Time of day analysis

Hamda Hassan

```
# Installing the libraries and loading packages
# install.packages("resampledata3")
# install.packages("tidyverse")
library(resampledata3)
```

```
##
## Attaching package: 'resampledata3'

## The following object is masked from 'package:datasets':
##
##     Titanic
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(nycflights13)
library(tibble)
```

Filtering for United Airlines (UA) Flights

```
ua_flights <- flights %>%
  filter(carrier == "UA" & !is.na(dep_delay) & !is.na(dep_time))
```

Creating a time of day category

```r
ua_flights <- ua_flights %>%
  mutate(dep_hour = dep_time %/% 100,
         time_of_day = case_when(
           dep_hour >= 0 & dep_hour < 12 ~ "Morning",       # 12 AM - 11:59 AM
           dep_hour >= 12 & dep_hour < 18 ~ "Afternoon",    # 12 PM - 5:59 PM
           dep_hour >= 18 & dep_hour <= 23 ~ "Night",       # 6 PM - 11:59 PM
           TRUE ~ NA_character_
         ))

ua_flights
```

```
## # A tibble: 57,979 x 21
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      554            558        -4      740            728
##  4  2013     1     1      558            600        -2      924            917
##  5  2013     1     1      558            600        -2      923            937
##  6  2013     1     1      559            600        -1      854            902
##  7  2013     1     1      607            607         0      858            915
##  8  2013     1     1      611            600        11      945            931
##  9  2013     1     1      623            627        -4      933            932
## 10  2013     1     1      628            630        -2     1016            947
## # i 57,969 more rows
## # i 13 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, dep_hour <dbl>,
## #   time_of_day <chr>
```

Counting the flights by the day time to see that everything is balanced

```r
ua_flights %>%
  count(time_of_day)
```

```
## # A tibble: 3 x 2
##   time_of_day     n
##   <chr>       <int>
## 1 Afternoon   21181
## 2 Morning     23940
## 3 Night       12858
```

Analyzing the average delays by their flight time (Morning, Afternoon, Night)

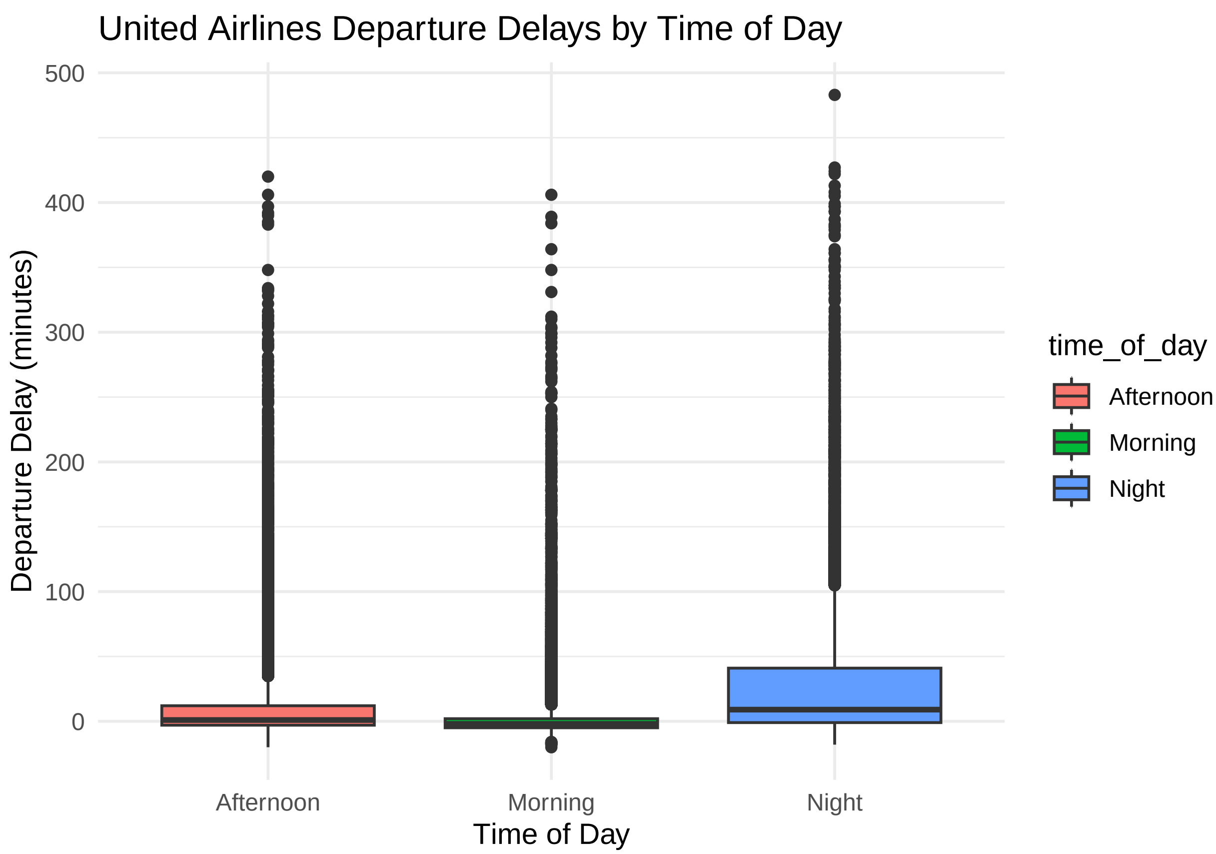
```
ua_flights %>%
  group_by(time_of_day) %>%
  summarize(
    mean_delay = mean(dep_delay, na.rm = TRUE),
    median_delay = median(dep_delay, na.rm = TRUE),
    sd_delay = sd(dep_delay, na.rm = TRUE),
    count = n()
  ) %>%
  arrange(time_of_day)
```

```
## # A tibble: 3 x 5
##   time_of_day mean_delay median_delay sd_delay count
##   <chr>            <dbl>        <dbl>    <dbl> <int>
## 1 Afternoon        10.8             1     30.3 21181
## 2 Morning           3.32           -2     22.0 23940
## 3 Night            30.6             9     53.4 12858
```

Visualizing the delays (Morning, Afternoon, Night)

To explore how the time of day affects departure delays, United Airlines (UA) flights were divided into three categories: morning (12 a.m.–12 p.m.), afternoon (12 p.m.–6 p.m.), and night (6 p.m.–12 a.m.). The boxplot shows that morning flights generally depart on time, with fewer and smaller delays compared to later flights. Afternoon flights show slightly longer and more variable delays, while night flights have the highest and most unpredictable delays, including several extreme cases. These results suggest that delays tend to build up as the day goes on, possibly due to scheduling bottlenecks and the ripple effects of earlier flight disruptions.

```
ggplot(ua_flights, aes(x = time_of_day, y = dep_delay, fill = time_of_day)) +
  geom_boxplot() +
  labs(title = "United Airlines Departure Delays by Time of Day",
       x = "Time of Day",
       y = "Departure Delay (minutes)") +
  theme_minimal()
```



Flight counts by Airport

```r
ua_flights %>%
  count(origin)
```

```
## # A tibble: 3 x 2
##   origin     n
##   <chr>  <int>
## 1 EWR    45652
## 2 JFK     4490
## 3 LGA     7837
```

Summarizing average delays by airport and time of day

When comparing average departure delays across New York City airports, the results show a consistent pattern: night flights experience the longest delays at every airport. At Newark (EWR), the mean delay rises sharply from about 4 minutes in the morning to nearly 29 minutes at night. A similar pattern appears at John F. Kennedy (JFK), where delays increase from less than 1 minute in the morning to around 30 minutes at night. LaGuardia (LGA) shows the most extreme difference, with average delays jumping from about 2 minutes in the morning to nearly 50 minutes at night. These results suggest that regardless of airport size or traffic volume, departure delays tend to accumulate as the day progresses, likely due to the compounding effects of earlier flight disruptions, heavier evening air traffic, and congestion in the national airspace system.

```
ua_airport_delay <- ua_flights %>%
  group_by(origin, time_of_day) %>%
  summarize(
    mean_delay = mean(dep_delay, na.rm = TRUE),
    median_delay = median(dep_delay, na.rm = TRUE),
    sd_delay = sd(dep_delay, na.rm = TRUE),
    flight_count = n()
  ) %>%
  arrange(origin, time_of_day)
```

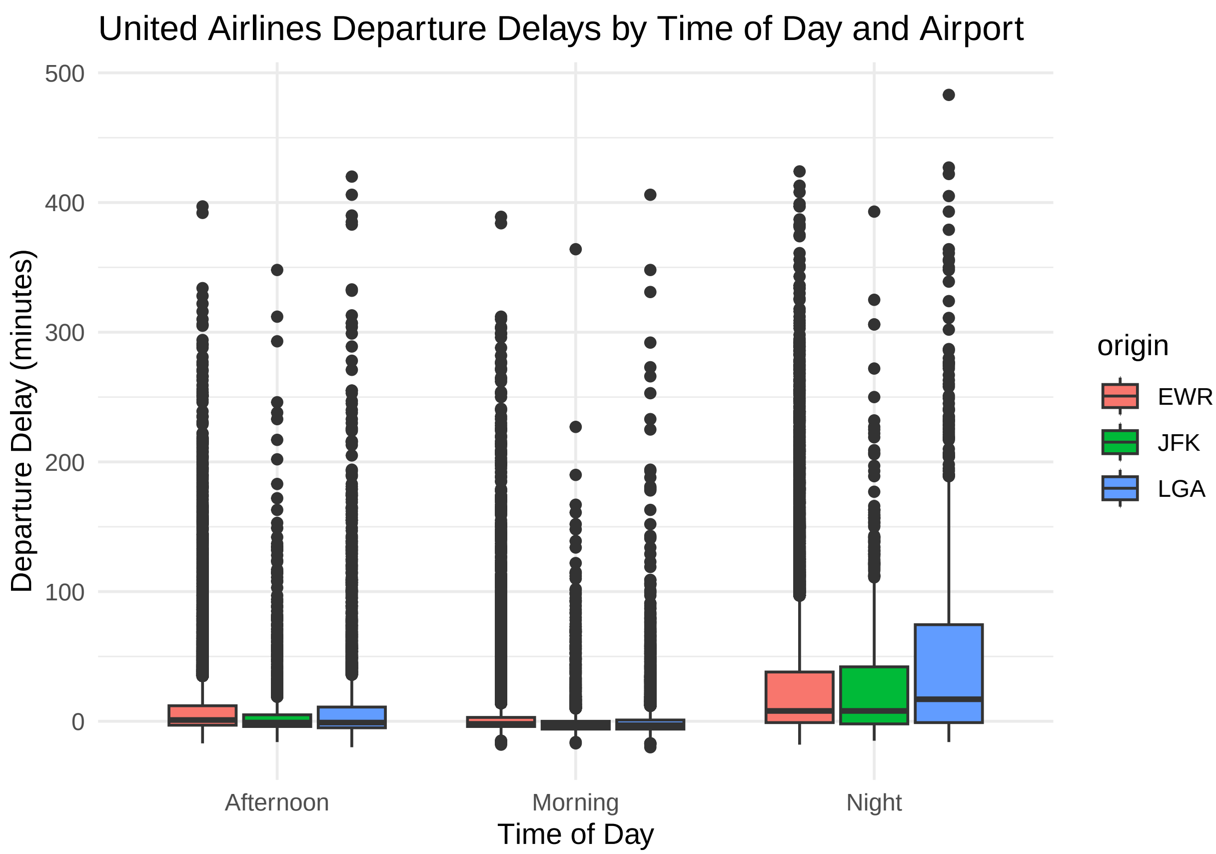
```
## `summarise()` has grouped output by 'origin'. You can override using the
## `.groups` argument.
```

```
ua_airport_delay
```

```
## # A tibble: 9 x 6
## # Groups:   origin [3]
##   origin time_of_day mean_delay median_delay sd_delay flight_count
##   <chr>  <chr>            <dbl>        <dbl>    <dbl>        <int>
## 1 EWR    Afternoon       10.8              1     28.1        17032
## 2 EWR    Morning          3.99            -2     22.3        17595
## 3 EWR    Night           28.8              8     50.6        11025
## 4 JFK    Afternoon        7.35            -1     30.5         1307
## 5 JFK    Morning          0.578           -3     18.5         2361
## 6 JFK    Night           29.8              8     52.1          822
## 7 LGA    Afternoon       12.8             -1     41.1         2842
## 8 LGA    Morning          2.03            -3     22.2         3984
## 9 LGA    Night           49.8             17     75.7         1011
```

Visualizing the delays by Airport

```
ggplot(ua_flights, aes(x = time_of_day, y = dep_delay, fill = origin)) +
  geom_boxplot() +
  labs(
    title = "United Airlines Departure Delays by Time of Day and Airport",
    x = "Time of Day",
    y = "Departure Delay (minutes)"
  ) +
  theme_minimal()
```



Permutation test: whether the mean departure delay differs between morning and night flights for United Airlines.

Only focusing on Morning and Night

```
ua_delay_subset <- ua_flights %>%
  filter(time_of_day %in% c("Morning", "Night")) %>%
  select(dep_delay, time_of_day) %>%
  drop_na()
```

Calculating the different in means

```
observed_diff <- ua_delay_subset %>%
  group_by(time_of_day) %>%
  summarize(mean_delay = mean(dep_delay)) %>%
  summarize(diff_means = diff(mean_delay)) %>%
  pull(diff_means)

observed_diff
```

```
## [1] 27.22953
```

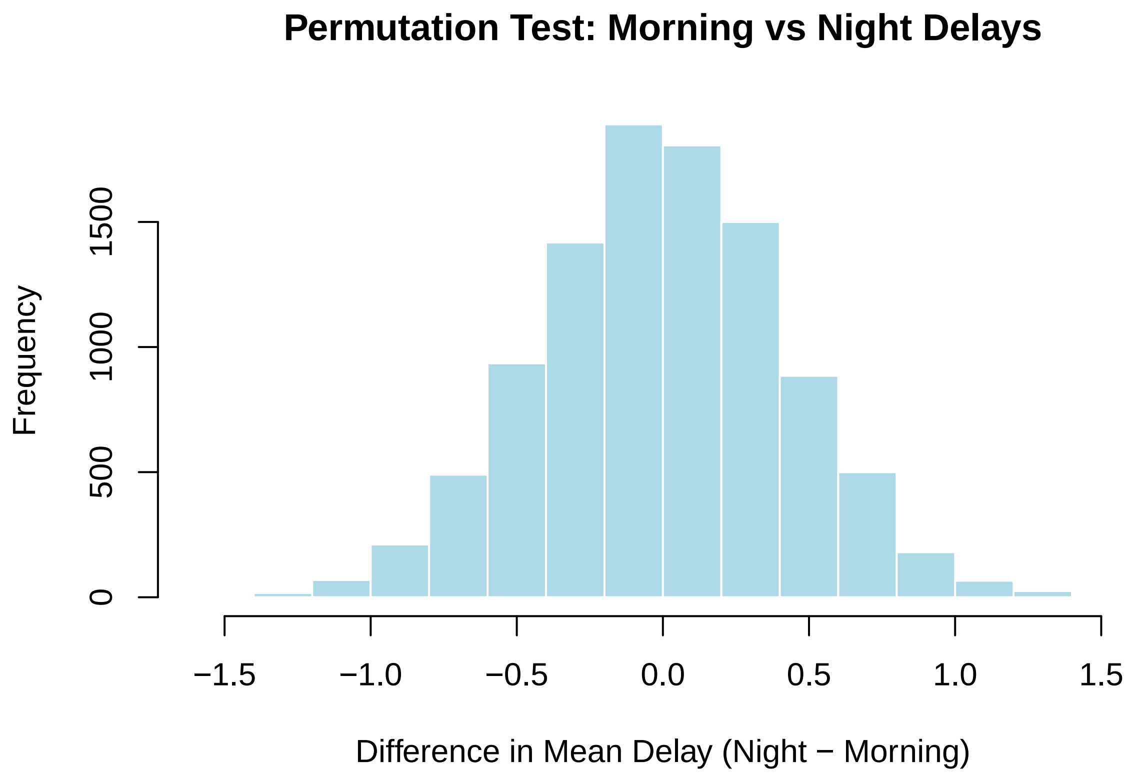
```r
set.seed(1)
N <- 10^4 - 1
perm_diffs <- numeric(N)

for (i in 1:N) {
  simulated_diff <- ua_delay_subset %>%
    mutate(time_of_day = sample(time_of_day)) %>%   # Randomly reassign Morning/Night labels
    group_by(time_of_day) %>%
    summarize(mean_delay = mean(dep_delay)) %>%
    summarize(diff_means = diff(mean_delay)) %>%
    pull(diff_means)

  perm_diffs[i] <- simulated_diff
}

p_value <- mean(perm_diffs >= observed_diff)
p_value
```

```
## [1] 0
```

Visualizing the permutation distribution

A permutation test was conducted to determine whether the difference in departure delays between morning and night United Airlines flights was statistically significant. The observed mean difference was approximately 27 minutes, with night flights departing later on average. The permutation distribution—representing what differences would be expected by chance if time of day had no effect—was centered near zero, while the observed value was far outside this range. Because none of the 9,999 simulated differences were as large as the observed one ($p < 0.001$), we reject the null hypothesis and conclude that night flights experience significantly greater departure delays than morning flights.

```r
hist(perm_diffs,
     main = "Permutation Test: Morning vs Night Delays",
     xlab = "Difference in Mean Delay (Night - Morning)",
     col = "lightblue", border = "white")
abline(v = observed_diff, col = "red", lwd = 2)
```



Conclusion

Overall, this analysis of United Airlines departure delays from the nycflights13 dataset reveals clear and consistent patterns related to both time of day and airport location. Across all New York City airports, morning flights experienced the shortest and most consistent departure times, while delays increased steadily throughout the afternoon and reached their peak at night. When comparing airports, Newark (EWR) had the highest overall number of flights and showed notable increases in delays during the evening hours. LaGuardia (LGA) and John F. Kennedy (JFK) displayed similar trends, though to a lesser extent.

To statistically confirm these observations, a permutation test comparing morning and night flights was performed. The observed mean difference in delays (approximately 27 minutes) was far greater than any of the differences generated under the null hypothesis, with a resulting p-value of less than 0.0001. This provides strong evidence that time of day has a significant effect on departure delays, particularly for flights departing at night.

Taken together, these results suggest that operational and scheduling factors likely contribute to the buildup of delays as the day progresses. Understanding these patterns could help airlines and airports optimize scheduling, allocate resources more efficiently, and reduce congestion-related delays in evening and nighttime flights.