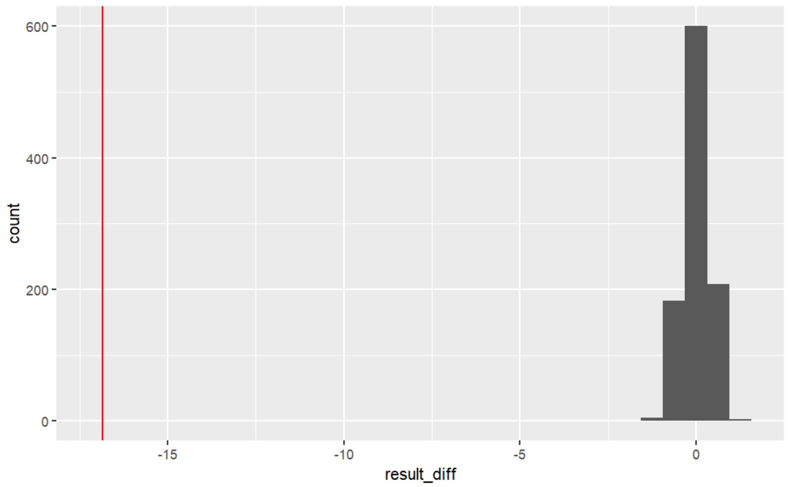
Relationship between departure delay and visibility



PERMUTATION TESTS

1. Permutation test between departure delay and time of day .

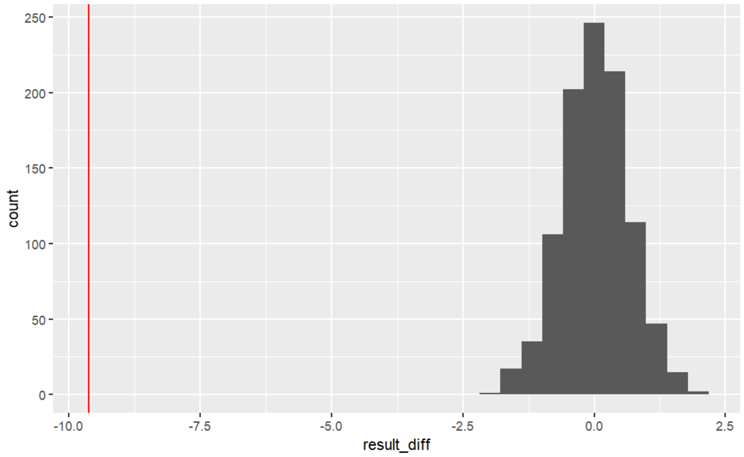
a) Permutation test is performed between departure delays for Morning and Evening.



The mean of observed difference value of departure delays between morning and evening is recorded as -16.84 and the p-value is 0.001.

From the above permutation test we observe that there is large differences between both groups. Finally, I conclude that on average flight delays are more during evening.

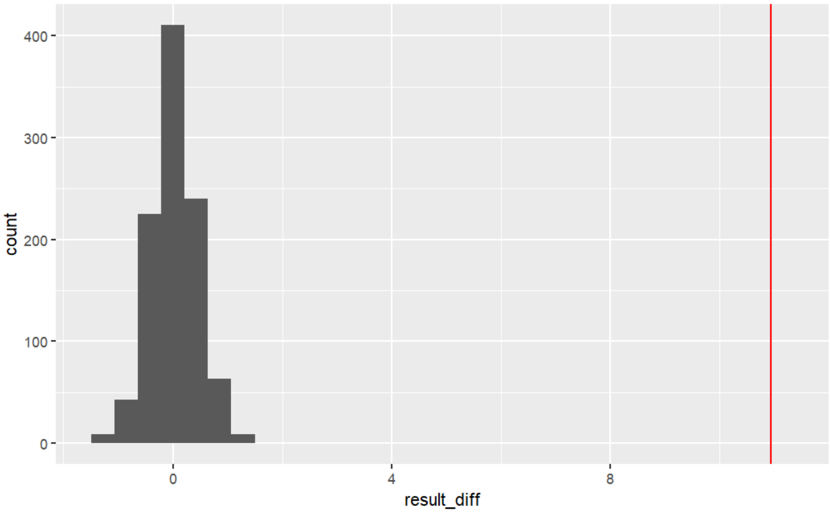
b) Permutation test is performed between departure delays for Noon and Night.



The mean observed difference value of departure delays of flights between Noon and Night is recorded as -9.61 and the p-value is 0.001 which will be statistically significant.

From the above permutation test we observe that there is a difference between Noon and Night. Finally, we can conclude that the average flight delays are more during Night than Noon.

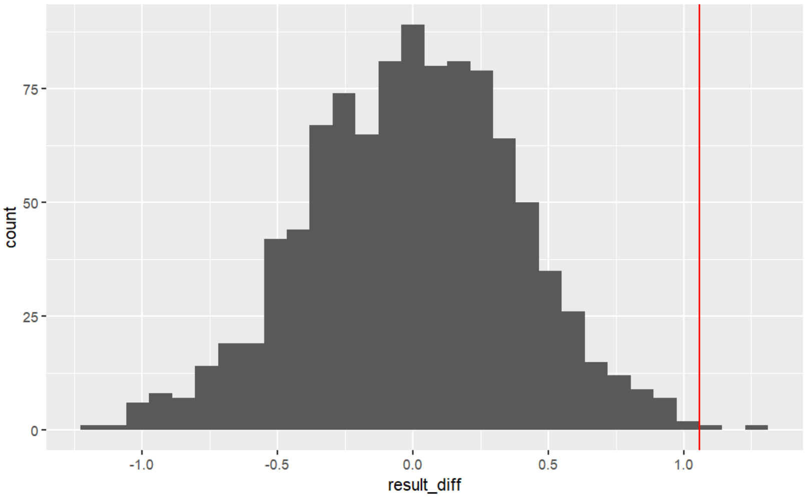
2. Permutation test is performed between departure delay and time of the year
    a. Summer and Fall



The mean observed difference of departure delay between Summer and Fall of year recorded as 10.94 and the p-value we got is 0.001 which is statistically significant.

From the permutation test it is observed that there is difference between both the groups. Finally, it can be concluded that on an average the flights delays are happening in Summer.

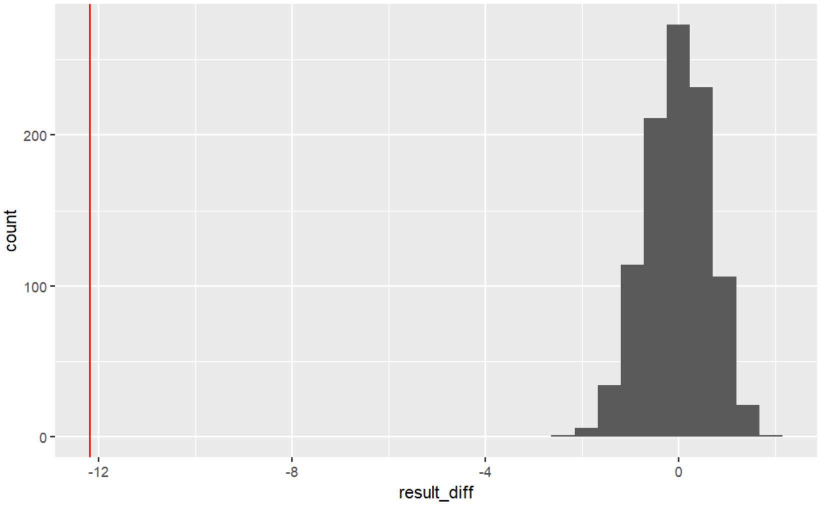
b. Spring and Winter



The mean of observed difference of departure delay between Spring and Winter Season is 1.05 and the p-value is 0.003 which is statistically significant.

From this permutation test it is observed that there is a small difference between both the groups. So, I finally conclude that on average there are some flight delays happening during Spring.

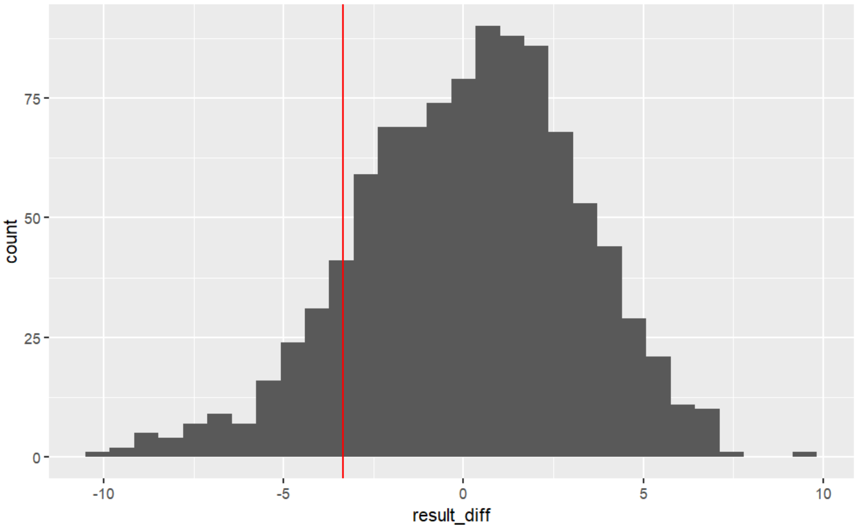
3. Permutation test between departure delays and temperature



The observed mean difference of departure delays of flights between cold and hot temperature is -12.18 and the p-value is 0.001 which is statistically significant.

From this permutation test it is observed that there is big difference between both the groups and on average the flight delays are happening during hot climate.

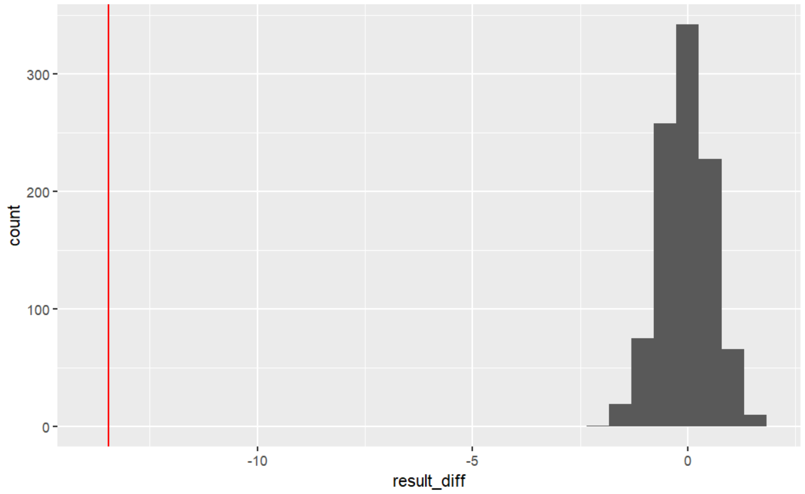
4. Permutation test between departure delays and windspeed



The observed mean difference between departure delays of low windspeed and strong windspeed is -3.35 and the p-value is recorded as 0.129 which cannot be statistically significant.

From the permutation test it is observed that there is some difference between the groups. Finally, from this we can conclude that the average flight delays are during strong windspeed.

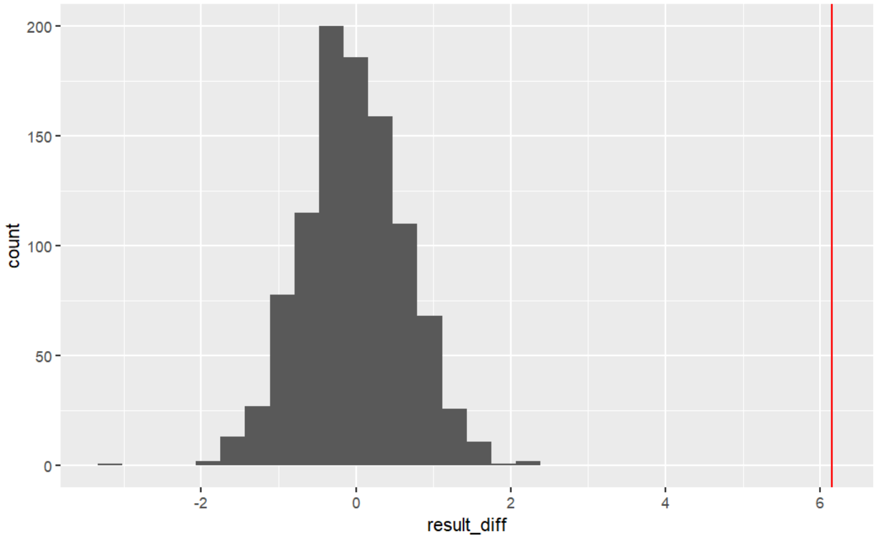
5. Permutation test between departure delays and precipitation



The observed mean difference value between departure delay of Non-rain and Raining is recorded as -13.46 and the p-value is recorded as 0.001 which is statistically significant.

From this permutation test it is observed that there is a huge difference between two groups. With this I conclude that the average flight delays is occurring during rain.

6. Permutation test is performed departure delay of flights and visibility



The observed mean difference of departure delays for flights of visibility length "0-4m" and "5-10m" is 6.155 and the p-value is 0.001 which is statistically significant.

From this permutation test we can say that there is some difference between two groups. With this I finally conclude that the average flight delays is also occurring due to less visibility length which is less than 5m.

APPENDIX

~~~
#loading the libraries
```{r}
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
library(nycflights13)
```

#loading the required datasets
```{r}
data("flights")
data("weather")
```

#joining two datasets into one dataframe
```{r}
flights_weather_joined <- flights %>%
  inner_join(weather, by = c("year", "month", "day", "hour", "origin"))
glimpse(flights_weather_joined)
```

#filtering datasets according to UA carrier and removing NA values wherever possible
```{r}
UA_flights <- flights_weather_joined %>%
  filter(carrier=="UA")%>%
~~~

~~~~
  filter(!is.na(dep_delay))%>%
  filter(!is.na(temp))%>%
  filter(!is.na(wind_speed))
```
#constructing dataframe of summaries for six objects
```{r}
table <- cbind(
summary(UA_flights$dep_delay),
summary(UA_flights$hour),
summary(UA_flights$temp),
summary(UA_flights$wind_speed),
summary(UA_flights$precip),
summary(UA_flights$visib))

columns <- c("dep_delay", "hour", "temp", "wind_speed", "precip", "visib")
colnames(table)<- columns
data.frame(table)
```


#Constructing the new categorical variables based on hour, month, temperature, wind_speed,
precip and visib
```{r}
new_df <- UA_flights %>%
  mutate(                        #mutate is used for creating new variables in dataset
    late = dep_delay > 0 ,
    very_late = dep_delay > 30,

    time_of_day = case_when(      #based on hour variable the new variable time_of_day is created
      hour>=5 & hour<12 ~ "Morning",        #the values are divided into “morning”, “noon”,
~~~~

```
  hour>=12 & hour<17 ~ "Noon",             #  “evening” and “night”
  hour>=17 & hour<20 ~"Evening",
  TRUE ~ "Night"
 ),
    time_of_year = case_when(     #time_of_year variable is created based on hour
     month %in% c(3, 4, 5) ~ "Spring",          # is divided into 4 values  “spring”, “summer”
     month %in% c(6, 7, 8) ~ "Summer",          # “fall” and “winter”
     month %in% c(9, 10, 11) ~ "Fall",
     month %in% c(12, 1, 2) ~ "Winter"
    ),

 temperature = case_when(    #temperature variable is created based on temperature values and
 temp < 55 ~ "Cold",            #is divided into Cold, Mild, and Hot
 temp >= 55 & temp < 85 ~ "Mild",
 temp >= 85 ~ "Hot"
),

 wind_speeds = case_when(
 wind_speed < 30 ~ "Low",
 wind_speed >= 30 ~ "Strong"
),

precipitation = case_when(
 precip<=0 ~ "Non-rain",
 precip>0 ~ "Raining"
),

visibility = case_when(
```

~~~~
    visib< 5.0 ~ "0 to 4m",
    visib>=5.0 ~ "5 to 10m"
  )
  )
```


#constructing histogram for departure delay(dep_delay) variable
```{r}
ggplot(data = new_df, mapping=aes(x=dep_delay))+
  geom_histogram(color="white", fill="orange")
```
#constructing barplots for dep_delay and time_of _day
```{r}
ggplot(data = new_df, mapping = aes(x = time_of_day, fill = late)) +
  geom_bar()
ggplot(data = new_df, mapping = aes(x = time_of_day, fill = very_late)) +
  geom_bar()
```


#constructing barplots for dep_delay and time_of_year
```{r}
ggplot(data = new_df, mapping = aes(x = time_of_year, fill = late)) +
  geom_bar()
ggplot(data = new_df, mapping = aes(x = time_of_year, fill = very_late)) +
  geom_bar()
```
~~~~

~~~
#constructing boxplot for dep_delay and temperaturue
```{r}
ggplot(data = new_df, mapping = aes(x = temperature, y = dep_delay)) +
  geom_boxplot()
```

#constructing boxplot for wind_speeds and dep_delay
```{r}
ggplot(data = new_df, mapping = aes(x = wind_speeds, y = dep_delay)) +
  geom_boxplot()
```

#constructing boxplot for precipitation and dep_delay
```{r}
ggplot(data = new_df, mapping = aes(x = precipitation, y = dep_delay)) +
  geom_boxplot()
```

#constructing boxplot for visibility and dep_delay
```{r}
ggplot(data = new_df, mapping = aes(x = visibility, y = dep_delay)) +
  geom_boxplot(color = "darkblue")
```

# permutation tests for dep_delay and time_of_day (Morning & Evening)
```{r}
hour <- new_df%>%
  filter(time_of_day=="Morning" | time_of_day=="Evening")
~~~

~~~~
observed_mean_hour <- mean(hour$dep_delay[hour$time_of_day=="Morning"]) - mean(hour$dep_delay[hour$time_of_day=="Evening"])
observed_mean_hour
N <- 10^3-1
sample.size = nrow(hour)
group.1.size = nrow(hour[hour$time_of_day=="Morning",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(hour$dep_delay[index]) - mean(hour$dep_delay[-index])
}
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_hour, color = "red")
#p-value
(sum(result_diff <= observed_mean_hour) + 1) / (N + 1)
```


# permutation tests for dep_delay and time_of_day (Noon&Night)
```{r}
hour2 <- new_df%>%
  filter(time_of_day=="Noon" | time_of_day=="Night")
observed_mean_hour2 <- mean(hour2$dep_delay[hour2$time_of_day=="Noon"]) - mean(hour2$dep_delay[hour2$time_of_day=="Night"])
observed_mean_hour2
N <- 10^3-1
sample.size = nrow(hour2)
group.1.size = nrow(hour2[hour2$time_of_day=="Noon",])
~~~~

~~~~
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(hour2$dep_delay[index]) - mean(hour2$dep_delay[-index])
}
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_hour2, color = "red")
#p-value
(sum(result_diff <= observed_mean_hour2) + 1) / (N + 1)
```


#permutation test for dep_delay and time_of_year (summer, fall)
```{r}
season <- new_df%>%
  filter(time_of_year=="Summer" | time_of_year=="Fall")
observed_mean_year <- mean(season$dep_delay[season$time_of_year=="Summer"]) - mean(season$dep_delay[season$time_of_year=="Fall"])
observed_mean_year
N <- 10^3-1
sample.size = nrow(season)
group.1.size = nrow(season[season$time_of_year=="Summer",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(season$dep_delay[index]) - mean(season$dep_delay[-index])
}
~~~~

~~~~
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_year, color = "red")
#p-value
(sum(result_diff >= observed_mean_year) + 1) / (N + 1)
```


#permutation test for dep_delay and time_of_year (Spring and Winter)
```{r}
season2 <- new_df%>%
  filter(time_of_year=="Spring" | time_of_year=="Winter")
observed_mean_year2 <- mean(season2$dep_delay[season2$time_of_year=="Spring"]) - mean(season2$dep_delay[season2$time_of_year=="Winter"])
observed_mean_year2
N <- 10^3-1
sample.size = nrow(season2)
group.1.size = nrow(season2[season2$time_of_year=="Spring",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(season2$dep_delay[index]) - mean(season2$dep_delay[-index])
}
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_year2, color = "red")
#p-value
(sum(result_diff >= observed_mean_year2) + 1) / (N + 1)
```
~~~~

~~~
#permutation of dep_delay and wind_speeds

```{r}

observed_mean_wind <- mean(new_df$dep_delay[new_df$wind_speeds=="Low"]) -
mean(new_df$dep_delay[new_df$wind_speeds=="Strong"])

observed_mean_wind

N <- 10^3-1

sample.size = nrow(new_df)

group.1.size = nrow(new_df[new_df$wind_speeds=="Low",])

result_diff <- numeric(N)

for(i in 1:N)

{

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

  result_diff[i] = mean(new_df$dep_delay[index]) - mean(new_df$dep_delay[-index])

}

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

  geom_histogram() +

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

#p-value

(sum(result_diff <= observed_mean_wind) + 1) / (N + 1)
```


#permutation test for dep_delay and temperature

```{r}

temp <- new_df%>%

  filter(temperature=="Cold" | temperature=="Hot")

observed_mean_temp <- mean(temp$dep_delay[temp$temperature=="Cold"]) -
mean(temp$dep_delay[temp$temperature=="Hot"])

observed_mean_temp

N <- 10^3-1
~~~

~~~
sample.size = nrow(temp)
group.1.size = nrow(temp[temp$temperature=="Cold",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(temp$dep_delay[index]) - mean(temp$dep_delay[-index])
}
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_temp, color = "red")
#p-value
(sum(result_diff <= observed_mean_temp) + 1) / (N + 1)
```


#permutation test for dep_delay and precipitation
```{r}
observed_mean_diff <- mean(new_df$dep_delay[new_df$precipitation=="Non-rain"]) - mean(new_df$dep_delay[new_df$precipitation=="Raining"])
observed_mean_diff
N <- 10^3-1
sample.size = nrow(new_df)
group.1.size = nrow(new_df[new_df$precipitation=="Non-rain",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(new_df$dep_delay[index]) - mean(new_df$dep_delay[-index])
}
~~~

~~~~
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_diff, color = "red")
#p-value
(sum(result_diff <= observed_mean_diff) + 1) / (N + 1)
```


#permutation test for dep_delay and visibility
```{r}
observed_mean_visib <- mean(new_df$dep_delay[new_df$visibility=="0 to 4m"]) - mean(new_df$dep_delay[new_df$visibility=="5 to 10m"])
observed_mean_visib
N <- 10^3-1
sample.size = nrow(new_df)
group.1.size = nrow(new_df[new_df$visibility=="0 to 4m",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(new_df$dep_delay[index]) - mean(new_df$dep_delay[-index])
}
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_visib, color = "red")
#p-value
(sum(result_diff >= observed_mean_visib) + 1) / (N + 1)
```
~~~~