map500_F416670

Introduction to Data Science: Capital Bikeshare Coursework


```
# Packages required for analysis, including "geosphere" for the haversine formula
library("tidyverse")
library("here")
library("janitor")
library("geosphere")

# Remove scientific notation
options(scipen = 100, digits = 4)
```

Introduction

The following report will look at data that Capital Bikeshare, the publicly-owned bicycle sharing system in Washington DC, USA, have uploaded on the rides undertaken by their users. Analysis of this data will identify key trends and patterns around the company's users and their behavior, and will enable improved, data-led decision making in the future based on the findings.

Data Prep

The data provided by Capital Bikeshare outlines each individual bike journey in the years 2020-2021, and is broken down by starting date & time, ending date & time, user type, bike type, station details, and the coordinates of when the journeys started and ended.

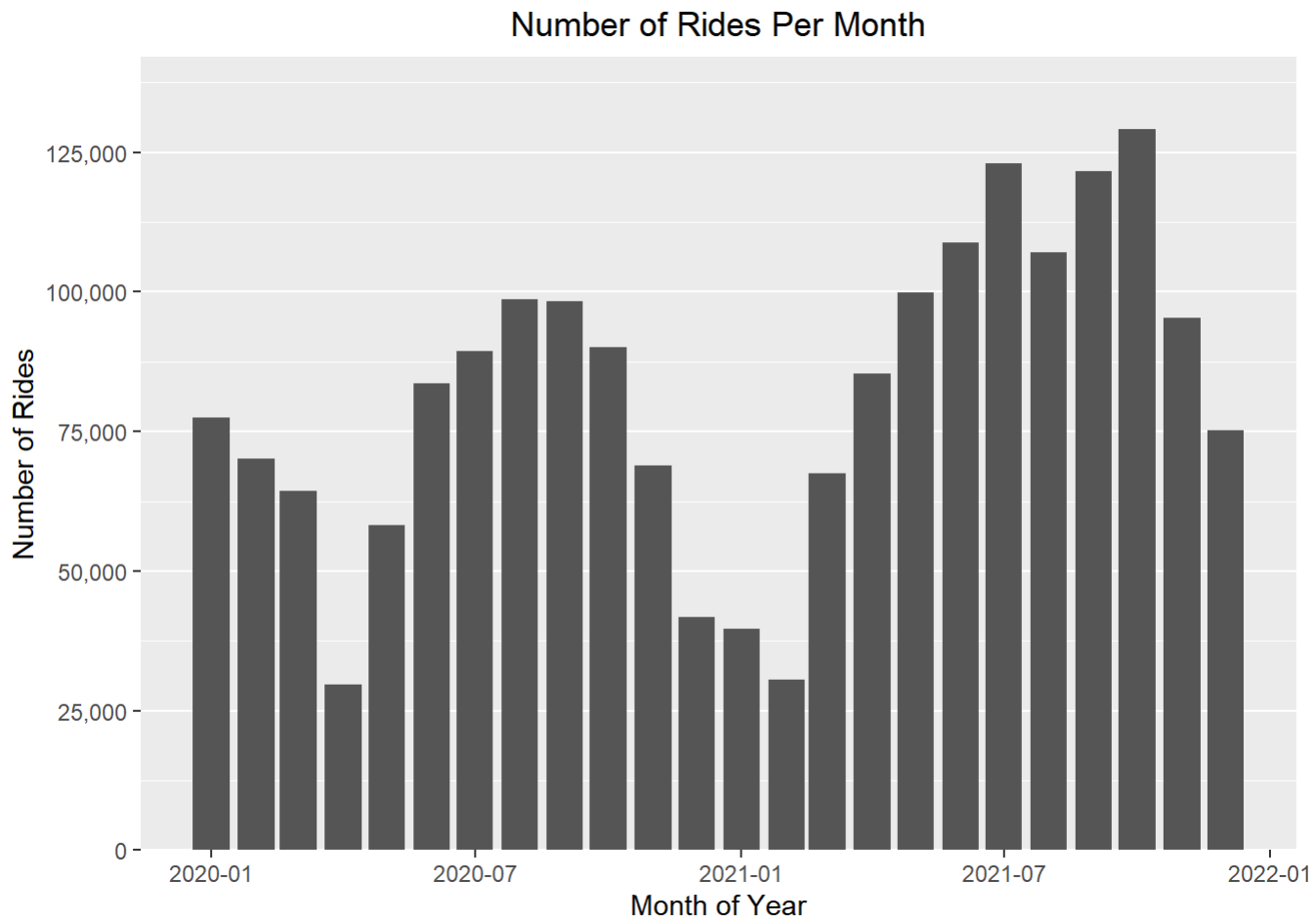
```r
# Import data from csv file
rides <- read_csv(here("data", "rides_2020_2021_extract.csv"))


# Get an initial preview of the dataset
glimpse(rides)
summary(rides)
head(rides)
tail(rides)


# Check for NAs in each column
# High amount of NAs in columns 'is_equity' and 'bike_number'
sapply(rides, function(x) sum(is.na(x)))


# Confirm if any of the rows are duplicated
# Found no duplicated rows
rides %>%
  group_by_all() %>%
  filter(n() > 1) %>%
  ungroup()


# Check is_equity column due to high NAs
# Data is only available for May 2020
rides %>%
  clean_names() %>%
  mutate(month = lubridate::floor_date(start_date, 'month')) %>%
  group_by(month, is_equity) %>%
  count() %>%
  arrange(month)



# Clean the data set and update the column names

# Removed bike_number column as 89% of the column is NA
# Removed is_equity column as 97% of the data is NA,
# Filter out rows where duration is negative, 0 or NA
rides_tidy <- rides %>%
  clean_names() %>%
  filter(
    !is.na(duration) &
    duration > 0
    ) %>%
  select(-c(is_equity, bike_number)) %>%
  mutate(duration = parse_double(duration)) %>%
  mutate(duration_mins = duration / 60) %>%
  mutate(duration_hours = duration / 3600) %>%
  mutate(month_year = floor_date(start_date, "month")) %>%
  mutate(member_casual = tolower(member_casual)) %>%
  rowwise() %>%
  mutate(distance_km = distHaversine(c(start_lng, start_lat), c(end_lng, end_lat))/ 1000) %>%
  mutate(speed_kmh = distance_km/duration_hours)
```

```
# Write cleaned data to a new csv for easy re-loading
write_csv(rides_tidy, here("data", "rides_tidy.csv"))
new_rides <- read_csv(here("data", "rides_tidy.csv"))
```

To clean the data, any columns with a high proportion of unusable data have been removed, as well as rows where the duration is 0 or less as these are likely to be errors in data. The duration of each journey in minutes and hours has been calculated, along with additional columns of distance and speed of the journey to support analyses.

As the dataset takes place during the COVID pandemic, an initial monthly view of the data is set out below just to identify any anomalies. Based on the graph, it shows there was a considerable drop in rides in April 2020, likely due to the stay-at-home order rolled out in Washington, DC (Custis, 2020). While this data will be kept in, it's worth highlighting should there be any further observations made on data by month either in this or in future reports. There was also a considerable drop in rides in January 2021 compared to the previous year, although the reason for this is less clear.

```
# Put the data into bar graph to identify any anomalies
new_rides %>%
  ggplot() +
  geom_bar(mapping = aes(x = month_year)) +
  scale_y_continuous(
    expand = expansion(mult = c(0, 0.1)),
    labels = scales::comma,
    breaks = c(0, 25000, 50000, 75000, 100000, 125000, 150000)
    ) +
  labs(
    title = "Number of Rides Per Month",
    x = "Month of Year",
    y = "Number of Rides") +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank()
    )
```



Questions

Question 1

Is there a discernible difference in peak travel times between casual users vs members?

To operationalise this question, this report will look at the frequency of bicycle journeys at each hour of the day in the last two years of data to see if there are any hours that have the highest frequency of travel, and compare the data between casual users and members to identify any differences in their patterns.

Question 2

Is there a difference in the length of the user's journeys when comparing type of bike and customer type?

To operationalise this question, this report will look at the mean and distribution of the duration of the bicycle journeys across the two years, and compare this between the types of bicycles used as well as whether there is any differences between members or casual users.

Question 3

What are the most popular journeys between the top stations?

To operationalise this question, this report will look at the frequency of rides between the top 10 start stations and the top 10 end stations to identify the most commonly taken journeys between these stations.

Analysis

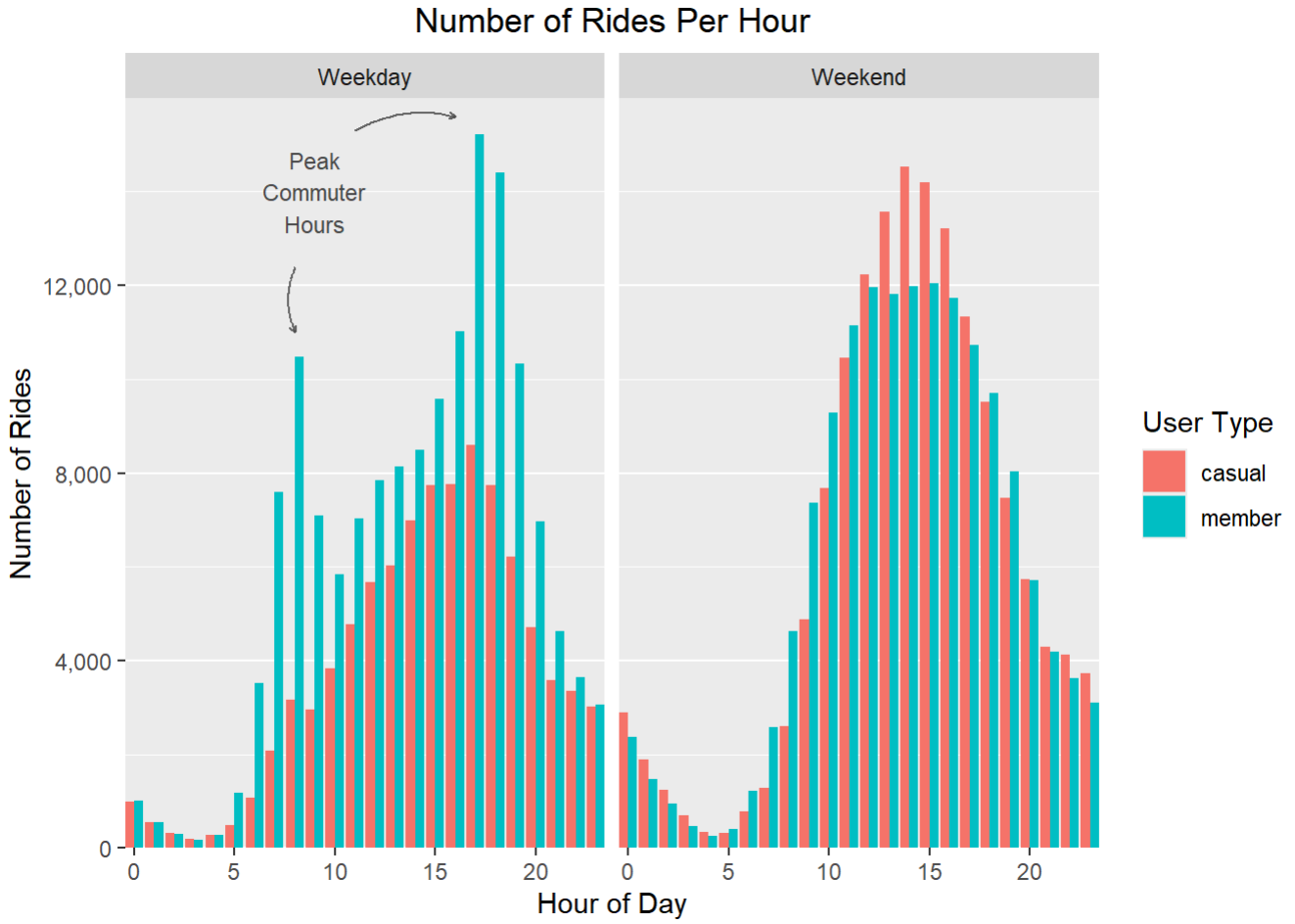
Analysis for Question 1

Is there a discernible difference in peak travel times between casual users vs members?

The figure below illustrates that the peak travel times for members on the weekdays are between 7-9am and 5-6pm, likely related to when members are traveling to and from work. Casual users, on the other hand, seem to follow a natural curve that peaks 5pm on weekdays with only a small morning uplift, potentially indicating that members are more likely to be commuters than casual users. On weekends however, both casual users and members follow a very similar curve, with casual users peaking higher than members in the early afternoon. It can be therefore concluded that the peak travel times are notably different by user type on weekdays, however on weekends the peak travel times are similar regardless of user type.

```
# Add in a column for time of day by hour
# Add a separate column to indicate if the day of the week is a weekday or weekend
# The data is split into a comparison between weekday and weekend to identify different patte
rns
# Only include journeys greater than 0

new_rides %>%
  filter(
    !is.na(distance_km) &
    distance_km > 0
  ) %>%
  mutate(day_of_week = wday(start_date, label = TRUE)) %>%
  mutate(week_day = ifelse(wday(start_date) %in% c(1, 7), "Weekend", "Weekday")) %>%
  mutate(time_of_day = hour(start_date)) %>%
  group_by(time_of_day, day_of_week, week_day, member_casual) %>%
  count() %>%
  ggplot(mapping = aes(x = time_of_day, y = n)) +
  geom_col(mapping = aes(fill = member_casual), position = position_dodge()) +
  facet_wrap(facets = vars(week_day)) +
  scale_y_continuous(
    expand = expansion(mult = c(0, 0.05)),
    labels = scales::comma
    ) +
  scale_x_continuous(expand = expansion(mult = c(0, 0))) +
  labs(
    title = "Number of Rides Per Hour",
    x = "Hour of Day",
    y = "Number of Rides",
    fill = "User Type"
    ) +
  theme(
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank(),
    plot.title = element_text(hjust = 0.5)
    ) +
  geom_text(
    x = 9,
    y = 14000,
    aes(label = label),
    data = data.frame(time_of_day = 6, n = 14, week_day = "Weekday", label = stringr::str_wra
p("Peak Commuter Hours", 5)),
    size = 3,
    colour = "grey28") +
  geom_curve(
    x = 11,
    y = 15300,
    xend = 16,
    yend = 15600,
    curvature = -0.2,
    arrow = arrow(length = unit(1, "mm")),
    alpha = 0.8,
    colour  = "grey28",
    data = data.frame(week_day = "Weekday")
  ) +
  geom_curve(
```

```
    x = 8,
    y = 12400,
    xend = 8,
    yend = 11000,
    curvature = 0.2,
    arrow = arrow(length = unit(1, "mm")),
    alpha = 0.8,
    colour  = "grey28",
    data = data.frame(week_day = "Weekday")
  )
```



Analysis for Question 2

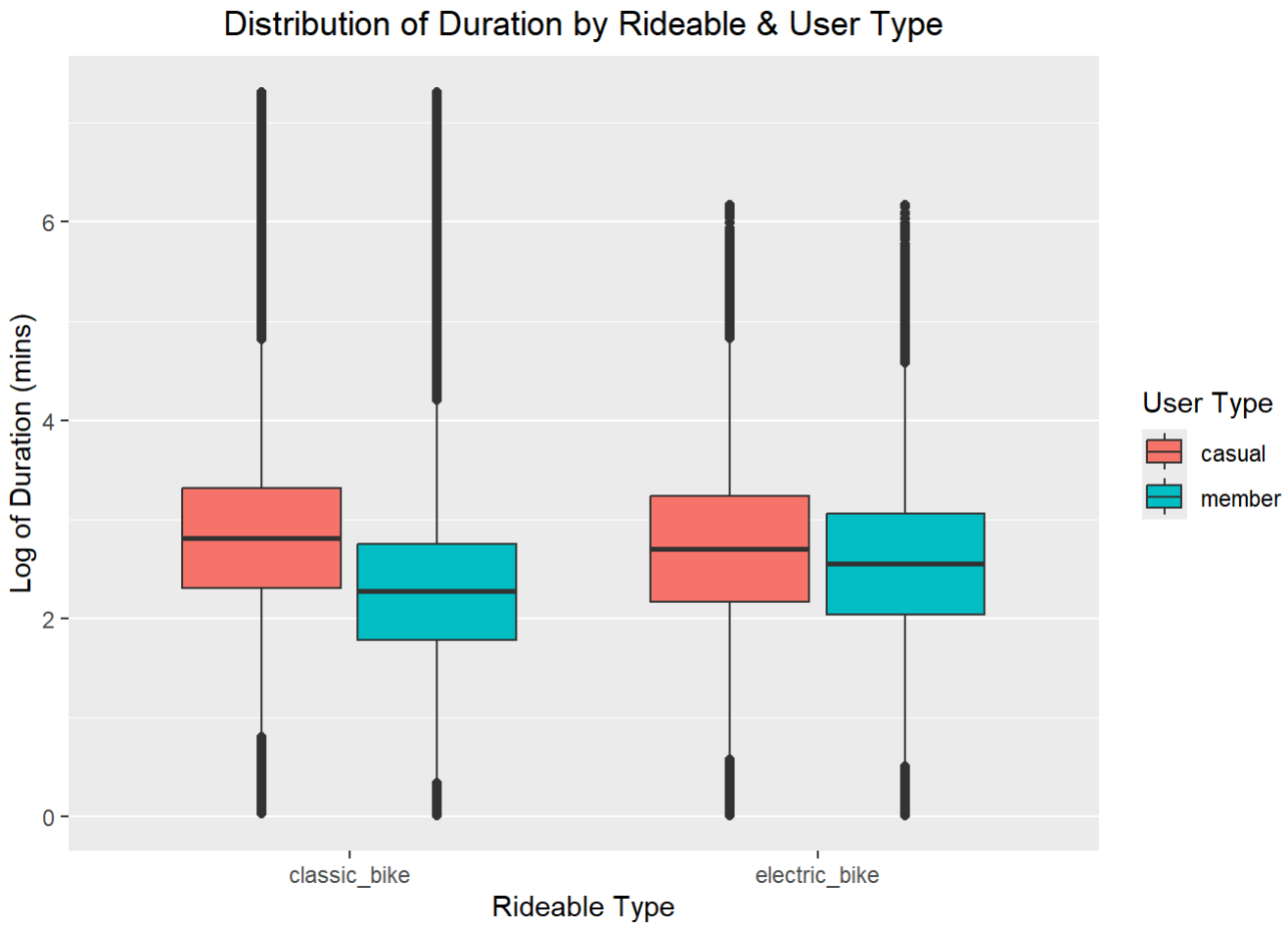
Is there a difference in the length of the user's journeys when comparing type of bike and customer type?

In the figure below, it can be seen that while the distribution of duration is greater with classic bikes than with electric bikes, their medians are relatively similar regardless of bicycle type, with a stronger difference coming from the customer type. For both types of bicycle, members' journeys have a shorter duration than casual users' journeys, regardless of bicycle type. The difference in duration between user types is wider with classic bikes than it is with electric bikes.

It can be concluded that the difference in the duration of the user's journey is minimal when looking at the type of bike used, however, when looking at user types it is clear that member's journeys are generally shorter on average than casual user's journeys.

It is currently unclear on what the rideable_type "docked_bike" encompasses even though it is quite a large proportion of the data. The Capital Bikeshare website does not provide enough information on this type to be usable for the analyses, so it has been removed.

```
# Remove docked_bike from the data
# Set duration_mins to greater than 1 to remove potential technical issues or "mistaken" ride
s
# Data from April 2020 onwards as rideable_type wasn't available before this month

new_rides %>%
  filter(!is.na(rideable_type) &
           !is.na(distance_km) &
           distance_km > 0 &
           duration_mins > 1 &
           !is.na(duration) &
           rideable_type != "docked_bike") %>%
  ggplot(mapping = aes(x = rideable_type, y = log(duration_mins))) +
  geom_boxplot(mapping = aes(fill = member_casual)) +
  labs(
    title = "Distribution of Duration by Rideable & User Type",
    x = "Rideable Type",
    y = "Log of Duration (mins)",
    fill = "User Type"
    ) +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.grid.major.x = element_blank()
    )
```



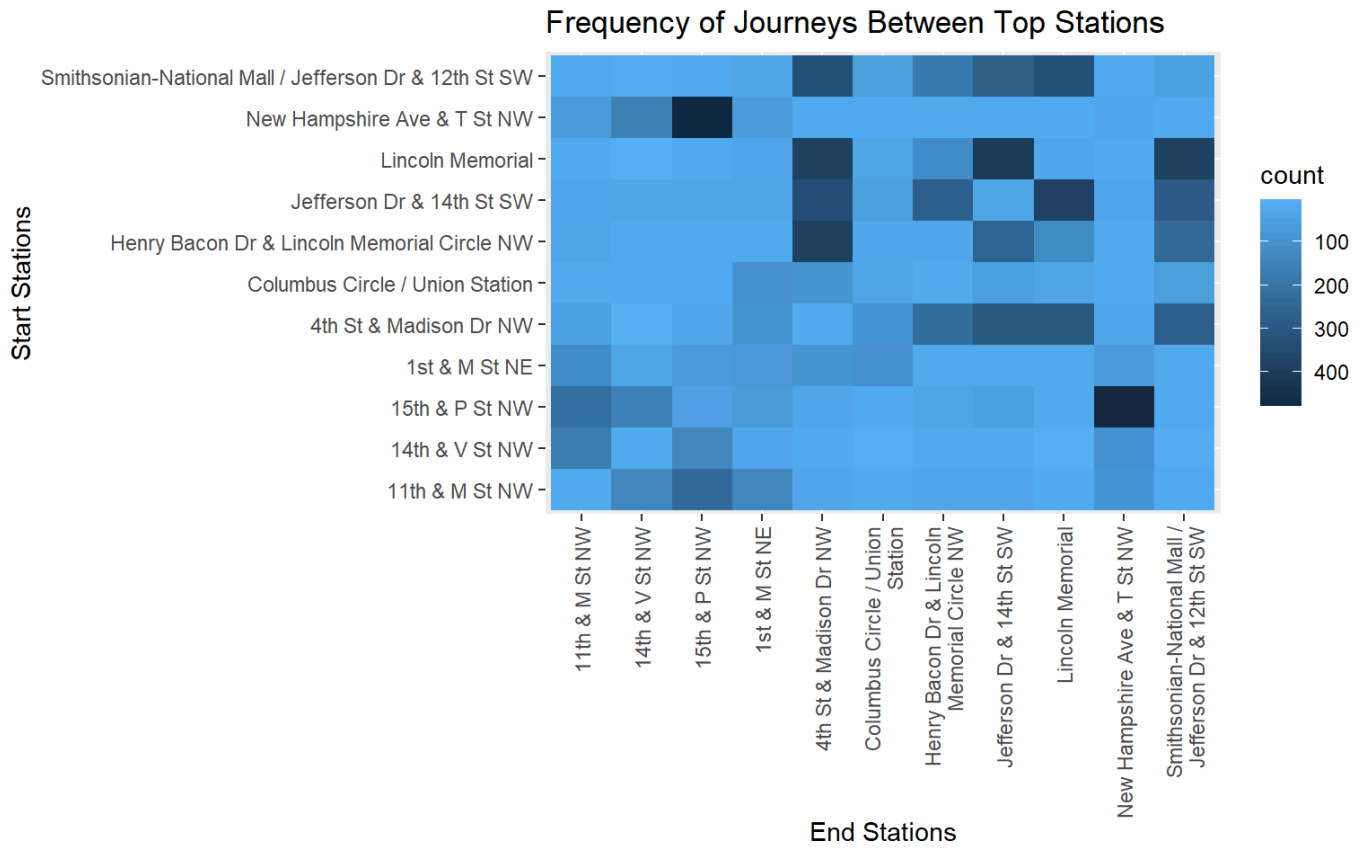
Analysis for Question 3

What are the most popular journeys between the top stations?

In the figure below, the popular stations for both starting and ending journeys seem to overlap considerably, and the journeys with the highest frequencies from the top starting stations are similarly frequent when the route is reversed, with these potentially being popular commuting or tourist location.

The top journeys between stations are identifiable within the heatmap below and provide an indicator of the most popular destinations and starting points for the company's users, and could support further analyses in a later report regarding whether there are enough bicycles and docking stations available to support the popularity of those stations and journeys. In particular, New Hampshire Ave & T St NW to 15th & P St NW (and vice versa) seems to be a particularly frequently undertaken route.

```
 # Identify the top 10 start stations
top_10_start <- new_rides %>%
  filter(!is.na(start_station_name)) %>%
  group_by(start_station_name) %>%
  count() %>%
  arrange(desc(n)) %>%
  head(n = 10) %>%
  rename(station = start_station_name)


# Identify the top 10 end stations
top_10_end <- new_rides %>%
  filter(!is.na(end_station_name)) %>%
  group_by(end_station_name) %>%
  count() %>%
  arrange(desc(n)) %>%
  head(n = 10) %>%
  rename(station = end_station_name)


# Merge the data together for top stations - results in 11 distinct stations
top_stations <- rbind(top_10_start, top_10_end)


# Remove NAs from data and set duration to higher than 1 min to remove mistaken rides
# Remove docked bikes due to limited information
new_rides %>%
  filter(!is.na(distance_km) &
           !is.na(duration_mins) &
           !is.na(rideable_type) &
           duration_mins > 1 &
           distance_km > 0 &
           rideable_type != "docked_bike" &
           start_station_name %in% top_stations$station &
           end_station_name %in% top_stations$station) %>%
  ggplot(mapping = aes(x = start_station_name, y = end_station_name)) +
  geom_bin2d() +
  scale_fill_continuous(trans = "reverse") +
  scale_x_discrete(labels = scales::label_wrap(28)) +
  labs(
      title = "Frequency of Journeys Between Top Stations",
      x = "End Stations",
      y = "Start Stations"
    ) +
  theme(
    axis.text.x  = element_text(
                     angle = 90,
                     hjust = 1,
                     vjust = 0.5
                   )
    )
```



Conclusion

The differences between members and casual users are notable both in time of day and with regards to the duration of the bike rides themselves, with members peak times during commuting hours during the weekdays, although weekends they follow a similar curve to casual users' rides. Unlike with users, rideable type seems to have little impact on the duration of the rides. The top starting and ending stations have a high frequency of users who travel back and forth between each station at a similar frequency.

For further study, we can dig further into user type trends beyond time of day/week and look into how they differ by season or month to identify additional factors that might impact customer behavior. There could also be more hypothesis testing to check any significance in differences between user types, as well as the impact the Oct 2021 increase in free cycling time had on Capital Bikeshare membership.

References

[1](Custis, A. (2020, April 22))

---

1. Custis, A. (2020, April 22) D.C. Policy Center. *A timeline of the D.C. region's COVID-19 pandemic* https://www.dcpolicycenter.org/publications/covid-19-timeline/ (https://www.dcpolicycenter.org/publications/covid-19-timeline/)(Accessed: 4 November 2024)↩