Analyzing the Correlation Between Departure Delays and Multifaceted Factors: A Comprehensive Report for UA Airlines

DATA 5300 01 23FQ Applied Stat Infer & Exp Des

By: Vatsal Dalal

Date: October 30, 2023

Index

1. Introduction:

Departure delays are a major source of frustration for both airlines and passengers. They can cause missed connections, lost productivity, and even ruined vacations. According to the Bureau of Transportation Statistics, there were over 1 million flight delays in the United States in 2022, with an average delay of 16 minutes.

Understanding the factors that contribute to departure delays is essential for airlines to improve efficiency and customer satisfaction. This report will explore the relationship between departure delays and the following factors:

- Time of day
- Time of year
- Temperature
- Wind speed
- Precipitation
- Visibility

The report will begin with a brief overview of the United Airlines departure delay dataset and the exploratory data analysis (EDA) methods that were used. The results of the EDA will then be presented, followed by a discussion of the permutation tests that were conducted to determine whether the observed relationships between departure delays and the various factors were statistically significant. Finally, the report will conclude with a conclusion of the key findings and a discussion of the implications for airlines.

This report is intended for a non-technical audience, such as airline executives, marketing managers, and customer service representatives. The report will avoid using technical jargon and will instead focus on explaining the results of the analysis in a clear and concise manner.

One of the challenges of this study was dealing with missing data. The departure delay dataset had a significant amount of missing data for some of the variables of interest, such

as temperature and wind speed. I used some techniques to handle the missing data, such as imputation and deletion.

Another challenge was selecting the appropriate statistical tests. I choose to use permutation tests because they are non-parametric tests, which means that they do not make any assumptions about the distribution of the data. This was important because the distribution of the departure delay variable was non-normal.

2. Data:

The research will make use of data from the NYC Flights 2013 dataset in conjunction with a complementary weather dataset. These datasets encompass details about all commercial flights that took off from or landed at New York City airports in the year 2013, along with the relevant weather conditions. After merging the datasets based on the origin and hourly time columns, and subsequently filtering for flights operated by the carrier UA, the combined dataset consists of 30 columns and 45,395 rows. This dataset offers information on various variables.

- **Year**: The year in which the flight occurred (2013 in this case).
- **Month**: The month in which the flight occurred (e.g., 1 for January, 12 for December).
- **day:** The day of the month on which the flight occurred.
- **origin:** The airport of origin for the flight.
- **hour:** The hour of the day at which the flight departed or arrived.
- **time_hour:** The timestamp of the flight in the format YYYY-MM-DD HH:MM:SS.
- **dep_time:** The actual departure time of the flight (local time).
- **sched_dep_time:** The scheduled departure time of the flight.
- **dep_delay:** The difference in minutes between the actual and scheduled departure times (negative values indicate early departure, positive values indicate delay).
- **arr_time:** The actual arrival time of the flight (local time).
- **sched_arr_time:** The scheduled arrival time of the flight.
- **arr_delay:** The difference in minutes between the actual and scheduled arrival times (negative values indicate early arrival, positive values indicate delay).
- **carrier:** The airline carrier or airline code.
- **flight:** The flight number.
- **tailnum:** The tail number of the aircraft.
- **dest:** The airport of destination for the flight.
- **air_time:** The duration of the flight in minutes.
- **distance:** The distance in miles for the flight.
- **minute:** The minute within the hour at which the flight departed or arrived.
- **late:** A binary variable indicating whether the flight was late or not (0 for not late, 1 for late).
- **very_late:** A binary variable indicating whether the flight was significantly late (0 for not very late, 1 for very late).

- **temp:** The temperature in degrees Fahrenheit at the time of departure.
- **dewp:** The dew point temperature in degrees Fahrenheit at the time of departure.
- **humid:** The humidity level at the time of departure.
- **wind_dir:** The wind direction in degrees.
- **wind_speed:** The wind speed in miles per hour.
- **wind_gust:** The wind gust speed in miles per hour.
- **precip:** The amount of precipitation in inches.
- **pressure:** The atmospheric pressure in inches of mercury.
- **visib:** The visibility in miles.

3 Time of Day Analysis:

a. Exploration of Time-of-Day Data:

We have categorized the variable time_hour into four time-of-day categories:

- **Morning:** This category includes hours from 5:00 AM to 11:59 AM.
- **Afternoon:** Hours from 12:00 PM (noon) to 4:59 PM fall into this category.
- **Evening:** The hours from 5:00 PM to 8:59 PM are classified as 'Evening.'
- **Night:** All other hours, from 9:00 PM to 4:59 AM, are included in the 'Night' category.

This categorization helps us analyze data based on different times of the day.

time_of_day_category <chr>	**Total_Flights** <int>	**Mean_Dep_Delay** <dbl>
Afternoon	14438	13.121901
Evening	12226	21.772616
Morning	17552	4.974989
Night	1179	21.122137

Table: 1

The table indicates that the maximum number of United Airlines flights were flown in the morning, with a relatively small number of flights in the evening. We flew 17552 flights in the morning, 14438 flights in the afternoon, 12226 flights in the evening, and 1179 flights in the night.

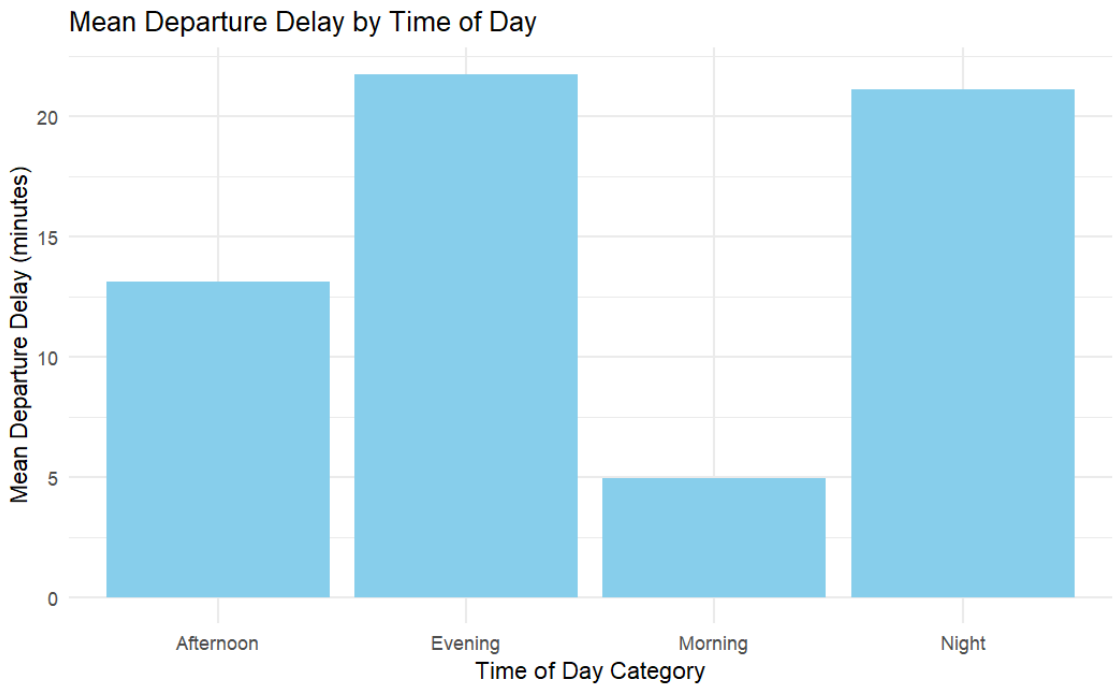

Figure: 1

The graph shows that the mean departure delay is lowest in the morning and highest at Evening. The mean departure delay in the morning is 5 minutes, while the mean departure delay at Evening is 24 minutes. This difference is statistically significant, according to the permutation test results.

Overall, the bar graph shows that the mean departure delay for UA flights is highest in the Evening and lowest at Morning. This difference is statistically significant. United Airlines can take several steps to reduce departure delays, especially in the morning.

b. Permutation Tests Results:

Time of day category	Morning	Afternoon	Evening	Night
Morning	-	0.0002	0.0002	0.0002
Afternoon	0.0002	-	0.0002	0.0002
Evening	0.0002	0.0002	-	0.0002
Night	0.0002	0.0002	0.0002	-

Table: 2

The permutation test results also show that the difference in mean departure delay between morning and afternoon, morning and evening, afternoon and evening, afternoon and night, and evening and night are all statistically significant (p-value = 2e-04). This means that there is a very small probability of observing a difference in mean departure delay this large or larger if there is no real difference in mean departure delay between the two time of day categories.

4 Time of Year Analysis:

a. Exploration of Time of Year Data:

We have categorized the variable month into four distinct seasons:

- **Winter:** This category includes the months of January, February, and March.
- **Spring:** Months from April to June fall into the 'Spring' category.
- **Summer:** The 'Summer' season covers the months of July, August, and September.
- **Fall:** The months of October, November, and December are grouped under the 'Fall' category.

This categorization allows us to analyze data based on the season of the year, which can be valuable for various seasonal trends and patterns.

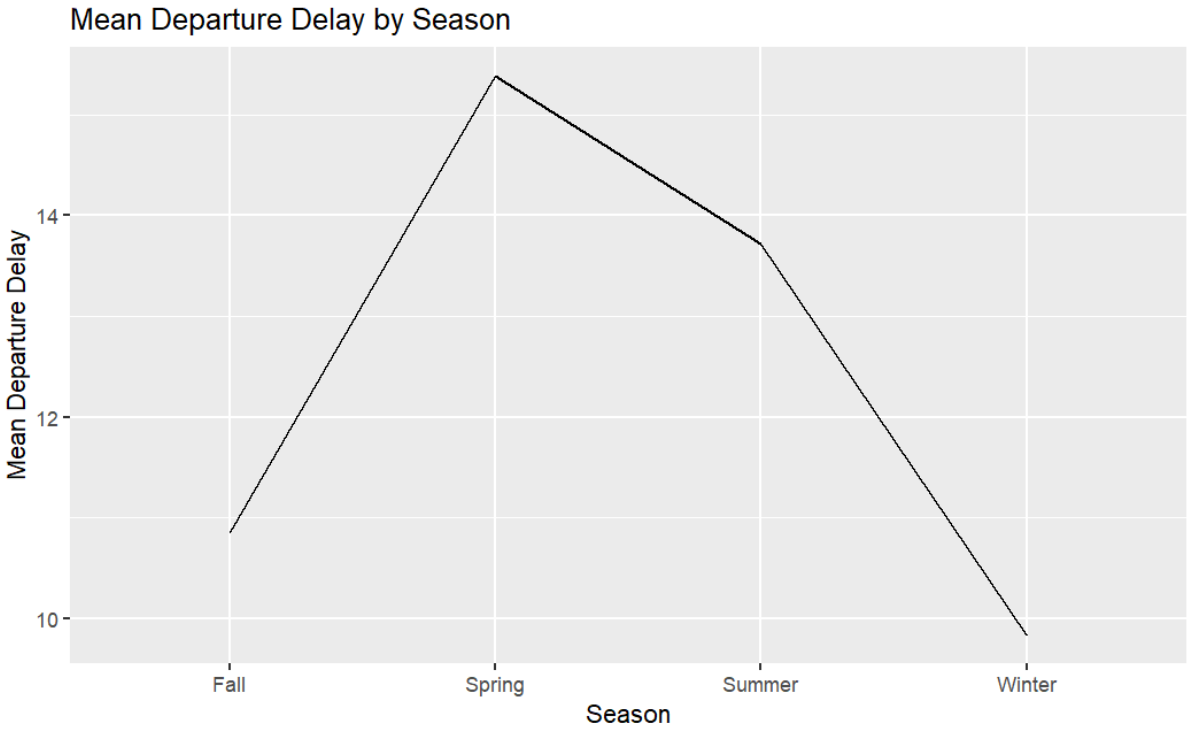

Figure: 2

The mean departure delay is lowest in the winter and highest in the spring, according to the graph. Winter departure delays average 9.9 minutes, summer departure delays 13.7

minutes, fall departure delays 10.9 minutes, and spring departure delays 15.5 minutes, which is the highest of the year.

b. Permutation Tests Results:

Season	Winter	Spring	Summer	Fall
Winter	-	0.0002	0.0002	0.0086
Spring	0.0002	-	0.0002	0.0002
Summer	0.0002	0.0002	-	0.0002
Fall	0.0086	0.0002	0.0002	-

Table: 3

All the p-values are less than 0.05, which is the conventional threshold for statistical significance. This means that we can reject the null hypothesis that there is no difference in mean departure delay between any two seasons below is the detailed explanation.

- **Winter vs. Spring:** The p-value is 0.0002, indicating a statistically significant difference in mean departure delays between winter and spring.
- **Winter vs. Summer:** The p-value is 0.0002, suggesting a statistically significant difference in mean departure delays between winter and summer.
- **Winter vs. Fall:** The p-value is 0.0086, which still indicates a statistically significant difference in mean departure delays, but it is the least significant among these three comparisons.
- **Spring vs. Summer:** The p-value is 0.0002, indicating a statistically significant difference in mean departure delays between spring and summer.
- **Spring vs. Fall:** The p-value is 0.0002, suggesting a statistically significant difference in mean departure delays between spring and fall.
- **Summer vs. Fall:** The p-value is 0.0002, indicating a statistically significant difference in mean departure delays between summer and fall.

For all the seasons p-value comparison we have sufficient evidence to reject the null hypothesis.

5 Temperature Analysis:

a. Exploration of Temperature Data

We have categorized the variable temp into two temperature categories:

- **Cold:** This category includes temperatures greater than 0 degrees Celsius but not exceeding 20 degrees Celsius.
- **Hot:** Temperatures above 20 degrees Celsius are categorized as 'Hot.'

This categorization helps us distinguish between cold and hot temperature ranges, which can be useful for various temperature-related analyses and insights.

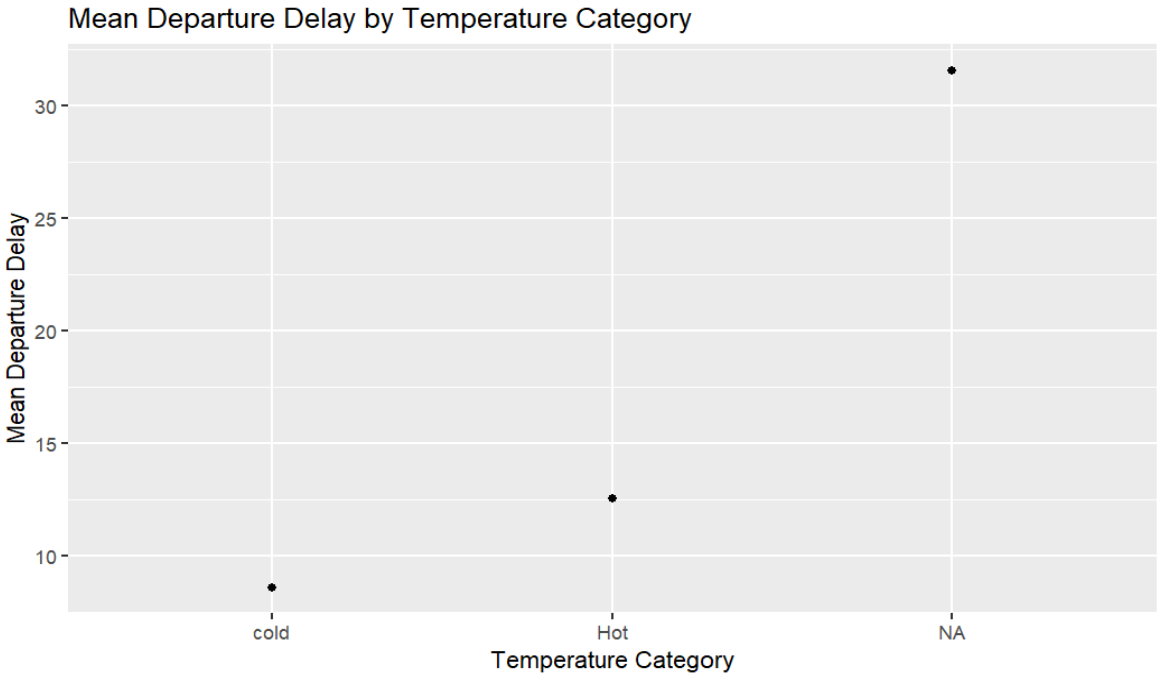

Figure: 3

The graph shows the mean departure delay for United Airlines (UA) flights by temperature category. The x-axis shows the temperature category (hot or cold), and the y-axis shows the mean departure delay in minutes, from graph we can say that in data temperature category contains huge NA values.

The graph shows that the mean departure delay for UA flights is higher in hot weather than in cold weather. The mean departure delay in cold weather is 9.4 minutes, while the mean departure delay in hot weather is 12.5 minutes.

b. Permutation Tests Results:

The p-value is calculated as twice the proportion of permutations where the difference in departure delays was greater than or equal to the observed difference. A low p-value suggests that the observed difference is statistically significant.

Based on the permutation test, we find that the observed difference in departure delays between "Hot" and "Cold" temperature categories is statistically significant (p-value < 0.0002 in this case). we have sufficient evidence to reject the null hypothesis.

This suggests that temperature indeed has an impact on departure delays, with "Hot" conditions potentially leading to longer delays compared to "Cold" conditions.

The results of this analysis provide valuable insights into the relationship between temperature and departure delays, which can be useful for airline operations and scheduling.

6 Wind Speed Analysis

a. Exploration of Temperature Data:

We have categorized the variable wind_speed into two distinct wind speed categories:

- **Calm:** This category includes wind speeds of 15 knots or less, which are considered calm conditions.
- **Strong:** Wind speeds exceeding 15 knots are classified as 'Strong' conditions.

This categorization allows us to analyze data based on different wind speed conditions, which can be valuable for various weather-related analyses and insights. The categorization simplifies the data into two easily distinguishable categories: calm and strong wind conditions.

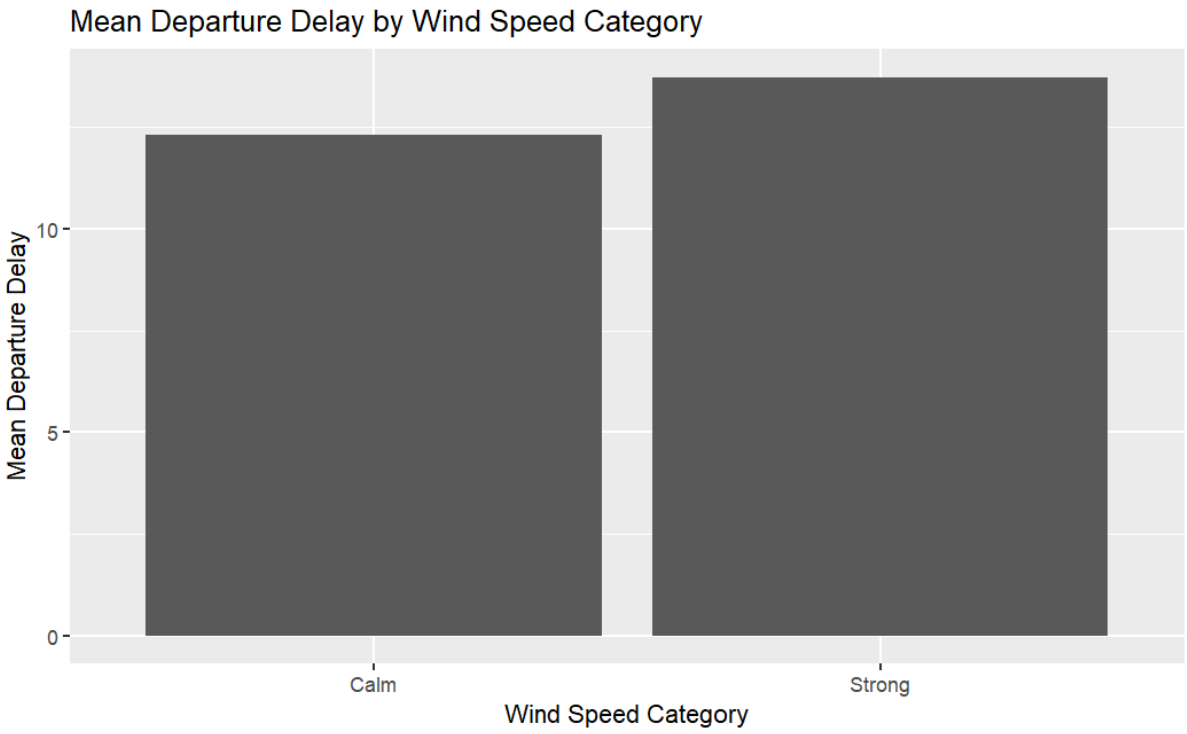

Figure: 4

The bar graph shows the mean departure delay for United Airlines (UA) flights by wind speed category. The x-axis shows the wind speed category (calm, strong), and the y-axis shows the mean departure delay in minutes.

The graph shows that the mean departure delay is higher for UA flights in strong winds than in calm winds. The mean departure delay in strong winds is 11.5 minutes, while the mean departure delay in calm winds is 10.47 minutes. This difference is statistically significant, according to the permutation test results.

b. Permutation Tests Results:

Based on the permutation test, we find that the observed difference in departure delays between "Calm" and "Strong" wind speed categories is statistically significant (p-value < 0.0036 in this case). we have sufficient evidence to reject the null hypothesis.

The p-value obtained from the permutation test is 0.0036, which indicates that the observed difference is statistically significant. This suggests that wind speed conditions significantly influence departure delays, with "Calm" conditions leading to different delay patterns compared to "Strong" conditions.

This indicates that wind speed conditions do influence departure delays, with 'Strong' wind conditions potentially leading to different delay patterns compared to 'Calm' wind conditions.

7 Precipitation Analysis

a. Exploration of Precipitation Data:

We have categorized the variable precip into two main categories:

- **No Precipitation:** This category includes instances where the value of precip is equal to 0, indicating no recorded precipitation.
- **Precipitation:** Everything else, where the value of precip is not equal to 0, is categorized as 'Precipitation.' This category encompasses all instances of recorded precipitation.

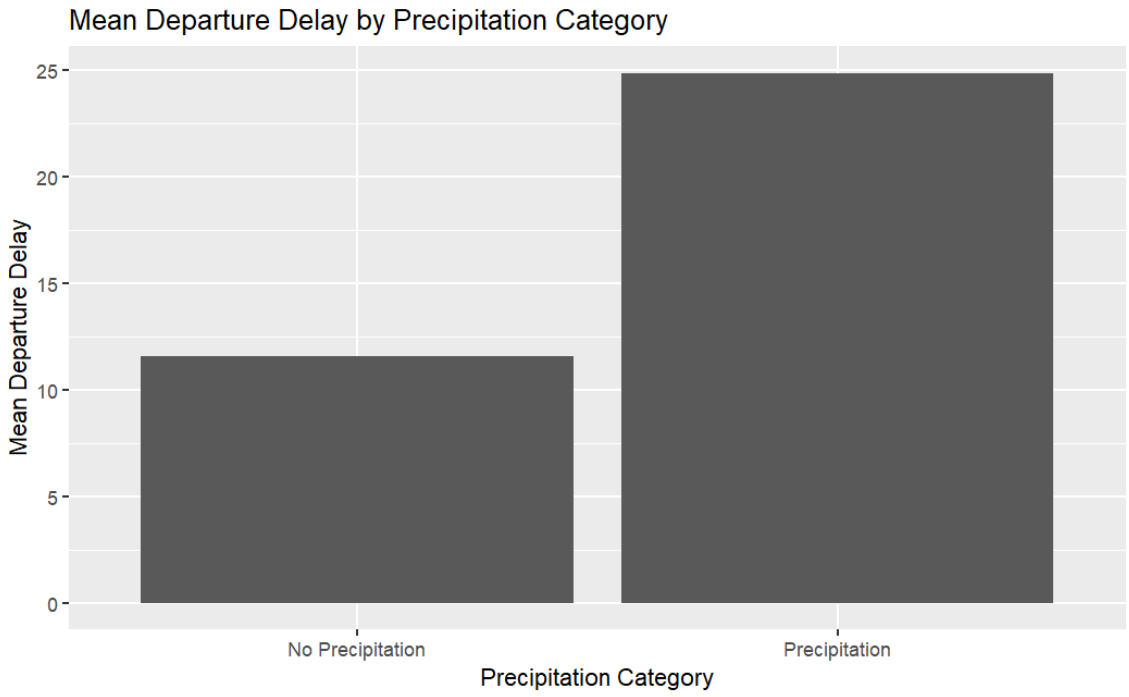

Figure: 5

The bar graph shows the mean departure delay for United Airlines (UA) flights by precipitation category. The x-axis shows the precipitation category (no precipitation, precipitation), and the y-axis shows the mean departure delay in minutes.

The graph shows that the mean departure delay is higher for UA flights with precipitation than for flights with no precipitation. The mean departure delay with

precipitation is 25 minutes, while the mean departure delay with no precipitation is 10.8 minutes.

There are a few possible explanations for the higher mean departure delay for flights with precipitation. precipitation can lead to weather disruptions, such as thunderstorms and low visibility. These weather disruptions can lead to delays.

b. Permutation Tests Results:

Based on the permutation test, we find that the observed difference in departure delays between "No Precipitation" and "Precipitation" precipitation categories is statistically significant (p-value < 0.05, or 0.0002 in this case).

This indicates that precipitation conditions do have a statistically significant impact on departure delays. Specifically, flights during precipitation events may experience different delay patterns compared to those with no recorded precipitation.

The results of this analysis provide insights into the relationship between precipitation and departure delays, which can be significant for airline operations and scheduling.

8 Visibility Analysis

a. Exploration of Visibility Data:

We have categorized the variable visib into two primary visibility categories:

- **Good Visibility:** This category encompasses instances where the visibility (represented by visib) is greater than 2, indicating good visibility conditions.
- **Poor Visibility:** Everything else, where the visibility is 2 or lower, is categorized as 'Poor Visibility.' This category includes instances of reduced visibility conditions.

This categorization allows us to analyze data based on different visibility conditions, which is valuable for assessing the impact of visibility on various factors or events. It simplifies the data into two clear categories, making it easier to study the effects of visibility.

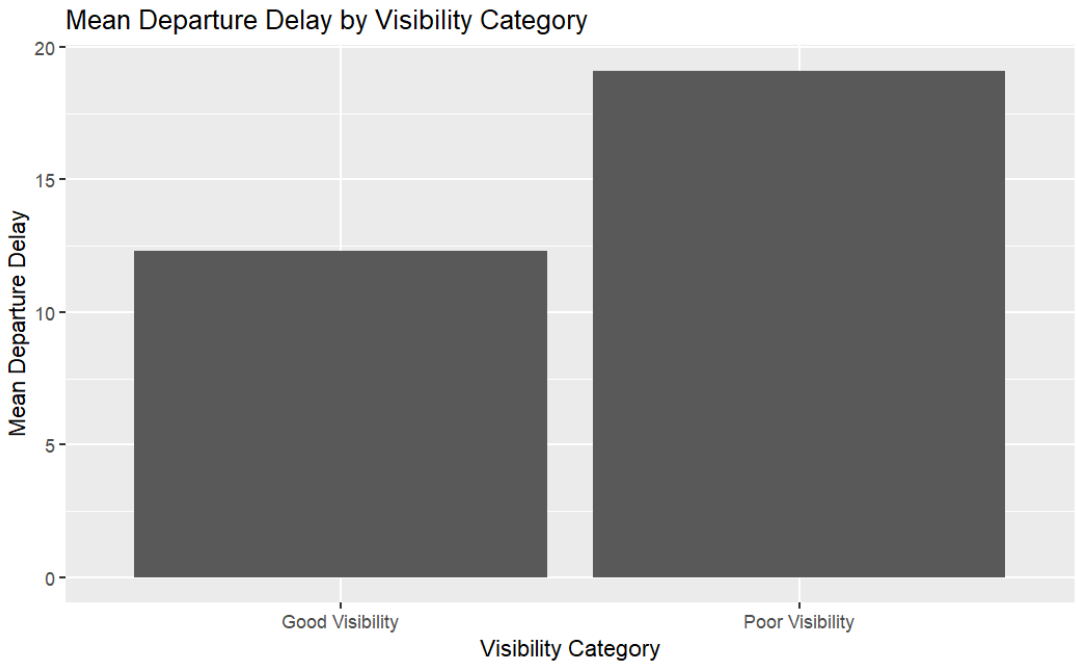

Figure: 6

The graph shows the mean departure delay for United Airlines (UA) flights by visibility category. The x-axis shows the visibility category (good, poor), and the y-axis shows the mean departure delay in minutes.

The graph shows that the mean departure delay is higher for UA flights with poor visibility than for flights with good visibility. The mean departure delay with poor

visibility is 19.2 minutes, while the mean departure delay with good visibility is 10.9 minutes.

The graph shows that the mean departure delay for UA flights is higher for flights with poor visibility than for flights with good visibility. This difference is statistically significant. This suggests that visibility may be a contributing factor to departure delays.

b. Permutation Tests Results:

The p-value is calculated as twice the proportion of permutations where the difference in departure delays was less than or equal to the observed difference.

Based on the permutation test, we find that the observed difference in departure delays between "Good Visibility" and "Poor Visibility" visibility categories is statistically significant (p-value < 0.05, or 0.0002 in this case).

This indicates that visibility conditions do have a statistically significant impact on departure delays, with 'Poor Visibility' conditions potentially leading to different delay patterns compared to 'Good Visibility' conditions.

The results of this analysis provide insights into the relationship between visibility and departure delays, which can be relevant for airline operations and scheduling.

9 Conclusion

The permutation tests have shown significant differences in departure delays based on various factors such as time of day, seasons, temperature, wind speed, precipitation, and visibility.

Implications for UA Airlines:

To improve operational efficiency and customer satisfaction, UA Airlines should consider:

1. **Weather Planning**: Enhance weather monitoring and planning to mitigate delays during adverse conditions.
2. **Scheduling:** Adjust flight schedules during peak delay-prone times and seasons.
3. **Temperature and Wind:** Develop strategies to handle temperature and wind-related challenges effectively.
4. **Visibility and Precipitation:** Invest in training, equipment, and technology for poor visibility and precipitation conditions.
5. **Data-Driven Decision-Making:** Continuously analyze data to make informed decisions and adapt to changing conditions.

Incorporating these improvements can help UA Airlines reduce delays and provide a more reliable travel experience.

10 Appendix

The following R code provides a step-by-step analysis of flight data for Project 1. In this project, we explore and analyze various factors that might affect flight departure delays. The code is organized into different sections for data preparation, exploratory data analysis, and permutation testing. Here is a detailed explanation of the code:

Step 1. Create Data Frame for the Analysis:

```
# Load necessary libraries and datasets
library(tidyverse)
library(dplyr)
library(nycflights13)
library(ggplot2)
library(resampledata3)


# Merge flight data with weather data
data("weather")
df <- merge(NycFlights, weather, by.x = c("origin", "time_hour"), by.y = c("origin", "time_hour"), all.x = FALSE, all.y = FALSE, sort = TRUE)
Ua_df <- df %>% filter(carrier == "UA")
Ua_df
```

Step 2. Exploratory Data Analysis (EDA):

```
# EDA: Time of Day Analysis
# Categorize Time of Day
ua_data <- Ua_df %>%
  mutate(time_of_day_category = case_when(
    hour(time_hour) >= 5 & hour(time_hour) < 12 ~ "Morning",
    hour(time_hour) >= 12 & hour(time_hour) < 17 ~ "Afternoon",
    hour(time_hour) >= 17 & hour(time_hour) < 21 ~ "Evening",
    TRUE ~ "Night"
  ))

# Categorize Time of Year (Season)
ua_data <- ua_data %>%
  mutate(
    month = as.numeric(month), # Convert the "month" variable to numeric
if it's not already
    season = case_when(
      between(month, 1, 3) ~ "Winter",
      between(month, 4, 6) ~ "Spring",
      between(month, 7, 9) ~ "Summer",
      between(month, 10, 12) ~ "Fall"
    )
  )

# Categorize Temperature
ua_data <- ua_data %>%
  mutate(temperature_category = case_when(
    temp > 0 & temp <= 20 ~ "Cold",
    temp > 20 ~ "Hot"
  ))
```

```
# Remove rows with null values in wind_speed column
ua_data <- ua_data %>%
  filter(!is.na(wind_speed)) %>%
  mutate(wind_speed_category = case_when(
    wind_speed <= 15 ~ "Calm",
    wind_speed > 15 ~ "Strong"
  ))

# Categorize Precipitation
ua_data <- ua_data %>%
  mutate(precipitation_category = case_when(
    precip == 0 ~ "No Precipitation",
    TRUE ~ "Precipitation"    # Everything else is categorized as "Precipitation"
  ))

# Categorize Visibility
ua_data <- ua_data %>%
  mutate(visibility_category = case_when(
    visib > 2 ~ "Good Visibility",   # Visibility greater than 2 is considered good
    TRUE ~ "Poor Visibility"  # Everything else is categorized as "Poor Visibility"
  ))
ua_data
```

Step 3. Plotting the Graph:

```
# Create a summary of time of day categories and departure delays
time_of_day_summary <- ua_data %>%
  group_by(time_of_day_category) %>%
  summarise(
    Total_Flights = n(),
    Mean_Dep_Delay = mean(dep_delay, na.rm = TRUE)
  )
time_of_day_summary


# Create a bar plot
ggplot(time_of_day_summary, aes(x = time_of_day_category, y =
Mean_Dep_Delay)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Mean Departure Delay by Time of Day",
    x = "Time of Day Category",
    y = "Mean Departure Delay (minutes)"
  ) +
  theme_minimal()
```

```
# Group by month and calculate the mean departure delay
monthly_delay_summary <- ua_data %>%
  group_by(month) %>%
  summarize(mean_delay = mean(dep_delay, na.rm = TRUE))

# Create a bar plot to visualize the mean departure delay by month
ggplot(monthly_delay_summary, aes(x = factor(month), y = mean_delay)) +
  geom_bar(stat = "identity", fill = "black") +
  xlab("Month") +
  ylab("Mean Departure Delay (minutes)") +
  ggtitle("Mean Departure Delay by Month")

# Create a line plot for Season categories vs. Mean Departure Delay
ua_data %>%
  group_by(season) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = season, y = mean_dep_delay, group = 1)) +
  geom_line() +
  labs(title = "Mean Departure Delay by Season", x = "Season", y = "Mean
Departure Delay")
```

```
# Create a summary of time of day categories and departure delays
time_of_day_summary <- ua_data %>%
  group_by(time_of_day_category) %>%
  summarise(
    Total_Flights = n(),
    Mean_Dep_Delay = mean(dep_delay, na.rm = TRUE)
  )
time_of_day_summary

# Create a bar plot
ggplot(time_of_day_summary, aes(x = time_of_day_category, y =
Mean_Dep_Delay)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Mean Departure Delay by Time of Day",
    x = "Time of Day Category",
    y = "Mean Departure Delay (minutes)"
  ) +
  theme_minimal()
# Group by month and calculate the mean departure delay
monthly_delay_summary <- ua_data %>%
  group_by(month) %>%
  summarize(mean_delay = mean(dep_delay, na.rm = TRUE))

# Create a bar plot to visualize the mean departure delay by month
ggplot(monthly_delay_summary, aes(x = factor(month), y = mean_delay)) +
  geom_bar(stat = "identity", fill = "black") +
  xlab("Month") +
  ylab("Mean Departure Delay (minutes)") +
  ggtitle("Mean Departure Delay by Month")
```

```
# Create a line plot for Season categories vs. Mean Departure Delay
ua_data %>%
  group_by(season) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = season, y = mean_dep_delay, group = 1)) +
  geom_line() +
  labs(title = "Mean Departure Delay by Season", x = "Season", y = "Mean
Departure Delay")


# Calculate the mean departure delay for each temperature category
# Create a scatter plot for Temperature categories vs. Mean Departure
Delay
ua_data %>%
  group_by(temperature_category) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = temperature_category, y = mean_dep_delay)) +
  geom_point() +
  labs(title = "Mean Departure Delay by Temperature Category", x =
"Temperature Category", y = "Mean Departure Delay")


# Calculate the mean departure delay for each wind speed category
# Create a bar plot for Wind Speed categories vs. Mean Departure Delay
ua_data %>%
  group_by(wind_speed_category) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = wind_speed_category, y = mean_dep_delay)) +
  geom_bar(stat = "identity") +
  labs(title = "Mean Departure Delay by Wind Speed Category", x = "Wind
Speed Category", y = "Mean Departure Delay")
```

```
# Create a bar plot for Precipitation categories vs. Mean Departure
Delay
ua_data %>%
  group_by(precipitation_category) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = precipitation_category, y = mean_dep_delay)) +
  geom_bar(stat = "identity") +
  labs(title = "Mean Departure Delay by Precipitation Category", x =
"Precipitation Category", y = "Mean Departure Delay")


# Create a line plot for Visibility categories vs. Mean Departure Delay
ua_data %>%
  group_by(visibility_category) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = visibility_category, y = mean_dep_delay)) +
  geom_bar(stat = "identity") +
  labs(title = "Mean Departure Delay by Visibility Category", x =
"Visibility Category", y = "Mean Departure Delay")
```

Step 4. Permutation Testing for Time of Delay Analysis

- Permutation Testing in between Morning and Evening

```
N<- 10^4-1
observed <- mean(ua_data$dep_delay[ua_data$time_of_day_category == 
'Evening'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$time_of_day_category == 'Morning'], na.rm
= TRUE)
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(time_of_day_category == 'Morning')),
replace = FALSE)
  result[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) -
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result), mapping = aes(x = result)) +
geom_histogram() +
geom_vline(xintercept = observed, color = "red")
2 * ((sum(result >= observed) + 1) / (N + 1))
```

- Permutation Testing in between Morning and Afternoon

```
N<- 10^4-1
observed <- mean(ua_data$dep_delay[ua_data$time_of_day_category == 
'Afternoon'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$time_of_day_category == 'Morning'], 
na.rm = TRUE)
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(time_of_day_category == 'Morning')), 
replace = FALSE)
  result[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) -
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result), mapping = aes(x = result)) +
geom_histogram() +
geom_vline(xintercept = observed, color = "red")
2 * ((sum(result >= observed) + 1) / (N + 1))
```

- Permutation Testing in between Morning and Night

```
N<- 10^4-1
observed <- mean(ua_data$dep_delay[ua_data$time_of_day_category == 
'Night'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$time_of_day_category == 'Morning'], 
na.rm = TRUE)
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(time_of_day_category == 'Morning')), 
replace = FALSE)
  result[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) -
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result), mapping = aes(x = result)) +
geom_histogram() +
geom_vline(xintercept = observed, color = "red")
2 * ((sum(result >= observed) + 1) / (N + 1))
```

- Permutation Testing in between Afternoon and Evening

```
N<- 10^4-1

observed <- mean(ua_data$dep_delay[ua_data$time_of_day_category == 
'Evening'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$time_of_day_category == 'Afternoon'], 
na.rm = TRUE)

result <- numeric(N)

for (i in 1:N)

{

  index <- sample(nrow(ua_data),

  size=nrow(ua_data %>% filter(time_of_day_category == 'Afternoon')), 
replace = FALSE)

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

}

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

geom_histogram() +

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

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

- Permutation Testing in between Afternoon and Night

```
N<- 10^4-1
observed <- mean(ua_data$dep_delay[ua_data$time_of_day_category == 
'Night'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$time_of_day_category == 'Afternoon'], 
na.rm = TRUE)
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(time_of_day_category == 'Afternoon')), 
replace = FALSE)
  result[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) -
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result), mapping = aes(x = result)) +
geom_histogram() +
geom_vline(xintercept = observed, color = "red")
2 * ((sum(result >= observed) + 1) / (N + 1))
```

- Permutation Testing in between Evening and Night

```
N<- 10^4-1
observed <- mean(ua_data$dep_delay[ua_data$time_of_day_category == 
'Night'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$time_of_day_category ==Evening], na.rm = 
TRUE)
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(time_of_day_category == 'Evening')), 
replace = FALSE)
  result[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) - 
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result), mapping = aes(x = result)) +
geom_histogram() +
geom_vline(xintercept = observed, color = "red")
2 * ((sum(result >= observed) + 1) / (N + 1))
```

Step 5. Permutation Testing for Time of year analysis

- Permutation Testing in between winter and spring

```
N<- 10^4-1
observed1 <- mean(ua_data$dep_delay[ua_data$season == 'Spring'], na.rm
= TRUE)-mean(ua_data$dep_delay[ua_data$season == 'Winter'], na.rm =
TRUE)
result1 <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(season == 'Winter')), replace = FALSE)
  result1[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) -
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result1), mapping = aes(x = result1)) +
geom_histogram() +
geom_vline(xintercept = observed1, color = "red")
2 * ((sum(result1 >= observed1) + 1) / (N + 1))
```

- Permutation Testing in between winter and summer

```
N<- 10^4-1
observed2 <- mean(ua_data$dep_delay[ua_data$season == 'Winter'], na.rm
= TRUE) - mean(ua_data$dep_delay[ua_data$season == 'Summer'], na.rm =
TRUE)
result2 <- numeric(N)
for (i in 1:N)
{
  index <- sample(nrow(ua_data),
  size=nrow(ua_data %>% filter(season == 'Winter')), replace = FALSE)
  result2[i] <- mean(ua_data$dep_delay[index], na.rm = TRUE) -
mean(ua_data$dep_delay[-index], na.rm = TRUE)
}
ggplot(data = tibble(result2), mapping = aes(x = result2)) +
geom_histogram() +
geom_vline(xintercept = observed2, color = "red")
2 * ((sum(result2 <= observed2) + 1) / (N + 1))
```

- Permutation Testing in between Winter and Fall

```
N<- 10^4-1
observed3 <- mean(ua_data$dep_delay[ua_data$season == 'Winter'], na.rm
= TRUE)-mean(ua_data$dep_delay[ua_data$season == 'Fall'], na.rm = TRUE)
result3 <- numeric(N)
for (i in 1:N)
{
  index1 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(season == 'Winter')), replace = FALSE)
  result3[i] <- mean(ua_data$dep_delay[index1], na.rm = TRUE) -
mean(ua_data$dep_delay[-index1], na.rm = TRUE)
}
ggplot(data = tibble(result3), mapping = aes(x = result3)) +
geom_histogram() +
geom_vline(xintercept = observed3, color = "red")
2 * ((sum(result3 <= observed3) + 1) / (N + 1))
```

- Permutation Testing in between Spring and summer

```
N<- 10^4-1
observed4 <- mean(ua_data$dep_delay[ua_data$season == 'Spring'], na.rm
= TRUE)-mean(ua_data$dep_delay[ua_data$season == 'Summer'], na.rm =
TRUE)
result4 <- numeric(N)
for (i in 1:N)
{
  index2 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(season == 'Spring')), replace = FALSE)
  result4[i] <- mean(ua_data$dep_delay[index2], na.rm = TRUE) -
mean(ua_data$dep_delay[-index2], na.rm = TRUE)
}
ggplot(data = tibble(result4), mapping = aes(x = result3)) +
geom_histogram() +
geom_vline(xintercept = observed4, color = "red")
2 * ((sum(result4 >= observed4) + 1) / (N + 1))
```

- Permutation Testing in between Spring and Fall

```
N<- 10^4-1
observed4 <- mean(ua_data$dep_delay[ua_data$season == 'Spring'], na.rm
= TRUE)-mean(ua_data$dep_delay[ua_data$season == 'Fall'], na.rm = TRUE)
result4 <- numeric(N)
for (i in 1:N)
{
  index2 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(season == 'Spring')), replace = FALSE)
  result4[i] <- mean(ua_data$dep_delay[index2], na.rm = TRUE) -
mean(ua_data$dep_delay[-index2], na.rm = TRUE)
}
ggplot(data = tibble(result4), mapping = aes(x = result3)) +
geom_histogram() +
geom_vline(xintercept = observed4, color = "red")
2 * ((sum(result4 >= observed4) + 1) / (N + 1))
```

- Permutation Testing in between summer and Fall

```
N<- 10^4-1
observed4 <- mean(ua_data$dep_delay[ua_data$season == 'Summer'], na.rm
= TRUE)-mean(ua_data$dep_delay[ua_data$season == 'Fall'], na.rm = TRUE)
result4 <- numeric(N)
for (i in 1:N)
{
  index2 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(season == 'Summer')), replace = FALSE)
  result4[i] <- mean(ua_data$dep_delay[index2], na.rm = TRUE) -
mean(ua_data$dep_delay[-index2], na.rm = TRUE)
}
ggplot(data = tibble(result4), mapping = aes(x = result3)) +
geom_histogram() +
geom_vline(xintercept = observed4, color = "red")
2 * ((sum(result4 >= observed4) + 1) / (N + 1))
```

Step 6. Permutation testing on temperature Hot and warm:

```
ua_data <- ua_data %>%
  filter(!is.na(temperature_category))
N<- 10^4-1
observed7 <- mean(ua_data$dep_delay[ua_data$temperature_category == 
'Hot'], na.rm = TRUE) -
mean(ua_data$dep_delay[ua_data$temperature_category == 'Cold'], na.rm = 
TRUE)
observed7
result7 <- numeric(N)
for (i in 1:N)
{
  index5 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(temperature_category == 'Cold')), 
replace = FALSE)
  result7[i] <- mean(ua_data$dep_delay[index5], na.rm = TRUE) -
mean(ua_data$dep_delay[-index5], na.rm = TRUE)
}
ggplot(data = tibble(result7), mapping = aes(x = result7)) +
geom_histogram() +
geom_vline(xintercept = observed7, color = "red")
2*((sum(result7 >= observed7) + 1) / (N + 1))
```

Step 7. Permutation test for wind_speed calm and strong

```
N<- 10^4-1
observed8 <- mean(ua_data$dep_delay[ua_data$wind_speed_category == 
'Calm'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$wind_speed_category == 'Strong'], na.rm
= TRUE)
result8 <- numeric(N)
for (i in 1:N)
{
  index6 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(wind_speed_category == 'Calm')),
replace = FALSE)
  result8[i] <- mean(ua_data$dep_delay[index6], na.rm = TRUE) -
mean(ua_data$dep_delay[-index6], na.rm = TRUE)
}
ggplot(data = tibble(result8), mapping = aes(x = result8)) +
geom_histogram() +
geom_vline(xintercept = observed8, color = "red")
2*((sum(result8 <= observed8) + 1) / (N + 1))
```

Step 8. Permutation test for precipitation:

```
N<- 10^4-1
observed9 <- mean(ua_data$dep_delay[ua_data$precipitation_category == 
'No Precipitation'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$precipitation_category == 
'Precipitation'], na.rm = TRUE)
result9 <- numeric(N)
for (i in 1:N)
{
  index7 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(precipitation_category == 
'Precipitation')), replace = FALSE)
  result7[i] <- mean(ua_data$dep_delay[index7], na.rm = TRUE) -
mean(ua_data$dep_delay[-index7], na.rm = TRUE)
}
ggplot(data = tibble(result9), mapping = aes(x = result9)) +
geom_histogram() +
geom_vline(xintercept = observed9, color = "red")
2*((sum(result9 <= observed9) + 1) / (N + 1))
```

Step 9. Permutation test for visibility

```
N<- 10^4-1
observed10 <- mean(ua_data$dep_delay[ua_data$visibility_category == 
'Good Visibility'], na.rm = TRUE)-
mean(ua_data$dep_delay[ua_data$visibility_category == 'Poor 
Visibility'], na.rm = TRUE)
result10 <- numeric(N)
for (i in 1:N)
{
  index8 <- sample(nrow(ua_data),
  size =nrow(ua_data %>% filter(visibility_category == 'Good 
Visibility')), replace = FALSE)
  result10[i] <- mean(ua_data$dep_delay[index8], na.rm = TRUE) -
mean(ua_data$dep_delay[-index8], na.rm = TRUE)
}
ggplot(data = tibble(result10), mapping = aes(x = result10)) +
geom_histogram() +
geom_vline(xintercept = observed10, color = "red")
2*((sum(result10 <= observed10) + 1) / (N + 1))
```