Explore Bike Share Data

Versions & Library used this project

1. R version - 4.4.0
2. ggplot2 3.5.2 (visualization)
3. dplyr 2.5.0 (data manipulation)
4. lubridate 1.9.4 (dates & times manipulation)

- OS: Windows 11
- IDE: RStudio

Install Libraries

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.4.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Load Data and process to single data set

```
# Load all csv files in data frame (df)
df_chicago <- read.csv("chicago.csv")
df_newyork <- read.csv("new-york-city.csv")
df_washington <- read.csv("washington.csv")
```

```r
# Add column to df for city name to group by in charts
# we use function so we can use for each of the 3 files or more if they are added
process_city <- function(df, city) {
    df %>% mutate(
        City.Name = city
    )
}

# Load into new df with city name
df_chicago <- process_city(df_chicago, "Chicago")
df_newyork <- process_city(df_newyork, "New York City")
df_washington <- process_city(df_washington, "Washington")

# Make 1 dataset with all 3 cities so we can compare in same chart or filter this down
combined_df <- bind_rows(df_chicago, df_newyork, df_washington)
```

Question 1:

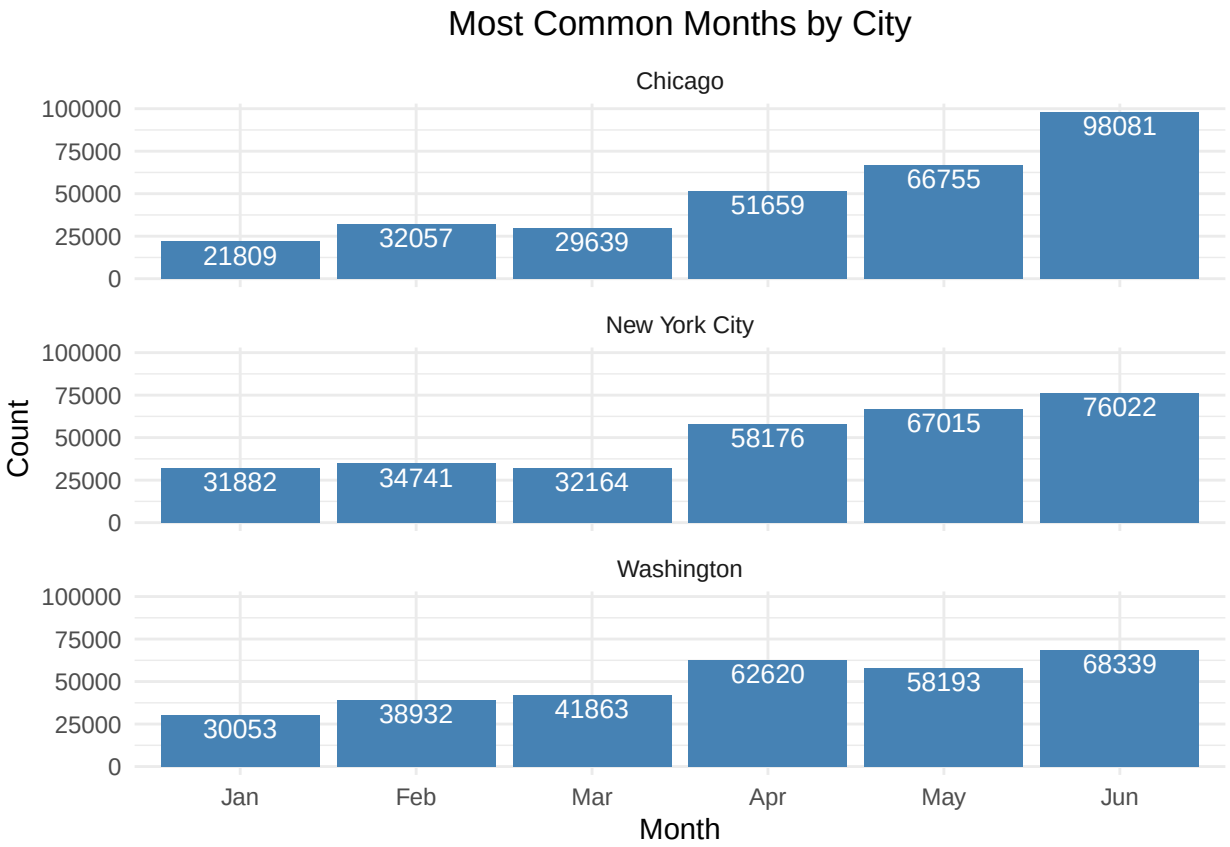
What is the most common month of usage across each of the cities

```r
# Objective: Display all cities and months into bar charts to compare against each other
# Find most common month fo a data frame
# Having function lets you use single city or multi-city data frame
find_most_common_month <- function(df) {
    df %>% mutate(
        month = month(Start.Time, label = TRUE, abbr = TRUE), # mutate to just month
        city = City.Name
    ) %>%
    count(city, month) # count city by month
}

# Make df for most common month so if I want to change plot I don't redue transformation
most_common_month_df <- find_most_common_month(combined_df)

# Plot all 3 cities in different bar charts for easier comparison among each other
# Doing this in case the months for each city is different

ggplot(most_common_month_df, aes(x = month, y = n)) +
    geom_col(fill = "steelblue") +
    geom_text( # add data label to each bar to show exact number
        aes(label = n),
        position = position_dodge(width = 0.8), # dodge overlaps
        vjust = 1.2, # Adjust label into bar
        color = "white", # Label color
        size = 3.5 # Size of label
    ) +
    facet_wrap(~ city, ncol = 1) + # Use facet_wrap to make bar chart for each city
    labs(title = "Most Common Months by City", x = "Month", y = "Count") +
    theme_minimal() + # Removes gray from chart and leaves white background
    theme(
        plot.title = element_text(hjust = 0.5) # Set title to middle
    )
```




```
most_common_month_df %>%
  group_by(city) %>%
  slice_max(n, n = 1, with_ties = FALSE)
```

```
## # A tibble: 3 x 3
## # Groups:   city [3]
##   city          month     n
##   <chr>         <ord> <int>
## 1 Chicago       Jun   98081
## 2 New York City Jun   76022
## 3 Washington    Jun   68339
```

Summary

I grouped the trips by city and by month, using a sum for total number of trips each month calculated. Using bar charts I showed 1 bar chart for each of the cities, with the x-axis being the month and y-axis showing the count per month. Across the 3 citites June had the highest usage for bikes, making it the most common month to take a bike out around town. Chicago with a total trip of 98,081, and had the largest increase from month to month usage being from May - June. New York City with a total trip count of 76,022 and Chicago had a drop from Feb - March then increasing over the following months with each to increase over the next month. Washington with a total ride count of 68,339, had a steady increase across the months until May, when they dipped from April and then climbing back up in June around 5000 more in June with May being less the April. With these cities all being in the northern part of the United States we can see that the usage is very seasonal with summer being a high usage since the weather is a lot nicer in the these locations than winter and spring seasons

Question 2

Find what these bikes are mainly used for, either for work commuters or fun to bike around. Let's assume they are mainly used for work, peak would be 7-9AM and 4-6PM the time average American commutes to work. I would like to find the number of rides each month that start during ON-PEAK times and compare the the OFF-PEAK usage.

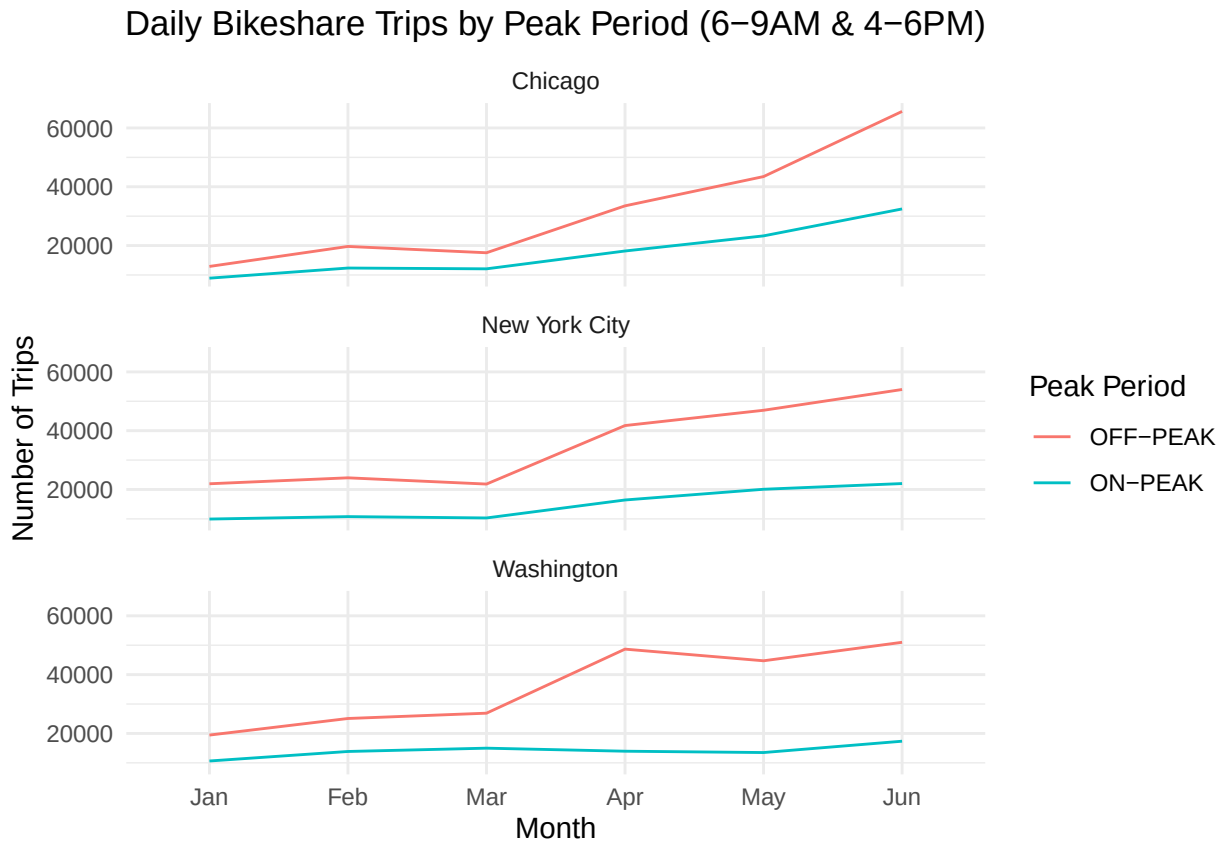
```
# Objective: Define the peak windows in a new column and set to ON-PEAK and OFF-PEAK depending on start
#            show a line graph for each city by month that has 2 lines for ON-PEAK and OFF-PEAK. Displa
#            separate graphs using facet_wrap to dig into each city. Doing this will not only show us e
#            if our bikes are possibly more used during these times we can assume the user is commuting

# Background of why I am curious about this. I work in the power industry and we control during peak ti
# we have found to be the peak usage of residents consuming power. Our peaks are 6-9AM and 5-8PM each d

combined_peak_df <- combined_df %>%
    mutate(
        start_hour = hour(Start.Time), # Get hour of start commute, do this once so we don't repeat in
        month = month(Start.Time, label = TRUE, abbr = TRUE), # mutate to just month
        peak_period = case_when(
            start_hour >= 7 & start_hour < 9   ~ "ON-PEAK", # Morning peak
            start_hour >= 16 & start_hour < 18 ~ "ON-PEAK", # Evening peak
            TRUE                               ~ "OFF-PEAK" # Off Peak
        )
    ) %>%
    count(City.Name, month, peak_period)


ggplot(combined_peak_df, aes(x = month, y = n, color = peak_period, group = peak_period)) +
    geom_line() +  # Line Chart
    facet_wrap(~ City.Name, ncol = 1) + # make chart for each city
    labs(
        title = "Daily Bikeshare Trips by Peak Period (6-9AM & 4-6PM)",
```

```
        x = "Month",
        y = "Number of Trips",
        color = "Peak Period"
    ) +
    theme_minimal() +
    theme (
        plot.title = element_text(hjust = 0.5) # Set title in middle of chart
    )
```




```
combined_peak_df %>%
  group_by(City.Name, month) %>%
  mutate(percent = n / sum(n))
```

```
## # A tibble: 36 x 5
## # Groups:   City.Name, month [18]
##    City.Name month peak_period     n percent
##    <chr>     <ord> <chr>       <int>   <dbl>
##  1 Chicago   Jan   OFF-PEAK    12899   0.591
##  2 Chicago   Jan   ON-PEAK      8910   0.409
##  3 Chicago   Feb   OFF-PEAK    19699   0.614
##  4 Chicago   Feb   ON-PEAK     12358   0.386
##  5 Chicago   Mar   OFF-PEAK    17551   0.592
##  6 Chicago   Mar   ON-PEAK     12088   0.408
##  7 Chicago   Apr   OFF-PEAK    33511   0.649
##  8 Chicago   Apr   ON-PEAK     18148   0.351
##  9 Chicago   May   OFF-PEAK    43464   0.651
```

```
## 10 Chicago   May   ON-PEAK     23291   0.349
## # i 26 more rows
```

Summary

To determine whether the bikes are used for commuting or leisure we defined what would be compute hours as ON-PEAK (7-9 AM & 4-6 PM) and OFF-PEAK periods which would occur outside the 4 hours. I then calculated the sum of usage for each city by month showing 2 lines for ON-PEAK and OFF-PEAK. Across all cities and months the OFF-PEAK is consitently higher the ON-PEAK. During the month of June the OFF-PEAK is about 2.0 times the amount of usage the ON-PEAK usage. In June the OFF-PEAK in Chicago is about 2.0x, New York City caw a increase about 2.45x and Washington with a 2.94x increase. Seeing this we can determine that the bikes are not dominated by individuals commuting to work. We could dive deeper into by grabbing the user home address where they start and where the end up each day. Then by checking if the user commutes multiple times a day and if the start and stop desinations match the other of previous trip.

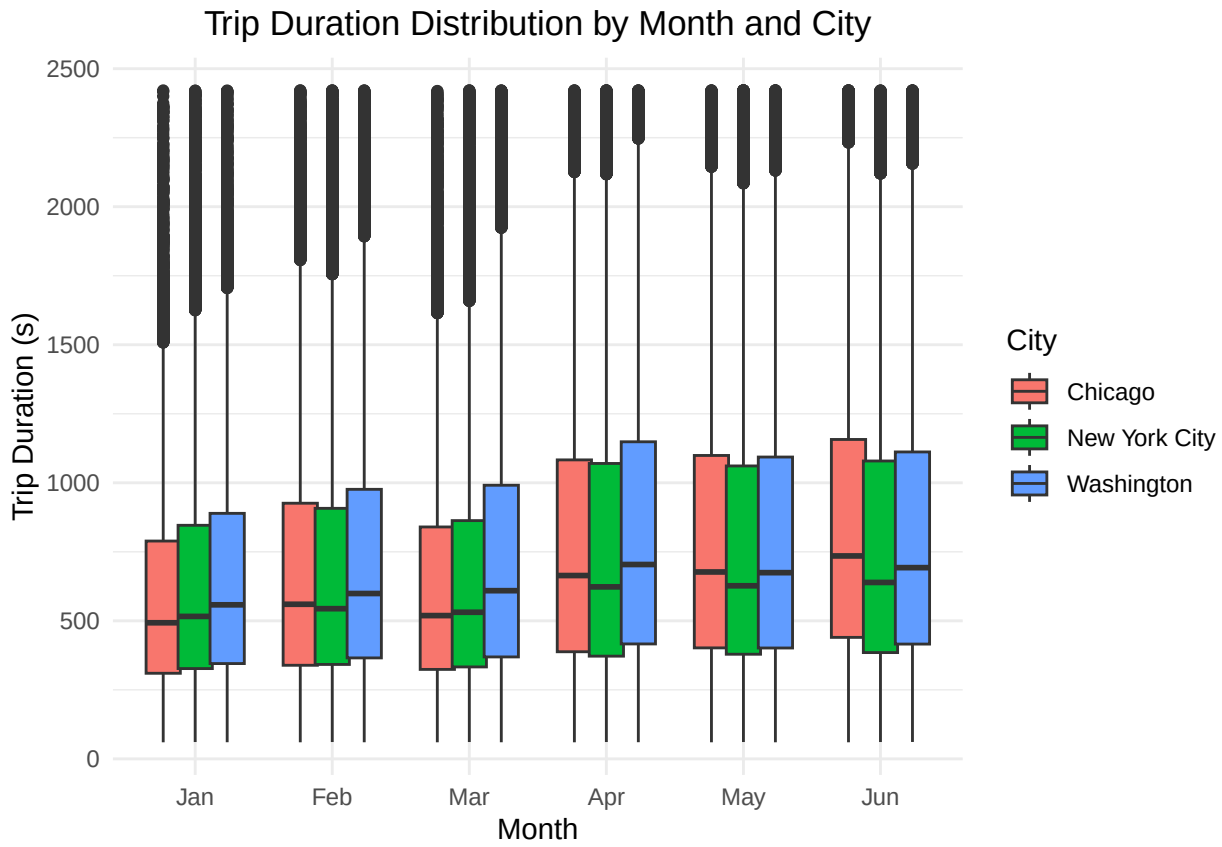
Question 3:

What is the minimum, median, and maximum ride duration for each of the city by month.

```r
# Objective: Create a chart that has 3 box-and-whisker for each month. Each box-and-whisker will repres
#            a city that the rides are accruing.

box_and_whisker_df <- combined_df %>%
    mutate(
        month = month(Start.Time, label = TRUE, abbr = TRUE),
        City.Name = City.Name
    ) %>%
    # I created it but each month as extreme outlier and I would like to get rid of it
    # trim anything above 95%
    filter(Trip.Duration <= quantile(Trip.Duration, 0.95, na.rm = TRUE))

ggplot(box_and_whisker_df, aes(x = month, y = Trip.Duration, fill = City.Name)) + # Color by city
    geom_boxplot(position = position_dodge(width = 0.7)) +  # Dodge for any overlap
    labs(
        title = "Trip Duration Distribution by Month and City",
        x = "Month",
        y = "Trip Duration (s)",
        fill = "City" # Define legend
    ) +
    theme_minimal() +
    theme(
        plot.title = element_text(hjust = 0.5) # Adjust chart title to middle of chart
    )
```




```
q95 <- quantile(combined_df$Trip.Duration, 0.95)

box_and_whisker_df %>%
  filter(Trip.Duration <= q95) %>%
  mutate(month = month(Start.Time, label = TRUE, abbr = TRUE)) %>%
  group_by(City.Name, month) %>%
  summarise(
    n = n(),
    min = min(Trip.Duration),
    q1 = quantile(Trip.Duration, 0.25),
    median = median(Trip.Duration),
    mean = mean(Trip.Duration),
    q3 = quantile(Trip.Duration, 0.75),
    max = max(Trip.Duration),
    .groups = "drop"
  )
```

```
## # A tibble: 18 x 9
##    City.Name     month     n   min    q1 median  mean    q3   max
##    <chr>         <ord> <int> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
##  1 Chicago       Jan   21559  60    310    493   599.  789  2421
##  2 Chicago       Feb   31106  60    339    560   687.  926  2421
##  3 Chicago       Mar   29158  60    324    519   634.  840  2419
##  4 Chicago       Apr   49105  60    388    664   782. 1083  2421
##  5 Chicago       May   63680  60    402    677   793. 1099  2421
##  6 Chicago       Jun   93041  60    440    735   839. 1157  2421
```

```
##  7 New York City Jan   31429  61    327    516   649.  846  2420
##  8 New York City Feb   34115  61    342    544   687.  908. 2421
##  9 New York City Mar   31716  61    333    531   664.  863  2421
## 10 New York City Apr   56227  61    372    623   775. 1070  2421
## 11 New York City May   65103  61    379    627   773. 1061  2421
## 12 New York City Jun   73645  61    385    639   785. 1079  2421
## 13 Washington    Jan   28770  60.2  345.   558.  671.  889. 2421.
## 14 Washington    Feb   36418  60.1  366.   599.  723.  976. 2421.
## 15 Washington    Mar   39016  60.1  369.   609.  734.  991. 2421.
## 16 Washington    Apr   55557  60.3  416.   704.  827. 1149. 2421.
## 17 Washington    May   53193  60.0  402.   674.  796. 1093. 2422.
## 18 Washington    Jun   62162  61.2  416.   693.  811. 1112. 2421.
```

Summary

I analyzed the distribution of trip duration for each city by month using a box-and-whisker plot, which shows the minimum, first quartile, median, third quartile, and maximum ride durations in seconds. After making one chart which showed major distortion due to a few rides I decided to only include the 95th percentile of Trip.Duration to improve the visualization rather than showing rare long trips. The median trip duration increased from winter to summer months across all 3 of the cities. After removing the 95th percentile we removed rare maximum ride distances and upper tail is more true max. We are able to see it is focused in the summer months with the winter months being more spread out for the maximum ride distance.