Project Topic: United Airlines departure delays study

TABLE OF CONTENT

Title	Page No
Introduction	3
Data Analysis and Permutation test	4
Appendix	10

Introduction

The project uses the data included in the nycflights13 package

The report addresses the relationship between departure delays and each of the following:

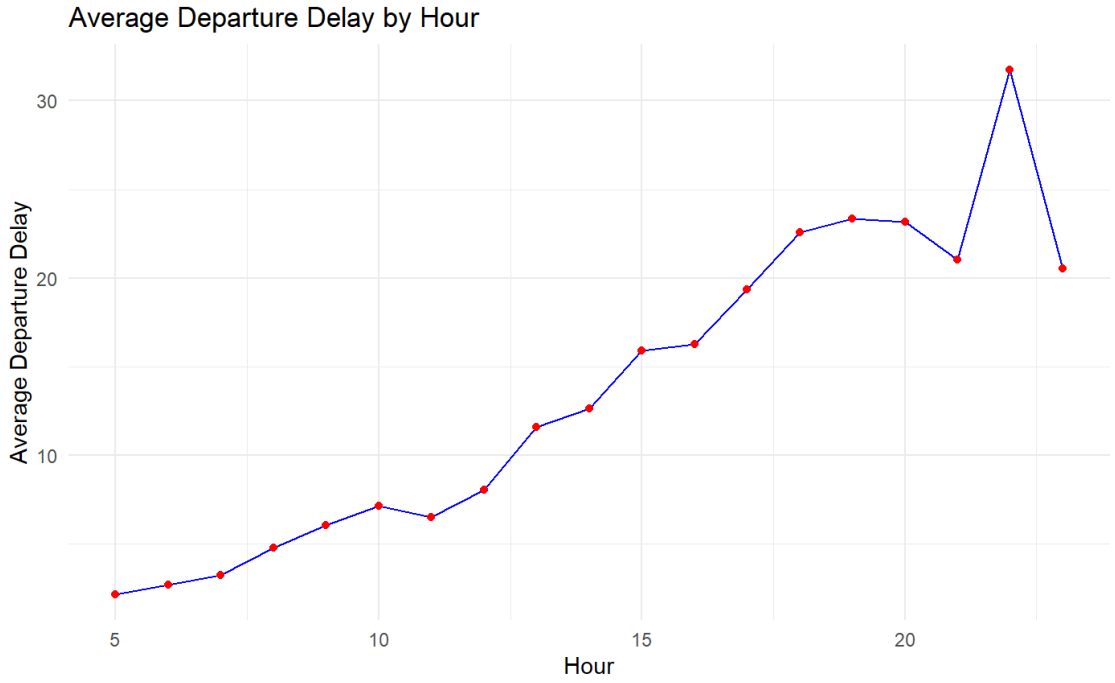
1. Time of day
2. Time of year
3. Temperature
4. Wind speed
5. Precipitation
6. Visibility

Data Analysis and Permutation test

Data Analysis and Permutation test to find out Relations between departure delays and other factors such as Time of day, Time of year, Temperature, Wind speed, Precipitation, Visibility

1)Time of day

Graphical analysis



The graph shows the average departure delay by hour, revealing an increasing trend in delays throughout the day, peaking in the late evening hours, followed by a sharp decline

Permutation Test

H0: Departure delay and Time of Day have no relation

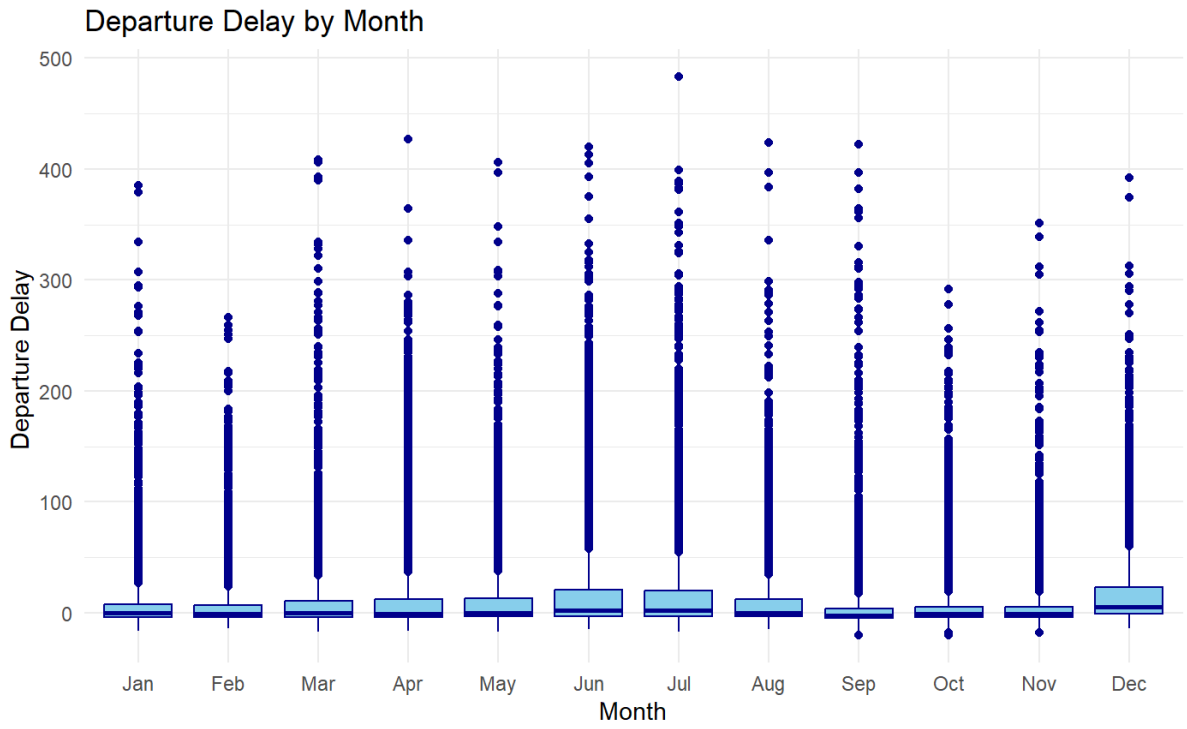
H1: Time of Day effects the departure delay

By dividing the hour variable into different categories and relabeling them, then the permutation test is conducted

The test is done for departure delay between different categories. We got p value very of 2e-04 which is less than alpha=0.05, so we reject null hypothesis and conclude that Time of Day effects the departure delay

2) Time of year

Graphical analysis



The boxplot shows the distribution of departure delays by month, indicating that delays are consistently present throughout the year, with some months like June and July experiencing higher maximum delays, as evident from the outliers.

Permutation test

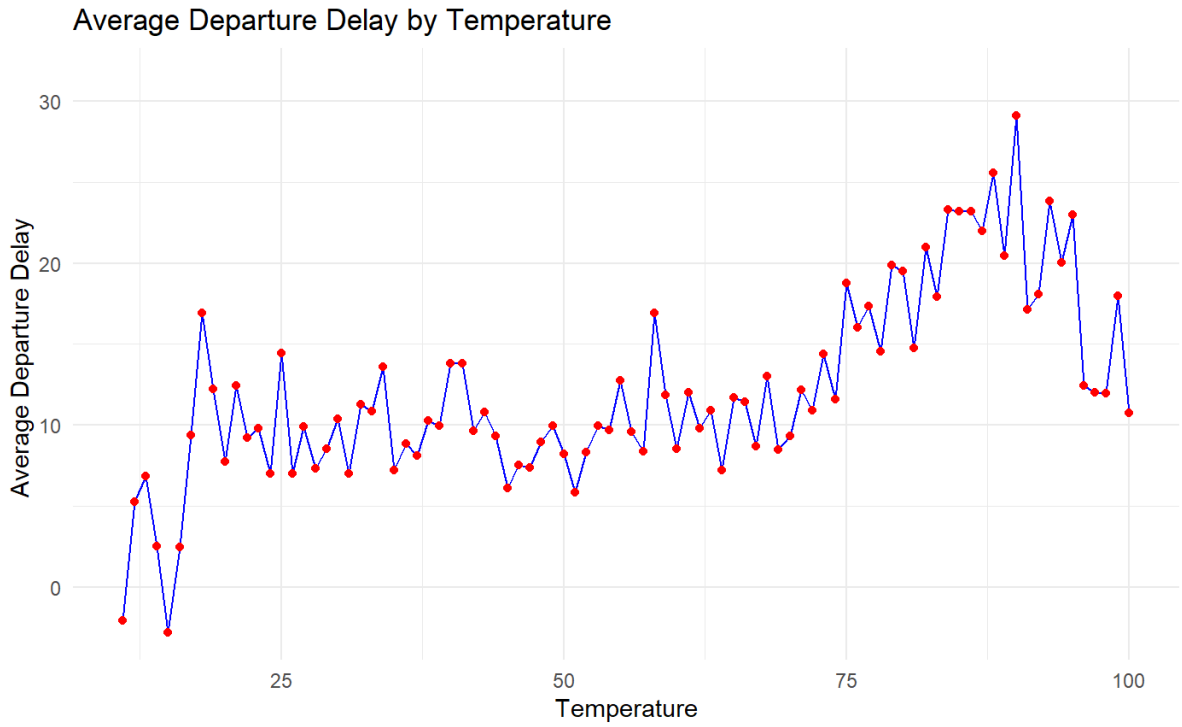
H0: Departure delay and Time of Year has no relation
H1: Time of Year effects departure delay

We divided the month variable into four categories based on seasons

The test is done for departure delay between different categories. We got p value for all and all are less than alpha=0.05, so we reject null hypothesis and conclude that Time of Year effects the departure delay

3)Temperature

Graphical analysis



The graph shows a positive correlation between temperature and average departure delay, indicating that as the temperature increases, departure delays generally become longer, peaking around 85-90°F before slightly declining.

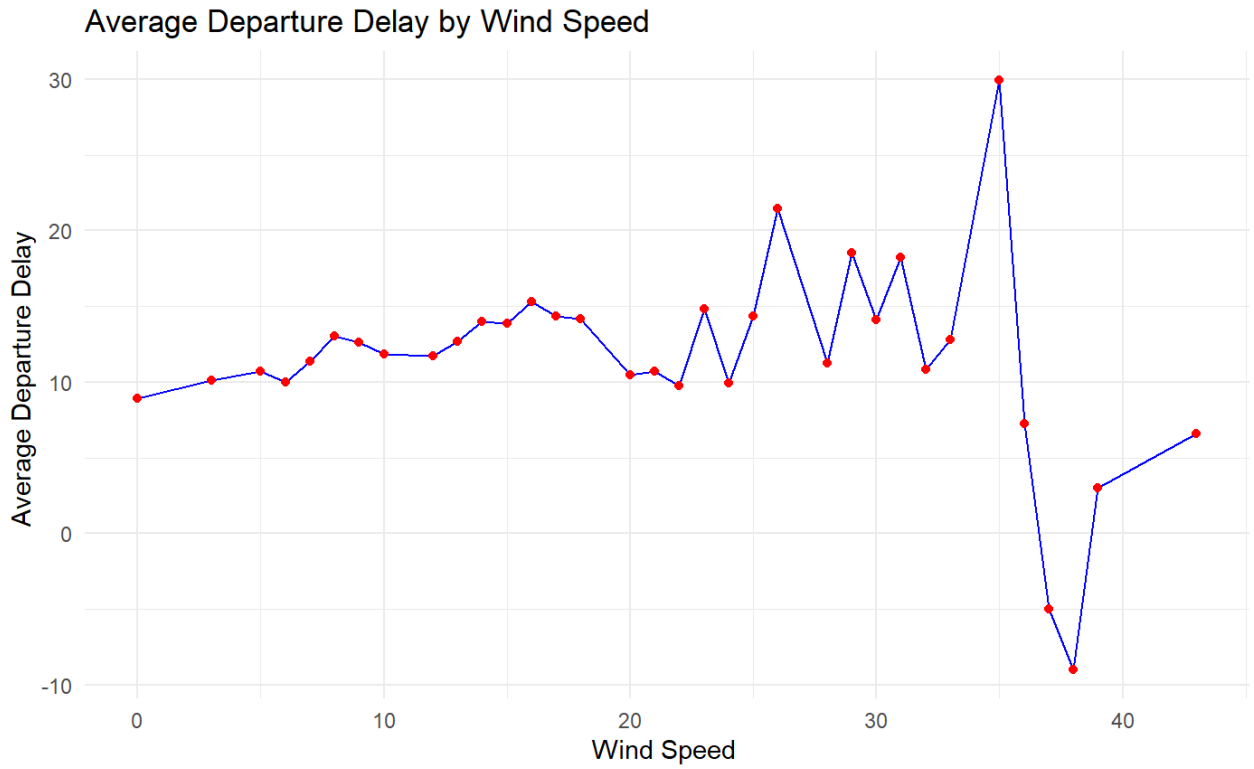
Permutation test
H0: Departure delay and Temperature have no relation
H1: Temperature effects departure delay

We divided the temperature variable into two categories

The test is done for departure delay between two categories. We got p value of 2e-04 which is less than alpha=0.05, so we reject null hypothesis and conclude that Temperature effects the departure delay

4)Wind speed

Graphical analysis



The graph shows that average departure delays generally increase with wind speed, peaking around 30 mph. However, there is a significant drop in delays when wind speed exceeds 35 mph, followed by a slight increase again.

Permutation test:
H0: Departure delay and Wind Speed have no relation
H1: Wind Speed effects the departure delay

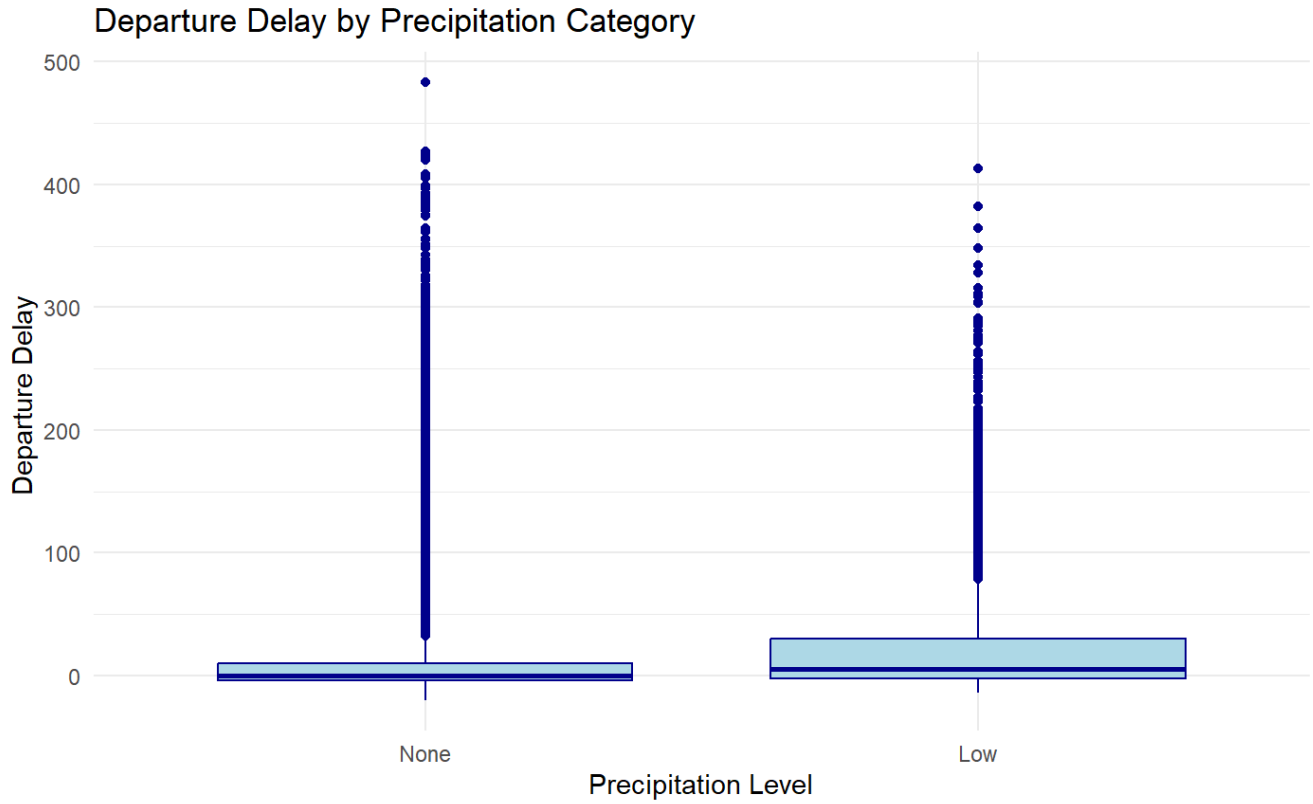
We divided the wind speed variable into two categories

The test is done for departure delay between two categories. We got p value of 2e-04

which is less than alpha=0.05, so we reject null hypothesis and conclude that Wind Speed effects the departure delay

5) Precipitation

Graphical analysis



The boxplot indicates that flights experience more departure delays when there is low precipitation compared to no precipitation, with a wider spread of delays and more extreme outliers under low precipitation conditions.

Permutation test

H0: Departure delay and Precipitation have no relation

H1: Precipitation effects the departure delay

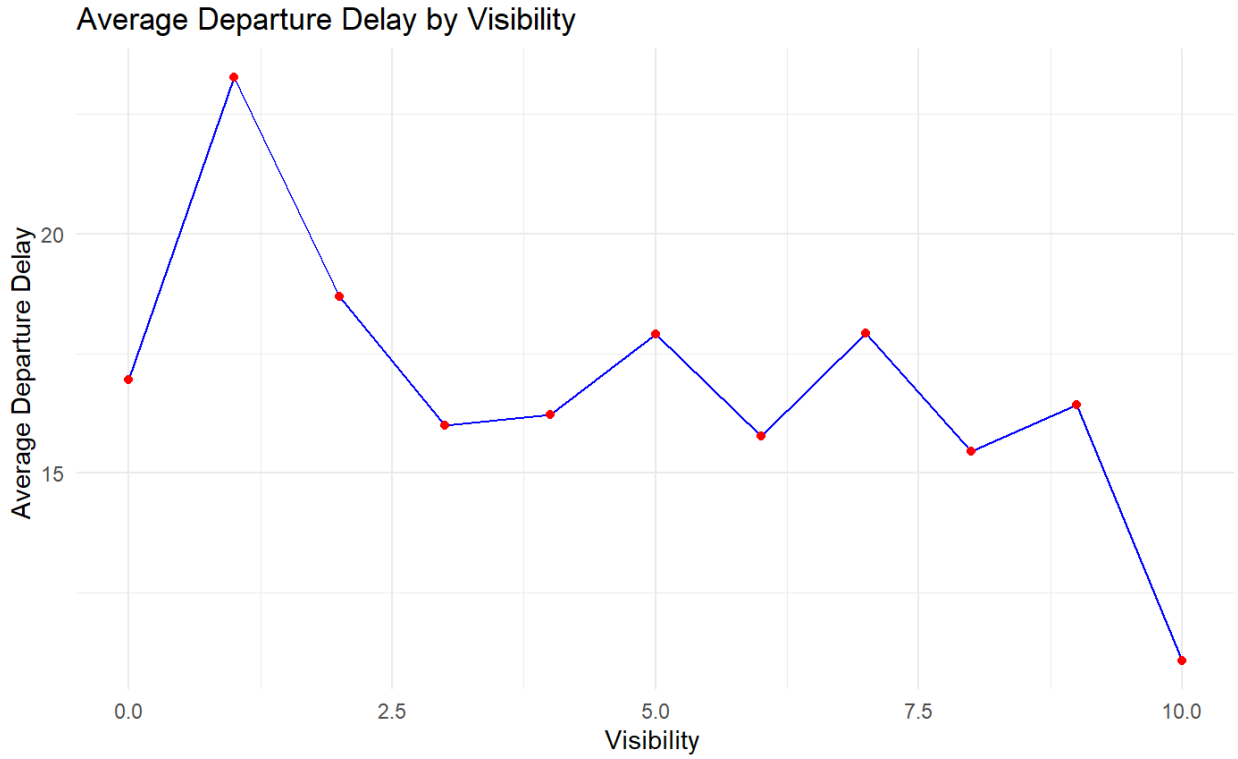
We divided the precipitation variable into two categories

The test is done for departure delay between two categories. We got p value of 2e-04

which is less than alpha=0.05, so we reject null hypothesis and conclude that precipitation effects the departure delay

6)Visibility

Graphical analysis



We can see a variation in average departure delays based on visibility values, where average departure delays decreases with visibility

Permutation test

H0: departure delay and Visibility have no relation

H1: Visibility effects the departure delay

We divided the visibility variable into two categories

The test is done for departure delay between two categories. We got p value of 2e-04

which is less than alpha=0.05, so we reject null hypothesis and conclude that visibility effects the departure delay

Appendix

Program Codes

Processing Data

```
library(tidyverse)

library(nycflights13)

UAf <- flights %>%

  filter(carrier=="UA")

UAf <- UAf %>%

  filter(!is.na(dep_delay))


m1 <-merge(flights, weather, flights = c("origin", "time_hour","hour"), by.weather = c("origin",
"time_hour","hour"), all.x = FALSE, all.y = FALSE, sort = TRUE)

m2 <- m1 %>%

  filter(carrier=="UA")

maindf <- m2 %>%

  filter(!is.na(dep_delay)) %>%


select(dep_time,year,month,day,dep_time,dep_delay,carrier,time_hour,hour,temp,wind_speed,pr
ecip,visib)
```

Data Visualization

```
ggplot(data = maindf, aes(x = hour, y =dep_delay ))+
  geom_point()
avg_delay_by_hour <- maindf %>%
  group_by(hour) %>%
  summarize(avg_delay = mean(dep_delay, na.rm = TRUE))

# Create the line plot
ggplot(avg_delay_by_hour, aes(x = hour, y = avg_delay)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average Departure Delay by Hour ",
       x = "Hour",
       y = "Average Departure Delay") +
  theme_minimal()



ggplot(data = maindf, aes(x = month, y =dep_delay ))+
  geom_point()

ggplot(maindf, aes(x = factor(month), y = dep_delay)) +
  geom_boxplot(fill = "skyblue", color = "darkblue") +
  labs(title = "Departure Delay by Month",
       x = "Month",
       y = "Departure Delay ") +
  scale_x_discrete(labels = month.abb) +
  theme_minimal()
```

```
ggplot(data = maindf, aes(x = temp, y =dep_delay ))+
  geom_point()
avg_delay_by_temp <- maindf %>%
  mutate(temp_rounded = round(temp)) %>%
  group_by(temp_rounded) %>%
  summarize(avg_delay = mean(dep_delay, na.rm = TRUE))

# Create the line plot
ggplot(avg_delay_by_temp, aes(x = temp_rounded, y = avg_delay)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average Departure Delay by Temperature",
       x = "Temperature",
       y = "Average Departure Delay ") +
  theme_minimal()

ggplot(data = maindf, aes(x = wind_speed, y =dep_delay ))+
  geom_point()
avg_delay_by_wind <- maindf %>%
  mutate(wind_speed_rounded = round(wind_speed)) %>%
  group_by(wind_speed_rounded) %>%
  summarize(avg_delay = mean(dep_delay, na.rm = TRUE))

# Create the line plot
ggplot(avg_delay_by_wind, aes(x = wind_speed_rounded, y = avg_delay)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average Departure Delay by Wind Speed",
       x = "Wind Speed ",
```

```
      y = "Average Departure Delay ") +
  theme_minimal()



ggplot(data = maindf, aes(x = precip, y =dep_delay ))+
  geom_point()
flights_data <- maindf %>%
  mutate(precip_category = cut(precip,
                               breaks = c(-Inf, 0, 5, 15, Inf),
                               labels = c("None", "Low", "Medium", "High")))

# Create the boxplot
ggplot(flights_data, aes(x = precip_category, y = dep_delay)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Departure Delay by Precipitation Category",
       x = "Precipitation Level",
       y = "Departure Delay ") +
  theme_minimal()

ggplot(data = maindf, aes(x = visib, y =dep_delay ))+
  geom_point()
avg_delay_by_visibility <- maindf %>%
  mutate(visibility_rounded = round(visib)) %>%
  group_by(visibility_rounded) %>%
  summarize(avg_delay = mean(dep_delay, na.rm = TRUE))

# Create the line plot
ggplot(avg_delay_by_visibility, aes(x = visibility_rounded, y = avg_delay)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
```

```
labs(title = "Average Departure Delay by Visibility",
     x = "Visibility ",
     y = "Average Departure Delay ") +
theme_minimal()
```

Permutation Test

1. Time of day

```
maindf <- maindf %>%
  mutate(time_of_day = case_when(
    between(hour, 5, 11) ~ 1,
    between(hour, 12, 17) ~ 2,
    TRUE ~ 3))
N<- 10^4-1
observed <- mean(maindf$dep_delay[maindf$time_of_day == 2], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$time_of_day == 1], na.rm = TRUE)
result <- numeric(N)
md<-maindf %>%
  filter(time_of_day == 1)
for (i in 1:N)
{
  index <- sample(nrow(maindf), size=nrow(md),replace = FALSE)
  result[i] <- mean(maindf$dep_delay[index], na.rm = TRUE) -mean(maindf$dep_delay[-
index], na.rm = TRUE)
}
ggplot(data = tibble(result), mapping = aes(x = result)) +
  geom_histogram() +
  geom_vline(xintercept = observed, color = "green")
2 * ((sum(result >= observed) + 1) / (N + 1))
```

```
N<- 10^4-1

observed <- mean(maindf$dep_delay[maindf$time_of_day == 3], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$time_of_day == 2], na.rm = TRUE)

result <- numeric(N)

mdt<-maindf %>%

  filter(time_of_day == 2)

for (i in 1:N)

{

  index1 <- sample(nrow(maindf), size=nrow(mdt),replace = FALSE)

  result[i] <- mean(maindf$dep_delay[index1], na.rm = TRUE) -
mean(maindf$dep_delay[-index1], na.rm = TRUE)

}

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

  geom_histogram() +

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

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

```
N<- 10^4-1

observedl <- mean(maindf$dep_delay[maindf$time_of_day == 3], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$time_of_day == 1], na.rm = TRUE)

resultl <- numeric(N)

mdt1<-maindf %>%

  filter(time_of_day == 1)

for (i in 1:N)

{

  index2 <- sample(nrow(maindf), size=nrow(mdt1),replace = FALSE)

  resultl[i] <- mean(maindf$dep_delay[index2], na.rm = TRUE) -
mean(maindf$dep_delay[-index2], na.rm = TRUE)

}

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

  geom_histogram() +

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

2 * ((sum(resultl >= observedl) + 1) / (N + 1))
```

2.Time of year

```
maindf <- maindf %>%
  mutate(Seasons = case_when(
    between(maindf$month, 1, 2) ~ "Wi",
    between(maindf$month, 3, 5) ~ "Sp",
    between(maindf$month, 6, 8) ~ "Su",
    between(maindf$month, 8, 12) ~ "Au",
    TRUE ~ "NA"))
N<- 10^5-1
observedu <- mean(maindf$dep_delay[maindf$Seasons == 'Sp'], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$Seasons == 'Wi'], na.rm = TRUE)
resultu <- numeric(N)
sample.size= length(maindf$Seasons)
group.1.size = nrow(tibble(maindf$dep_delay[maindf$Seasons == "Wi"]))
for (i in 1:N)
{
  indexNew <- sample(sample.size, size= group.1.size, replace = FALSE)
  resultu[i] <- mean(maindf$dep_delay[indexNew], na.rm = TRUE) -mean(maindf$dep_delay[-
indexNew], na.rm = TRUE)
}
ggplot(data = tibble(resultu), mapping = aes(x = resultu)) +
  geom_histogram() +
  geom_vline(xintercept = observedu, color = "red")
2 * ((sum(resultu >= observedu) + 1) / (N + 1))
```

```
N<- 10^5-1

observedu1 <- mean(maindf$dep_delay[maindf$Seasons == 'Su'], na.rm=TRUE)-
mean(maindf$dep_delay[maindf$Seasons == 'Wi'], na.rm = TRUE)

resultu2 <- numeric(N)

sample.size= length(maindf$Seasons)

group.1.size = nrow(tibble(maindf$dep_delay[maindf$Seasons == "Wi"]))

for (i in 1:N)

{

  indexNew1 <- sample(sample.size, size= group.1.size, replace = FALSE)

  resultu2[i] <- mean(maindf$dep_delay[indexNew1], na.rm = TRUE) -mean(maindf$dep_delay[-
indexNew1], na.rm = TRUE)

}

ggplot(data = tibble(resultu2), mapping = aes(x = resultu2)) +

  geom_histogram() +

  geom_vline(xintercept = observedu1, color = "red")

2 * ((sum(resultu2 >= observedu1) + 1) / (N + 1))
```

```
N<- 10^5-1

observedu2 <- mean(maindf$dep_delay[maindf$Seasons == 'Au'], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$Seasons == 'Wi'], na.rm = TRUE)

resultu3 <- numeric(N)

sample.size= length(maindf$Seasons)

group.1.size = nrow(tibble(maindf$dep_delay[maindf$Seasons == "Wi"]))

for (i in 1:N)

{

  indexNew2 <- sample(sample.size, size= group.1.size, replace = FALSE)

  resultu3[i] <- mean(maindf$dep_delay[indexNew2], na.rm = TRUE) -mean(maindf$dep_delay[-
indexNew2], na.rm = TRUE)

}

ggplot(data = tibble(resultu3), mapping = aes(x = resultu3)) +

  geom_histogram() +

  geom_vline(xintercept = observedu2, color = "red")

2 * ((sum(resultu3 >= observedu2) + 1) / (N + 1))
```

```
N<- 10^5-1

observedu3 <- mean(maindf$dep_delay[maindf$Seasons == 'Su'], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$Seasons == 'Sp'], na.rm = TRUE)

resultu4 <- numeric(N)

sample.size= length(maindf$Seasons)

group.1.size = nrow(tibble(maindf$dep_delay[maindf$Seasons == "Sp"]))

for (i in 1:N)

{

  indexNew3 <- sample(sample.size, size= group.1.size, replace = FALSE)

  resultu4[i] <- mean(maindf$dep_delay[indexNew3], na.rm = TRUE) -mean(maindf$dep_delay[-
indexNew3], na.rm = TRUE)

}

ggplot(data = tibble(resultu4), mapping = aes(x = resultu4)) +

  geom_histogram() +

  geom_vline(xintercept = observedu3, color = "red")

2 * ((sum(resultu4 >= observedu3) + 1) / (N + 1))
```

```
N<- 10^5-1

observedu4 <- mean(maindf$dep_delay[maindf$Seasons == 'Sp'], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$Seasons == 'Au'], na.rm = TRUE)

resultu5 <- numeric(N)

sample.size= length(maindf$Seasons)

group.1.size = nrow(tibble(maindf$dep_delay[maindf$Seasons == "Sp"]))

for (i in 1:N)

{

  indexNew4 <- sample(sample.size, size= group.1.size, replace = FALSE)

  resultu5[i] <- mean(maindf$dep_delay[indexNew4], na.rm = TRUE) -mean(maindf$dep_delay[-
indexNew4], na.rm = TRUE)

}

ggplot(data = tibble(resultu5), mapping = aes(x = resultu5)) +

  geom_histogram() +

  geom_vline(xintercept = observedu4, color = "red")

2 * ((sum(resultu5 >= observedu4) + 1) / (N + 1))
```

```
N<- 10^5-1

observedu5 <- mean(maindf$dep_delay[maindf$Seasons == 'Su'], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$Seasons == 'Au'], na.rm = TRUE)

resultu6 <- numeric(N)

sample.size= length(maindf$Seasons)

group.1.size = nrow(tibble(maindf$dep_delay[maindf$Seasons == "Su"]))

for (i in 1:N)

{

  indexNew5 <- sample(sample.size, size= group.1.size, replace = FALSE)

  resultu6[i] <- mean(maindf$dep_delay[indexNew5], na.rm = TRUE) -mean(maindf$dep_delay[-
indexNew5], na.rm = TRUE)

}

ggplot(data = tibble(resultu6), mapping = aes(x = resultu6)) +

  geom_histogram() +

  geom_vline(xintercept = observedu5, color = "red")

2 * ((sum(resultu6 >= observedu5) + 1) / (N + 1))
```

3.Temperature

```
maindf <- maindf %>%
  mutate(tempNew = case_when(
    between(temp, 0 , 58) ~ 0,
    TRUE ~ 1))
N<- 10^4-1
observedy <- mean(maindf$dep_delay[maindf$tempNew == 1], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$tempNew == 0], na.rm = TRUE)
resulty <- numeric(N)
for (i in 1:N)
{
  index3 <- sample(nrow(maindf), size=nrow(maindf %>% filter(tempNew == 1)),replace =
FALSE)
  resulty[i] <- mean(maindf$dep_delay[index3], na.rm = TRUE) -mean(maindf$dep_delay[-
index3], na.rm = TRUE)
}
ggplot(data = tibble(resulty), mapping = aes(x = resulty)) +
  geom_histogram() +
  geom_vline(xintercept = observedy, color = "red")
2 * ((sum(resulty >= observedy) + 1) / (N + 1))
```

4.Wind Speed

```
maindf <- maindf %>%
  mutate(WindNew = case_when(
    between(wind_speed, 0 , 10.31242) ~ 0,
    TRUE ~ 1))


N<- 10^4-1
observedi <- mean(maindf$dep_delay[maindf$WindNew == 1], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$WindNew == 0], na.rm = TRUE)
resulti <- numeric(N)
for (i in 1:N)
{
  index4 <- sample(nrow(maindf), size=nrow(maindf %>% filter(WindNew == 1)),replace =
FALSE)
  resulti[i] <- mean(maindf$dep_delay[index4], na.rm = TRUE) -mean(maindf$dep_delay[-
index4], na.rm = TRUE)
}
ggplot(data = tibble(resulti), mapping = aes(x = resulti)) +
  geom_histogram() +
  geom_vline(xintercept = observedi, color = "red")
2 * ((sum(resulti >= observedi) + 1) / (N + 1))
```

5.Precipitation

```
maindf <- maindf %>%
  mutate(PrecipNew = case_when(
  between(precip, 0 , 0.005091357) ~ 0,
  TRUE ~ 1))

N<- 10^4-1
observedq <- mean(maindf$dep_delay[maindf$PrecipNew == 1], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$PrecipNew == 0], na.rm = TRUE)
resultq <- numeric(N)
for (i in 1:N)
{
  index5 <- sample(nrow(maindf), size=nrow(maindf %>% filter(PrecipNew == 1)),replace =
FALSE)
  resultq[i] <- mean(maindf$dep_delay[index5], na.rm = TRUE) -mean(maindf$dep_delay[-
index5], na.rm = TRUE)
}
ggplot(data = tibble(resultq), mapping = aes(x = resultq)) +
  geom_histogram() +
  geom_vline(xintercept = observedq, color = "red")
2 * ((sum(resultq >= observedq) + 1) / (N + 1))
```

6.Visibility

```
maindf <- maindf %>%
  mutate(VisibNew = case_when(
    between(visib, 0 , 9.266209           ) ~ 0,
    TRUE ~ 1))


N<- 10^4-1
observedg <- mean(maindf$dep_delay[maindf$VisibNew == 0], na.rm =TRUE)-
mean(maindf$dep_delay[maindf$VisibNew == 1], na.rm = TRUE)
resultg <- numeric(N)
for (i in 1:N)
{
  index6 <- sample(nrow(maindf), size=nrow(maindf %>% filter(VisibNew == 1)),replace =
FALSE)
  resultg[i] <- mean(maindf$dep_delay[index6], na.rm = TRUE) -mean(maindf$dep_delay[-
index6], na.rm = TRUE)
}
ggplot(data = tibble(resultg), mapping = aes(x = resultg)) +
  geom_histogram() +
  geom_vline(xintercept = observedg, color = "red")
2 * ((sum(resultg >= observedg) + 1) / (N + 1))
```