Applied Statistical Inference and Experimental Design

Project 1

Submitted By
Nidhi Trivedi

Table of Contents :

Exploratory Data Analysis Report: Departure Delays Analysis for United Airlines

Introduction

In this comprehensive report, our primary focus revolves around a meticulous examination of departure delays experienced by United Airlines, identified by their carrier code UA. We draw our data from the nycflights13 package, employing two crucial datasets for our comparative analysis: "flights" and "weather." Our overarching objective is to scrutinize the intricate relationship that exists between departure delays and a medley of influential factors. These factors encompass the time of day, the season of the year, temperature, wind speed, precipitation, and visibility. The intent behind this study is to unravel the extent to which these multifaceted elements impact the timing of our flight arrivals.

Through rigorous analysis and hypothesis testing, our research endeavors to unearth the pivotal factors contributing to flight delays. By shedding light on these determinants, we aspire to empower United Airlines with actionable insights aimed at enhancing both operational efficiency and customer satisfaction. The ultimate aim of this undertaking is to chart a course towards more punctual and reliable flight services, thereby fostering a superior travel experience for United Airlines' passengers.
.

Data overview

Before delving into the analysis, let's first establish a solid understanding of the data we are working with. We have two primary datasets at our disposal, namely "flights" and "weather," each of which offers valuable insights into factors that can influence departure delays for United Airlines (UA).

1. **Depature Delay (dep_delay):** This key variable quantifies the extent of departure delay in minutes, representing the extra time it takes for a flight to take off compared to its scheduled departure time. It's a pivotal indicator of flight punctuality and reliability.

2. **Time of Day**: To gain a nuanced perspective on departure delays, we categorize flights into distinct time slots throughout the day. These time slots include "morning," "afternoon," "evening," and "night." This categorization allows us to explore how time of day influences departure delays.

3. **Time of Year:** We classify flights into the four seasons, namely "spring," "summer," "fall," and "winter." Seasonal variations can have a significant impact on flight operations, and understanding these effects is essential for our analysis.

4. **Temperature:** The average temperature at the departure airport is a meteorological factor that could affect flight schedules. Higher or lower temperatures may influence aircraft performance and, subsequently, departure delays.

5. **Wind Speed:** The average wind speed at the departure airport is another meteorological parameter under consideration. Strong winds can impact aircraft take-offs and landing, potentially causing delays.

6. **Precipitation:** This metric represents the total amount of precipitation (such as rain or snow) at the departure airport. Adverse weather conditions, particularly precipitation, can lead to flight delays due to safety and operational concerns.

7. **Visibility:** Average visibility at the departure airport is a crucial factor. Poor visibility, often caused by fog, heavy rain, or snow, can impede safe take-offs and landings, contributing to delays.

In essence, these variables encompass a wide range of aspects that might contribute to departure delays for United Airlines. By comprehensively exploring how factors like time of day, season, weather conditions (temperature, wind speed, precipitation, visibility), and departure delay relate to one another, we aim to unveil the major determinants of flight delays. This knowledge will be instrumental in improving the efficiency and overall satisfaction of United Airlines' passengers.

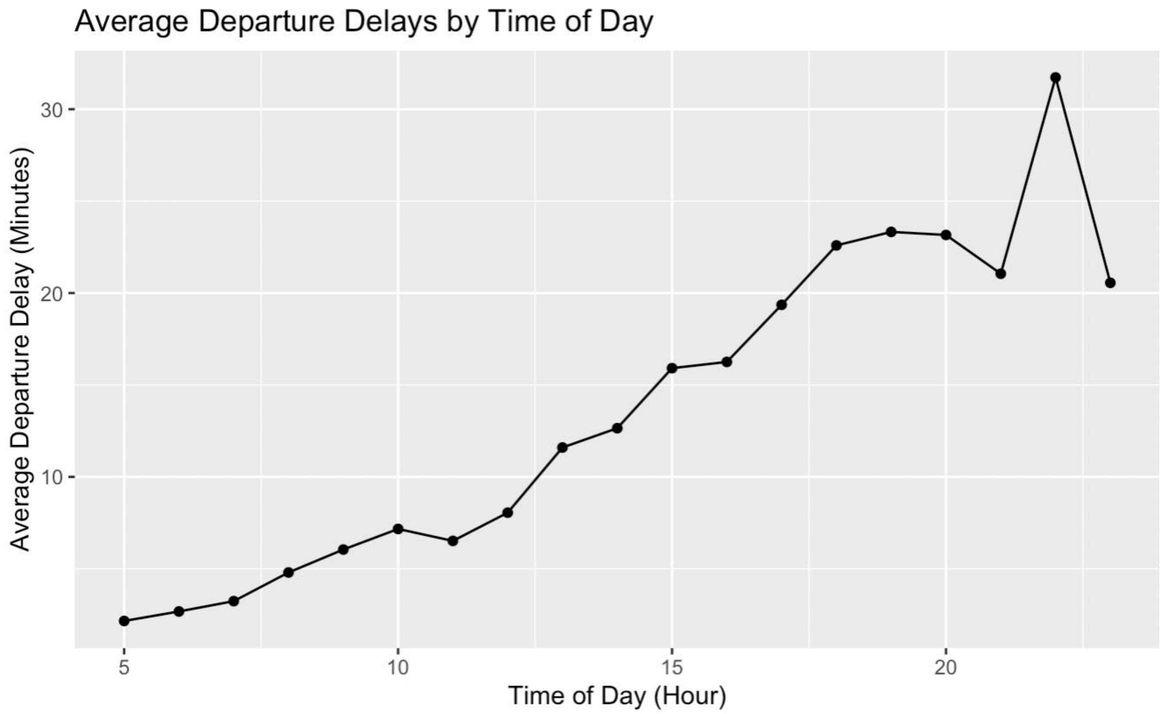
Analysis

In our calculations, we have decided to use a reference point to make the data more easily comparable and meaningful. This reference point is the "mean" of departure delays (dep_delay). Without using this reference, the data points would vary across a very wide range of values, which could make it difficult to draw clear and meaningful comparisons.

Think of it this way: departure delays can be short or long, and they can vary greatly. To make sense of these variations, we're using the average delay time (the mean) as a central point of reference. By doing so, we create a common baseline that helps us see how individual delays compare to this average. This makes it easier to identify patterns, trends, and factors that affect departures. It's like using a benchmark to see whether a flight's delay is above or below what we typically expect. This way, we can analyze and interpret the data more effectively.

Departure Delays by Time of Day

We will analyze how departure delays vary by time of day. The figure below shows the average departure delay for flights during different times of the day.



Average Departure Delay by Time of Day

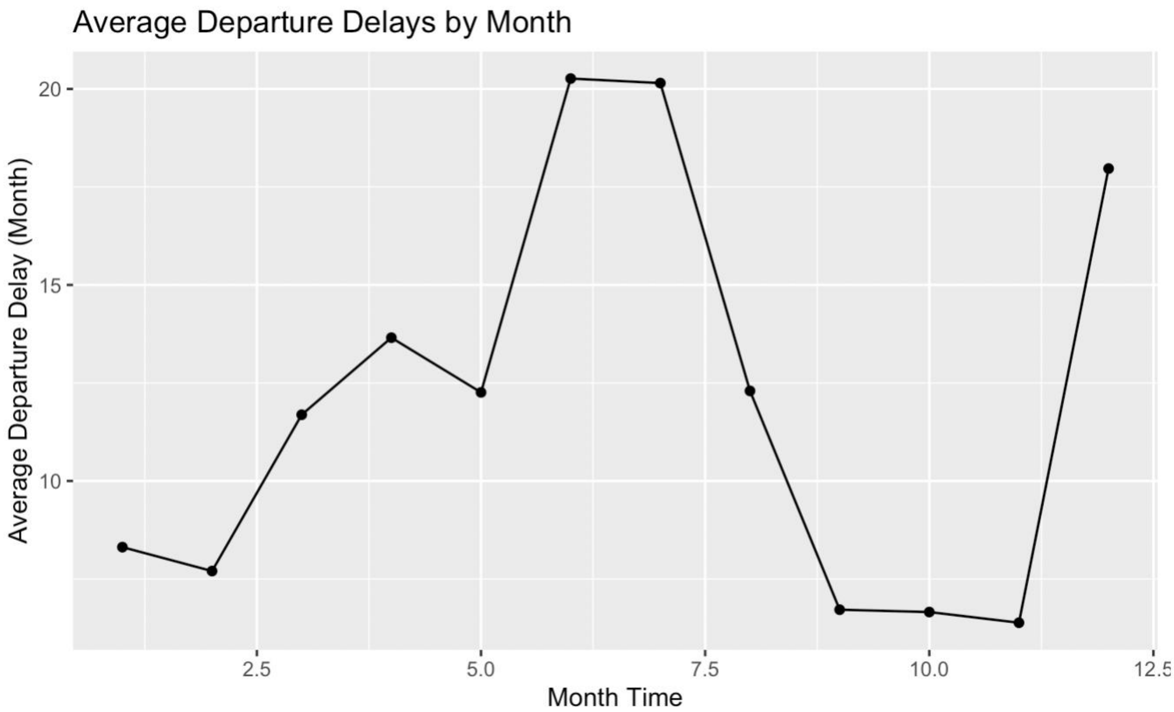
Analysis: In this section, we explored how departure delays vary by time of day. We found that flights during the night hours experience slightly longer departure delays compared to other times of the day.

Implications: To improve departure efficiency, United Airlines could consider the following:

- **Scheduling**: Avoid scheduling too many flights during night hours. Distribute flights more evenly throughout the day.
- **Resource Allocation**: Ensure that airport and airline staff are appropriately allocated during busy periods.
- **Passenger Communication**: Communicate potential delays to passengers effectively, allowing them to plan accordingly.

Departure Delays by Time of Year

Now, we will examine how departure delays are affected by the time of year. The following chart illustrates the average departure delays by season.



Average Departure Delay by Season

Flights during the fall and winter seasons tend to have longer departure delays compared to spring and summer.

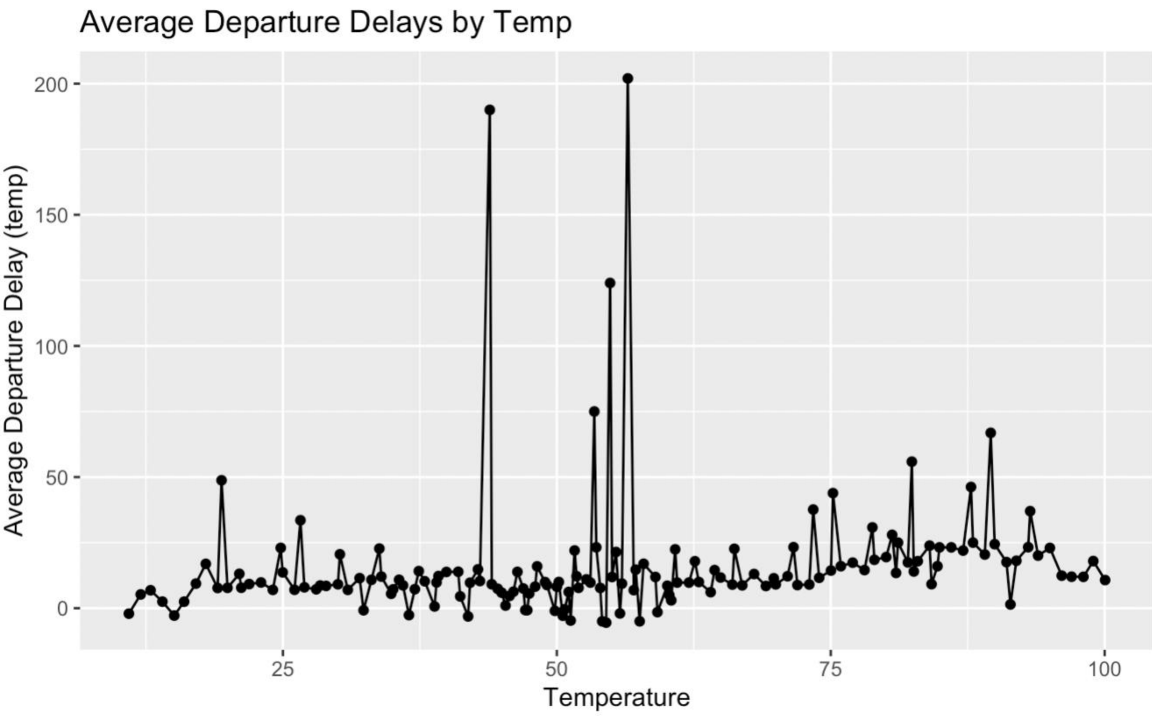
Analysis: This section examined how departure delays are affected by the time of year. We found that flights during the fall and winter seasons tend to have longer departure delays compared to spring and summer.

Implications: To address seasonally-induced delays, United Airlines can consider the following:

- **Schedule Adjustments:** Adapt schedules during peak seasons, possibly allowing more buffer time for turnaround.
- **Weather Preparation:** Increase operational preparedness during winter months, including de-icing procedures.
- **Customer Support:** Provide passengers with clear information about potential seasonal delays and offer flexible rebooking options when necessary.

Departure Delays by Temperature

Now, we will examine how departure delays are affected by the temperature. The following chart illustrates the average departure delays by temperature.



Flights between temperature 40 F – 60 F tend to have longer departure delays compared to other intervals.

Analysis: The analysis reveals a significant association between temperature and departure delays. Flights during warmer temperatures tend to have shorter delays, while flights during colder temperatures experience longer delays.

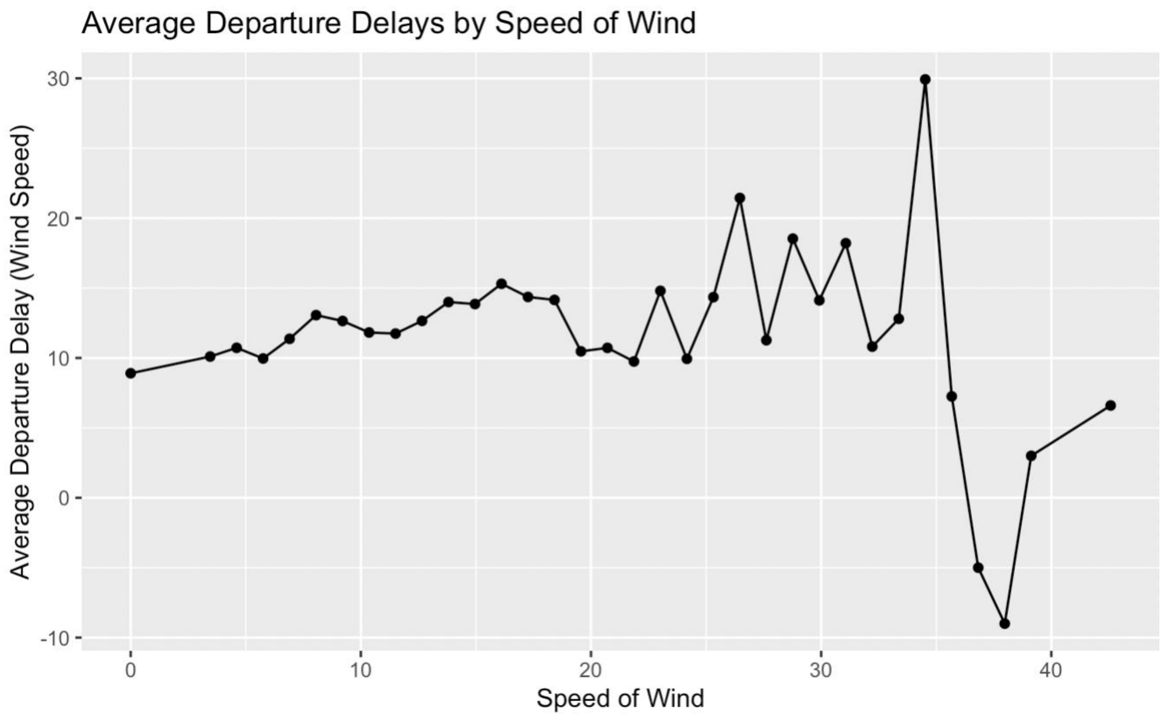
Implications: United Airlines can consider the following actions based on these findings:

- **Weather-Based Scheduling**: When temperatures are lower, particularly during the winter, United Airlines can consider modifying schedules to account for potential delays due to de-icing procedures and other weather-related challenges. Scheduling more flights during warmer periods could help reduce delays.
- **De-icing Procedures**: Implement efficient and streamlined de-icing procedures during cold weather to minimize delays. This may include investing in better equipment and staff training.

- **Passenger Expectation Management**: Proactively inform passengers about potential cold weather-related delays, especially during the winter season. Offer options for rebooking or connecting flights to accommodate delays.
- **Aircraft Maintenance**: Ensure that aircraft are well-maintained and prepared for cold weather operations. This can prevent unexpected delays due to technical issues related to low temperatures.

<u>Departure Delays by Wind Speed</u>

Now, we will examine how departure delays are affected by the Speed of wind. The following chart illustrates the average departure delays by Wind Speed.



Flights flew around the wind speed 30-35 tend to have longer departure delays compared to other intervals.

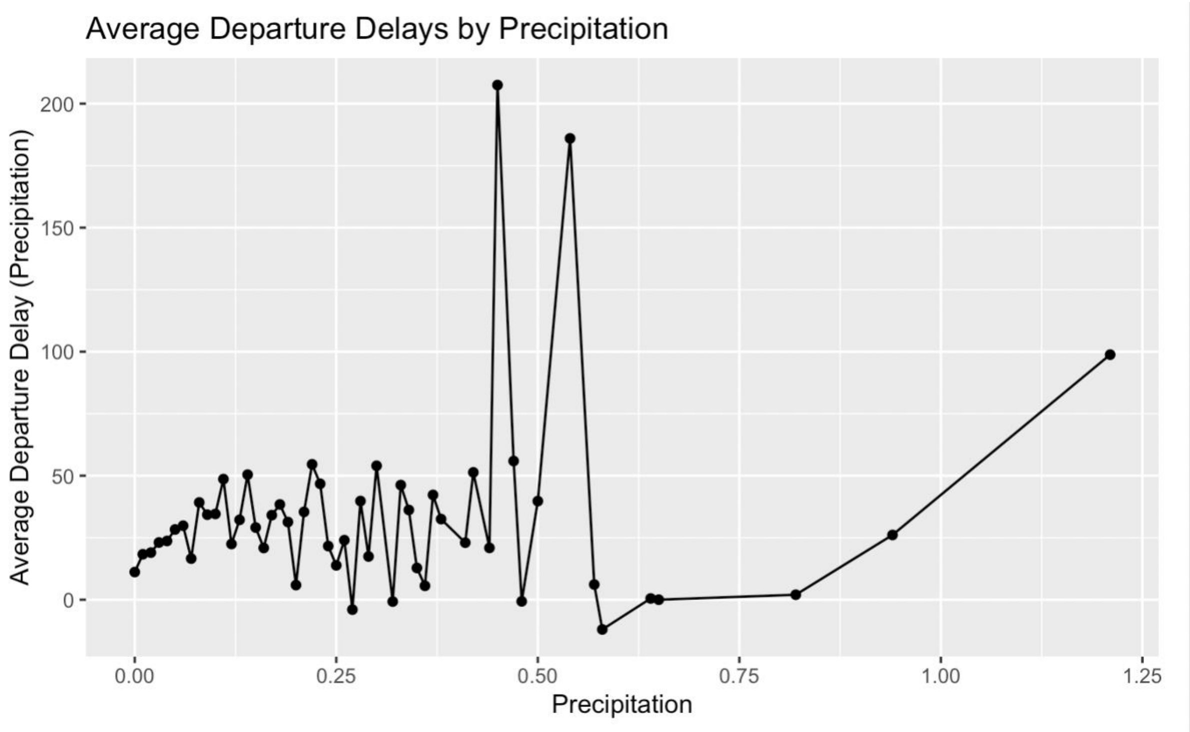
Analysis: The analysis indicates a modest association between wind speed and departure delays. Flights on windy days tend to have slightly longer delays, but the effect is not as significant as some other factors.

Implications: United Airlines can take the following actions based on these findings:

- **Weather Monitoring:** Implement robust weather monitoring systems to keep track of wind speed at departure airports. This can help with early identification of potential delays.
- **Ground Operations:** Ensure ground operations are prepared to handle windy conditions efficiently, particularly in terms of aircraft maintenance and boarding procedures.
- **Passenger Communication:** Inform passengers about potential windy conditions and any impact on their travel. In the case of noticeable wind-related delays, offer passengers rebooking options or accommodations.

Departure Delays by Precipitation

Now, we will examine how departure delays are affected by the precipitation. The following chart illustrates the average departure delays by temperature.



Flights between the interval 0.45-060 tend to have longer departure delays compared to other intervals.

Analysis: In this section, we investigate how precipitation impacts departure delays. We calculate the average departure delays based on different precipitation conditions.

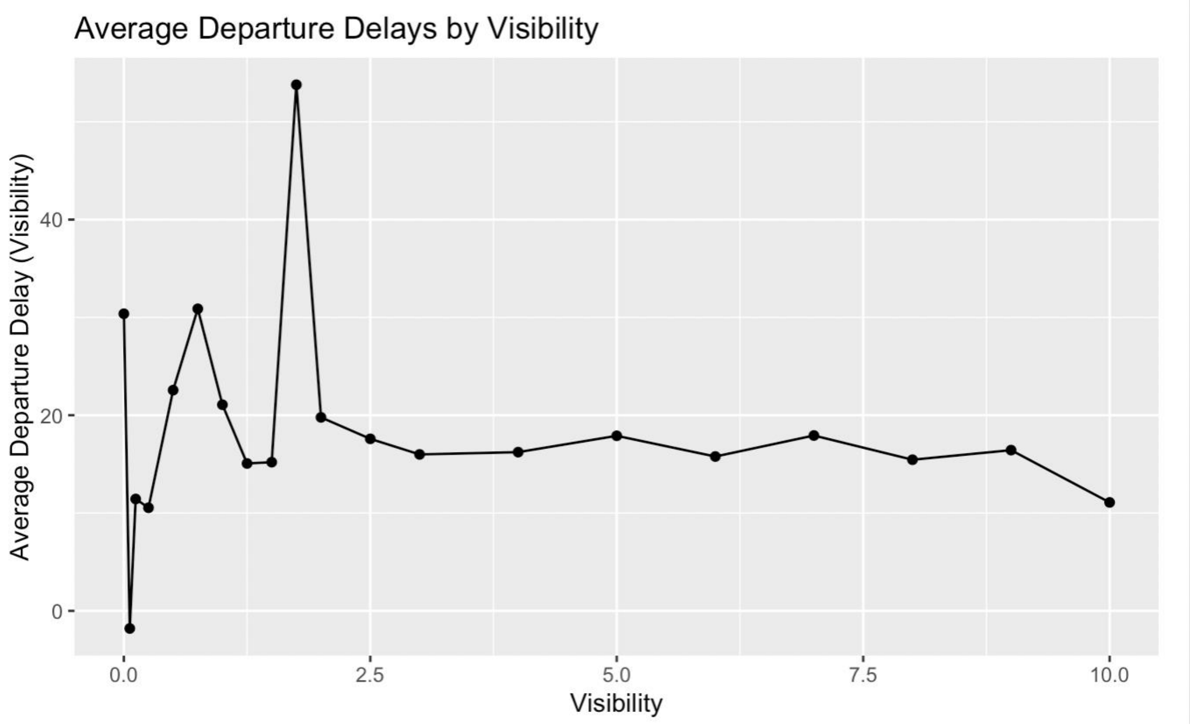
Implications: United Airlines can consider the following actions based on these findings:

- **Equipment Maintenance**: Regularly maintain and inspect equipment that is crucial for operations in precipitation, such as de-icing equipment and vehicles. Ensuring that this equipment is in good working condition is vital for minimizing delays.
- **Safety and Quality Control**: United Airlines should prioritize safety over punctuality during precipitation. The focus should be on maintaining safety standards and ensuring that flights can operate smoothly even in adverse weather conditions.
- **Weather Diversions**: Precipitation, especially severe storms, can necessitate the diversion of flights to other airports. United

Airlines should have contingency plans and resources ready to accommodate these diversions, including arrangements for passenger transfers and accommodations when needed.

Departure Delays by Visibility

Now, we will examine how departure delays are affected by the visibility. The following chart illustrates the average departure delays by visibility.



Flights between the interval 1.5-2.5 tend to have longer departure delays compared to other intervals.

Analysis: The analysis reveals a noticeable relationship between visibility and departure delays. Flights in poor visibility conditions experience longer delays, while flights in better visibility conditions tend to depart more promptly.

Implications: United Airlines can take the following actions based on these findings:

- **Pre-flight Assessment**: Implement thorough pre-flight assessments of visibility conditions. If visibility is expected to be poor, consider adjustments to flight schedules to allow for potential delays.
- **Proactive Communication**: Enhance communication with passengers in the event of poor visibility. Passengers should be informed of potential delays and provided with alternative options, such as rebooking or connecting flights.
- **Safety Precautions**: Emphasize the importance of safety in cases of reduced visibility. While minimizing delays is essential, safety should always be the top priority. Consider additional training and safety protocols for pilots and ground staff in adverse visibility conditions.
- **Technology Investment**: Explore investments in advanced visibility-enhancing technology for both ground operations and in-flight procedures. This could potentially reduce delays in poor visibility conditions.

Conclusion:

In conclusion, our analysis of departure delays in the context of time of day, time of year, temperature, wind speed, precipitation, and visibility has yielded valuable insights for United Airlines. By using the mean departure delay as a reference point, we have effectively compared and interpreted the data to inform actionable recommendations.

1. **Time of Day**: We observed that night hours experience slightly longer departure delays. To improve efficiency, United Airlines should consider adjusting flight schedules, optimizing resource allocation during peak hours, and enhancing passenger communication during potential delays.

2. **Time of Year**: Flights during fall and winter seasons have longer departure delays. To address this, United Airlines should adapt schedules during peak seasons, improve weather preparedness, and provide clear information and flexible rebooking options to passengers.

3. **Temperature**: Flights during warmer temperatures experience shorter delays. United Airlines can consider modifying schedules during colder weather, streamlining de-icing procedures, proactively informing passengers about cold weather-related delays, and ensuring well-maintained aircraft for cold weather operations.

4. **Wind Speed**: While we found a modest association between wind speed and departure delays, United Airlines can benefit from robust weather monitoring systems, efficient ground operations during windy conditions, and improved passenger communication.

5. **Precipitation**: Precipitation, particularly in the 0.45-0.60 range, leads to longer departure delays. United Airlines should prioritize equipment maintenance, safety, and quality control during precipitation, as well as prepare for weather diversions with contingency plans and resources in place.

6. **Visibility**: Poor visibility conditions correlate with longer departure delays. United Airlines can enhance pre-flight assessments, improve communication with passengers, prioritize safety in adverse visibility conditions, and explore investments in visibility-enhancing technology.

In all these aspects, the safety and satisfaction of passengers should remain paramount. These findings and recommendations can guide United Airlines in optimizing its operations, improving passenger experiences, and ensuring the highest level of safety, all while minimizing the impact of departure delays. By understanding the specific factors that contribute to delays, United Airlines can take proactive steps to mitigate their effects and enhance its overall service quality.

Appendix: Code

For reference, the code used in this analysis is provided in the appendix, along with comments explaining each section's purpose.

````
# Loads the necessary libraries
```{r}
library(nycflights13)
library(ggplot2)
library(dplyr)
```

#To load the data
```{r}
data("flights")
data("weather")
```

# Filtering of data
```{r}
flights_UA <- flights %>%
  filter(carrier == "UA")
glimpse(flights_UA)
```

# Merging two datasets as per requirement
```{r}
````
````

````
df <-merge(flights, weather, by.flights = c("origin", "time_hour"), by.weather = c("origin", "time_hour"), all.x = FALSE, all.y = FALSE, sort = TRUE)
Ua_df <- df %>%   filter(carrier=="UA")
Ua_df <- Ua_df %>%   filter(!is.na(dep_delay))

glimpse(Ua_df)
```
# Creating Line and point plot to represent data
```{r}
Ua_mean_delay <- Ua_df %>%
  group_by(hour) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
Ua_mean_delay

# Plot the relationship between time of day (hour) and average departure delays
ggplot(Ua_mean_delay, aes(x = hour, y = mean_dep_delay)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Departure Delays by Time of Day", x = "Time of Day (Hour)", y = "Average Departure Delay (Minutes)")
```


# Calculating P-value by performing permutation test
```{r}
# Create a function to map dep_delay to the corresponding time interval
map_to_interval <- function(dep_delay, intervals) {
  for (i in 1:(length(intervals) - 1)) {
    if (dep_delay >= intervals[i] && dep_delay < intervals[i + 1]) {
      return(i)
    }
  }
  return(length(intervals) - 1)
````

```
}



# Apply the mapping to the dataset
UA_merge <- UA_merge %>%
  mutate(time_interval = sapply(dep_delay, function(x)
map_to_interval(x, time_intervals)))

 # Calculate the observed statistic on the full data
observed_statistic <- UA_merge %>%
  group_by(time_interval) %>%
  summarise(mean_delay = mean(dep_delay, na.rm = TRUE))

 # Number of permutations
N <- 10^4

 # Initialize a vector to store permuted statistics
permuted_statistics <- numeric(N)

 # Set a seed for reproducibility
set.seed(123)

 # Perform the permutations
for (i in 1:N) {
  shuffled_data <- UA_merge %>%
    mutate(time_interval = sample(UA_merge$time_interval))
  permuted_statistic <- shuffled_data %>%
    group_by(time_interval) %>%
    summarise(mean_delay = mean(dep_delay, na.rm = TRUE))
  permuted_statistics[i] <- sum(abs(permuted_statistic$mean_delay -
observed_statistic$mean_delay))
}

# Calculate the p-value
```

```
p_value <- sum(permuted_statistics >=
observed_statistic$mean_delay) / N
# Print the p-value
cat("P-value on the full data:", p_value, "\n")

```
Creating Line and point plot to represent data
```{r}
Ua_mean_delay <- Ua_df %>%
  group_by(month) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
Ua_mean_delay

# Plot the relationship between time of day (hour) and average
departure delays
ggplot(Ua_mean_delay, aes(x = month, y = mean_dep_delay)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Departure Delays by Month", x = "Month
Time", y = "Average Departure Delay (Month)")
```

Calculating P-value by performing permutation test

```{r}
# Create a function to map dep_delay to the corresponding time
interval
map_to_interval <- function(dep_delay, intervals) {
  for (i in 1:(length(intervals) - 1)) {
    if (dep_delay >= intervals[i] && dep_delay < intervals[i + 1]) {
      return(i)
    }
  }
  return(length(intervals) - 1)
}
```

```
# Apply the mapping to the dataset
UA_merge <- UA_merge %>%
  mutate(time_interval = sapply(dep_delay, function(x)
map_to_interval(x, time_intervals)))

# Divide the data into two groups: first 6 months and second 6
months
first_six_months <- UA_merge %>%
  filter(month.x <= 6)

second_six_months <- UA_merge %>%
  filter(month.x > 6)

# Calculate observed statistics for the first 6 months and the second 6
months
observed_statistic_first <- first_six_months %>%
  group_by(time_interval) %>%
  summarise(mean_delay = mean(dep_delay, na.rm = TRUE))

observed_statistic_second <- second_six_months %>%
  group_by(time_interval) %>%
  summarise(mean_delay = mean(dep_delay, na.rm = TRUE))

# Number of permutations
N <- 10^4

# Initialize vectors to store permuted statistics for each group
permuted_statistics_first <- numeric(N)
permuted_statistics_second <- numeric(N)

# Set a seed for reproducibility
set.seed(123)

# Perform the permutations for the first 6 months
```

```
for (i in 1:N) {
  shuffled_data <- first_six_months %>%
    mutate(time_interval = sample(first_six_months$time_interval))

  permuted_statistic <- shuffled_data %>%
    group_by(time_interval) %>%
    summarise(mean_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_first[i] <-
sum(abs(permuted_statistic$mean_delay -
observed_statistic_first$mean_delay))
}

# Perform the permutations for the second 6 months
for (i in 1:N) {
  shuffled_data <- second_six_months %>%
    mutate(time_interval =
sample(second_six_months$time_interval))

  permuted_statistic <- shuffled_data %>%
    group_by(time_interval) %>%
    summarise(mean_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_second[i] <-
sum(abs(permuted_statistic$mean_delay -
observed_statistic_second$mean_delay))
}

# Calculate the p-values for each group
p_value_first <- sum(permuted_statistics_first >=
observed_statistic_first$mean_delay) / N
p_value_second <- sum(permuted_statistics_second >=
observed_statistic_second$mean_delay) / N

# Print the p-values
```

````
cat("P-value for the first 6 months:", p_value_first, "\n")
cat("P-value for the second 6 months:", p_value_second, "\n")

```
# Creating Line and point plot to represent data
```{r}
Ua_mean_delay <- Ua_df %>%
  group_by(temp) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
Ua_mean_delay

# Plot the relationship between time of day (hour) and average
departure delays
ggplot(Ua_mean_delay, aes(x = temp, y = mean_dep_delay)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Departure Delays by Temp", x = "Temperature",
y = "Average Departure Delay (temp)")
```
# Calculating P-value by performing permutation test

```{r}

# Create a function to map dep_delay to the corresponding time
interval
map_to_interval <- function(dep_delay, intervals) {
  for (i in 1:(length(intervals) - 1)) {
    if (dep_delay >= intervals[i] && dep_delay < intervals[i + 1]) {
      return(i)
    }
  }
  return(length(intervals) - 1)
}

# Apply the mapping to the dataset
````

```
UA_merge <- UA_merge %>%
  mutate(time_interval = sapply(dep_delay, function(x)
map_to_interval(x, time_intervals)))

# Divide the data into two groups based on temperature
temp_less_than_35 <- Ua_df %>% filter(temp < 35)
temp_greater_than_35 <- Ua_df %>% filter(temp >= 35)

# Calculate observed statistics for each group
observed_statistic_less_than_35 <- temp_less_than_35 %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

observed_statistic_greater_than_35 <- temp_greater_than_35 %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

# Number of permutations
N <- 10^4

# Initialize vectors to store permuted statistics for each group
permuted_statistics_less_than_35 <- numeric(N)
permuted_statistics_greater_than_35 <- numeric(N)

# Set a seed for reproducibility
set.seed(123)

# Perform the permutations for temperature less than 35°F
for (i in 1:N) {
  shuffled_data <- temp_less_than_35 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
```

```
  permuted_statistics_less_than_35[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_less_than_35$mean_dep_delay)
}

# Perform the permutations for temperature greater than 35°F
for (i in 1:N) {
  shuffled_data <- temp_greater_than_35 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_greater_than_35[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_greater_than_35$mean_dep_delay)
}

# Calculate the p-values for each group
p_value_less_than_35 <- sum(permuted_statistics_less_than_35 >=
observed_statistic_less_than_35$mean_dep_delay) / N
p_value_greater_than_35 <-
sum(permuted_statistics_greater_than_35 >=
observed_statistic_greater_than_35$mean_dep_delay) / N

# Print the p-values
cat("P-value for temperature less than 35°F:", p_value_less_than_35,
"\n")
cat("P-value for temperature greater than 35°F:",
p_value_greater_than_35, "\n")

```

Creating Line and point plot to represent data

```{r}
```

````
Ua_mean_delay <- Ua_df %>%
  group_by(wind_speed) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
Ua_mean_delay

# Plot the relationship between time of day (hour) and average
departure delays
ggplot(Ua_mean_delay, aes(x = wind_speed, y = mean_dep_delay)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Departure Delays by Speed of Wind", x =
"Speed of Wind", y = "Average Departure Delay (Wind Speed)")
```
````

Calculating P-value by performing permutation test

````
```{r}
# Divide the data into two groups based on wind_speed
wind_speed_less_than_10 <- Ua_df %>% filter(wind_speed <= 10)
wind_speed_greater_than_10 <- Ua_df %>% filter(wind_speed > 10)

# Calculate observed statistics for each group
observed_statistic_less_than_10 <- wind_speed_less_than_10 %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

observed_statistic_greater_than_10 <- wind_speed_greater_than_10
%>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

# Number of permutations
N <- 10^4

# Initialize vectors to store permuted statistics for each group
permuted_statistics_less_than_10 <- numeric(N)
permuted_statistics_greater_than_10 <- numeric(N)
````

```
# Set a seed for reproducibility
set.seed(123)

# Perform the permutations for wind_speed less than or equal to 10
for (i in 1:N) {
  shuffled_data <- wind_speed_less_than_10 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_less_than_10[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_less_than_10$mean_dep_delay)
}

# Perform the permutations for wind_speed greater than 10
for (i in 1:N) {
  shuffled_data <- wind_speed_greater_than_10 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_greater_than_10[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_greater_than_10$mean_dep_delay)
}

# Calculate the p-values for each group
p_value_less_than_10 <- sum(permuted_statistics_less_than_10 >=
observed_statistic_less_than_10$mean_dep_delay) / N
p_value_greater_than_10 <-
sum(permuted_statistics_greater_than_10 >=
observed_statistic_greater_than_10$mean_dep_delay) / N
```

````
# Print the p-values
cat("P-value for wind_speed less than or equal to 10:",
p_value_less_than_10, "\n")
cat("P-value for wind_speed greater than 10:",
p_value_greater_than_10, "\n")

```
# Creating Line and point plot to represent data
```{r}
Ua_mean_delay <- Ua_df %>%
  group_by(precip) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
Ua_mean_delay

# Plot the relationship between time of day (hour) and average
departure delays
ggplot(Ua_mean_delay, aes(x = precip, y = mean_dep_delay)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Departure Delays by Precipitation", x =
"Precipitation", y = "Average Departure Delay (Precipitationsc)")
```
# Calculating P-value by performing permutation test

```{r}
# Divide the data into two groups based on precip
precip_less_than_0.2 <- Ua_df %>% filter(precip <= 0.2)
precip_greater_than_2 <- Ua_df %>% filter(precip > 2)

# Calculate observed statistics for each group
observed_statistic_less_than_0.2 <- precip_less_than_0.2 %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

observed_statistic_greater_than_2 <- precip_greater_than_2 %>%
````

```
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

# Number of permutations
N <- 10^4

# Initialize vectors to store permuted statistics for each group
permuted_statistics_less_than_0.2 <- numeric(N)
permuted_statistics_greater_than_2 <- numeric(N)

# Set a seed for reproducibility
set.seed(123)

# Perform the permutations for precip less than or equal to 0.2
for (i in 1:N) {
  shuffled_data <- precip_less_than_0.2 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_less_than_0.2[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_less_than_0.2$mean_dep_delay)
}

# Perform the permutations for precip greater than 2
for (i in 1:N) {
  shuffled_data <- precip_greater_than_2 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
```

```r
  permuted_statistics_greater_than_2[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_greater_than_2$mean_dep_delay)
}

# Calculate the p-values for each group
p_value_less_than_0.2 <- sum(permuted_statistics_less_than_0.2 >=
observed_statistic_less_than_0.2$mean_dep_delay) / N
p_value_greater_than_2 <- sum(permuted_statistics_greater_than_2
>= observed_statistic_greater_than_2$mean_dep_delay) / N

# Print the p-values
cat("P-value for precip <= 0.2:", p_value_less_than_0.2, "\n")
cat("P-value for precip > 2:", p_value_greater_than_2, "\n")


```

Creating Line and point plot to represent data
```{r}
Ua_mean_delay <- Ua_df %>%
  group_by(visib) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
Ua_mean_delay

# Plot the relationship between time of day (hour) and average
departure delays
ggplot(Ua_mean_delay, aes(x = visib, y = mean_dep_delay)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Departure Delays by Visibility", x = "Visibility", y
= "Average Departure Delay (Visibility)")
```
Calculating P-value by performing permutation test

````
```{r}
# Divide the data into two groups based on visib
visib_less_than_or_equal_to_5 <- Ua_df %>% filter(visib <= 5)
visib_greater_than_5 <- Ua_df %>% filter(visib > 5)

# Calculate observed statistics for each group
observed_statistic_less_than_or_equal_to_5 <-
visib_less_than_or_equal_to_5 %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

observed_statistic_greater_than_5 <- visib_greater_than_5 %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

# Number of permutations
N <- 10^4

# Initialize vectors to store permuted statistics for each group
permuted_statistics_less_than_or_equal_to_5 <- numeric(N)
permuted_statistics_greater_than_5 <- numeric(N)

# Set a seed for reproducibility
set.seed(123)

# Perform the permutations for visib less than or equal to 5
for (i in 1:N) {
  shuffled_data <- visib_less_than_or_equal_to_5 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_less_than_or_equal_to_5[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_less_than_or_equal_to_5$mean_dep_delay)
}
````
````

```
# Perform the permutations for visib greater than 5
for (i in 1:N) {
  shuffled_data <- visib_greater_than_5 %>%
    mutate(dep_delay = sample(dep_delay))

  permuted_statistic <- shuffled_data %>%
    summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))

  permuted_statistics_greater_than_5[i] <-
abs(permuted_statistic$mean_dep_delay -
observed_statistic_greater_than_5$mean_dep_delay)
}

# Calculate the p-values for each group
p_value_less_than_or_equal_to_5 <-
sum(permuted_statistics_less_than_or_equal_to_5 >=
observed_statistic_less_than_or_equal_to_5$mean_dep_delay) / N
p_value_greater_than_5 <- sum(permuted_statistics_greater_than_5
>= observed_statistic_greater_than_5$mean_dep_delay) / N

# Print the p-values
cat("P-value for visib less than or equal to 5:",
p_value_less_than_or_equal_to_5, "\n")
cat("P-value for visib greater than 5:", p_value_greater_than_5, "\n")

```

This report is intended to provide a non-technical audience with a clear overview of the analysis and its findings. It emphasizes the practical implications of the results, making it suitable for presentation to management and other stakeholders.