02 Analyze Data

S A Nawash Akhtar

Data analysis processes of Cyclistic bike-share case study

1. Introduction

This documentation describes the data analysis processes of the Cyclistic bike-share case study. The specific task was to answer “How do annual members and casual riders use”Cyclistic” bikes differently?“. Hence the analysis was based on this question.

2. Load packages

```
library(tidyverse)
library(data.table)
library(leaflet)
library(sp)
library(leaflegend)
library(ggrepel)
library(ggpubr)
```

3. Create directory

```
my_path <- "E:\\Nawash\\Works\\Datasets\\Google data analysis\\Track 1\\CS1_Bike_share\\Csv files"
```

4. Load cleaned dataset

```
setwd(my_path)
dir()
all_trips_cleaned <- fread("all_trips_4.csv")
```

5. Arrange weekdays in order

```
all_trips_cleaned$day_of_week <- ordered(all_trips_cleaned$day_of_week,
                                         levels = c("Monday", "Tuesday",
                                                    "Wednesday", "Thursday",
                                                    "Friday", "Saturday",
                                                    "Sunday"))
```

5. Change month name from numeric to character

```
all_trips_cleaned$month <- month.abb[all_trips_cleaned$month] %>%
  ordered(levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                     "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
```

6. Popular days of week

6.1 Number of trips analysis

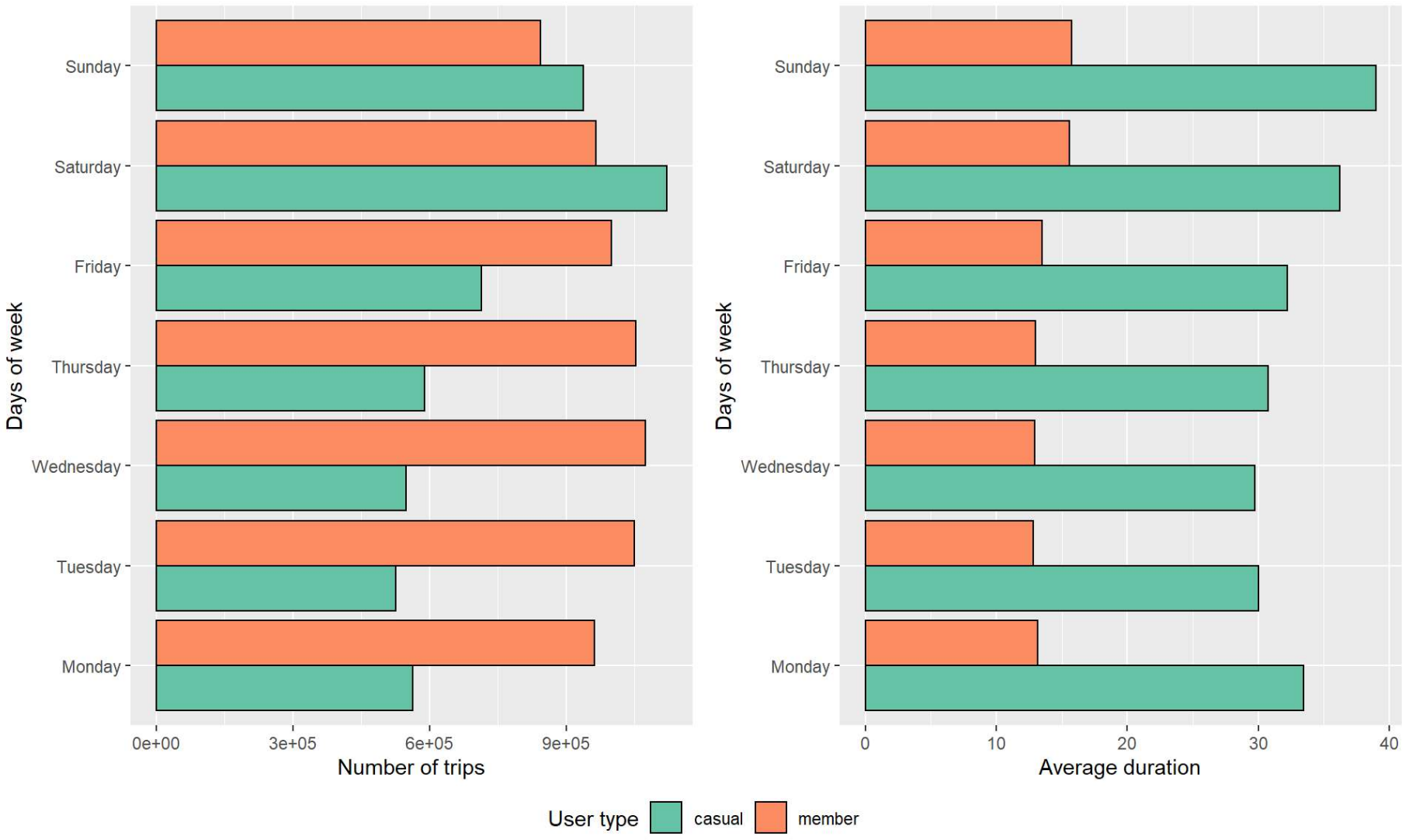
```
df1 <- all_trips_cleaned %>%
  group_by(member_casual,
           day_of_week) %>%
  summarise(number_of_trips = n()) %>%
  arrange (-number_of_trips) %>%
  ggplot(aes(x = number_of_trips,
             y = day_of_week, fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Number of trips", y = "Days of week")
```

6.2 Average ride duration analysis

```
df2 <- all_trips_cleaned %>%
  group_by(member_casual,
           day_of_week) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  arrange(-average_duration) %>%
  ggplot(aes(x = average_duration,
             y = day_of_week,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Average duration", y = "Days of week")
```

6.3 Create visualization comparing number of trips and average duration

```
ggarrange(df1, df2, ncol = 2,
          common.legend = TRUE, legend = "bottom")
```



7. Popular month

7.1 Number of trips analysis

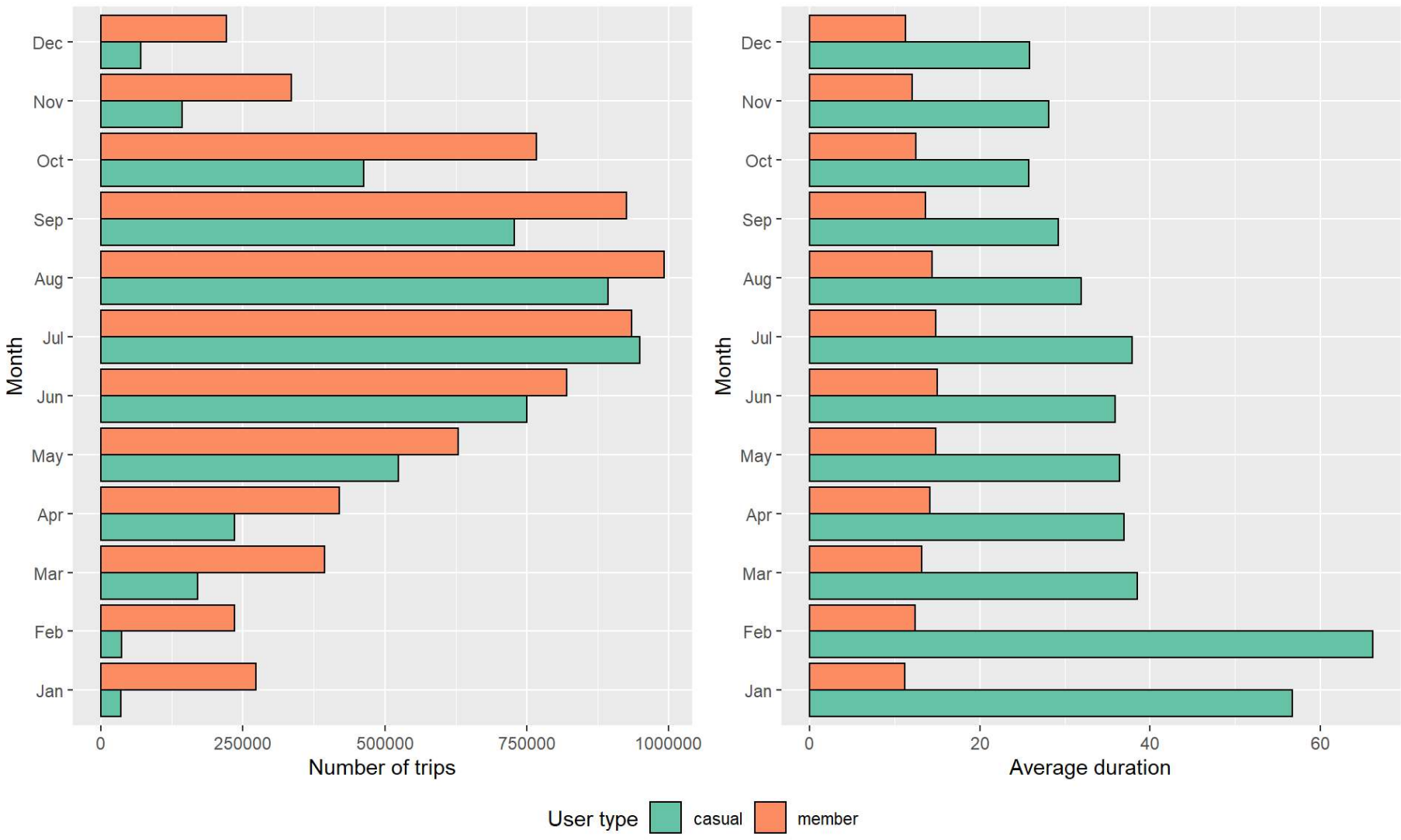
```
df3 <- all_trips_cleaned %>%
  group_by(member_casual,
           month) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x = number_of_trips,
             y = month,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Number of trips", y = "Month")
```

7.2 Average ride duration analysis

```
df4 <- all_trips_cleaned %>%
  group_by(member_casual,
           month) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  ggplot(aes(x = average_duration,
             y = month,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Average duration", y = "Month")
```

7.3 Create visualization comparing number of trips and average duration

```
ggarrange(df3, df4, ncol = 2,
          common.legend = TRUE, legend = "bottom")
```



8. Popular days of month

8.1 Number of trips analysis

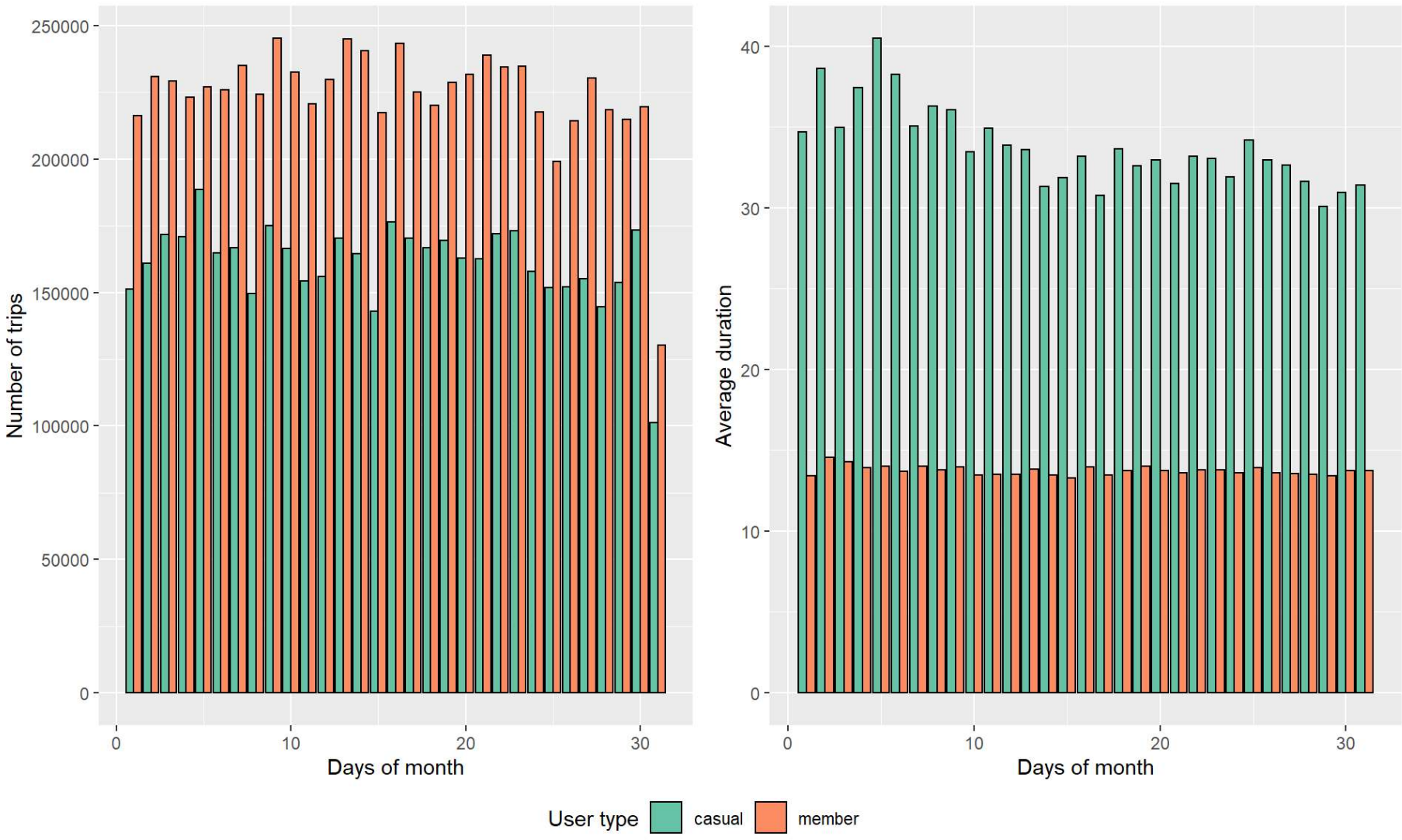
```
df5 <- all_trips_cleaned %>%
  group_by(member_casual,
           day) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x = day,
             y = number_of_trips,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Days of month", y = "Number of trips")
```

8.2 Average duration analysis

```
df6 <- all_trips_cleaned %>%
  group_by(member_casual,
           day) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  ggplot(aes(x = day,
             y = average_duration,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Days of month", y = "Average duration")
```

8.3 Create visualization comparing number of trips and avearge duration

```
ggarrange(df5, df6, ncol = 2,
          common.legend = TRUE, legend = "bottom")
```



9. Yearly ride data analysis

9.1 Number of trips analysis

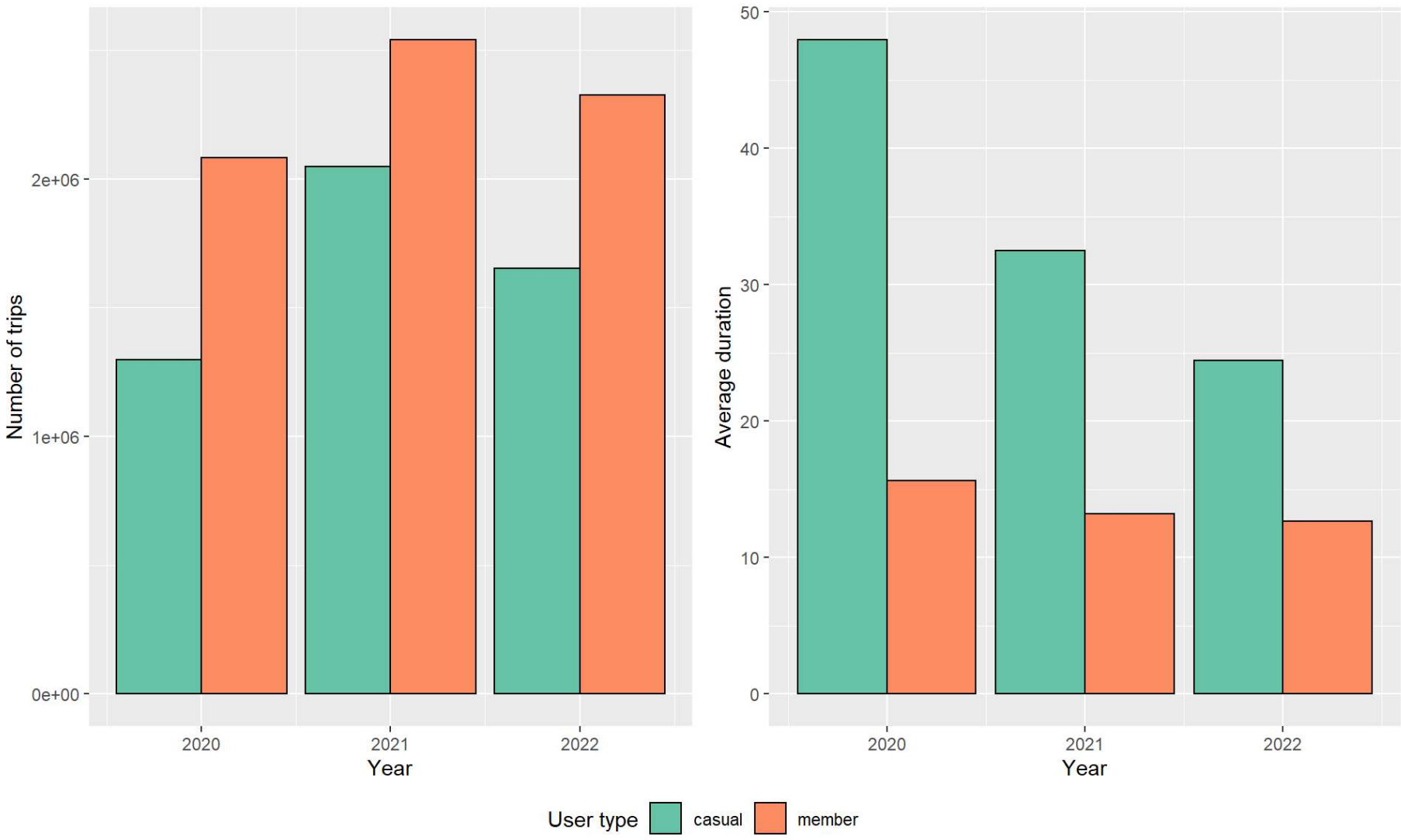
```
df7 <- all_trips_cleaned %>%
  group_by(member_casual,
           year) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x = year,
             y = number_of_trips,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Year", y = "Number of trips")
```

9.2 Average duration analysis

```
df8 <- all_trips_cleaned %>%
  group_by(member_casual,
           year) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  ggplot(aes(x = year,
             y = average_duration,
             fill = member_casual)) +
  geom_col(position = "dodge",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "User type") +
  labs(x = "Year", y = "Average duration")
```

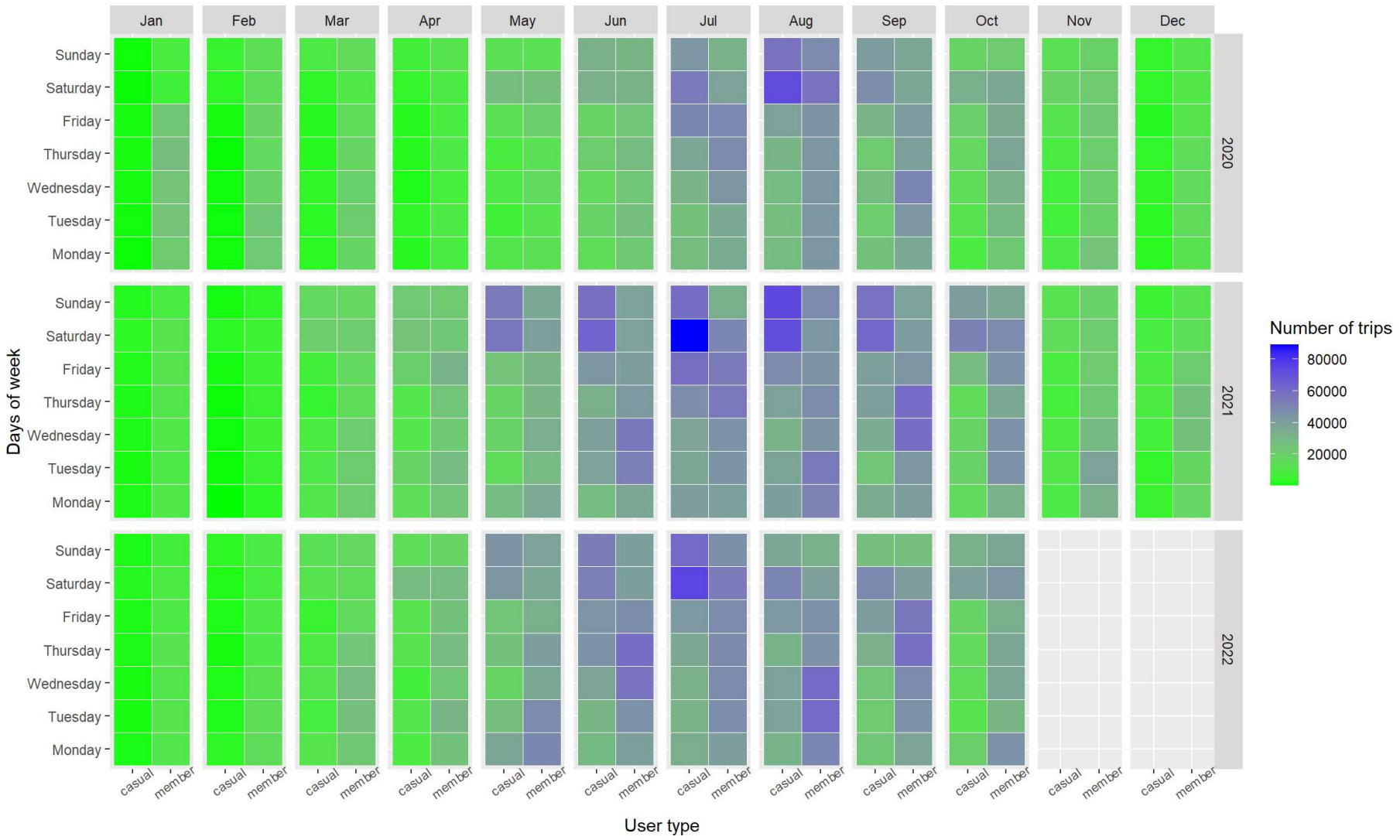
9.3 Create visualization comparing number of trips and avearage duration

```
ggarrange(df7, df8, ncol = 2,
          common.legend = TRUE, legend = "bottom")
```



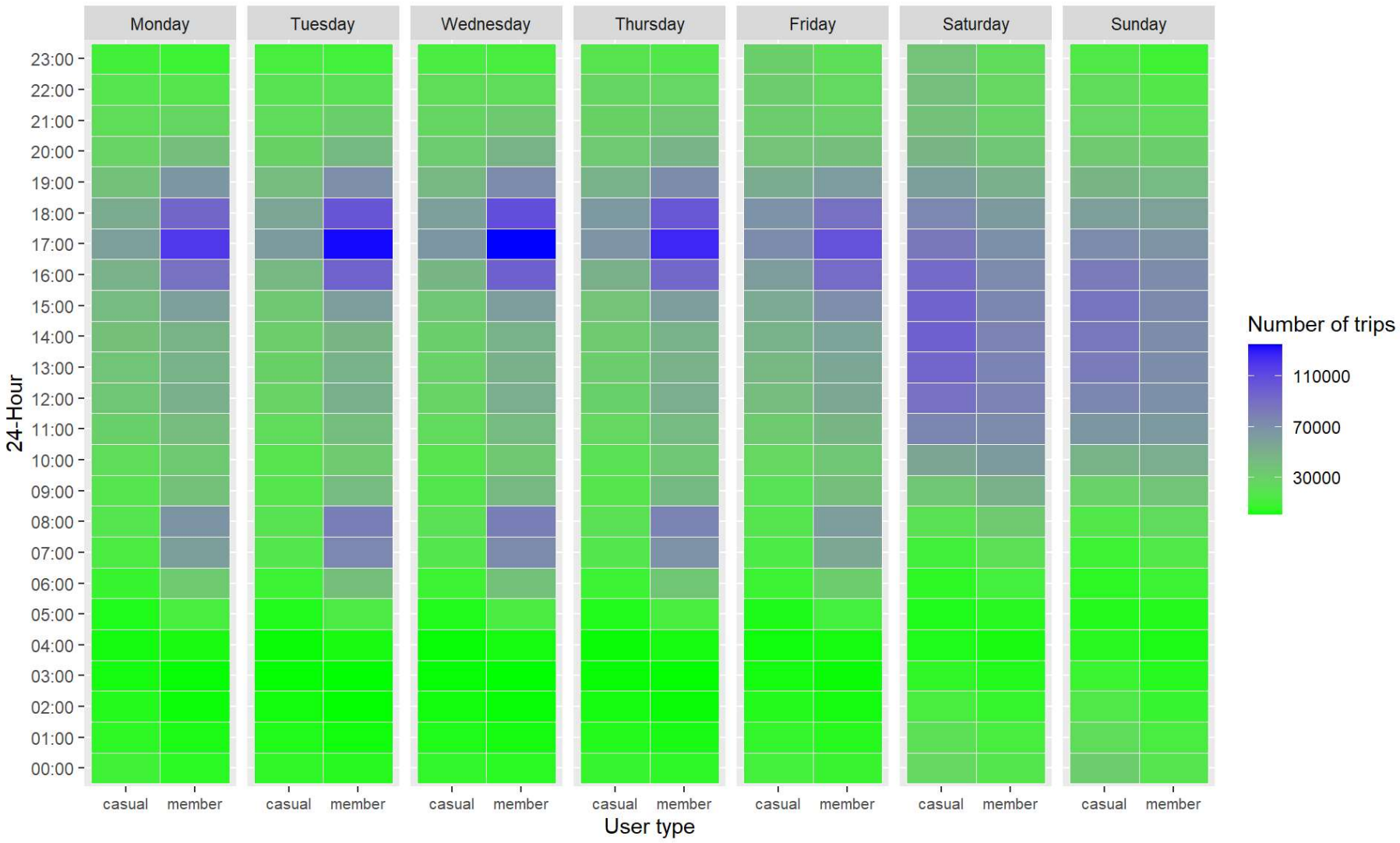
10. Bike usage heatmap for days of week and months (2020 - 2022)

```
df9 <- all_trips_cleaned %>%
  group_by(member_casual,
           year,
           month,
           day_of_week) %>%
  summarise(number_of_trips = n()) %>%
  arrange(member_casual,
          year,
          month,
          day_of_week) %>%
  ggplot(aes(x = member_casual,
             y = day_of_week,
             fill = number_of_trips)) +
  geom_tile(color = "white") +
  facet_grid(year ~ month) +
  scale_fill_gradient(low = "green",
                      high = "blue",
                      name = "Number of trips") +
  theme(axis.text.x = element_text(size = 8,
                                   angle = 35)) +
  labs( x = "User type", y = "Days of week")

df9
```



11. Popular time of the day

```
all_trips_cleaned$hour <- paste(format(
  all_trips_cleaned$started_at, "%H"),
  ":00", sep = "")

df10 <- all_trips_cleaned %>%
  group_by(member_casual,
           day_of_week,
           hour) %>%
  summarise(number_of_trips = n()) %>%
  arrange(member_casual,
          day_of_week,
          hour) %>%
  ggplot(aes(x = member_casual,
             y = hour,
             fill = number_of_trips)) +
  geom_tile(color = "white") +
  facet_grid(. ~ day_of_week) +
  scale_fill_gradient(low = "green",
                      high = "blue",
                      name = "Number of trips",
                      breaks = c(30000, 70000, 110000)) +
  theme(axis.text.x = element_text(size = 8)) +
  labs(x = "User type", y = "24-Hour")

df10
```

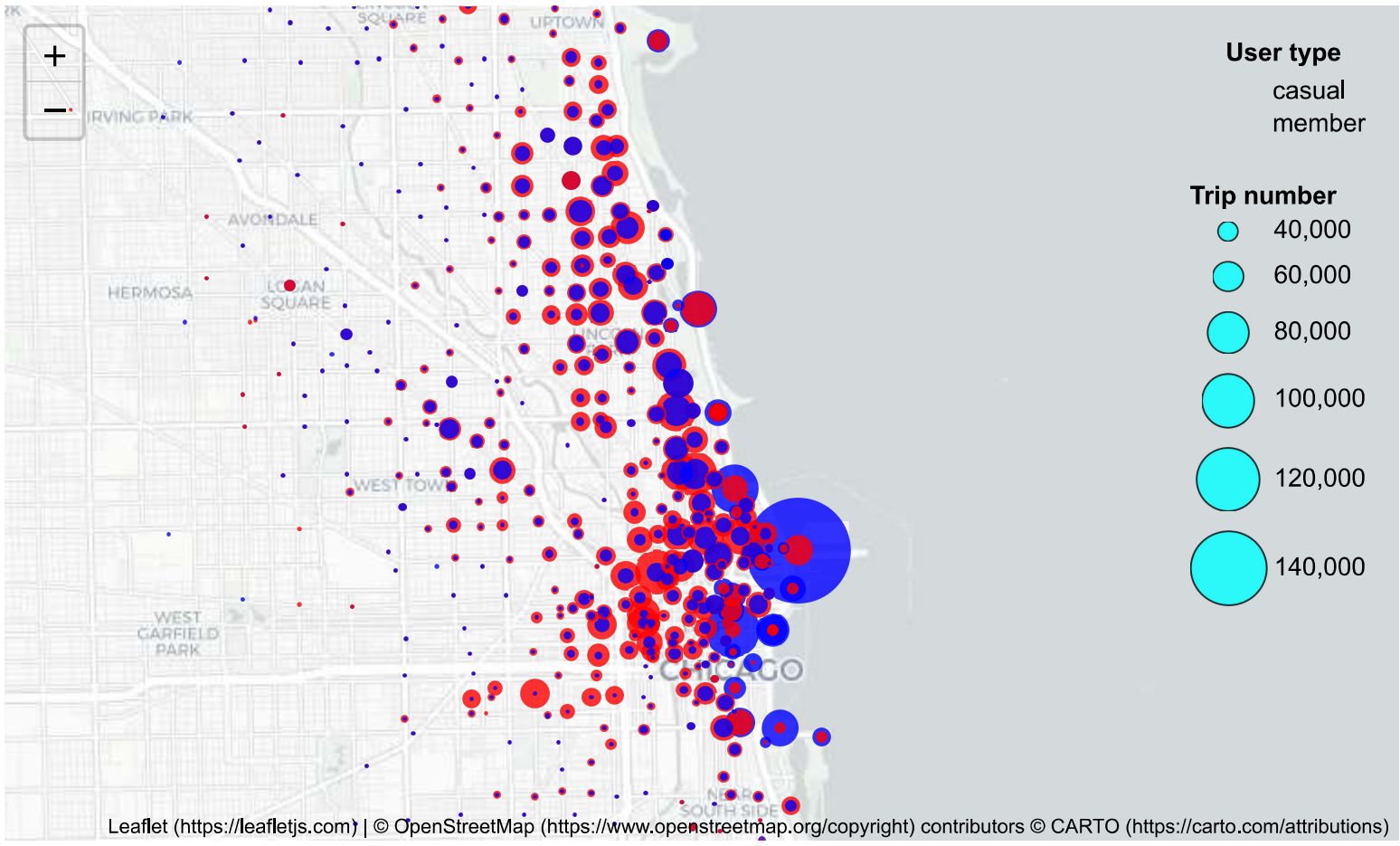


12. Most popular station

```
df11 <- all_trips_cleaned %>%
  select(start_station_name,
         member_casual,
         start_lat, start_lng) %>%
  group_by(start_station_name,
           member_casual) %>%
  summarise(number_of_trips = n(),
            across(start_lat:start_lng)) %>%
  arrange(-number_of_trips) %>%
  distinct(start_station_name,
           .keep_all = TRUE) %>%
  head(1000)

df11
```

```
## # A tibble: 1,000 × 5
## # Groups:   start_station_name, member_casual [1,000]
##    start_station_name       member_casual number_of_trips start_lat start_lng
##    <chr>                    <chr>                  <int>     <dbl>     <dbl>
##  1 Streeter Dr & Grand Ave  casual                143289      41.9     -87.6
##  2 Millennium Park          casual                 73360      41.9     -87.6
##  3 Michigan Ave & Oak St    casual                 64908      41.9     -87.6
##  4 Clark St & Elm St        member                 62380      41.9     -87.6
##  5 Kingsbury St & Kinzie St member                 59873      41.9     -87.6
##  6 Wells St & Concord Ln    member                 55886      41.9     -87.6
##  7 Theater on the Lake      casual                 52149      41.9     -87.6
##  8 Shedd Aquarium           casual                 50196      41.9     -87.6
##  9 Wells St & Elm St        member                 49550      41.9     -87.6
## 10 Dearborn St & Erie St    member                 48167      41.9     -87.6
## # … with 990 more rows
```

```
lon <- df11$start_lng
lat <- df11$start_lat
station_position <- as.data.frame(cbind(lon,lat))

color_pal <- colorFactor(palette = c("blue", "red"),
                         domain = c("casual", "member"))
station_label <- paste(df11$start_station_name,
                       paste("Trip number: ", df11$number_of_trips, sep = ""),
                       sep = " # ")

coordinates(station_position) <- ~ lon + lat
leaflet(station_position) %>%
  addCircleMarkers(fillColor = color_pal(df11$member_casual),
                   fillOpacity = 0.8,
                   stroke = FALSE,
                   label = station_label,
                   radius = df11$number_of_trips/5000) %>%
  addProviderTiles(providers$CartoDB.Positron) %>%
  setView(lng = -87.6300, lat = 41.9100, zoom = 12.45) %>%
  addLegend(pal = color_pal,
            values = df11$member_casual,
            title = "User type") %>%
  addLegendSize(values = df11$number_of_trips[1:10],
                color = "black", fillColor = "cyan",
                shape = "circle", position = "topright",
                title = "Trip number", opacity = 0.8)
```



13. Most popular bike route

Average duration and maximum duration are shown in minutes.

```
df12 <- all_trips_cleaned %>%
  group_by(member_casual, start_station_name,
           end_station_name) %>%
  summarise(number_of_trips = n(),
            average_duration = round(mean(ride_duration), 2),
            max_duration = round(max(ride_duration), 2)) %>%
  arrange(-number_of_trips) %>%
  head(1000)
```

13.1 Popular route for casual riders

```
df12 %>%
  filter(member_casual == "casual") %>%
  head(10)
```

```
## # A tibble: 10 × 6
## # Groups:   member_casual, start_station_name [9]
##    member_casual start_station_name              end_s…¹ numbe…² avera…³ max_d…⁴
##    <chr>         <chr>                           <chr>     <int>   <dbl>   <dbl>
##  1 casual        Streeter Dr & Grand Ave         Street…   27716    48.2   4036.
##  2 casual        Millennium Park                 Millen…   14612    49.3  10245.
##  3 casual        Michigan Ave & Oak St           Michig…   13878    51.8   1329.
##  4 casual        Lake Shore Dr & Monroe St       Lake S…   10337    50.5   2874.
##  5 casual        Buckingham Fountain             Buckin…    9425    67.3   8656.
##  6 casual        DuSable Lake Shore Dr & Monroe… DuSabl…    9387    37.1   1518.
##  7 casual        Indiana Ave & Roosevelt Rd      Indian…    8458    53.9   1607.
##  8 casual        Theater on the Lake             Theate…    8229    50     2828.
##  9 casual        DuSable Lake Shore Dr & Monroe… Street…    7584    28.5   1432.
## 10 casual        Michigan Ave & 8th St           Michig…    7386    55.8   2487.
## # … with abbreviated variable names ¹end_station_name, ²number_of_trips,
## #   ³average_duration, ⁴max_duration
```

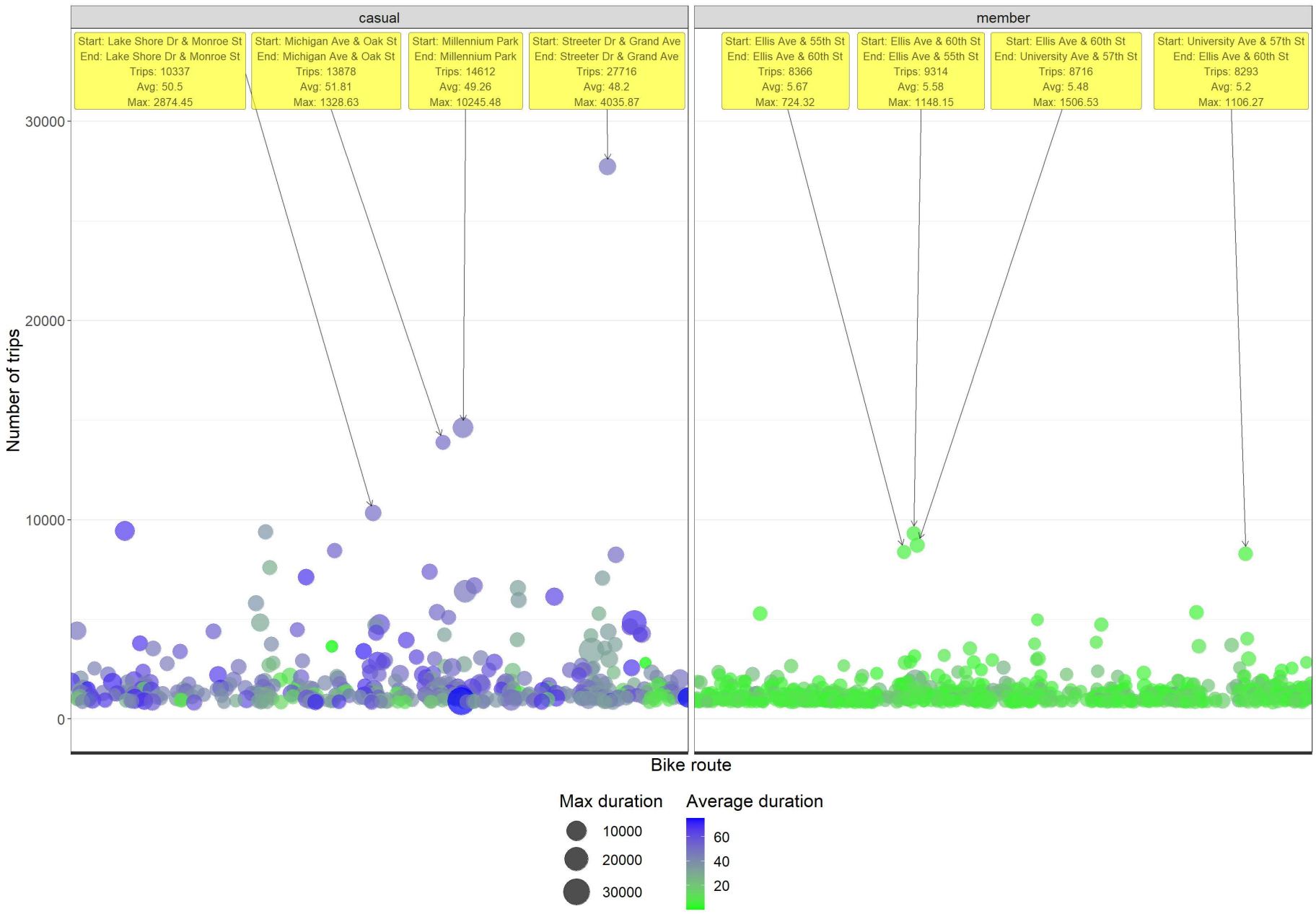
13.2 Popular route for members

```
df12 %>%
  filter(member_casual == "member") %>%
  head(10)
```

```
## # A tibble: 10 × 6
## # Groups:   member_casual, start_station_name [8]
##    member_casual start_station_name       end_station_…¹ numbe…² avera…³ max_d…⁴
##    <chr>         <chr>                    <chr>            <int>   <dbl>   <dbl>
##  1 member        Ellis Ave & 60th St      Ellis Ave & 5…    9314    5.58  1148.
##  2 member        Ellis Ave & 60th St      University Av…    8716    5.48  1507.
##  3 member        Ellis Ave & 55th St      Ellis Ave & 6…    8366    5.67   724.
##  4 member        University Ave & 57th St Ellis Ave & 6…    8293    5.2   1106.
##  5 member        State St & 33rd St       Calumet Ave &…    5341    4.7   1011.
##  6 member        Calumet Ave & 33rd St    State St & 33…    5295    4.34  1145.
##  7 member        Loomis St & Lexington St Morgan St & P…    4966    5.6     86.6
##  8 member        Morgan St & Polk St      Loomis St & L…    4728    6.17   783.
##  9 member        University Ave & 57th St Kimbark Ave &…    4025    7.79   455.
## 10 member        MLK Jr Dr & 29th St      State St & 33…    3842    7.82   180.
## # … with abbreviated variable names ¹end_station_name, ²number_of_trips,
## #   ³average_duration, ⁴max_duration
```

13.3 Bike route visualization

```
plt <- df12 %>%
  mutate(bike_route = paste(paste("Start:", start_station_name, sep = " "),
                            paste("End:", end_station_name, sep = " "),
                            sep = "\n")) %>%
  ggplot(aes(x = bike_route,
             y = number_of_trips)) +
  geom_point(aes(color = average_duration,
                 size = max_duration),
             alpha = 0.7) +
  geom_label_repel(aes(
    label = ifelse(
      number_of_trips > 9500 & member_casual == "casual" |
        number_of_trips > 6000 & member_casual == "member",
      paste(bike_route,
            paste("Trips:", number_of_trips, sep = " "),
            paste("Avg:", average_duration, sep = " "),
            paste("Max:", max_duration, sep = " "),
            sep = "\n"), "")),
    size = 3.5, nudge_y = 30000, force = 100, direction = "x",
    fill = "yellow", alpha = 0.6,
    arrow = arrow(length = unit(0.01, "npc")),
    point.padding = 0.7) +
  ylim(0, 33000) +
  scale_color_gradient(low = "green", high = "blue") +
  scale_size(range = c(5, 12)) +
  theme_bw() +
  theme(panel.grid.major.x = element_blank(),
        axis.text.x = element_blank(),
        text = element_text(size = 17),
        legend.position = "bottom",
        legend.direction = "vertical") +
  facet_grid(. ~ member_casual) +
  labs(x = "Bike route", y = "Number of trips",
       size = "Max duration", color = "Average duration")

plt
```



14. Bike type preference

14.1 Number of trips analysis

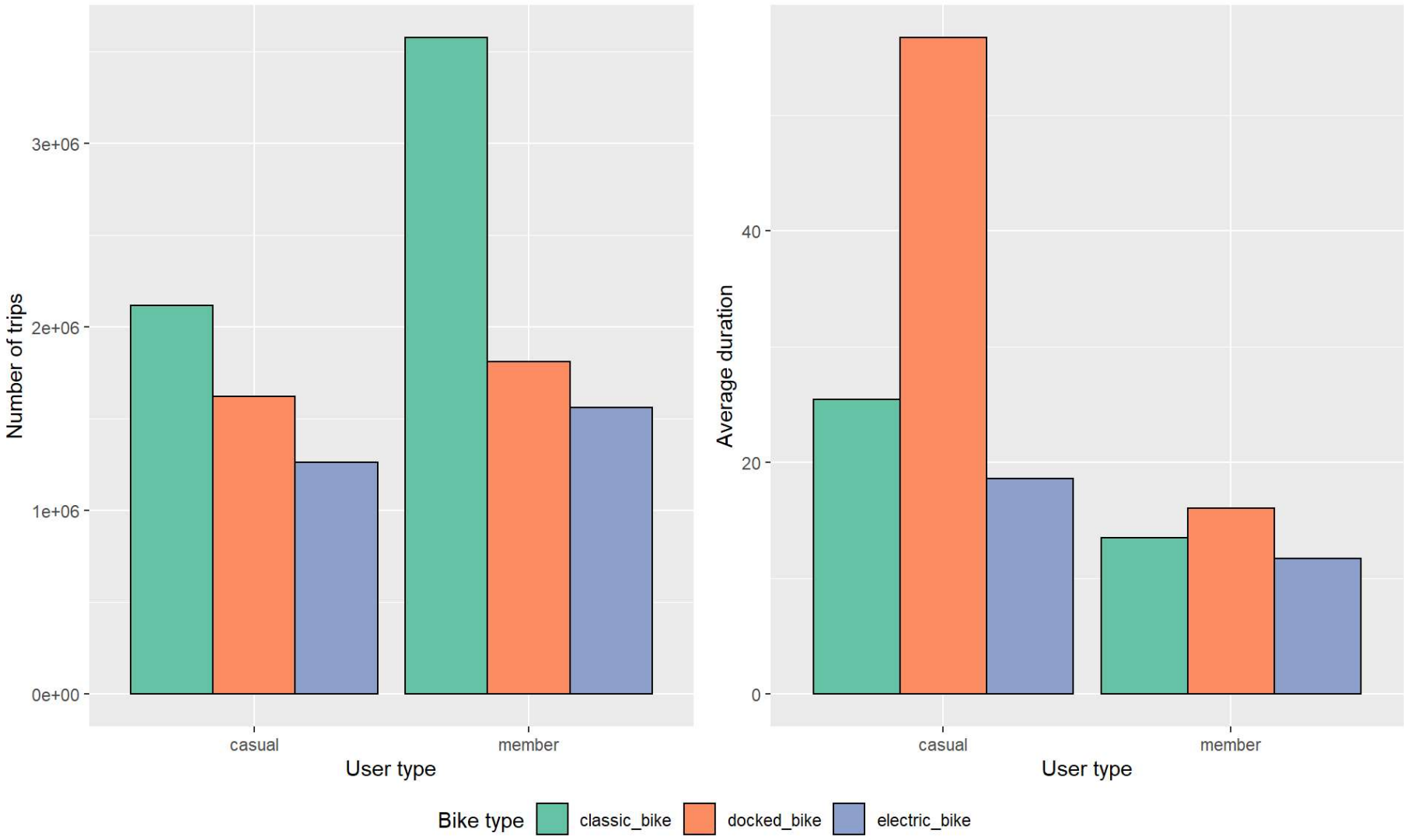
```
df13 <- all_trips_cleaned %>%
  group_by(member_casual,
           rideable_type) %>%
  summarise(number_of_trips = n()) %>%
  arrange(-number_of_trips) %>%
  ggplot(aes(x = member_casual,
             y = number_of_trips,
             fill = rideable_type)) +
  geom_bar(position = "dodge",
           stat = "identity",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "Bike type") +
  labs(x = "User type", y = "Number of trips")
```

14.2 Average duration analysis

```
df15 <- all_trips_cleaned %>%
  group_by(member_casual,
           rideable_type) %>%
  summarise(average_duration = mean(ride_duration)) %>%
  arrange(-average_duration) %>%
  ggplot(aes(x = member_casual,
             y = average_duration,
             fill = rideable_type)) +
  geom_bar(position = "dodge",
           stat = "identity",
           color = "black") +
  scale_fill_brewer(palette = "Set2",
                    name = "Bike type") +
  labs(x = "User type", y = "Average duration")
```

14.3 Create visualization comparing number of trips and average duration

```
ggarrange(df13, df15, ncol = 2,
          common.legend = TRUE, legend = "bottom")
```


Number of trips
3e+06
2e+06
1e+06
0e+00
casual
member
User type
Average duration
40
20
0
casual
member
User type
Bike type
classic_bike
docked_bike
electric_bike