Analysis of Capital Bikeshare Data

Investigating bike riding data from 2020 and 2021


```
library("here") # for relative instead of absolute paths to, e.g., data files
library("tidyverse") # includes ggplot2, dplyr, readr, tidyr, tibble and more
library("lubridate") # for dealing with date and time objects
library("ggrepel") # adds improved labeling compared to standard ggplot2
library("janitor") # provides tools for cleaning data

knitr::opts_chunk$set( # This sets chunk options globally.
  echo = TRUE,
  warning = FALSE,
  message = FALSE,
  fig.show = "hold"
)
```

Introduction

This report analyses a dataset containing information about bike rides for Capital Bikeshare, a publicly-owned, American company that operates a bike rental service around Washington DC.

Understanding the patterns of how users interact with this service is important for improving the operational efficiency and customer satisfaction. Increased operational efficiency may lead to higher profit margins, allowing the company to expand its operations. Biking is a sustainable form of transportation, so expanding the company allows sustainable transportation in more areas.

Data

The data set is sourced directly from Capital Bikeshare (CaBi) who publish downloadable files of their bike trip data regularly.

The original data set contains 16 variables, however, some of these variables contain no relevant information and others can be combined together in the cleaning of the data.

We begin the cleaning by excluding the variables `bike_number` and `ride_id` because there is no relevant information contained in either of these columns that can be used for data analysis. The four columns consisting of latitude and longitude values can be combined into one column via the Haversine formula which finds the shortest distance between the two sets of coordinates.

The Haversine formula is $d = 2r \arcsin\left(\sqrt{\frac{1-\cos(\Delta\phi)+\cos(\phi_1)*\cos(\phi_2)*(1-\cos(\Delta\lambda))}{2}}\right)$ where ϕ_1 and ϕ_2 are the starting and ending latitude and λ_1 and λ_2 are the starting and ending longitude. Additionally, $\Delta\phi$ and $\Delta\lambda$ refer to the absolute value of the difference between the starting and ending latitude and longitude. Finally, r refers to the radius of Earth which is 6378137 metres. The units of the distance, d, is the same as the units of r, therefore, in our case, distance is measured in metres.

We also coerce columns into the correct data types using `parse_*` functions. The `member_casual` and `rideable_type` columns are changed to factors whilst `duration`, `start_station_id` and `end_station_id` become integers.

Finally, after using the `summary()` function, we observe that the lowest duration value is negative. The website states that all rides under 60 seconds have been removed from the data set; this evidently has not happened so we exclude any values which are below 60.

This leaves us with 11 variables:

- `duration` - Duration of the ride in seconds
- `start_date` - Starting date and time of the ride
- `end_date` - Ending date and time of the ride
- `start_station_id` - Integer corresponding to the starting station of the ride
- `start_station_name` - Name of the starting station of the ride
- `end_station_id` - Integer corresponding to the end station of the ride
- `end_station_name` - Name of the end station of the ride
- `member_casual` - Whether the customer is a member (annual member, 30-day member or day key member) or a casual user (single trip user or a 24-hour, 3-day or 5-day pass user)
- `rideable_type` - Type of bike used
- `is_equity` - Whether a user has a discount by qualifying for a state or federal assistance program or they are a key worker as part of a scheme during the Covid-19 pandemic where key workers were granted 30 days of free membership as stated in this blog post (https://web.archive.org/web/20200522190305/https://www.capitalbikeshare.com/blog/covid19) from 2020. Data for this column is only available for May of 2020.
- `distance` - Shortest distance between the start and end point in metres

```
clean_rides_data <- read_csv(here("data", "rides_2020_2021_extract.csv")) %>%
  clean_names() %>% # converts column names to snake case
  drop_na(start_lat) %>% # Removing NAs
  drop_na(start_lng) %>%
  drop_na(end_lat) %>%
  drop_na(end_lng) %>%
  drop_na(duration) %>%
  select(!ride_id & !bike_number) %>% # No meaningful data in these columns.
  mutate(start_lat = start_lat * pi / 180) %>% # Converting coordinates from
  mutate(start_lng = start_lng * pi / 180) %>% # degrees into radians which is
  mutate(end_lat = end_lat * pi / 180) %>% # necessary in order to use trig
  mutate(end_lng = end_lng * pi / 180) %>% # functions in r.
  mutate(
    distance = # Haversine formula for calculating distance
      2 * 6378137 * asin(sqrt( # 6378137 is radius of earth (metres).
        (1 - cos(abs(start_lat - end_lat)) +
            cos(start_lat) *
              cos(end_lat) *
              (1 - cos(abs(start_lng - end_lng)))
        ) / 2
      ))
  ) %>%
  select(!start_lat & !start_lng & !end_lat & !end_lng) %>%
  mutate(member_casual = parse_factor( # Correcting data types
    member_casual,
    levels = c("member", "casual")
  )) %>%
  mutate(rideable_type = parse_factor(
    rideable_type,
    levels = c("classic_bike", "docked_bike", "electric_bike")
  )) %>%
  mutate(start_station_id = as.integer(start_station_id)) %>%
  mutate(end_station_id = as.integer(end_station_id)) %>%
  mutate(duration = parse_integer(duration)) %>%
  filter(duration > 60) # Website says all trips < 60 seconds are removed
```

Questions

This report will investigate the following three questions.

Question 1

What are the most popular times of day for bike usage, and how does this vary by which day of the week it is and the type of bike used?

In order to operationalise this question we begin by grouping the starting time of rides by each hour of the day and considering the total number of rides that started within each of these hours. We also split each day of the week into its own graph using faceting to allow comparisons to be made between different days of the week. Both hour of the day and day of the week can be extracted from the `start_date` column.

Question 2

How does ride duration vary between members and casual users, and how was it affected by the 2021 policy change that extended the no-additional-charge ride time from 30 to 45 minutes for members and day pass holders?

To operationalise this question, we begin by finding when exactly this policy change occurred. Using old snapshots of the Capital Bikeshare website from September 2021 (https://web.archive.org/web/20210924063308/https://www.capitalbikeshare.com/) and October 2021 (https://web.archive.org/web/20211004104659/https://www.capitalbikeshare.com/), we find that the policy change only came into affect on the 1st October 2021. We will therefore only consider 2021 data where the start date is post 1st October. In order to eliminate the chance seasonal variability causing bias, we also only consider 2020 data post 1st October. This leaves us with two time periods to compare; October 1st - December 31st 2020 and October 1st - December 31st 2021. We can compare the duration of rides between both members and non-members for both of these periods. We will convert the duration from seconds to minutes for better clarity and ease of understanding and only consider rides lasting less than 2 hours, as longer rides are likely due to users forgetting to return the bike, or using it for multiple trips within a single rental.

Question 3

How do the distance and duration of bike rides change over time for both equity and non-equity members, and what patterns or trends are there between these two groups?

To operationalise this question, we first observe that, as mentioned earlier, the `is_equity` column only has values from the month of May 2020 so our analysis will only consider this period. We will compare general trend of how the distance per trip changes throughout the month of May between equity and non-equity members. We are therefore excluding casual users from this analysis.

Analysis

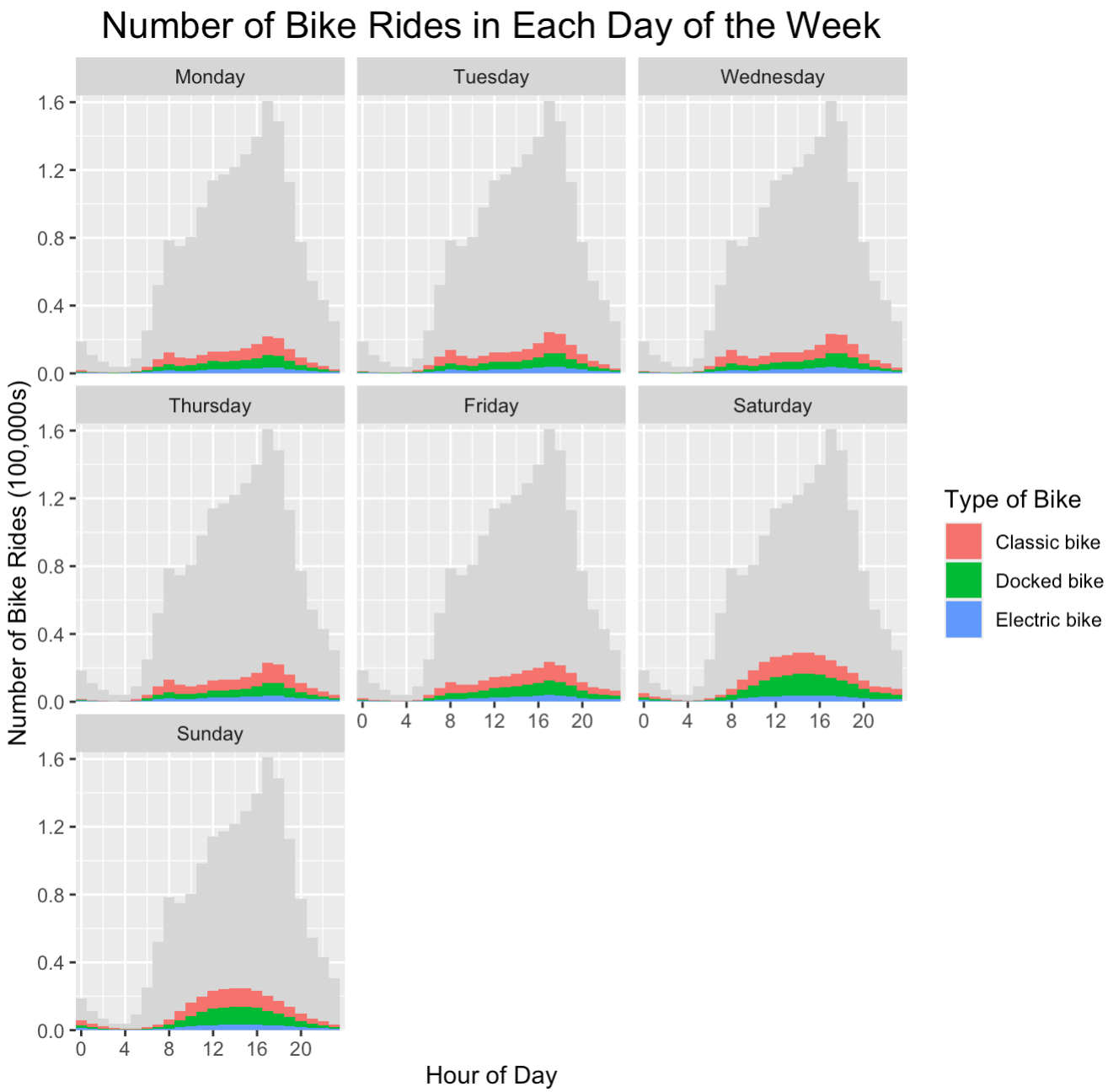
Answer to Question 1

The seven figures produced here clearly demonstrate that the most popular times of day for bike usage is very similar for each of the 5 weekdays. The peak number of bike rides occurs at 5 pm (hour 17) whilst the morning peak is 8 am for each of the weekdays. There is a lower ride count during the middle of each day and late evening. The general shape for each graph of the weekdays is the same; a steady increase from 6 am until 8 am, followed by a small decrease until 10 am and then a steady increase up until 5 pm. From 5pm until midnight, there is a steady decline. Friday has a slightly larger number of rides through the evening and up to midnight relative to each of the other weekdays.

Meanwhile, the weekend days have a very different shape and trend compared to the weekdays, but, both Saturday and Sunday are very similar. Rides are more evenly distributed throughout the day with a smaller morning peak. The peak occurs at 2 and 3 pm for both Saturday and Sunday. Prior to the peak, there is a gradual increase starting at 6 am. After the peak, there is a gradual decrease but there is still a significant number of rides starting at midnight and even through till 2 am Saturday and Sunday mornings.

Classic bikes are the most popular choice, though docked bikes are also commonly used. Electric bikes have significantly lower popularity, perhaps because they cost notably more than the classic/docked bikes. The proportion of each bike type used remains relatively consistent throughout each hour and each day with the electric bike significantly less used than the classic and docked bikes.

```
df1 <- clean_rides_data %>%
  mutate(hour_of_day = parse_integer(format(start_date, "%H"))) %>%
  mutate(day_of_week = factor(
    weekdays(start_date),
    levels = c(
      "Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
      "Saturday", "Sunday"
    )
  )) # Ordering the days of the week
df1 %>%
  ggplot(
    mapping = aes(
      x = hour_of_day,
      y = after_stat(count) / 10^5, # count in 100,000s
      fill = rideable_type
    ),
    bins = 24
  ) +
  geom_histogram(
    data = select(df1, !day_of_week), # Full histogram in background
    fill = "grey85",
    bins = 24
  ) +
  geom_histogram(bins = 24) +
  facet_wrap(facets = vars(day_of_week)) +
  labs(
    title = "Number of Bike Rides in Each Day of the Week",
    x = "Hour of Day",
    y = "Number of Bike Rides (100,000s)",
    fill = "Type of Bike",
    caption =
      "The grey histogram shows the overall number of bike rides per hour."
  ) +
  scale_x_continuous(
    breaks = c(0, 4, 8, 12, 16, 20, 24),
    expand = expansion(mult = c(0, 0.02))
  ) +
  scale_fill_discrete(
    labels = c(
      "classic_bike" = "Classic bike",
      "electric_bike" = "Electric bike",
      "docked_bike" = "Docked bike"
    )
  ) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.02))) +
  theme(plot.title = element_text(
    hjust = 0.5, # centers the title
    size  = rel(1.5) # changes the title font size
  ))
```



The grey histogram shows the overall number of bike rides per hour.

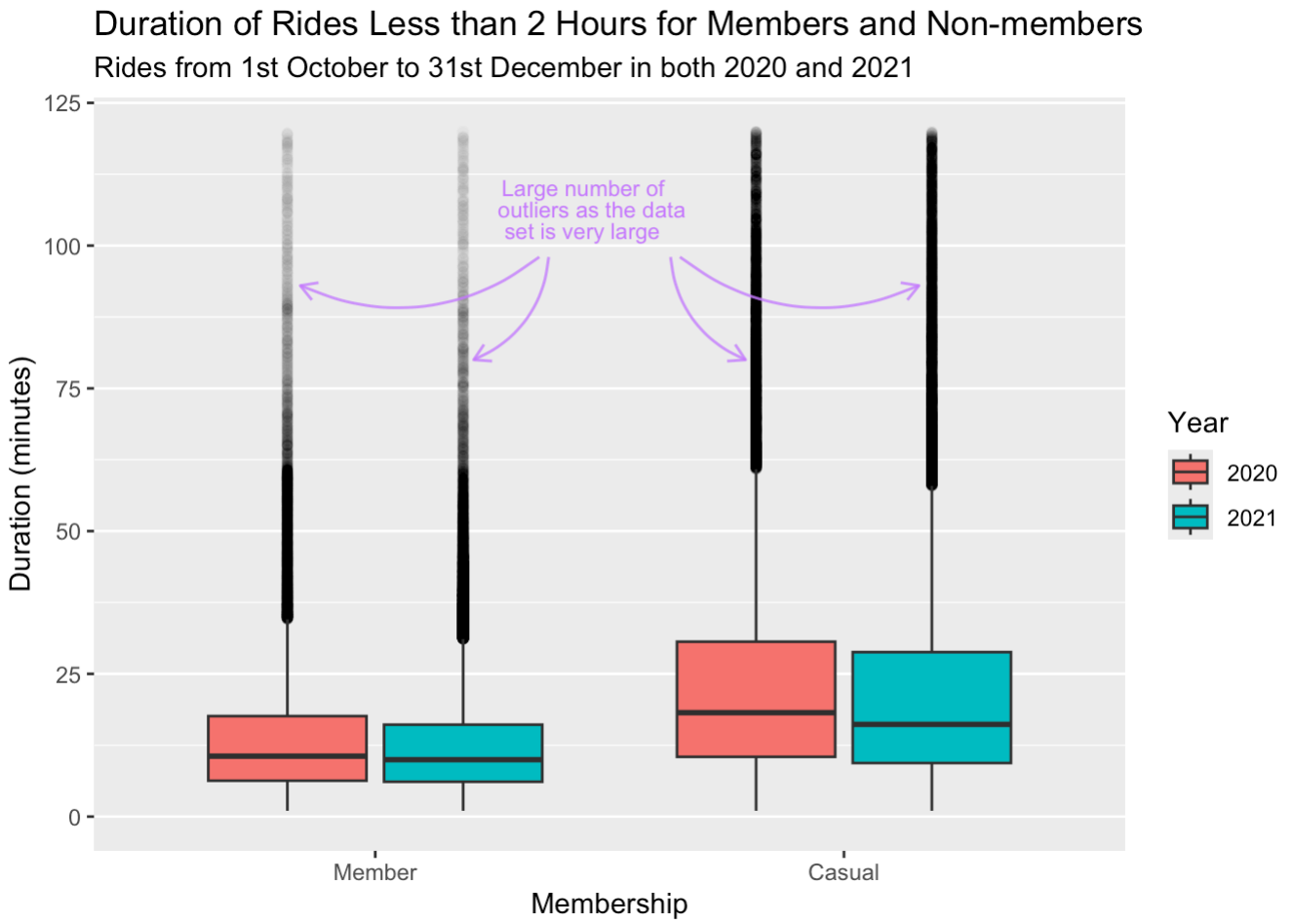
Answer to Question 2

The boxplot clearly demonstrates the average duration of ride is significantly higher for casual users than members as the median, lower quartile and upper quartile are all noticeably higher for casual users in both 2020 and 2021. Casual members also have a higher spread of durations evidenced by the length of the boxes (interquartile range).

As stated in the question, the no-additional-charge ride time increased from 30 to 45 minutes in 2021, so we might expect the average duration of ride times to increase from 2020 to 2021. However, this is not the case. For both members and casual users, the median, lower quartile and upper quartile duration of ride times has slightly decreased from 2020 to 2021.

```
df2 <- clean_rides_data %>%
  mutate(year = parse_factor(format(start_date, "%Y"),
    levels = c("2020", "2021")
  )) %>%
  filter(
    start_date < ymd_hms("2021-01-01 00:00:00") | # Filtering the dates to
      start_date > ymd_hms("2021-10-01 00:00:00")
  ) %>% # only include the two
  filter(start_date > ymd_hms("2020-10-01 00:00:00")) %>% # periods mentioned.
  filter(duration < 120 * 60) %>% # Filters to rides under 2 hours only
  ggplot() +
  geom_boxplot(
    mapping = aes(
      x = member_casual,
      y = duration / 60, # Duration in minutes
      fill = year
    ),
    outlier.alpha = 0.01
  ) +
  labs(
    title = "Duration of Rides Less than 2 Hours for Members and Non-members",
    subtitle =
      "Rides from 1st October to 31st December in both 2020 and 2021",
    x = "Membership",
    y = "Duration (minutes)",
    fill = "Year"
  ) +
  scale_y_continuous(limits = c(0, 120)) +
  scale_x_discrete(labels = c("member" = "Member", "casual" = "Casual")) +
  theme(panel.grid.major.x = element_blank()) + # Removing vertical lines
  annotate(
    geom = "text",
    x = 1.3,
    y = 100,
    label = "
    Large number of \n
    outliers as the data\n
    set is very large",
    hjust = 0.2, vjust = -0.1,
    lineheight = 0.4,
    colour = "#C77CFF",
    size = 3
  ) +
  annotate(
    geom      = "curve",
    x         = 1.35,
    y         = 98,
    xend      = 0.84,
    yend      = 93,
    curvature = -0.3,
    arrow     = arrow(length = unit(2.5, "mm")),
    alpha     = 0.8,
    colour    = "#C77CFF"
  ) +
  annotate(
```

```
    geom      = "curve",
    x         = 1.37,
    y         = 98,
    xend      = 1.21,
    yend      = 80,
    curvature = -0.3,
    arrow     = arrow(length = unit(2.5, "mm")),
    alpha     = 0.8,
    colour    = "#C77CFF"
  ) +
  annotate(
    geom      = "curve",
    x         = 1.65,
    y         = 98,
    xend      = 2.16,
    yend      = 93,
    curvature = 0.3,
    arrow     = arrow(length = unit(2.5, "mm")),
    alpha     = 0.8,
    colour    = "#C77CFF"
  ) +
  annotate(
    geom      = "curve",
    x         = 1.63,
    y         = 98,
    xend      = 1.79,
    yend      = 80,
    curvature = 0.3,
    arrow     = arrow(length = unit(2.5, "mm")),
    alpha     = 0.8,
    colour    = "#C77CFF"
  )
df2
```

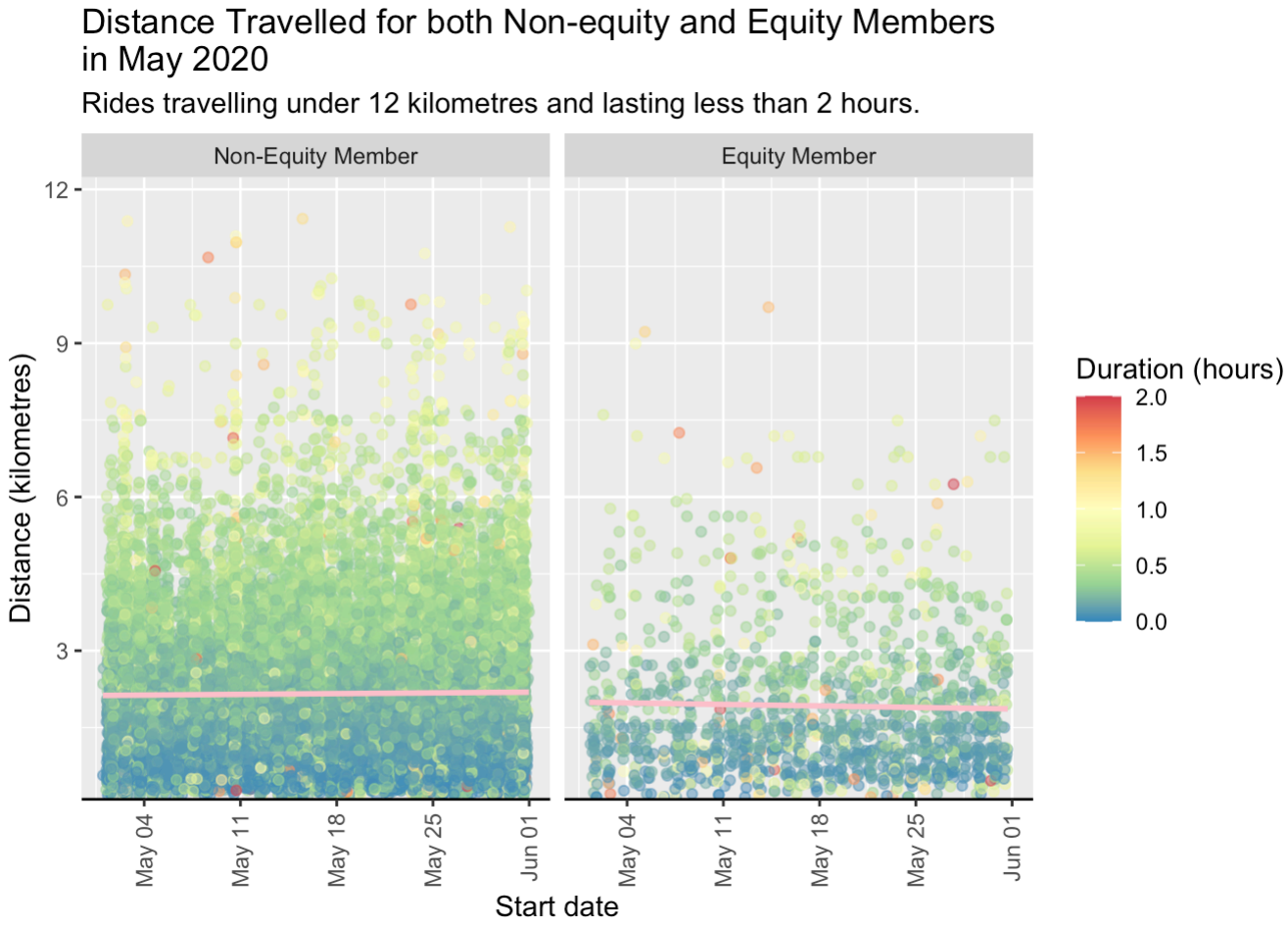


Answer to Question 3

Firstly, the figure demonstrates that there is clearly far more non-equity members than equity members. Throughout the month, the pink line of best fit shows that the average distance travelled per ride increases very slightly for non-equity members but decreases very slightly for equity members.

The duration generally increases with distance but there are some high duration points with very small distances travelled. This is likely users who rent the bike for a round trip; returning to the same position as where they started as the distance values in this data set only consider the start and end points of the ride.

```
df3 <- clean_rides_data %>%
  drop_na(is_equity) %>%
  filter(distance > 100 & distance < 12000) %>%
  filter(duration < 2 * 3600) %>% # Duration less than 2 hours
  filter(member_casual == "member") %>%
  ggplot(mapping = aes(
    x = start_date,
    y = distance / 1000, # Distance in km
    colour = duration / 3600 # Duration in hours
  )) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, colour = "pink") +
  scale_y_continuous(
    limits = c(0.1, 12),
    expand = expansion(mult = c(0, 0.02)) # Removing empty space
  ) +
  scale_colour_distiller(palette = "Spectral", limits = c(0, 2)) +
  facet_wrap(
    facets = vars(is_equity),
    labeller = labeller(is_equity = c(
      "TRUE" = "Equity Member",
      "FALSE" = "Non-Equity Member"
    ))
  ) +
  theme(
    axis.text.x = element_text(
      angle = 90, # make labels vertical
      hjust = 1, # right-adjust horizontally
      vjust = 0.5 # center vertically
    ),
    axis.line.x = element_line(colour = "black")
  ) +
  labs(
    title =
      "Distance Travelled for both Non-equity and Equity Members \nin May 2020",
    subtitle =
      "Rides travelling under 12 kilometres and lasting less than 2 hours.",
    x = "Start date",
    y = "Distance (kilometres)",
    colour = "Duration (hours)"
  )
df3
```



Conclusions

From our analysis of the first question, we can conclude that the most popular times of day for bike usage are consistent for each of the weekdays. The weekend days are similar to each other but very different to the weekdays. Classic bikes are the most popular choice, followed by docked bikes. Electric bikes have significantly lower popularity and the proportions of each bike used for each hour of the day and day of the week appears to remain consistent.

Our analysis from question two demonstrates that the 2021 policy change in which the free ride time was increased from 30 to 45 minutes had no effect on the average duration of a ride; the durations were actually slightly lower in 2021. We also found that causal users have a much higher average duration than members in both 2020 and 2021. The spread of durations was greater for casual users.

Finally, the analysis of our third question showed that there were substantially more non-equity members than equity members. Throughout May, the distances that non-equity members travelled slighlty increased whilst the distances that equity members travelled slighlty decreased.

One further question of interest raised by the results of our analysis is related to question three; how does the distance travelled vary throughout May 2020 for casual users? It would be interesting to see if the trend was similar to equity members or non-equity members.