United Airlines

Departure Delays Analysis Report

**Prepared by: Aarushi Kotwani**

United Airlines (UA)

Table of Contents

1. Introduction
2. Data Set Overview
3. Time of Day Analysis
   1. Data and Methodology
   2. Exploration of Data Analysis
   3. Permutation Test
4. Time of Year Analysis

4.1 Data and Methodology

4.2 Exploration of Data Analysis

* 1. Permutation Test

5. Temperature Analysis

5.1 Data and Methodology

5.2 Exploration of Data Analysis

* 1. Permutation Test

6. Wind Speed Analysis

6.1 Data and Methodology

* 1. Exploration of Data Analysis
  2. Permutation Test

1. Precipitation Analysis

7.1 Data and Methodology

7.2 Exploration of Data Analysis

* 1. Permutation Test

8. Visibility Analysis

* 1. Data and Methodology
  2. Exploration of Data Analysis

8.3 Permutation Test

1. Appendix:

9.1 Data & Permutation Test Code

INTRODUCTION

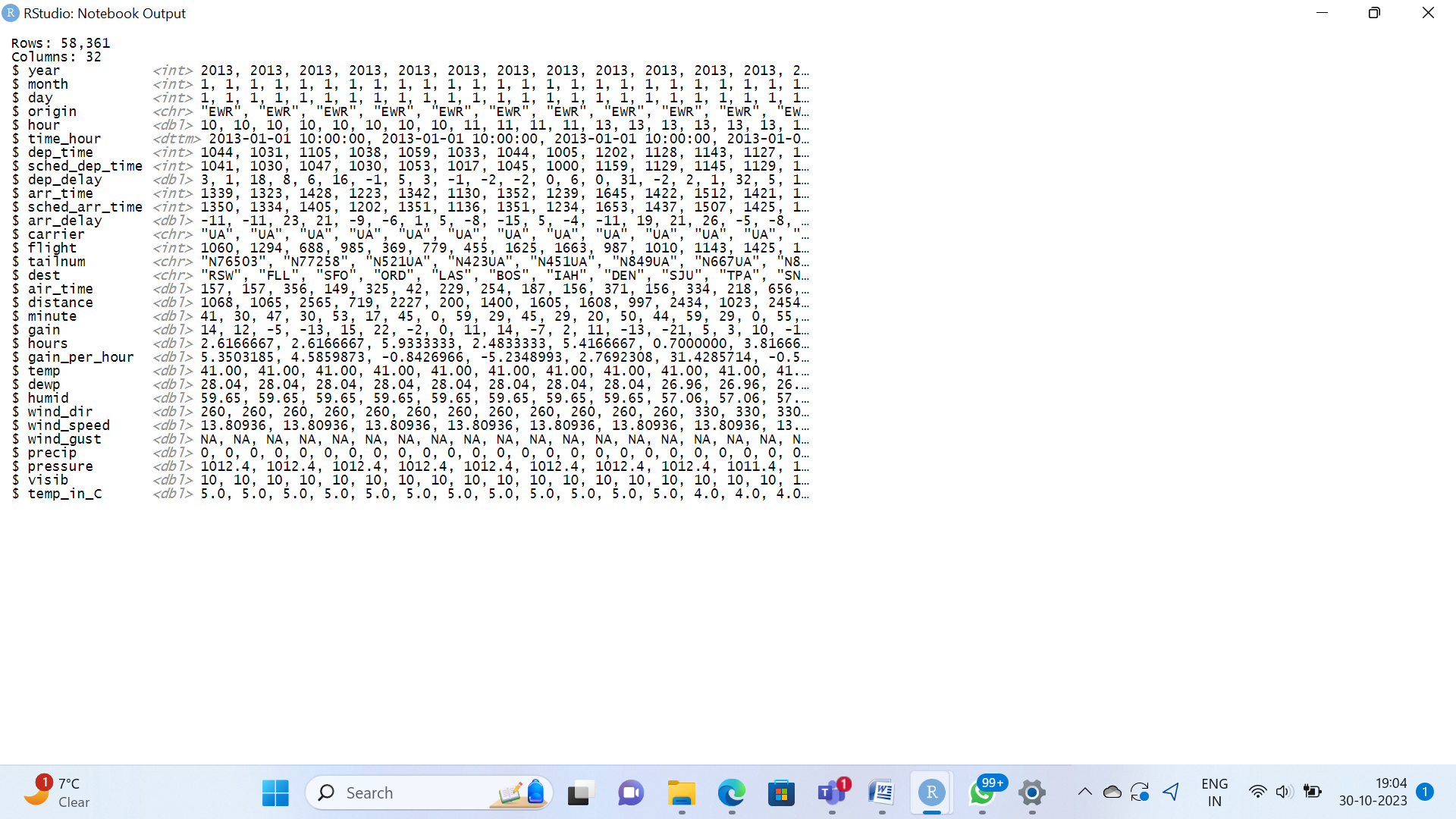
In the fast-paced realm of aviation, ensuring timely departures is of utmost importance. Departure delays can be influenced by several factors, including time of day, time of year, temperature, wind speed, precipitation, and visibility.

This project endeavors to explore the intricate dynamics of departure delays, using data from the nycflights13 package, and apply exploratory data analysis (EDA) techniques and permutation tests to derive meaningful insights. By conducting a comprehensive analysis of the factors, we aim to uncover the relationship between departure delays and these variables. Throughout the report, we will discuss our findings in a non-technical manner

**DATA SET OVERVIEW**

The dataset used in this project is sourced from the nycflights13 package and comprises information related to flights.

The dataset is rich in information and will serve as the foundation for our exploration of departure delays and their relationships with different factors. In the subsequent sections of this report, we will delve into these variables, applying exploratory data analysis techniques and permutation tests to extract valuable insights.



Time of Day Analysis

**Data and Methodology**

**Data Sources**:

* For our analysis of departure delays with respect to time of day, we utilized the nycflights13 dataset, which contains detailed flight information for carrier United Airlines.
* The primary variables used for this analysis are hour and dep\_delay.

**Methodology:**

To investigate the relationship between departure delays and time of day, we have conducted the following steps:

* **Data** **Extraction**: We extracted relevant flight information from the nycflights13 dataset.
* **Categorizing** **time**:
* The time data is provided in hour.
* The following hour range are proposed as starting point:
* Night: 12 am to 6 am
* Morning: 6 am to 12 pm
* Afternoon: 12 pm to 6pm
* Evening: 6pm to 12 am
* Creating a Categorical variable:

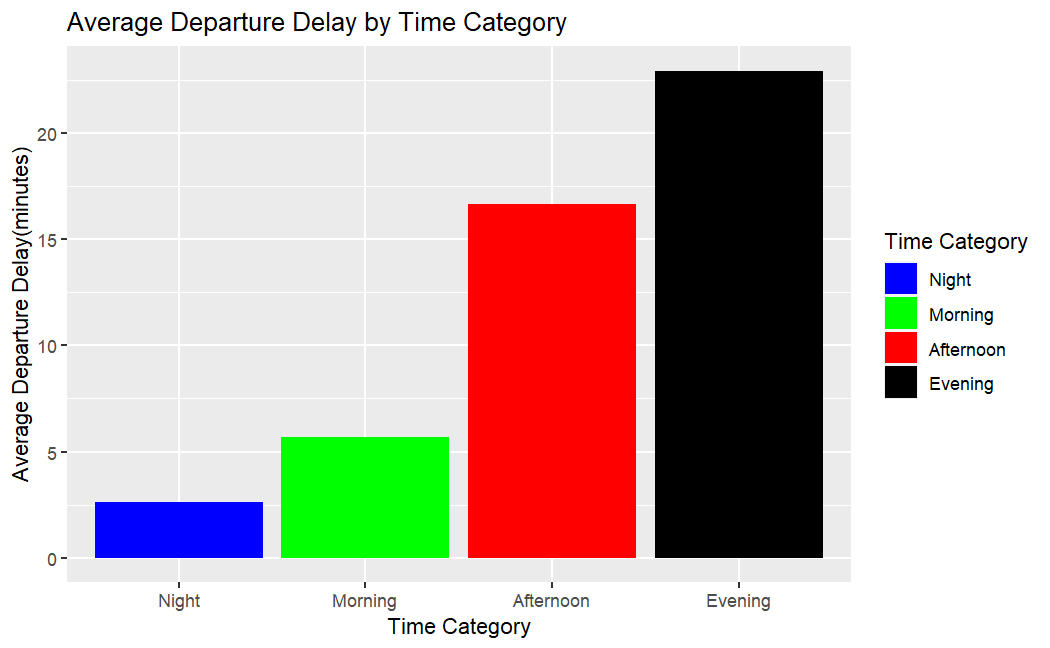
A new categorical variable is created to store the time categories. This variable will store one of the four category labels based on the hour value associated.

* EDA involved summary statistics, bar plots to assess the distributions and relationships between the variables.
* Statistical Analysis: We performed statistical tests, including a permutation test, to determine the impact of time on departure delays.

Our analysis was conducted using the R programming language.

**EXPLORATORY DATA ANALYSIS**

**Bar Plot**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Category | Average Departure Delay |  |  |  |
| Night | 2.609411 |  |  |  |
| Morning | 5.701128 |  |  |  |
| Afternoon | 16.633412 |  |  |  |
| Evening | 22.915291 |  |  |  |

**Notable Findings:**

In our analysis of departure delays with respect to time of day, several notable findings emerged:

* **Night (2.609411 minutes):**
* Flights departing during the "Night" category experience the shortest average departure delay, with an average delay of approximately 2.61 minutes.
* This suggests that, on average, flights during the nighttime period have minimal delays.
* **Morning (5.701128 minutes):**
* Flights departing in the "Morning" experience a slightly longer average delay, around 5.70 minutes.
* Morning flights tend to have somewhat longer delays compared to nighttime flights but are still relatively short.
* **Afternoon (16.633412 minutes):**
* The "Afternoon" time category has a significantly higher average departure delay of approximately 16.63 minutes.
* Flights departing in the afternoon experience longer delays compared to both night and morning flights.
* **Evening (22.915291 minutes):**
* The "Evening" time category has the highest average departure delay, with flights delayed by approximately 22.92 minutes on average.
* This indicates that flights departing in the evening have the longest average delays compared to the other time categories.

**Permutation Test**

To assess the effect of time of day on departure delays, we conducted permutation tests for each time category (Night, Morning, Afternoon and Evening).

Hypotheses:

Null Hypothesis (H0): There is no significant difference in departure delays between time categories.

Alternative Hypothesis (H1): Time category lead to greater departure delays.

Significance Level:

A significance level of α = 0.05 was used for hypothesis testing.

P-value:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Night | Morning | Afternoon | Evening |
| Night | ------ | 0.02 | 0.02 | 0.02 |
| Morning | 0.02 | ------- | 0.02 | 0.02 |
| Afternoon | 0.02 | 0.02 | ------- | 0.02 |
| Evening | 0.02 | 0.02 | 0.02 | ------- |

**“Analysis of Departure Delays and Time of Day"**

In our analysis of departure delays and time of day, we observed a p-value of 0.002. This p-value represents the result of a hypothesis test that examined the relationship between time of day conditions and departure delays for flights.

**Interpretation:**

A p-value of 0.002 is below the significance level (alpha) of 0.05, suggesting strong evidence against the null hypothesis. In practical terms, this means that the observed relationship between time of day and departure delays is unlikely to have occurred by random chance alone.

**Conclusion:**

Based on this analysis, we can conclude that there is a statistically significant relationship between time of day conditions and departure delays. Flights operating under different time of day conditions experience departure delays that are significantly different from one another.

Time of Year Analysis

**Data and Methodology**

**Data Sources**:

* For our analysis of departure delays with respect to time of year, we utilized the nycflights13 dataset, which contains detailed flight information for carrier United Airlines.
* The primary variables used for this analysis are month and dep\_delay.

**Methodology:**

To investigate the relationship between departure delays and time of year, we conducted the following steps:

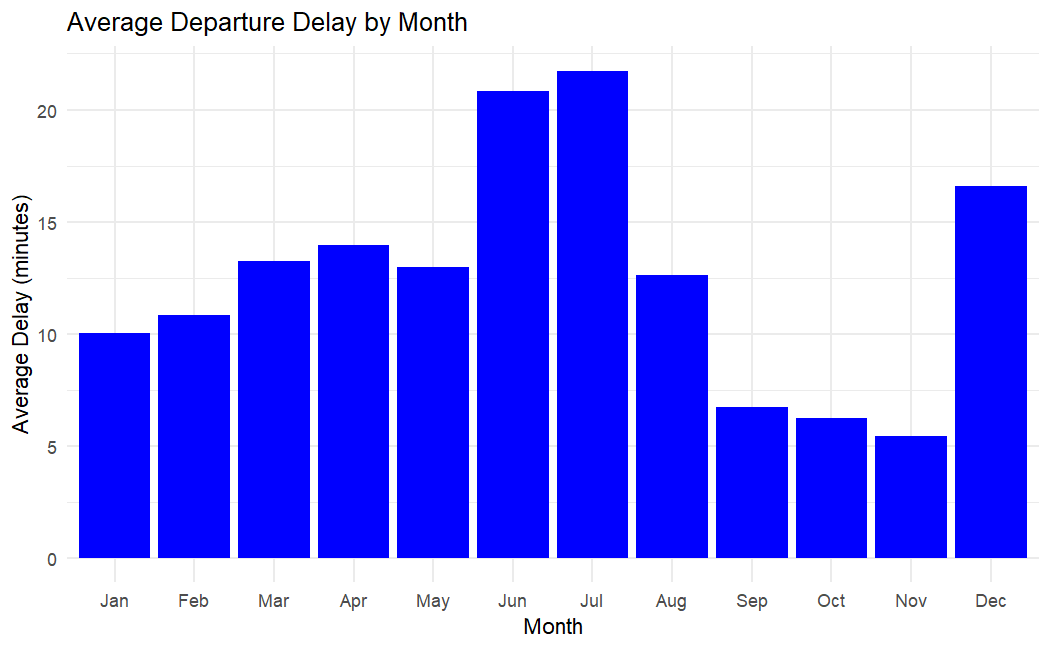
* Data Extraction: We extracted relevant flight information from the nycflights13 dataset.
* Categorizing time: The time data is provided in hour.
* The following month range are proposed as starting point:
* Quarter 1: January, February, March
* Quarter 2: April, May, June
* Quarter 3: July, August, September
* Quarter 4: October, November, December
* Creating a Categorical variable:

A new categorical variable is created to store the month categories. This variable will store one of the four category labels based on the month value associated.

* EDA involved summary statistics, bar plots to assess the distributions and relationships between the variables.
* Statistical Analysis: We performed statistical tests, including a permutation test, to determine the significance of departure delays across different months.

Our analysis was conducted using the R programming language.

**EXPLORATORY DATA ANALYSIS**

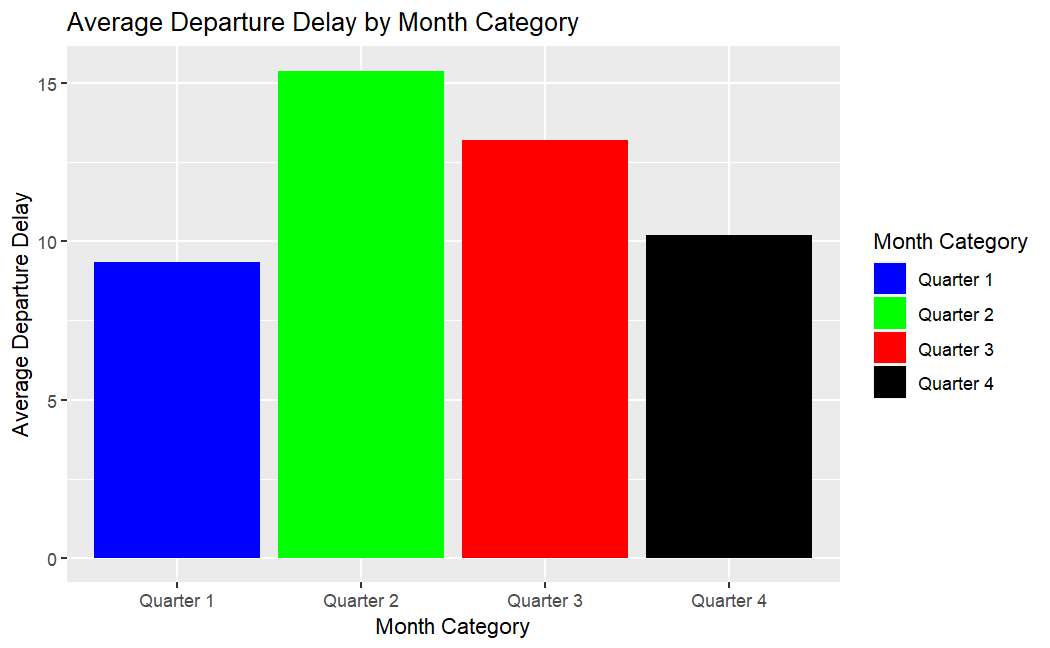


**Notable Findings:**

In our analysis of departure delays with respect to time of year, several notable findings emerged:

**Month with Highest Average Delays:** June and July exhibited the highest average delays.

**Month with Lowest Average Delays:** October and November exhibited the lowest average departure delays.



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Month Category   |  | | --- | |  | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  | | Average Departure Delay |
| Quarter 1 | 9.342813 |
| Quarter 2 | 15.383634 |
| Quarter 3 | 13.196951 |
| Quarter 4 | 10.212163 |

**Notable Findings:**

**Quarter 1:** The average departure delay during the first quarter of the year is approximately 9.34 minutes.

**Quarter 2:** The average departure delay during the second quarter of the year is approximately 15.38 minutes.

**Quarter 3:** The average departure delay during the third quarter of the year is approximately 13.20 minutes.

**Quarter 4:** The average departure delay during the fourth quarter of the year is approximately 10.21 minutes.

**Permutation Test**

To assess the effect of time of the year on departure delays, we conducted permutation tests for each time of the year category (Quarter 1, Quarter 2, Quarter 3, and Quarter4).

Hypotheses:

Null Hypothesis (H0): There is no significant difference in departure delays between times of the year category.

Alternative Hypothesis (H1): Time of the year category lead to greater departure delays.

Significance Level:

A significance level of α = 0.05 was used for hypothesis testing.

P-value:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4 |
| Quarter 1 | **--------** | 0.02 | 0.02 | 0.42 |
| Quarter 2 | 0.02 | **--------** | 0.02 | 0.02 |
| Quarter 3 | 0.02 | 0.02 | **-------** | 0.02 |
| Quarter 4 | 0.042 | 0.02 | 0.02 | **------** |

**"Analysis of Departure Delays and Time of Year"**

In our analysis of departure delays and time of year, we observed a p-value of 0.002. This p-value represents the result of a hypothesis test that examined the relationship between time of year conditions and departure delays for flights.

**Interpretation:**

A p-value of 0.002 is below the significance level (alpha) of 0.05, suggesting strong evidence against the null hypothesis. In practical terms, this means that the observed relationship between time of year and departure delays is unlikely to have occurred by random chance alone.

**Conclusion:**

Based on this analysis, we can conclude that there is a statistically significant relationship between time of year conditions and departure delays. Flights operating under different time of year conditions experience departure delays that are significantly different from one another.

Temperature Analysis

**Data and Methodology**

**Data Sources**:

* For our analysis of departure delays with respect to temperature, we utilized the merged data of nycflights13 and weather dataset, which contains detailed flight information for carrier United Airlines.
* The primary variables used for this analysis are temp and dep\_delay.

**Methodology:**

To investigate the relationship between departure delays and temperature, we conducted the following steps:

* Data Extraction: We extracted relevant flight information data from the nycflights13 dataset and weather dataset.
* Data Preprocessing: There were few rows with missing data, used na.omit() function to remove the rows from the dataset.
* Categorizing temperatures:
* The temperature data is provided in degrees Fahrenheit and Celsius ranges. For our analysis, temperature in degrees Fahrenheit is used.
* The following temperature ranges are proposed as starting point:
* Cold: Temperatures below 500F.
* Moderate: Temperature between 500F and 800F.
* Hot: Temperatures above 800F.
* Creating a Categorical variable:

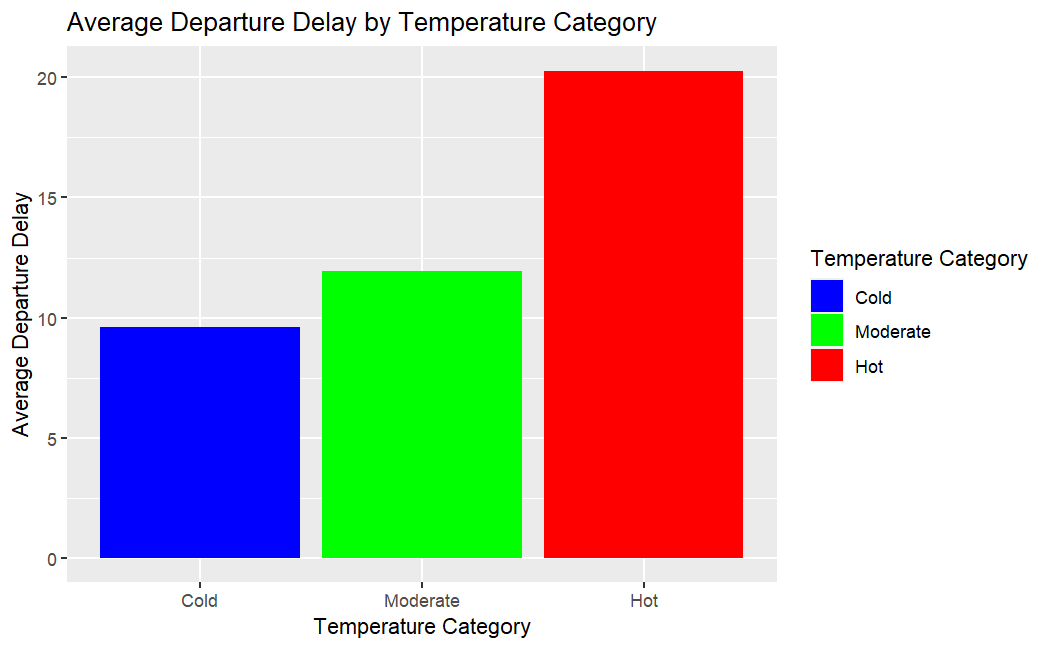
A new categorical variable is created to store the temperatures categories. This variable will store one of the three category labels based on the temperature value associated.

* EDA involved summary statistics, bar plots and box plots to assess the distributions and relationships between the variables.
* Statistical Analysis: We performed statistical tests, including a permutation test, to determine the impact of temperature on departure delays.

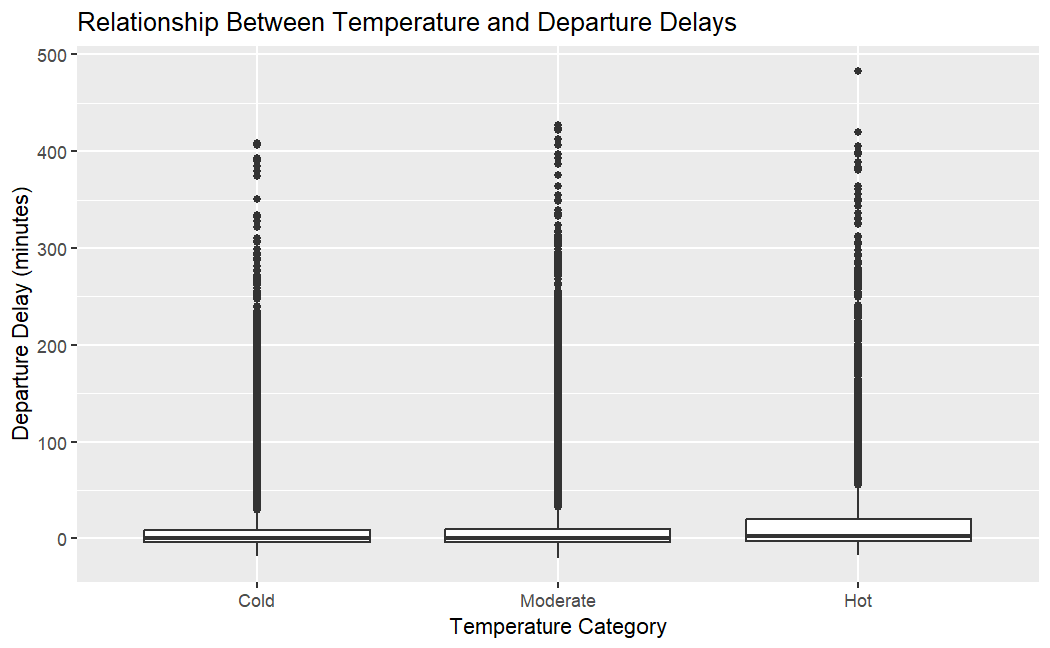
Our analysis was conducted using the R programming language.

**EXPLORATORY DATA ANALYSIS**

**Bar plot**



**Box plot**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Temperature Category | Average Departure Delay |  |  |  |
| Cold | 9.608418 |  |  |  |
| Moderate | 11.934135 |  |  |  |
| Hot | 20.274318 |  |  |  |

**Notable Findings:**

* **Cold (9.608418 minutes):**
* The average departure delay for flights in the "Cold" temperature category is approximately 9.6 minutes.
* This suggests that, on average, flights during colder weather conditions experience a delay of around 9.6 minutes.
* **Moderate (11.934135 minutes):**
* The average departure delay for flights in the "Moderate" temperature category is approximately 11.9 minutes.
* This indicates that flights during moderate or temperate weather conditions experience a slightly longer average delay of around 11.9 minutes compared to colder weather.
* **Hot (20.274318 minutes):**
* The average departure delay for flights in the "Hot" temperature category is approximately 20.3 minutes.
* This suggests that flights during hot weather conditions experience the longest average delay, with an average delay of around 20.3 minutes.

**Permutation Test**

To assess the effect of temperature on departure delays, we conducted permutation tests for each temperature of the year category (Cold, Moderate, Hot).

Hypotheses:

Null Hypothesis (H0): There is no significant difference in departure delays between temperature categories.

Alternative Hypothesis (H1): Temperature category lead to greater departure delays.

Significance Level:

A significance level of α = 0.05 was used for hypothesis testing.

P-value:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cold | Moderate | Hot |
| Cold | **------** | 0.02 | 0.02 |
| Moderate | 0.02 | **-----** | 0.02 |
| Hot | 0.02 | 0.02 | **-----** |

**"Analysis of Departure Delays and Time of Year"**

In our analysis of departure delays and temperature, we observed a p-value of 0.002. This p-value represents the result of a hypothesis test that examined the relationship between temperature conditions and departure delays for flights.

**Interpretation:**

A p-value of 0.002 is below the significance level (alpha) of 0.05, suggesting strong evidence against the null hypothesis. In practical terms, this means that the observed relationship between temperature and departure delays is unlikely to have occurred by random chance alone.

**Conclusion:**

Based on this analysis, we can conclude that there is a statistically significant relationship between temperature conditions and departure delays. Flights operating under different temperature conditions experience departure delays that are significantly different from one another.

**Wind Speed** Analysis

**Data and Methodology**

**Data Sources**:

* For our analysis of departure delays with respect to wind speed, we merged the data of nycflights13 and weather, which contains detailed flight information for carrier United Airlines.
* The primary variables used for this analysis are wind\_speed and dep\_delay.

**Methodology:**

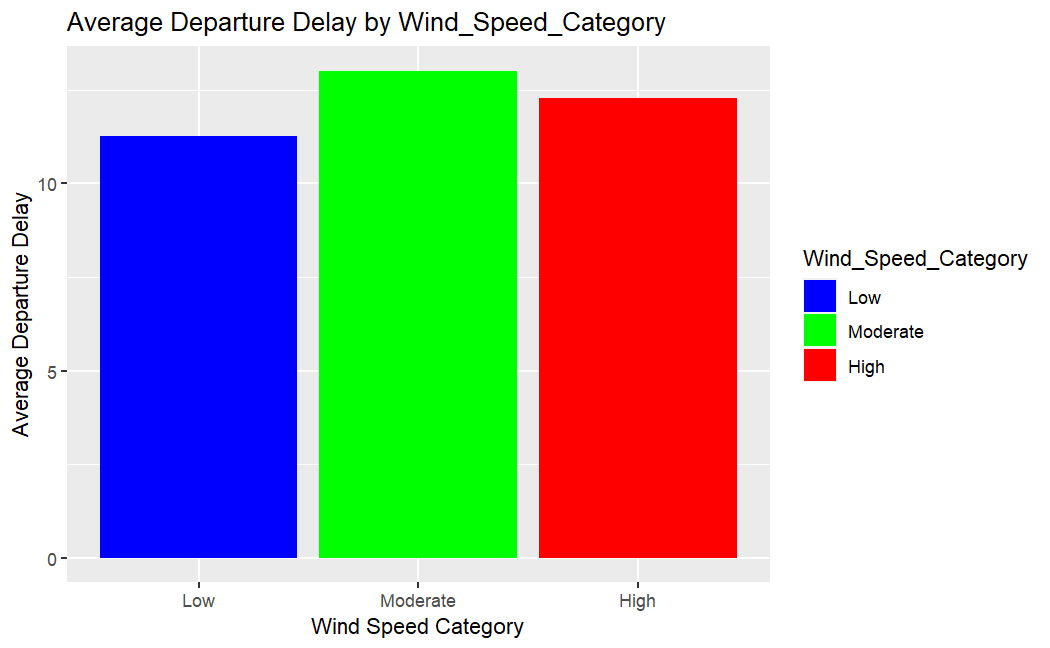
To investigate the relationship between departure delays and wind speed, we conducted the following steps:

* Data Extraction:
* We extracted relevant flight information data from the nycflights13 dataset and weather dataset.
* Data Preprocessing:
* Missing values in the "dep\_delay" and “wind\_speed” variable were removed from the dataset.
* Categorization: Wind speed was categorized into :
* Low : Below10 mph
* Moderate: Between 11-20 mph
* High: Above 21 mph
* EDA involved summary statistics, histograms, box plots and scatter plots to assess the distributions and relationships between the variables.
* Statistical Analysis: We performed statistical tests, including a permutation test, to determine the impact of temperature on departure delays.

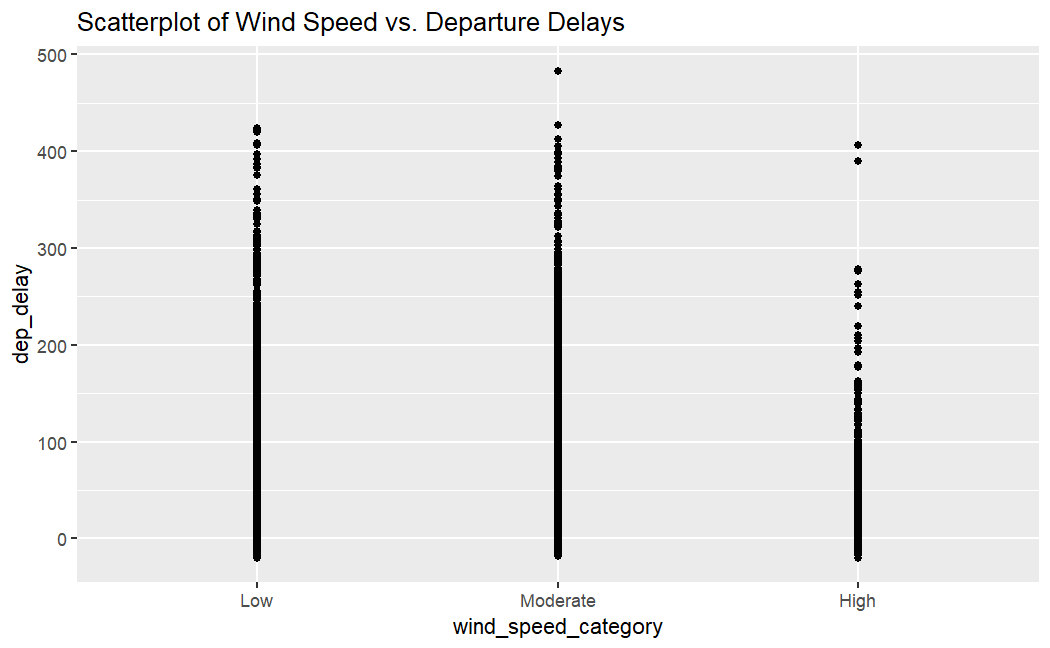
Our analysis was conducted using the R programming language.

**EXPLORATORY DATA ANALYSIS**

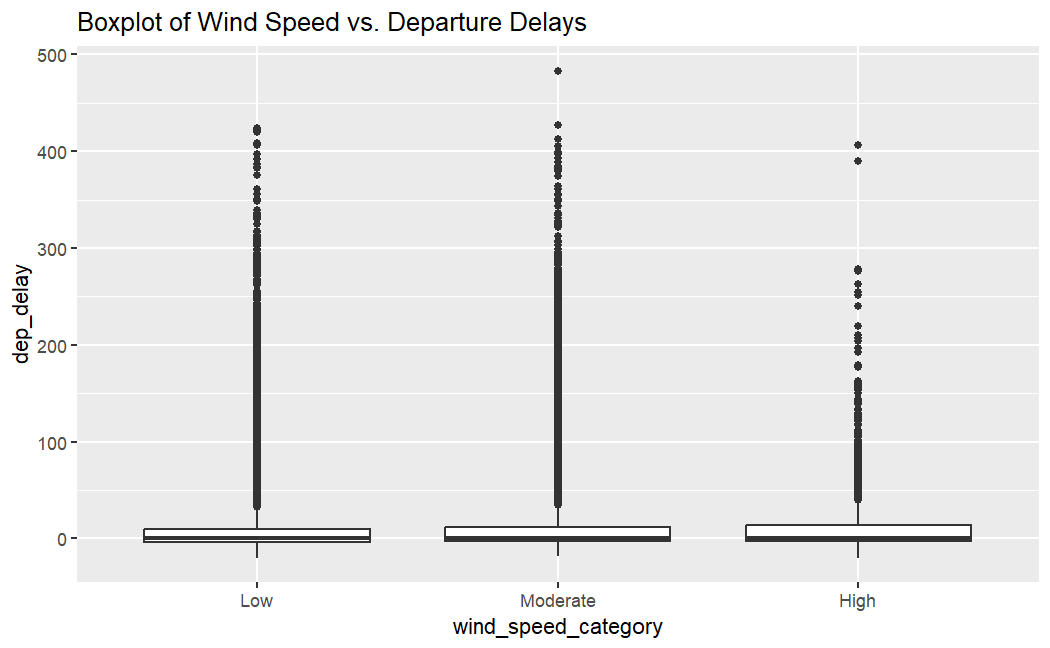
**Bar plot**



***Scatter plot***



***Box plot***



|  |  |
| --- | --- |
| Wind\_Speed\_Category | Average\_Departure\_Delay |
| Low | 11.26283 |
| Moderate | 13.00936 |
| High | 12.29033 |

**Interpretations:**

* **Low Wind Speed Category (Average Departure Delay: 11.26 minutes):**
* The average departure delay for flights in the "low" wind speed category is approximately 11.26 minutes.
* This suggests that during periods of low wind speed, the impact on departure delays is relatively minimal.
* **Moderate Wind Speed Category (Average Departure Delay: 13.01 minutes):**
* The average departure delay for flights in the "moderate" wind speed category is approximately 13.01 minutes.
* This indicates that there may be a slightly more noticeable effect on departure delays when wind speeds are in the "Moderate" category compared to "Low.
* **High Wind Speed Category (Average Departure Delay: 12.29 minutes):**
* The average departure delay for flights in the "High" wind speed category is approximately 12.29 minutes.
* This suggests that, while "High" wind speed categories do have an impact on delays, the average delay is somewhat lower than in the "Moderate" category.

**Permutation Test**

To assess the effect of wind speed on departure delays, we conducted permutation tests for each wind speed category (Low, Moderate, High) separately.

**Hypotheses:**

Null Hypothesis (H0): There is no significant difference in departure delays between wind speed categories.

Alternative Hypothesis (H1): Wind speed categories lead to greater departure delays.

**Significance Level:**

A significance level of α = 0.05 was used for hypothesis testing.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Low | Moderate | High |
| Low | ----- | 0.02 | 0.138 |
| Moderate | 0.02 | ------- | 0.358 |
| High | 0.138 | 0.358 | -------- |

**P-Value**

**"Analysis of Departure Delays and Wind speed"**

In our analysis of departure delays and wind speed, we observed a p-value of 0.002. This p-value represents the result of a hypothesis test that examined the relationship between wind conditions and departure delays for flights.

**Interpretation:**

* The p-value for the "Low" vs. "Moderate" comparison is 0.02, which is less than 0.05. This suggests that there is statistical evidence to reject the null hypothesis for this specific comparison, indicating that there is a significant relationship between departure delay and wind speed when comparing Low to Moderate conditions.
* The p-value for the "Moderate" vs. "High" comparison is 0.358, which is greater than 0.05. This suggests that there is not enough evidence to reject the null hypothesis for this comparison, indicating that there may not be a significant relationship between departure delay and wind speed when comparing Moderate to High conditions.
* The p-values for the "Low" vs. "High" and the symmetric comparison ("High" vs. "Low") are both 0.138, which is greater than 0.05, indicating that there may not be a significant relationship between departure delay and wind speed when comparing Low to High conditions.

**Conclusion:** Based on this analysis, we can conclude that there is a significant relationship between departure delay and wind speed when comparing Low to Moderate conditions, while there may not be a significant relationship in other comparisons.

**Precipitation** Analysis

**Data and Methodology**

**Data Sources**:

* For our analysis of departure delays with respect to precipitation, we merged the data of nycflights13 and weather, which contains detailed flight information for carrier United Airlines.
* The primary variables used for this analysis are precip and dep\_delay.

**Methodology:**

To investigate the relationship between departure delays and precipitation, we conducted the following steps:

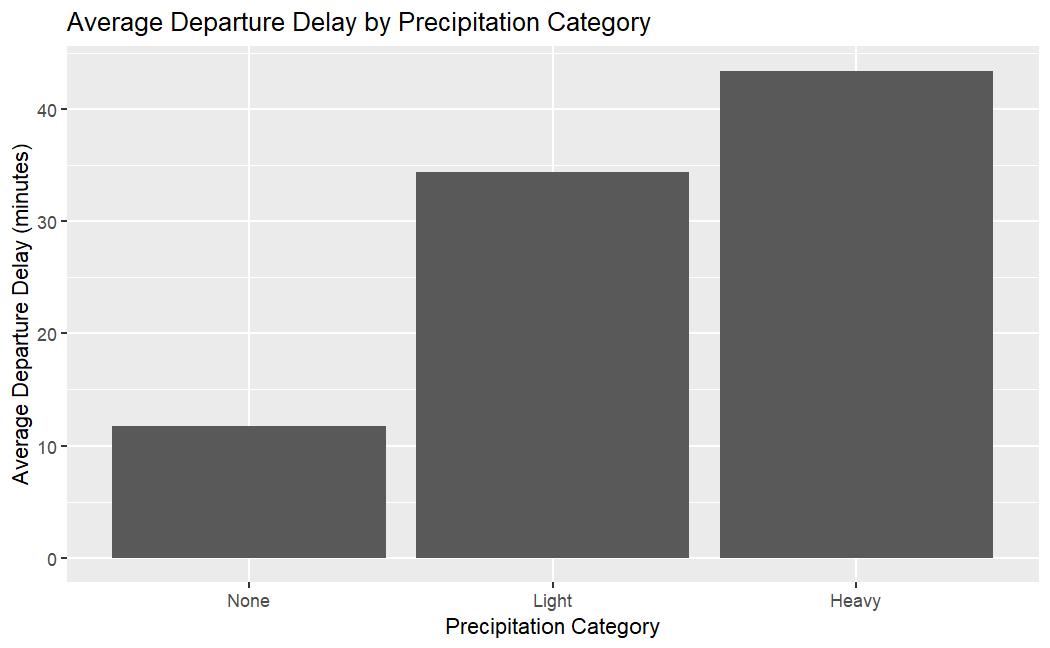
* Data Extraction:
* We extracted relevant flight information data from the nycflights13 dataset and weather dataset.
* Data Preprocessing:
* Missing values in the "dep\_delay" variable were removed from the dataset.
* Categorization: Precipitation was categorized into :
* None : Between 0 – 0.1 inches
* Light: Between 0.1-0.5 inches
* Heavy: 0.5 – Infinity inches

* EDA involved summary statistics, histograms and box plots to assess the distributions and relationships between the variables.
* Statistical Analysis: We performed statistical tests, including a permutation test, to determine the impact of temperature on departure delays.

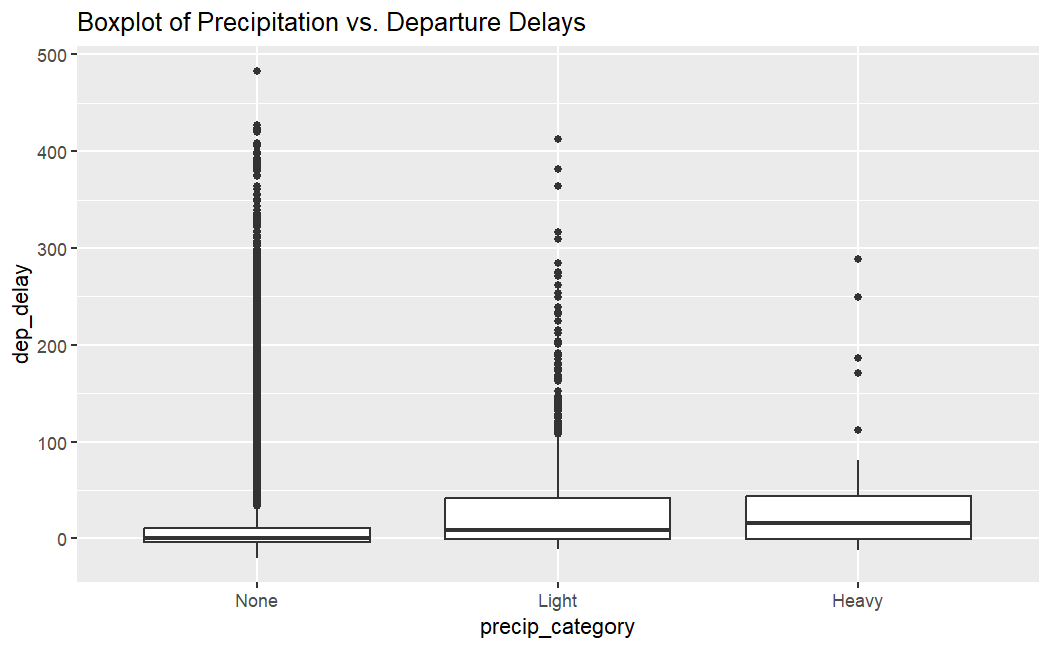
Our analysis was conducted using the R programming language.

**EXPLORATORY DATA ANALYSIS**

**Bar plot**



**Box plot**



|  |  |
| --- | --- |
| Precipitation\_Category | Average\_Departure\_Delay |
| None | 11.77067 |
| Light | 34.40026 |
| Heavy | 43.38235 |

**Interpretations:**

* **None:**
* The average departure delay when there is no precipitation is approximately 11.77 minutes.
* **Light:**
* The average departure delay during light precipitation is significantly higher, at approximately 34.40 minutes.
* This suggests that flights tend to experience longer delays during light precipitation.
* **Heavy:**
* The average departure delay during heavy precipitation is even higher, at approximately 43.38 minutes.
* This indicates that flights experience the longest delays when heavy precipitation is present.

**Permutation Test**

To assess the effect of precipitation on departure delays, we conducted permutation tests for each precipitation category (None, Light, Heavy).

Hypotheses:

Null Hypothesis (H0): There is no significant difference in departure delays between precipitation categories.

Alternative Hypothesis (H1): Precipitation leads to greater departure delays.

Significance Level:

A significance level of α = 0.05 was used for hypothesis testing.

P-value:

|  |  |  |  |
| --- | --- | --- | --- |
|  | None | Light | Heavy |
| None | **------** | 0.02 | 0.02 |
| Light | 0.02 | **-------** | 0.156 |
| Heavy | 0.02 | 0.156 | **-------** |

**"Analysis of Departure Delays and Precipitation"**

In our analysis of departure delays and precipitation, we observed a p-value of 0.002. This p-value represents the result of a hypothesis test that examined the relationship between visibility conditions and departure delays for flights.

**Interpretation:** The p-value for the "None" vs. "Light" comparison is 0.02, which is less than 0.05. This suggests that there is statistical evidence to reject the null hypothesis for this specific comparison, indicating that there is a significant relationship between departure delay and precipitation when comparing None to Light conditions.

The p-value for the "Light" vs. "Heavy" comparison is 0.156, which is greater than 0.05. This suggests that there is not enough evidence to reject the null hypothesis for this comparison, indicating that there may not be a significant relationship between departure delay and precipitation when comparing Light to Heavy conditions.

The p-values for the "None" vs. "Heavy" is 0.02, which is less than 0.05, indicating that there is a significant relationship between departure delay and precipitation when comparing None to Heavy conditions.

**Conclusion:** Based on this analysis, we can conclude that there is a significant relationship between departure delay and precipitation when comparing None to Light or None to Heavy conditions, while there may not be a significant relationship when comparing Light to Heavy conditions.

**Visibility** Analysis

**Data and Methodology**

**Data Sources**:

* For our analysis of departure delays with respect to visibility, we merged the data of nycflights13, flights and weather, which contains detailed flight information for carrier United Airlines.
* The primary variables used for this analysis are visib and dep\_delay.

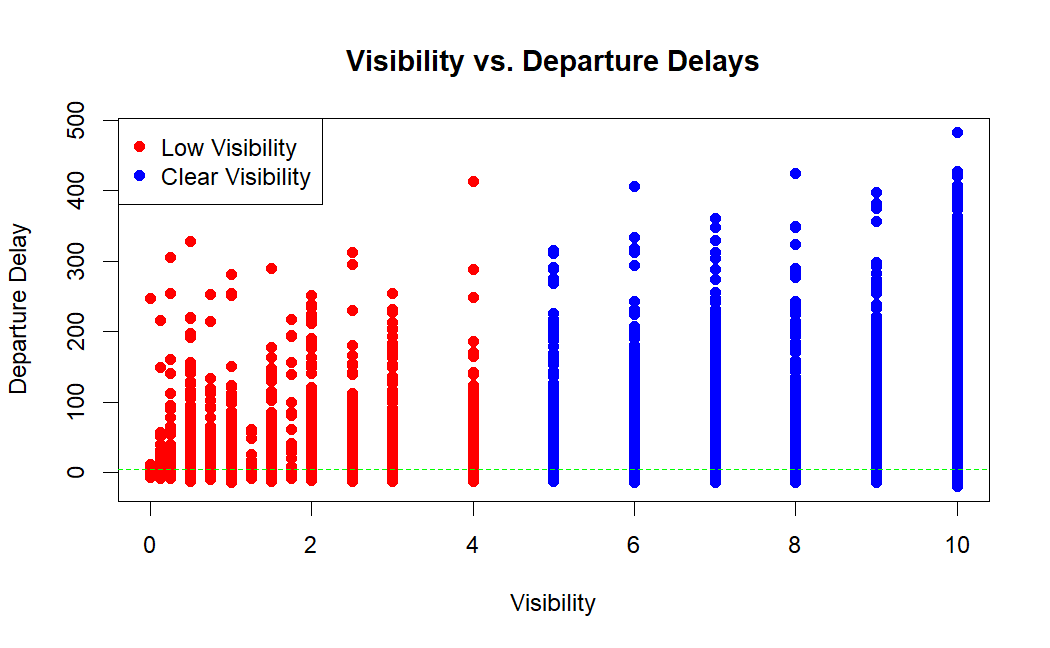
**Methodology:**

To investigate the relationship between departure delays and visibility, we conducted the following steps:

* Data Extraction:
* We extracted relevant flight information data from the nycflights13, flights and weather package.
* Data Preprocessing:
* Missing values in the "dep\_delay" variable were removed from the dataset.
* Categorization: Visibility was categorized into :
* Low: Below 5
* Clear: Above 5 ( 5 inclusive)
* EDA involved summary statistics, scatter plots to assess the distributions and relationships between the variables.
* Statistical Analysis: We performed statistical tests, including a permutation test, to determine the impact of temperature on departure delays.

Our analysis was conducted using the R programming language.

**EXPLORATORY DATA ANALYSIS**



|  |
| --- |
|  |
| Visibility\_Category | **Average\_Departure\_Delay** |  |  |  |
| Clear | 11.74591 |  |  |  |
| Low | 17.89801 |  |  |  |

**Interpretations:**

* **Average Departure Delay for "Clear" Visibility:**
* The average departure delay for flights with "clear" visibility is approximately 11.75 minutes.
* On average, flights with clear visibility conditions experience a departure delay of about 11.75 minutes.
* **Average Departure Delay for "Low" Visibility:**
* The average departure delay for flights with "low" visibility is approximately 17.90 minutes.
* This suggests that flights operating in conditions of low visibility experience a higher average departure delay of about 17.90 minutes.

**Permutation Test**

To assess the effect of visibility on departure delays, we conducted permutation tests for each visibility category (Low, Clear).

Hypotheses:

Null Hypothesis (H0): There is no significant difference in departure delays between visibilities categories.

Alternative Hypothesis (H1): Visible category lead to greater departure delays.

Significance Level:

A significance level of α = 0.05 was used for hypothesis testing.

P-value:

|  |  |  |
| --- | --- | --- |
|  | Low | Clear |
| Low | ------- | 0.02 |
| Clear | 0.02 | ------- |

**"Analysis of Departure Delays and Visibility"**

In our analysis of departure delays and visibility, we observed a p-value of 0.002. This p-value represents the result of a hypothesis test that examined the relationship between visibility conditions and departure delays for flights.

**Interpretation:**

A p-value of 0.002 is below the significance level (alpha) of 0.05, suggesting strong evidence against the null hypothesis. In practical terms, this means that the observed relationship between visibility and departure delays is unlikely to have occurred by random chance alone.

**Conclusion:**

Based on this analysis, we can conclude that there is a statistically significant relationship between visibility conditions and departure delays. Flights operating under different visibility conditions experience departure delays that are significantly different from one another.

**Appendix**

**Data Preprocessing Code**

**Load necessary libraries**:

library(nycflights13)

library(dplyr)

library(ggplot2)

**Filter flights data for United Airlines**:

UA\_flights <- flights %>%

filter(carrier==“UA")

**Merge flight and weather package based on common columns:**

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

**Filter out missing values in departure delays**

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

**Relationship between Departure Delays and Time of Day**

**EDA**

# Create a new categorical variable 'temp\_category' based on time bins

UA\_flights$time\_category <- cut(UA\_flights$hour,

breaks = c(0, 6, 12, 18,24),

labels = c("Night", "Morning", "Afternoon", "Evening"))

# Calculate the average departure delay for each category

avg\_delay\_by\_time <- aggregate(UA\_flights$dep\_delay, by = list(UA\_flights$time\_category), mean, na.rm = TRUE)

avg\_delay\_by\_time

# Rename the columns for clarity

colnames(avg\_delay\_by\_time) <- c("Time Category", "Average Departure Delay")

avg\_delay\_by\_time

# Create a bar plot

ggplot(avg\_delay\_by\_time, aes(x = `Time Category`, y = `Average Departure Delay`, fill = `Time Category`)) +

geom\_bar(stat = "identity") +

labs(title = "Average Departure Delay by Time Category", x = "Time Category", y = "Average Departure Delay(minutes)") +

scale\_fill\_manual(values = c("Night" = "blue", "Morning" = "green", "Afternoon" = "red", "Evening" = "black"))

**PERMUTATION TEST**

# Data:

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

# Categorised Data

UA\_flights$time\_category <- cut(UA\_flights$hour,

breaks = c(0, 6, 12, 18,24),

labels = c("Night", "Morning", "Afternoon", "Evening"))

1. **Night and Morning**

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Night"]) -

mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Morning"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$time\_category == "Night"]) -

mean(shuffled\_delay[UA\_flights$time\_category == "Morning"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. **Morning and Afternoon**

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Afternoon"]) -

mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Morning"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$time\_category == "Afternoon"]) -

mean(shuffled\_delay[UA\_flights$time\_category == "Morning"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. **Afternoon and Evening**

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Afternoon"]) -

mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Evening"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$time\_category == "Afternoon"]) -

mean(shuffled\_delay[UA\_flights$time\_category == "Evening"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. **Evening and Night**

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Night"]) -

mean(UA\_flights$dep\_delay[UA\_flights$time\_category == "Evening"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$time\_category == "Night"]) -

mean(shuffled\_delay[UA\_flights$time\_category == "Evening"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

**Relationship between Departure Delays and Time of Year**

**EDA**

# Data:

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

Bar Plot for Average Departure Delay By Month

# Calculate the average departure delay for each month

avg\_delay\_by\_month <- tapply(flights$dep\_delay, factor(flights$month), mean, na.rm = TRUE)

avg\_delay\_by\_month

# Define the order in which the months appear

month\_order <- c(1,2,3,4,5,6,7,8,9,10,11,12)

# Reorder the 'month' variable using the defined order

delay\_data$month <- factor(delay\_data$month, levels = month\_order)

# Create a bar chart

ggplot(delay\_data, aes(x = month, y = avg\_delay)) +

geom\_bar(stat = "identity", fill = "blue") +

labs(title = "Average Departure Delay by Month", x = "Month", y = "Average Delay (minutes)")+

scale\_x\_discrete(labels = month.abb) +

theme\_minimal()

Bar Plot for Average Departure Delay By Month Category

UA\_flights$month\_category <- cut(UA\_flights$month,

breaks = c(0, 3, 6, 9,12),

labels = c("Quarter 1", "Quarter 2", "Quarter 3", "Quarter 4"))

# Calculate the average departure delay for each category

avg\_delay\_by\_quarter <- aggregate(UA\_flights$dep\_delay, by = list(UA\_flights$month\_category), mean, na.rm = TRUE)

avg\_delay\_by\_quarter

# Rename the columns for clarity

colnames(avg\_delay\_by\_quarter) <- c("Month Category", "Average Departure Delay")

avg\_delay\_by\_quarter

# Create a bar plot

ggplot(avg\_delay\_by\_quarter, aes(x = `Month Category`, y = `Average Departure Delay`, fill = `Month Category`)) +

geom\_bar(stat = "identity") +

labs(title = "Average Departure Delay by Month Category", x = "Month Category", y = "Average Departure Delay") +

scale\_fill\_manual(values = c("Quarter 1" = "blue", "Quarter 2" = "green", "Quarter 3" = "red", "Quarter 4" = "black"))

**PERMUTATION TEST**

#Data

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

# Categorised Data

UA\_flights <- na.omit(UA\_flights[, c("dep\_delay", "month")])

UA\_flights$month\_category <- cut(UA\_flights$month,

breaks = c(0, 3, 6, 9,12),

labels = c("Quarter 1", "Quarter 2", "Quarter 3", "Quarter 4"))

1. Quarter 1 and Quarter 2

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 1"]) - mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 2"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 1"]) - mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 2"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Quarter 2 and Quarter 3

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 2"]) - mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 3"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 2"]) -

mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 3"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Quarter 3 and Quarter 4

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 3"]) - mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 4"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 3"]) - mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 4"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Quarter 4 and Quarter 1

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 4"]) - mean(UA\_flights$dep\_delay[UA\_flights$month\_category == "Quarter 1"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 4"]) - mean(shuffled\_delay[UA\_flights$month\_category == "Quarter 1"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

**Relationship between Departure Delays and Temperature**

**EDA**

# Data:

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

# Remove rows with NAs in dep\_delay and temp

UA\_flights <- na.omit(UA\_flights[, c("dep\_delay", "temp")])

# Create a new categorical variable 'temp\_category' based on temperature bins

UA\_flights$temp\_category <- cut(UA\_flights$temp,

breaks = c(-Inf, 50, 80, Inf),

labels = c("Cold", "Moderate", "Hot"))

# Calculate the average departure delay for each temperature category

avg\_delay\_by\_temp <- aggregate(UA\_flights$dep\_delay, by = list(UA\_flights$temp\_category), mean, na.rm = TRUE)

avg\_delay\_by\_temp

# Rename the columns for clarity

colnames(avg\_delay\_by\_temp) <- c("Temperature Category", "Average Departure Delay")

avg\_delay\_by\_temp

# Create a bar plot

ggplot(avg\_delay\_by\_temp, aes(x = `Temperature Category`, y = `Average Departure Delay`, fill = `Temperature Category`)) +

geom\_bar(stat = "identity") +

labs(title = "Average Departure Delay by Temperature Category", x = "Temperature Category", y = "Average Departure Delay") +

scale\_fill\_manual(values = c("Cold" = "blue", "Moderate" = "green", "Hot" = "red"))

**PERMUTATION TEST**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**DATA**

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

#Remove null values from dep\_delay and temp variable:

UA\_flights <- na.omit(UA\_flights[, c("dep\_delay", "temp")])

# Categorized the data

UA\_flights$temp\_category <- cut(UA\_flights$temp, breaks = c(-Inf, 50, 80, Inf),

labels = c("Cold","Moderate","Hot"))

1. Cold and Moderate

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$temp\_category == "Cold"]) -

mean(UA\_flights$dep\_delay[UA\_flights$temp\_category == "Moderate"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$temp\_category == "Cold"]) -

mean(shuffled\_delay[UA\_flights$temp\_category == "Moderate"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Moderate and Hot

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$temp\_category == "Hot"]) -

mean(UA\_flights$dep\_delay[UA\_flights$temp\_category == "Moderate"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$temp\_category == "Hot"]) -

mean(shuffled\_delay[UA\_flights$temp\_category == "Moderate"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Hot and Cold

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$temp\_category == "Cold"]) -

mean(UA\_flights$dep\_delay[UA\_flights$temp\_category == "Hot"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$temp\_category == "Cold"]) -

mean(shuffled\_delay[UA\_flights$temp\_category == "Hot"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

**Relationship between Departure Delays and Wind Speed**

**EDA**

# Data:

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

# Remove rows with NAs in dep\_delay and wind\_speed

UA\_flights <- na.omit(UA\_flights[, c("dep\_delay", "wind\_speed")])

UA\_flights$wind\_speed\_category <- cut(UA\_flights$wind\_speed,

breaks = c(-Inf, 10, 20, Inf),

labels = c("Low", "Moderate", "High"))

delay\_summary <- aggregate(UA\_flights$dep\_delay, by = list(UA\_flights$wind\_speed\_category),

mean,na.rm=TRUE)

colnames(delay\_summary) <- c("Wind\_Speed\_Category", "Average\_Departure\_Delay")

delay\_summary

# Scatterplot

ggplot(UA\_flights, aes(x = wind\_speed\_category, y = dep\_delay)) +

geom\_point() +

labs(title = "Scatterplot of Wind Speed vs. Departure Delays")

# Boxplot

ggplot(UA\_flights, aes(x = wind\_speed\_category, y = dep\_delay)) +

geom\_boxplot() +

labs(title = "Boxplot of Wind Speed vs. Departure Delays")

#Barplot

ggplot(delay\_summary, aes(x = Wind\_Speed\_Category, y = Average\_Departure\_Delay, fill = Wind\_Speed\_Category)) +

geom\_bar(stat = "identity") +

labs(title = "Average Departure Delay by Wind\_Speed\_Category", x = "Wind Speed Category", y = "Average Departure Delay") +

scale\_fill\_manual(values = c("Low" = "blue", "Moderate" = "green", "High" = "red")

**PERMUTATION TEST**

1. Low and Moderate

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$wind\_speed\_category == "Low"]) - mean(UA\_flights$dep\_delay[UA\_flights$wind\_speed\_category == "Moderate"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$wind\_speed\_category == "Low"]) - mean(shuffled\_delay[UA\_flights$wind\_speed\_category == "Moderate"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Moderate and High

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$wind\_speed\_category == "High"]) - mean(UA\_flights$dep\_delay[UA\_flights$wind\_speed\_category == "Moderate"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$wind\_speed\_category == "High"]) - mean(shuffled\_delay[UA\_flights$wind\_speed\_category == "Moderate"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. High and Low

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$wind\_speed\_category == "Low"]) - mean(UA\_flights$dep\_delay[UA\_flights$wind\_speed\_category == "High"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$wind\_speed\_category == "Low"]) - mean(shuffled\_delay[UA\_flights$wind\_speed\_category == "High"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

**Relationship between Departure Delays and Precipitation**

**EDA**

# Data:

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

UA\_flights$precip\_category <- cut(UA\_flights$precip,

breaks = c(0, 0.1, 0.5, Inf),

labels = c("None", "Light", "Heavy"),

include.lowest = TRUE)

average\_delays <- UA\_flights %>%

group\_by(precip\_category) %>%

summarize(avg\_delay = mean(dep\_delay))

ggplot(average\_delays, aes(x = precip\_category, y = avg\_delay)) +

geom\_bar(stat = "identity") +

xlab("Precipitation Category") +

ylab("Average Departure Delay (minutes)") +

ggtitle("Average Departure Delay by Precipitation Category")

**PERMUTATION TEST**

1. None and Light

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$precip\_category == "None"]) - mean(UA\_flights$dep\_delay[UA\_flights$precip\_category == "Light"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$precip\_category == "None"]) -

mean(shuffled\_delay[UA\_flights$precip\_category == "Light"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Light and Heavy

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$precip\_category == "Heavy"]) - mean(UA\_flights$dep\_delay[UA\_flights$precip\_category == "Light"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$precip\_category == "Heavy"]) - mean(shuffled\_delay[UA\_flights$precip\_category == "Light"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

1. Heavy and None

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$precip\_category == "None"]) -

mean(UA\_flights$dep\_delay[UA\_flights$precip\_category == "Heavy"])

observed\_diff

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$precip\_category == "None"]) -

mean(shuffled\_delay[UA\_flights$precip\_category == "Heavy"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs <= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

**Relationship between Departure Delays and Visibility**

**EDA**

# Data:

flights <- flights %>%

filter(carrier=="UA")

UA\_flights <- merge(flights, weather, by=c("year", "month", "day", "origin", "hour", "time\_hour"))

UA\_flights <- UA\_flights %>%

filter(!is.na(dep\_delay))

# Create visibility categories

threshold = 5

UA\_flights$visibility\_category <- ifelse(UA\_flights$visib < threshold, "low", "clear")

visib\_summary <- aggregate(UA\_flights$dep\_delay, by = list(UA\_flights$visibility\_category), mean,na.rm=TRUE)

colnames(visib\_summary) <- c("Visibility\_Category", "Average\_Departure\_Delay")

visib\_summary

# Create a scatter plot with color-coding

plot(UA\_flights$visib, UA\_flights$dep\_delay,

pch = 16, # Use filled circles as data points

col = ifelse(UA\_flights$visibility\_category == "low", "red", "blue"), # Color points based on visibility

xlab = "Visibility",

ylab = "Departure Delay",

main = "Visibility vs. Departure Delays")

# Add a legend

legend("topleft", legend = c("Low Visibility", "Clear Visibility"),

col = c("red", "blue"), pch = 16)

# Add a horizontal line to mark the visibility threshold

abline(h = threshold, col = "green", lty = 2)

**PERMUTATION TEST**

# Perform permutation test

observed\_diff <- mean(UA\_flights$dep\_delay[UA\_flights$visibility\_category == "low"]) -

mean(UA\_flights$dep\_delay[UA\_flights$visibility\_category == "clear"])

# Permutation

num\_permutations <- 10^3-1

permutation\_diffs <- numeric(num\_permutations)

for (i in 1:num\_permutations) {

shuffled\_delay <- sample(UA\_flights$dep\_delay, replace = FALSE)

permutation\_diffs[i] <- mean(shuffled\_delay[UA\_flights$visibility\_category == "low"]) -

mean(shuffled\_delay[UA\_flights$visibility\_category == "clear"])

}

# Calculate the p-value

p\_value <- 2 \* (sum(permutation\_diffs >= observed\_diff) + 1) / (num\_permutations + 1)

p\_value

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*