PROJECT – 1

Introduction: We worked on the data of departure delays of United Airlines to improve efficiency and customer satisfaction. We have used Exploratory data analysis methods and permutation tests to compare and get the more affected part of a particular parameter. We are addressing the relationship between time of day, time of year, temperature, wind speed, precipitation, visibility and departure delay. This study would give us how these following factors are affecting the delay.

Methodology: The relationship between departure delay and the 6 parameters is represented in different types of graphs for a better understanding. We are conducting permutation tests by dividing each parameter differently as each of the parameter has different things to be considered.

Results:

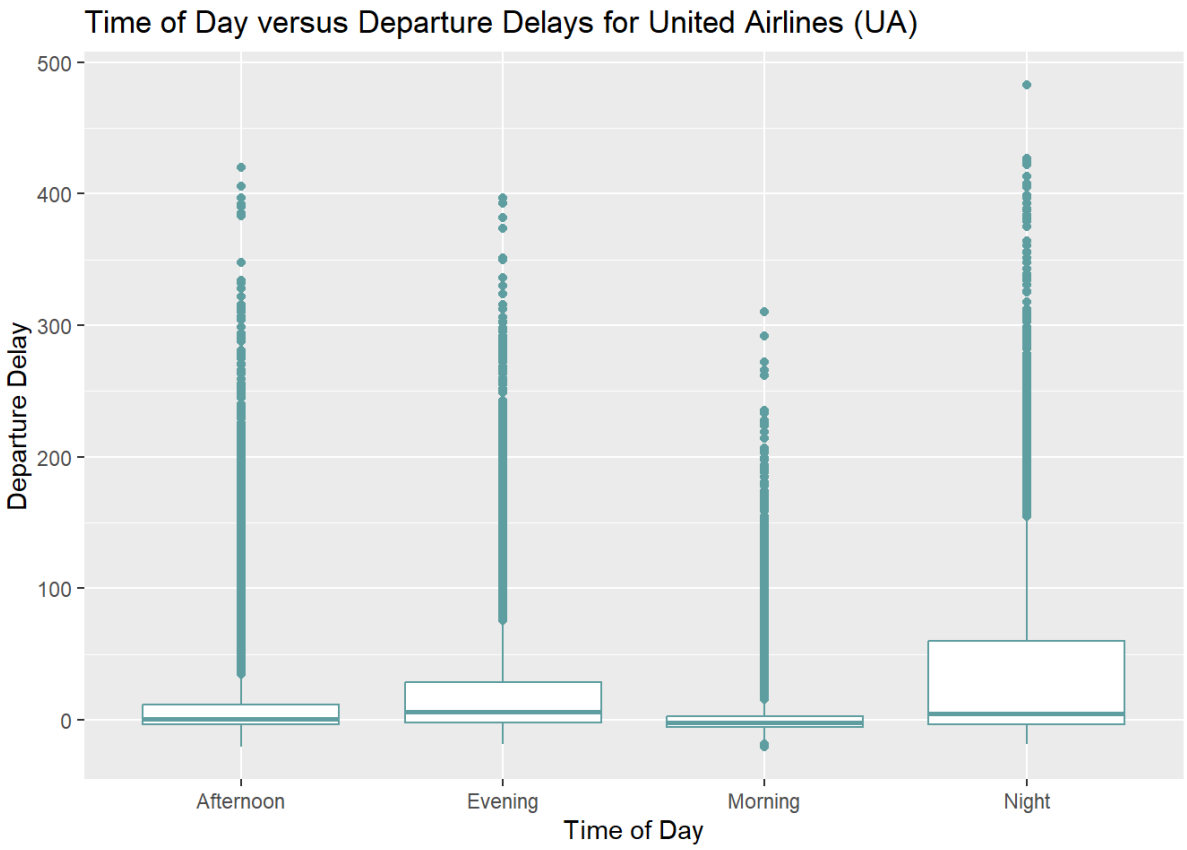
1. **Time of Day:**
   The time of day and departure delay is being compared to check at what time there are more departure delays. The time of day is divided into 4 parts.
   Morning: 6:00 AM to 12:00 noon
   Afternoon: 12:00 PM to 18:00 PM
   Evening: 18:00 PM to 21:00 PM
   Night: the rest of the time (21:00 PM to 6:00 AM)
   In the below table, dep_time is the departure time without being divided by ":" Example: 517 is 05:17 AM

```
## # A tibble: 58,361 × 3
##    dep_time dep_delay time_of_day
##       <int>     <dbl> <chr>
##  1      517         2 Night
```

```
##  2       533          4 Night
##  3       554         -4 Night
##  4       558         -2 Night
##  5       558         -2 Night
##  6       559         -1 Night
##  7       607          0 Morning
##  8       611         11 Morning
##  9       623         -4 Morning
## 10       628         -2 Morning
## # i 58,351 more rows
```

As per our suggested time of day the time_of_day is classified to Morning, Afternoon, Evening, Night.

Graph:



We have taken a boxplot in this case because it would give us a visual summary of values in the dataset.

Afternoon and Evening, the departure delays are almost same except for a few outliers. In the morning, the departure delays

are less than any other time of the day. They are the most during night.
Overall, during all the times there is significant departure delays.

Permutation Test:

Permutation test between Morning and Afternoon:

While working on the test, we have removed NA values for the study to be more accurate and understandable.
The observed difference is **-7.68191** which means the departure delays in the morning are less than the departure delays in the afternoon. The p-value is found to be **0.0002** which is more than 0.05 implying that it is statistically significant. It means that the time of day is affecting the departure delays. We cannot say if it is positive or negative because it is just the time of a day.

Permutation test between Afternoon and Evening:

While working on the test, we have removed NA values for the study to be more accurate and understandable.
The observed difference is **-11.63008** which means the departure delays in the afternoon are less than the departure delays in the evening. The p-value is found to be **0.0002** which is more than 0.05 implying that it is statistically significant. It means that the time of day is affecting the departure delays. We cannot say if it is positive or negative because it is just the time of a day.

Permutation test between Evening and Night:

While working on the test, we have removed NA values for the study to be more accurate and understandable.
The observed difference is **-18.06599** which means the departure delays in the evening are less than the departure delays in the night. The p-value is found to be **0.0002** which is more than 0.05 implying that it is statistically significant. It means

that the time of day is affecting the departure delays. We cannot say if it is positive or negative because it is just the time of a day.

Permutation test between Night and morning:

While working on the test, we have removed NA values for the study to be more accurate and understandable.

The observed difference is **-37.37798** which means the departure delays in the morning are less than the departure delays in the night. The p-value is found to be **0.0002** which is more than 0.05 implying that it is statistically significant. It means that the time of day is affecting the departure delays. We cannot say if it is positive or negative because it is just the time of a day.

Permutation test between Morning and Night:

In this case, we have divided the whole day into two parts Morning and Night.

Morning: 6:00 AM to 18:00 PM

Night: the rest of the time (18:00 PM to 6:00 AM)

```
## # A tibble: 57,686 × 3
##    dep_time dep_delay time_of_day
##       <int>     <dbl> <chr>
##  1      517         2 Night
##  2      533         4 Night
##  3      554        -4 Night
##  4      558        -2 Night
##  5      558        -2 Night
##  6      559        -1 Night
##  7      607         0 Morning
##  8      611        11 Morning
##  9      623        -4 Morning
## 10      628        -2 Morning
## # ℹ 57,676 more rows
```

While working on the test, we have removed NA values for the study to be more accurate and understandable.

The observed difference is **-20.77069** which means the departure delays in the morning are less than the departure delays in the night. The p-value is found to be **0.0002** which is more than 0.05 implying that it is statistically significant. It means that the time of day is affecting the departure delays. We cannot say if it is positive or negative because it is just the time of a day.

Discussion:

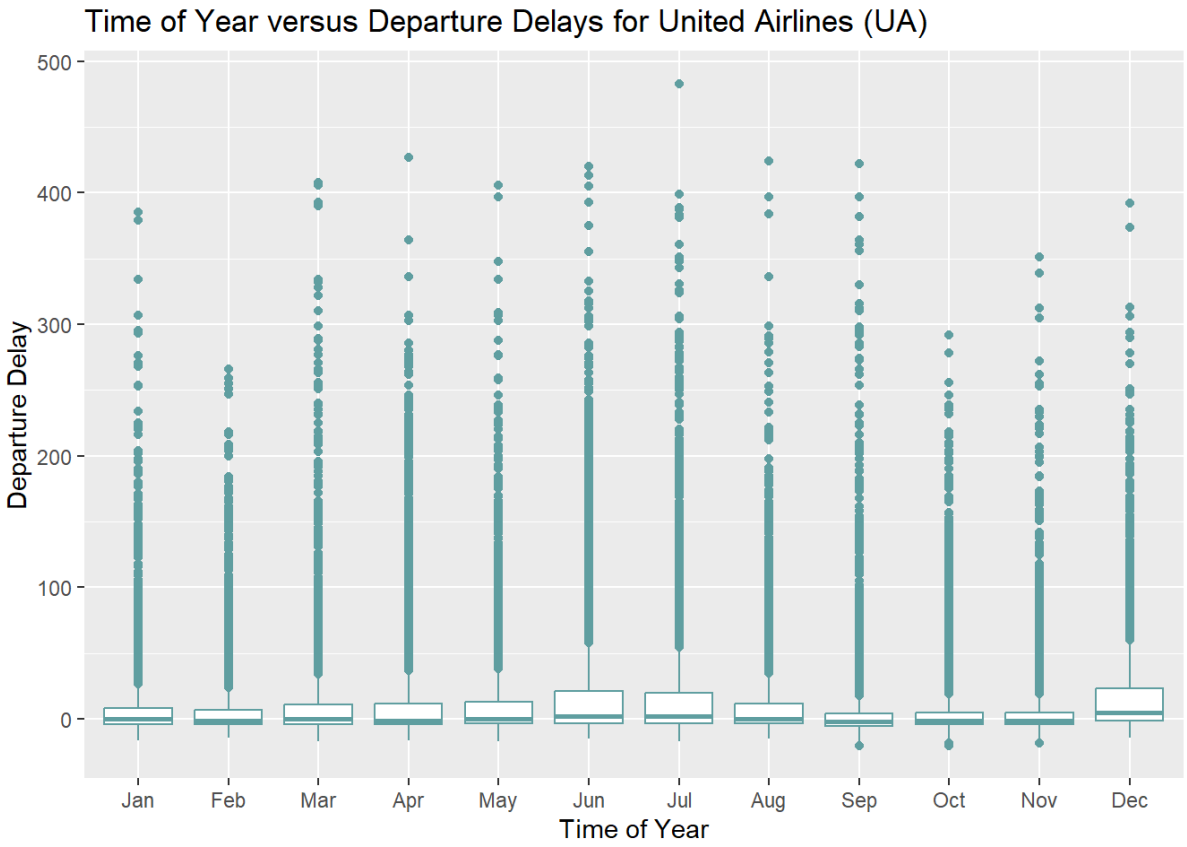
Overall, during the nights there is more departure delay than any other time in the day.

2. **Time of Year:**

   The time of Year and departure delay is being compared to check at what time of the year there are more departure delays. The time of year is divided into the usual 12 months.

```
## # A tibble: 58,361 × 2
##    month dep_delay
##    <fct>     <dbl>
##  1 Jan           2
##  2 Jan           4
##  3 Jan          -4
##  4 Jan          -2
##  5 Jan          -2
##  6 Jan          -1
##  7 Jan           0
##  8 Jan          11
##  9 Jan          -4
## 10 Jan          -2
## # i 58,351 more rows
```

Graph:



We have plotted a box plot so that it can give us a view of mean and median during those particular months. The graph is between the 12 months and departure delay. The months February and October has low departure delays. Both the months are right after the holidays (winter and summer). It might also be due to low number of flights scheduled during this month as the holiday season is finished. All other months look the same when compared with departure delays. Overall, there is no significant affect of months on the departure delay.

There are few outliers, it might be because of wrong data entry or unusual observations.

Permutation Test:

While working on the test, firstly we have divided all the months to two categories, summer and other seasons. As summer is when the people travel a lot. We have removed NA values for the study to be more accurate and understandable. Hence the study is between summer and other seasons.

The observed difference is **0** which is unusual. Means the departure delays in summer and other seasons didn't have any significant difference. The p-value is found to be **2** which is more than 0.05 implying that it is statistically insignificant. It means that the time of year is not affecting the departure delays.

Discussion:

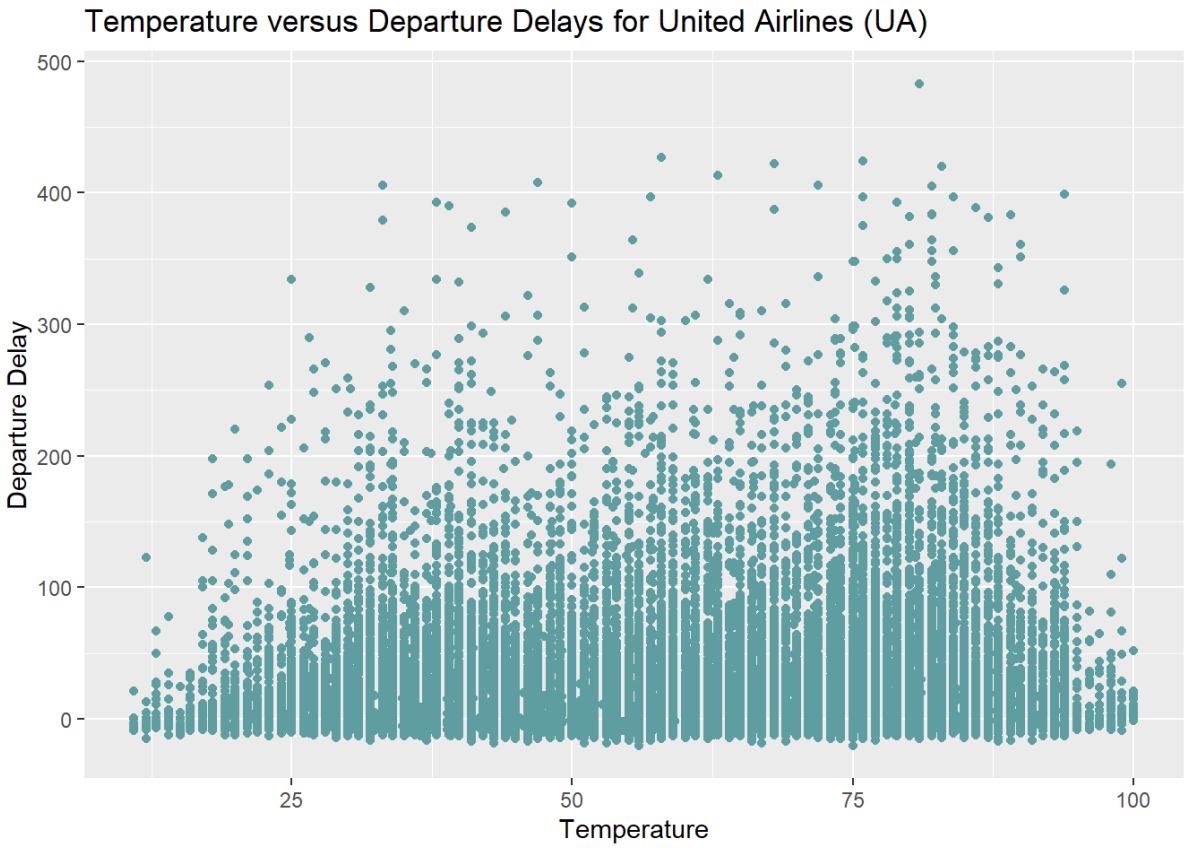
To conclude the time of year's effect on departure delay, there is no significant affect.

3. **Temperature:**

   Weather, is mainly determined by temperature (in F) is one of the main reasons for departure delays. The minimum temperature is recorded to be 10.94 and maximum temperature to be 100.04. Having the median of 57.92 and mean of 57.30.

```
## # A tibble: 58,361 × 2
##     temp dep_delay
##    <dbl>     <dbl>
##  1  39.0         2
##  2  39.9         4
##  3  39.0        -4
##  4  37.9        -2
##  5  37.9        -2
##  6  37.9        -1
##  7  37.9         0
##  8  37.9        11
##  9  39.9        -4
## 10  37.9        -2
## # i 58,351 more rows
```

Graph:



We have plotted a point plot to get a nearer view to which temperature affects the departure delays more. Each point represents the temperature and how many delays were caused during it. Very low temperature and very high temperature show less delays compared to others, this might be because there wouldn't be more instances where these temperatures would occur. This highest number of departure delays are seemed to be between the temperature 75 and 85, giving that temperature affects the delays most. There are few outliers, it might be because of wrong data entry or unusual observations.

Permutation Test:

While working on the test, we have removed NA values for the study to be more accurate and understandable. We have divided the data by median value. Below median and above median.

The observed difference is **-4.855276** which means the temperature above the median, which is higher temperature was causing more departure delays. The p-value is found to be

0.0002 which is less than 0.05 implying that it is statistically significant. It means that the temperature is affecting the departure time causing delays.

Discussion:

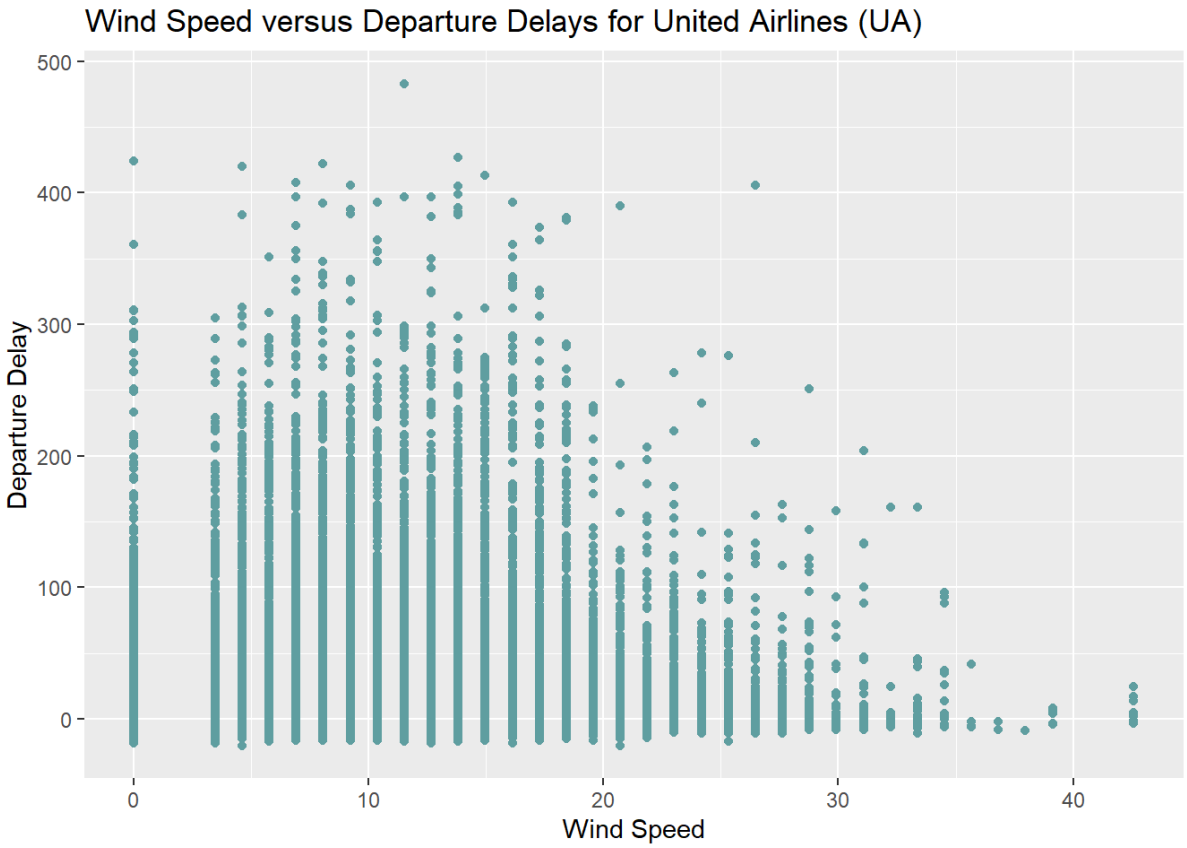
To conclude temperature's effect on departure delay, on an average medium to high temperature is causing more delay.

4. **Wind Speed:**

   Wind speed is one of the factors for departure delays. Wind speed goes from 0 to 42.579 having the median of 9.206 and mean of 10.34.

```
## # A tibble: 58,361 × 2
##    wind_speed dep_delay
##         <dbl>     <dbl>
##  1       12.7         2
##  2       15.0         4
##  3       12.7        -4
##  4       13.8        -2
##  5       11.5        -2
##  6       11.5        -1
##  7       11.5         0
##  8       13.8        11
##  9       16.1        -4
## 10       11.5        -2
## # ℹ 58,351 more rows
```

Graph:



We have plotted a point plot to get a nearer view to which wind speed affects the departure delays more. Each point represents the wind speed and how many delays were caused during it.

When the wind speed is 0 the departure delays are quite significant. As the wind speed increased the departure delays reduced. This shows that the wind speed has positively affected the departure delays. The wind speed is from 0 to 40. During the first half, low wind speed gave more departure delays. During the second half, high wind speed gave less departure delays. The few outliers would be either wrong data entry or unusual observation.

Permutation Test:

While working on the test, we have removed NA values for the study to be more accurate and understandable. We have divided the data by median value. Below median and above median as low and high.

The observed difference is **-1.947203** which means the wind speed above the median, which is wind speed from 10 and above was causing more departure delays than from 0 to 10. The p-value is found to be **0.0002** which is less than 0.05 implying that it is statistically significant. It means that the wind speed is affecting the departure time causing delays.

Discussion:

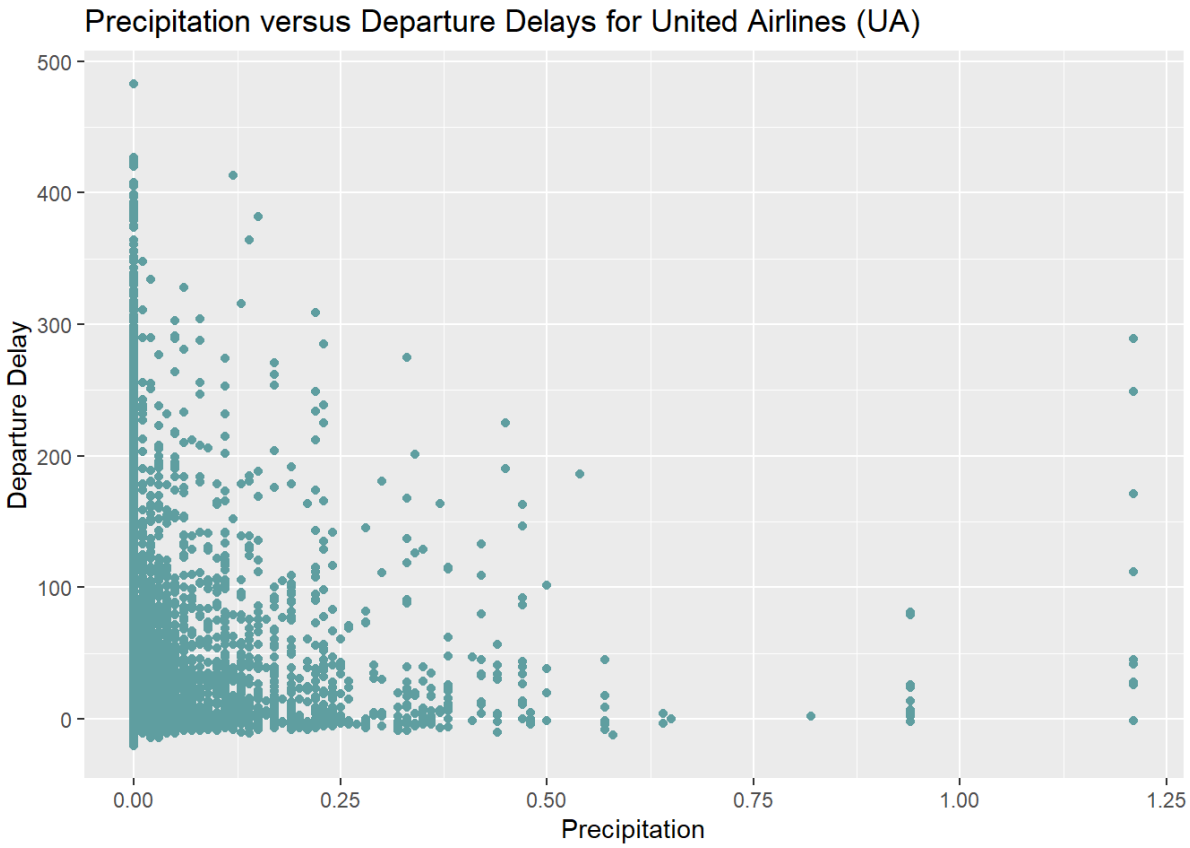
To conclude wind speed's effect on departure delay, on an average low to medium wind speed is causing more delay.

5. **Precipitation:**

   Precipitation negatively affects departure time by giving delays. Precipitation values go from 0 to 1.21 having the median of 0 and mean of 0.005218.

```
## # A tibble: 58,361 × 2
##    precip dep_delay
##     <dbl>     <dbl>
##  1      0         2
##  2      0         4
##  3      0        -4
##  4      0        -2
##  5      0        -2
##  6      0        -1
##  7      0         0
##  8      0        11
##  9      0        -4
## 10      0        -2
## # i 58,351 more rows
```

Graph:



We have plotted a point plot to get a nearer view to which precipitation affects the departure delays more. Each point represents the precipitation and how many delays were caused during it.

When the precipitation is 0 the departure delays are very significant. As the precipitation increased the departure delays reduced. This might be due to the flights being rescheduled and or halted. When the precipitation is almost 0 means the overall weather is normal, most departure delays are recorded at that time. The few outliers would be either wrong data entry or unusual observation.

Permutation Test:

While working on the test, we have removed NA values for the study to be more accurate and understandable. We have divided the data by zero and non-zero values. Below 0 and 0 as zero and above zero as non-zero.

The observed difference is **-13.47865** which means the precipitation when zero is causing more departure delays. The p-value is found to be **0.0002** which is less than 0.05 implying that it is statistically significant. It means that the precipitation is affecting the departure delays by not actually having lot of flights departing.

Discussion:

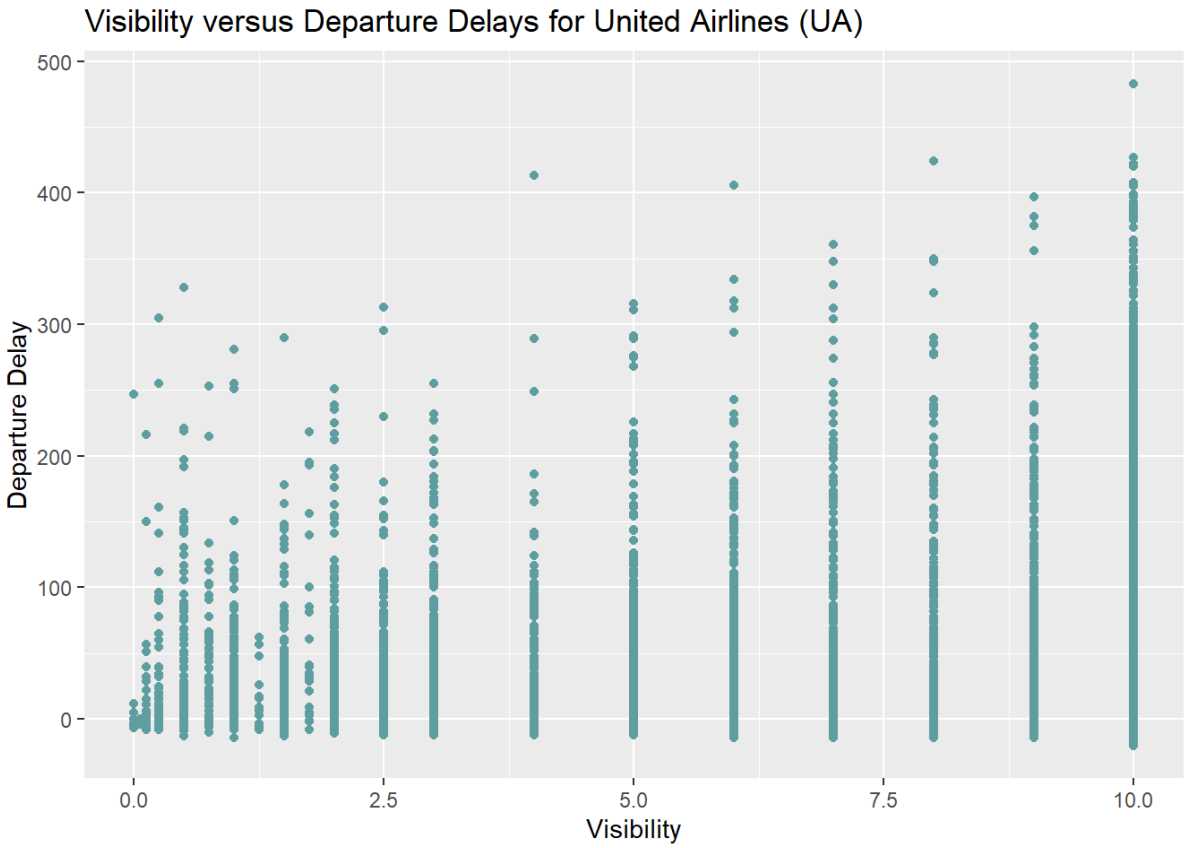
To conclude if there is more precipitation the flights are being rescheduled or cancelled giving no departure delays except in rare cases where there is data which might be a miss entry.

6. **Visibility:**

   Visibility is one of the issues for departure delays. Visibility values go from 0 to 10 having the median of 10 and mean of 9.251.

```
## # A tibble: 58,361 × 2
##    visib dep_delay
##    <dbl>     <dbl>
##  1    10         2
##  2    10         4
##  3    10        -4
##  4    10        -2
##  5    10        -2
##  6    10        -1
##  7    10         0
##  8    10        11
##  9    10        -4
## 10    10        -2
## # i 58,351 more rows
```

Graph:



We have plotted a point plot to get a nearer view to which precipitation affects the departure delays more. Each point represents the visibility and how many delays were caused during it.

When the visibility is 0 the departure delays are almost null. Because there is no chance of a flight departing when there is no visibility. As the visibility increased the departure delays increased. Because when the visibility increases there are more flights working and hence more departure delays. When the visibility is 10 means the overall weather is normal and good to fly, most departure delays are recorded at that time. As all the flights would be running during that time. The few outliers would be either wrong data entry or unusual observation.

Permutation Test:

While working on the test, we have removed NA values for the study to be more accurate and understandable. We have divided

the data by mean because median is 10 and it is the maximum value.

The observed difference is **5.941855** which means (when the visibility is less than mean) it is greater than (when the visibility is more than mean). The p-value is found to be **0.0002** which is less than 0.05 implying that it is statistically significant. It means that the visibility is affecting the departure delays by not actually having a lot of flights departing during low visibility.

Discussion:

To conclude if there is less visibility the flights are being rescheduled or cancelled giving no departure delays except in rare cases where there is data which might be a miss entry. When the visibility is increasing in the same order the departure delays are also increasing.

Conclusion:

In conclusion, departure delays for United Airlines (UA) are influenced by carious factors. Time of day, time of year, temperature, wind speed, precipitation and visibility each one of these factors affect differently. We can use this study to make decisions and incorporate strategies to improve both efficiency and customer satisfaction.

Appendix:

Code used for all the analysis:

---

title: 'Project #1'

author: "Tejaswi Neelapu"

output: html_document

date: "2023-10-29"

```{r}

library(nycflights13)

library(tidyverse)

United_Airlines <- flights %>% filter(carrier == "UA") %>% inner_join(weather, by = c("year", "month", "day", "hour", "origin"))

glimpse(United_Airlines)

```

1. Time of day

```{r}

Time_of_day <- United_Airlines %>% select(dep_time, dep_delay) %>%

  mutate(time_of_day = case_when(

    dep_time >= 600 & dep_time < 1200 ~ "Morning",

    dep_time >= 1200 & dep_time < 1800 ~ "Afternoon",

    dep_time >= 1800 & dep_time < 2100 ~ "Evening",

    TRUE ~ "Night"))

Time_of_day
```

```
ggplot(Time_of_day, aes(x = time_of_day, y = dep_delay)) +
  geom_boxplot(color = "cadetblue") +
  labs(
    x = "Time of Day",
    y = "Departure Delay",
    title = "Time of Day versus Departure Delays for United Airlines (UA)"
  )
summary(Time_of_day)
```

Comparing by dividing the day into four categories:

Permutation test between Morning and Afternoon:

```{r}
Time_of_day <- United_Airlines %>% select(dep_time, dep_delay) %>%
  mutate(time_of_day = case_when(
    dep_time >= 600 & dep_time < 1200 ~ "Morning",
    dep_time >= 1200 & dep_time < 1800 ~ "Afternoon"))

Time_of_day <- na.omit(Time_of_day)
Time_of_day
```

```
observed <- mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Morning"]) - mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Afternoon"])

observed


N <- 10^4-1


sample.size = nrow(Time_of_day)

group.1.size = nrow(Time_of_day[Time_of_day$time_of_day == "Morning",])

result <- numeric(N)


for (i in 1:N)

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

 result[i] = median(Time_of_day$dep_delay[index]) - median(Time_of_day$dep_delay[-index])

}


p_value <- 2*(sum(result <= observed) + 1) / (N + 1)

p_value
```


Permutation test between Afternoon and Evening:

```{r}
Time_of_day <- United_Airlines %>% select(dep_time, dep_delay) %>%
  mutate(time_of_day = case_when(
    dep_time >= 1200 & dep_time < 1800 ~ "Afternoon",
    dep_time >= 1800 & dep_time < 2100 ~ "Evening"))

Time_of_day <- na.omit(Time_of_day)
Time_of_day

observed <- mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Afternoon"]) - mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Evening"])
observed

N <- 10^4-1

sample.size = nrow(Time_of_day)
group.1.size = nrow(Time_of_day[Time_of_day$time_of_day == "Afternoon",])
result <- numeric(N)
```

```
for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Time_of_day$dep_delay[index]) - median(Time_of_day$dep_delay[-index])
}

p_value <- 2*(sum(result <= observed) + 1) / (N + 1)
p_value
```

Permutation test between Evening and Night:

```{r}
Time_of_day <- United_Airlines %>% select(dep_time, dep_delay) %>%
 mutate(time_of_day = case_when(
   dep_time >= 600 & dep_time < 1200 ~ "Morning",
   dep_time >= 1200 & dep_time < 1800 ~ "Afternoon",
   dep_time >= 1800 & dep_time < 2100 ~ "Evening",
   TRUE ~ "Night"))

Time_of_day <- na.omit(Time_of_day)
Time_of_day
```

```
observed <- mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Evening"]) - mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Night"])

observed


N <- 10^4-1


sample.size = nrow(Time_of_day)

group.1.size = nrow(Time_of_day[Time_of_day$time_of_day == "Evening",])

result <- numeric(N)


for (i in 1:N)

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

 result[i] = median(Time_of_day$dep_delay[index]) - median(Time_of_day$dep_delay[-index])

}


p_value <- 2*(sum(result <= observed) + 1) / (N + 1)

p_value
```


Permutation test between Morning and Night:

```{r}
Time_of_day <- United_Airlines %>% select(dep_time, dep_delay) %>%
  mutate(time_of_day = case_when(
    dep_time >= 600 & dep_time < 1200 ~ "Morning",
    dep_time >= 1200 & dep_time < 1800 ~ "Afternoon",
    dep_time >= 1800 & dep_time < 2100 ~ "Evening",
    TRUE ~ "Night"))

Time_of_day <- na.omit(Time_of_day)
Time_of_day

observed <- mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Morning"]) - mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Night"])
observed

N <- 10^4-1

sample.size = nrow(Time_of_day)
group.1.size = nrow(Time_of_day[Time_of_day$time_of_day == "Morning",])
```

```
result <- numeric(N)

for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Time_of_day$dep_delay[index]) - median(Time_of_day$dep_delay[-index])
}

p_value <- 2*(sum(result <= observed) + 1) / (N + 1)
p_value
```

Comparing by dividing the day into two categories:

Permutation test between Morning and Night:

```{r}
Time_of_day <- United_Airlines %>% select(dep_time, dep_delay) %>%
 mutate(time_of_day = case_when(
  dep_time >= 600 & dep_time < 1800 ~ "Morning",
  TRUE ~ "Night"))

Time_of_day <- na.omit(Time_of_day)
```

Time_of_day

```
observed <- mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Morning"]) - mean(Time_of_day$dep_delay[Time_of_day$time_of_day == "Night"])
observed

N <- 10^4-1

sample.size = nrow(Time_of_day)
group.1.size = nrow(Time_of_day[Time_of_day$time_of_day == "Morning",])
result <- numeric(N)

for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Time_of_day$dep_delay[index]) - median(Time_of_day$dep_delay[-index])
}

p_value <- 2*(sum(result <= observed) + 1) / (N + 1)
p_value
```

2. Time of year

```{r}
Time_of_year <- United_Airlines %>% select(month, dep_delay) %>%
  mutate(month = factor(month, labels = month.abb))

Time_of_year
ggplot(Time_of_year, aes(x = month, y = dep_delay)) +
  geom_boxplot(color = "cadetblue") +
  labs(x = "Time of Year",
       y = "Departure Delay",
       title = "Time of Year versus Departure Delays for United Airlines (UA)",
       )
summary(Time_of_year)
```

Permutation Test:

```{r}
Time_of_year <- United_Airlines %>% select(month, dep_delay) %>%
  mutate(month = case_when( month >5 & month < 10 ~ "Summer",
```

```
                month <=5 & month >= 10 ~ "Other seasons"))

Time_of_year <- na.omit(Time_of_year)
Time_of_year


observed <- mean(Time_of_year$dep_delay[Time_of_year$month ==
"Summer"]) - mean(Time_of_year$dep_delay[Time_of_year$month
!= "Other seasons"])
observed


N <- 10^4-1


sample.size = nrow(Time_of_year)
group.1.size = nrow(Time_of_year[Time_of_year$month ==
"Summer",])
result <- numeric(N)


for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Time_of_year$dep_delay[index]) -
median(Time_of_year$dep_delay[-index])
}


p_value <- 2*(sum(result <= observed) + 1) / (N + 1)
p_value
```

```

3. Temperature

```{r}
Temperature <- United_Airlines %>% select(temp, dep_delay)

Temperature
ggplot(Temperature, aes(x = temp, y = dep_delay)) +
  geom_point(color = "cadetblue") +
  labs(
    x = "Temperature",
    y = "Departure Delay",
    title = "Temperature versus Departure Delays for United Airlines (UA)"
  )
summary(Temperature)
```

Permutation Test between lower temperature and higher temperature:

```{r}

```
Temperature <- United_Airlines %>% select(temp, dep_delay)

Temperature <- na.omit(Temperature)
Temperature

observed <- mean(Temperature$dep_delay[Temperature$temp <= 57.92]) - mean(Temperature$dep_delay[Temperature$temp > 57.92])
observed

N <- 10^4-1

sample.size = nrow(Temperature)
group.1.size = nrow(Temperature[Temperature$temp <= 57.92,])
result <- numeric(N)

for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Temperature$dep_delay[index]) - median(Temperature$dep_delay[-index])
}

p_value <- 2*(sum(result <= observed) + 1) / (N + 1)
p_value
```

4. Wind Speed

```{r}
Wind_Speed <- United_Airlines %>% select(wind_speed, dep_delay)

Wind_Speed
ggplot(Wind_Speed, aes(x = wind_speed, y = dep_delay)) +
  geom_point(color = "cadetblue") +
  labs(
    x = "Wind Speed",
    y = "Departure Delay",
    title = "Wind Speed versus Departure Delays for United Airlines (UA)"
  )
summary(Wind_Speed)
```

Permutation Test between lower wind speed and higher wind speed:

```{r}
Wind_Speed <- United_Airlines %>% select(wind_speed, dep_delay)
```

```
Wind_Speed <- na.omit(Wind_Speed)
Wind_Speed

observed <- mean(Wind_Speed$dep_delay[Wind_Speed$wind_speed <= 9.206]) - mean(Wind_Speed$dep_delay[Wind_Speed$wind_speed > 9.206])
observed

N <- 10^4-1

sample.size = nrow(Wind_Speed)
group.1.size = nrow(Wind_Speed[Wind_Speed$wind_speed <= 9.206,])
result <- numeric(N)

for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Wind_Speed$dep_delay[index]) - median(Wind_Speed$dep_delay[-index])
}

p_value <- 2*(sum(result <= observed) + 1) / (N + 1)
p_value
```

5. Precipitation

```{r}
Precipitation <- United_Airlines %>% select(precip, dep_delay)

Precipitation
ggplot(Precipitation, aes(x = precip, y = dep_delay)) +
  geom_point(color = "cadetblue") +
  labs(x = "Precipitation",
       y = "Departure Delay",
       title = "Precipitation versus Departure Delays for United Airlines (UA)",
       )
summary(Precipitation)
```

Permutation test between zero and non-zero Precipitation:

```{r}
Precipitation <- United_Airlines %>% select(precip, dep_delay)

Precipitation <- na.omit(Precipitation)
Precipitation
```

```
observed <- mean(Precipitation$dep_delay[Precipitation$precip <= 0]) - mean(Precipitation$dep_delay[Precipitation$precip > 0])

observed


N <- 10^4-1


sample.size = nrow(Precipitation)

group.1.size = nrow(Precipitation[Precipitation$precip <= 0,])

result <- numeric(N)


for (i in 1:N)

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

 result[i] = median(Precipitation$dep_delay[index]) - median(Precipitation$dep_delay[-index])

}


p_value <- 2*(sum(result <= observed) + 1) / (N + 1)

p_value
```

6. Visibility

```{r}
Visibility <- United_Airlines %>% select(visib, dep_delay)

Visibility
ggplot(Visibility, aes(x = visib, y = dep_delay)) +
  geom_point(color = "cadetblue") +
  labs(x = "Visibility",
       y = "Departure Delay",
       title = "Visibility versus Departure Delays for United Airlines (UA)",
       )
summary(Visibility)
```

Permutation Test between <10 visibility and >10 visibility:

```{r}
Visibility <- United_Airlines %>% select(visib, dep_delay)

Visibility <- na.omit(Visibility)
Visibility

observed <- mean(Visibility$dep_delay[Visibility$visib <= 9.251]) - mean(Visibility$dep_delay[Visibility$visib > 9.251])
```

```
observed

N <- 10^4-1

sample.size = nrow(Visibility)
group.1.size = nrow(Visibility[Visibility$visib <= 9.251,])
result <- numeric(N)

for (i in 1:N)
 {index = sample(sample.size, size=group.1.size, replace = FALSE)
 result[i] = median(Visibility$dep_delay[index]) - median(Visibility$dep_delay[-index])
}

p_value <- 2*(sum(result >= observed) + 1) / (N + 1)
p_value
```