Cyclistic Bike-share Case Study

Rongrong Xu 2022-07-01

1. Collect Data

- I imported 12 month's csv. files of the Cyclistic trip data for past 12 months in 2021, and combine them into one dataset.

- The dataset I'm going to use for the analysis is Cyclistic's historical trip data I obtained from google certificate program provided which is a public data and has been available by Motivate International Inc. under this license. The data-privacy issue prohibits you from using riders' personally identifiable information, but it won't be a problem since the dataset does not include the riders' personal identifiable information.

2. Clean up data and process it for analysis

In this step, I cleaned and prepared the data for analysis.

- I dropped the columns that will not be used in the later analysis.
- Since the data frame does not have enough information to help me do the analysis thoroughly, I added columns to get enough information: ride_length, day, month, and day_of_week.
- I removed 147 rows with the ride_length smaller than 0 and 0 duplicated row.

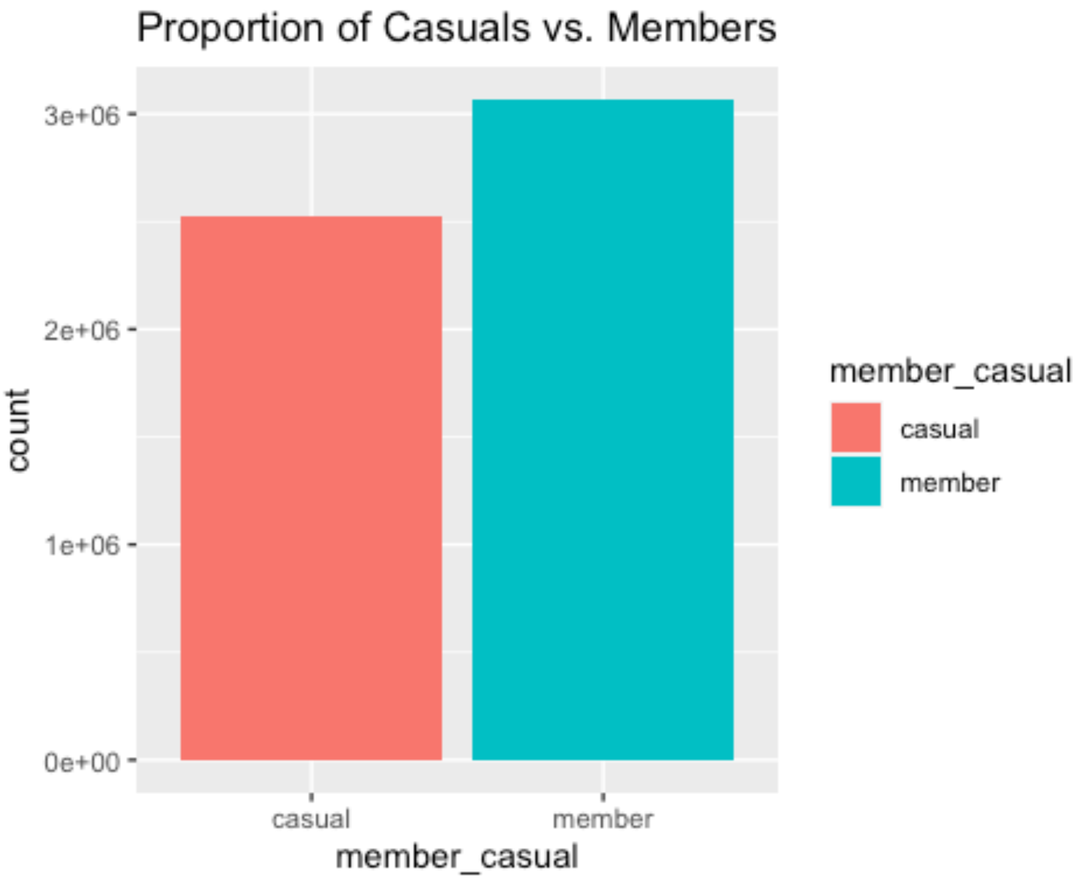
3. Conduct descriptive analysis

In this step, I focused on comparing and analyzing the different behavior between member and casual riders.

```r
# compare the proportion of member riders and casual riders
trip2021_v3 %>%
    group_by(member_casual) %>%
    summarise(count = length(ride_id),
              '%' = (length(ride_id) / nrow(trip2021_v3)) * 100)
```

```
## # A tibble: 2 × 3
##   member_casual   count   `%`
##   <chr>           <int> <dbl>
## 1 casual        2528946  45.2
## 2 member        3065970  54.8
```

```
ggplot(trip2021_v3, aes(member_casual, fill=member_casual)) +
    geom_bar() +
    labs(title="Proportion of Casuals vs. Members")
```



(1) I plotted the total number of member riders and casual riders and found that member riders 9% higher proportion than casual riders.

```
# Compare members and casual riders (aggregate(sum_var ~ group_var, data = df
, FUN = mean))
aggregate(trip2021_v3$ride_length ~ trip2021_v3$member_casual, FUN = mean)

##   trip2021_v3$member_casual trip2021_v3$ride_length
## 1                    casual                32.00221
## 2                    member                13.63355

aggregate(trip2021_v3$ride_length ~ trip2021_v3$member_casual, FUN = median)

##   trip2021_v3$member_casual trip2021_v3$ride_length
## 1                    casual                15.96667
## 2                    member                 9.60000
```

```
aggregate(trip2021_v3$ride_length ~ trip2021_v3$member_casual, FUN = max)
```

```
##   trip2021_v3$member_casual trip2021_v3$ride_length
## 1                    casual               55944.150
## 2                    member                1559.933
```

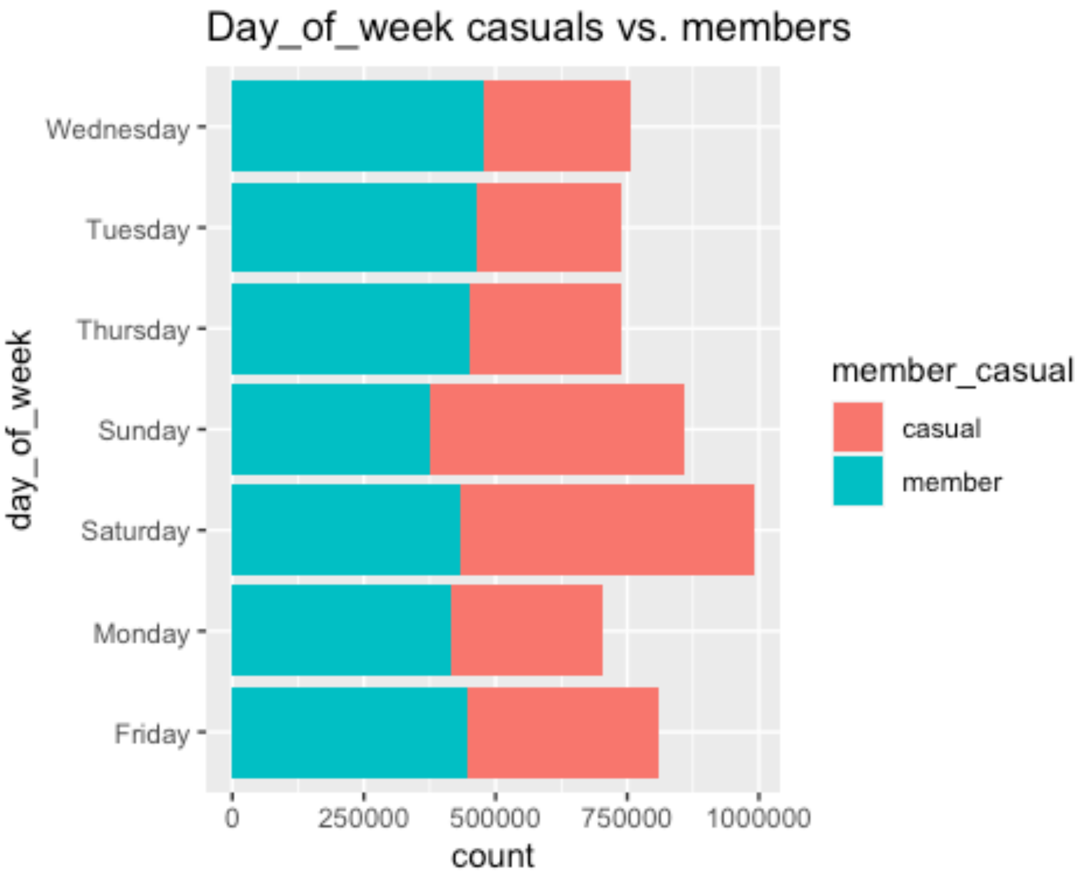
```
aggregate(trip2021_v3$ride_length ~ trip2021_v3$member_casual, FUN = min)
```

```
##   trip2021_v3$member_casual trip2021_v3$ride_length
## 1                    casual                       0
## 2                    member                       0
```

```
# Compare the average ride length on day of the week for members vs casual ri
ders
aggregate(trip2021_v3$ride_length ~ trip2021_v3$member_casual + trip2021_v3$d
ay_of_week, FUN = mean)
```

```
##    trip2021_v3$member_casual trip2021_v3$day_of_week trip2021_v3$ride_leng
th
## 1                     casual                  Friday                30.348
60
## 2                     member                  Friday                13.324
92
## 3                     casual                  Monday                31.875
45
## 4                     member                  Monday                13.247
53
## 5                     casual                Saturday                34.706
23
## 6                     member                Saturday                15.264
57
## 7                     casual                  Sunday                37.566
58
## 8                     member                  Sunday                15.657
94
## 9                     casual                Thursday                27.703
26
## 10                    member                Thursday                12.776
18
## 11                    casual                 Tuesday                27.972
33
## 12                    member                 Tuesday                12.788
12
## 13                    casual               Wednesday                27.657
31
## 14                    member               Wednesday                12.819
16
```

```
ggplot(trip2021_v3,aes(day_of_week,fill=member_casual)) + geom_bar() + labs(t
itle = "Day_of_week casuals vs. members") + coord_flip()
```



(2) The mean of ride length for casual riders is around 3 times longer than member riders.

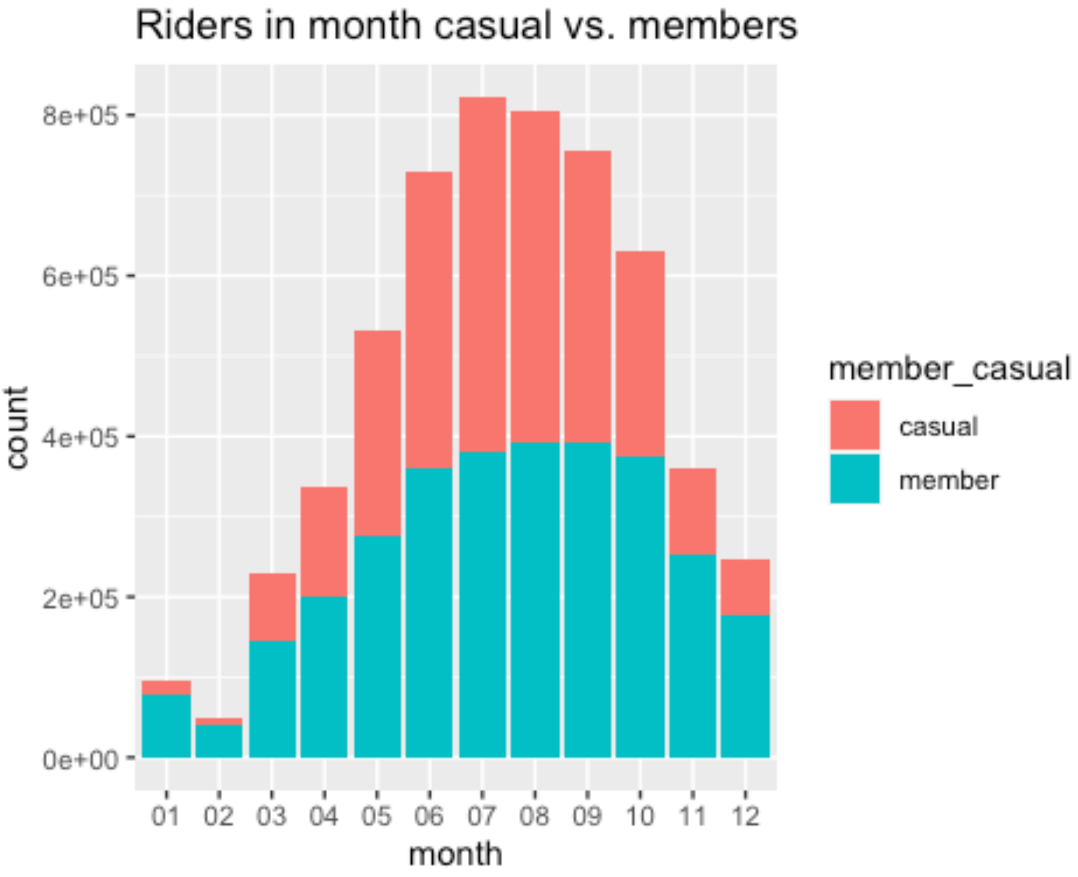
(3) The longest ride length for casual riders is 35 times longer than member riders.

(4) I compared the count of riders on different days of the week. There are the most riders using our products on weekends, Saturday has the biggest volume. Member riders are the main proportion of all the users, but casual riders take place having the most data during the weekend. The number of casual riders starts increasing from Friday every week. There was about the same number of member riders using the product each day of the week.

```
# Compare the average ride length on each month for members vs casuals
trip2021_v3 %>%
    group_by(month) %>%
    summarise(count = length(ride_id),
              '%' = (length(ride_id) / nrow(trip2021_v3)) * 100,
              'members%' = (sum(member_casual == "member") / length(ride_id))
* 100,
              'casual%' = (sum(member_casual == "casual") / length(ride_id))
* 100
              ) %>%
     arrange(desc(count))

## # A tibble: 12 × 5
##    month  count    `%` `members%` `casual%`
##    <chr>  <int>  <dbl>      <dbl>     <dbl>
##  1 07    822397 14.7         46.2      53.8
##  2 08    804323 14.4         48.7      51.3
##  3 09    756111 13.5         51.9      48.1
##  4 06    729590 13.0         49.2      50.8
##  5 10    631226 11.3         59.2      40.8
##  6 05    531631  9.50        51.7      48.3
##  7 11    359925  6.43        70.3      29.7
##  8 04    337225  6.03        59.5      40.5
##  9 12    247540  4.42        71.8      28.2
## 10 03    228494  4.08        63.2      36.8
## 11 01     96832  1.73        81.3      18.7
## 12 02     49622  0.887       79.6      20.4

ggplot(trip2021_v3,aes(month,fill=member_casual)) + geom_bar() + labs(title =
"Riders in month casual vs. members")
```



(5) I compared the count of riders using our products on each month.

And found that the high volume happened in June, July, August, and September, which had around 14% of the total number of riders in 2021.

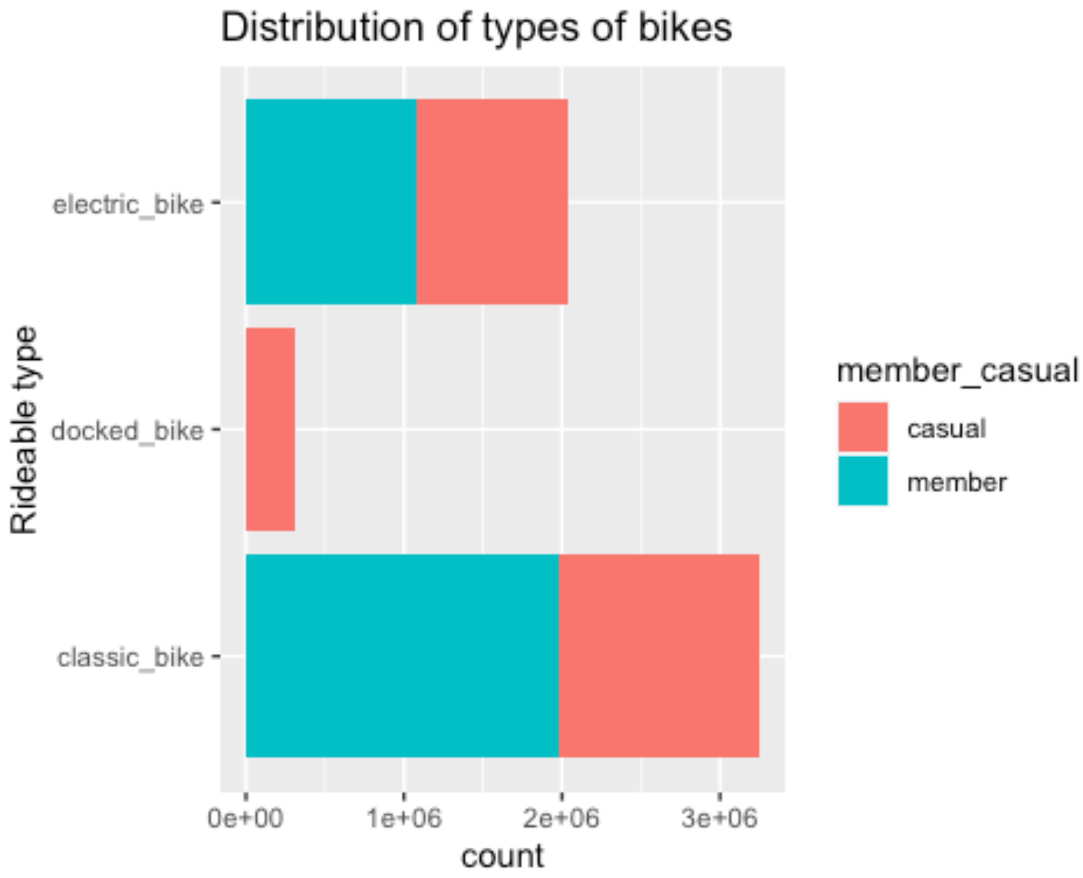
In most months, count of member riders were more than casual riders except July, August, and June.

Casual riders were 7% more than member users in July, 3% more than member users in August, and 1% more than member users in June.

```
# Rideable type
trip2021_v3 %>%
    group_by(rideable_type) %>%
    summarise(count = length(ride_id),
          '%' = (length(ride_id) / nrow(trip2021_v3)) * 100,
          'members%' = (sum(member_casual == "member") / length(ride_id)) * 1
00,
          'casual%' = (sum(member_casual == "casual") / length(ride_id)) * 10
0)

## # A tibble: 3 × 5
##   rideable_type   count   `%` `members%` `casual%`
##   <chr>           <int> <dbl>      <dbl>     <dbl>
## 1 classic_bike  3250943 58.1   61.0           39.0
## 2 docked_bike    312338  5.58   0.000320     100.
## 3 electric_bike 2031635 36.3   53.2           46.8

ggplot(trip2021_v3, aes(rideable_type, fill=member_casual)) +
    labs(x="Rideable type", title="Distribution of types of bikes") +
    geom_bar() +
    coord_flip()
```

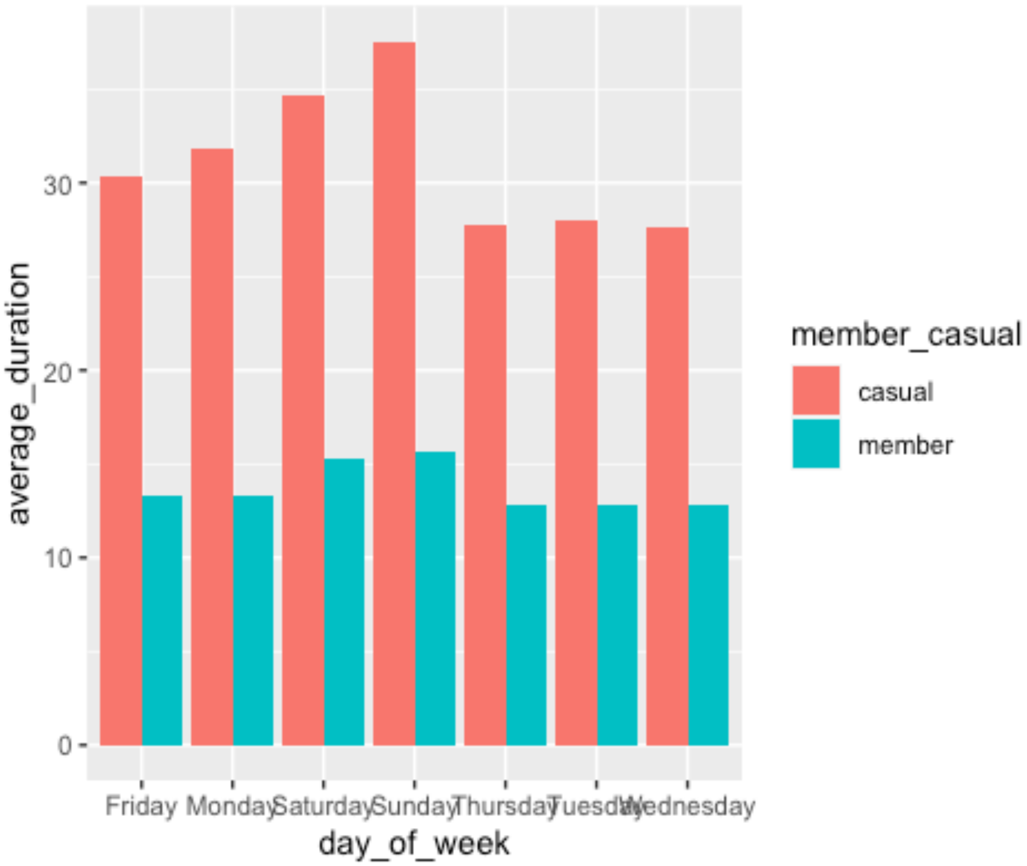


(6) I compared the count of riders by different types of bikes provided by the company. Classic bike has the most volume of riders, member riders were 23% more than casual riders. Same as electric_bike.

Only casual riders used docked bike.

```
# Average riding length(duration)
trip2021_v3 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
           ,average_duration = mean(ride_length)) %>%
  arrange(member_casual,day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")

## `summarise()` has grouped output by 'member_casual'. You can override usin
g the
## `.groups` argument.
```

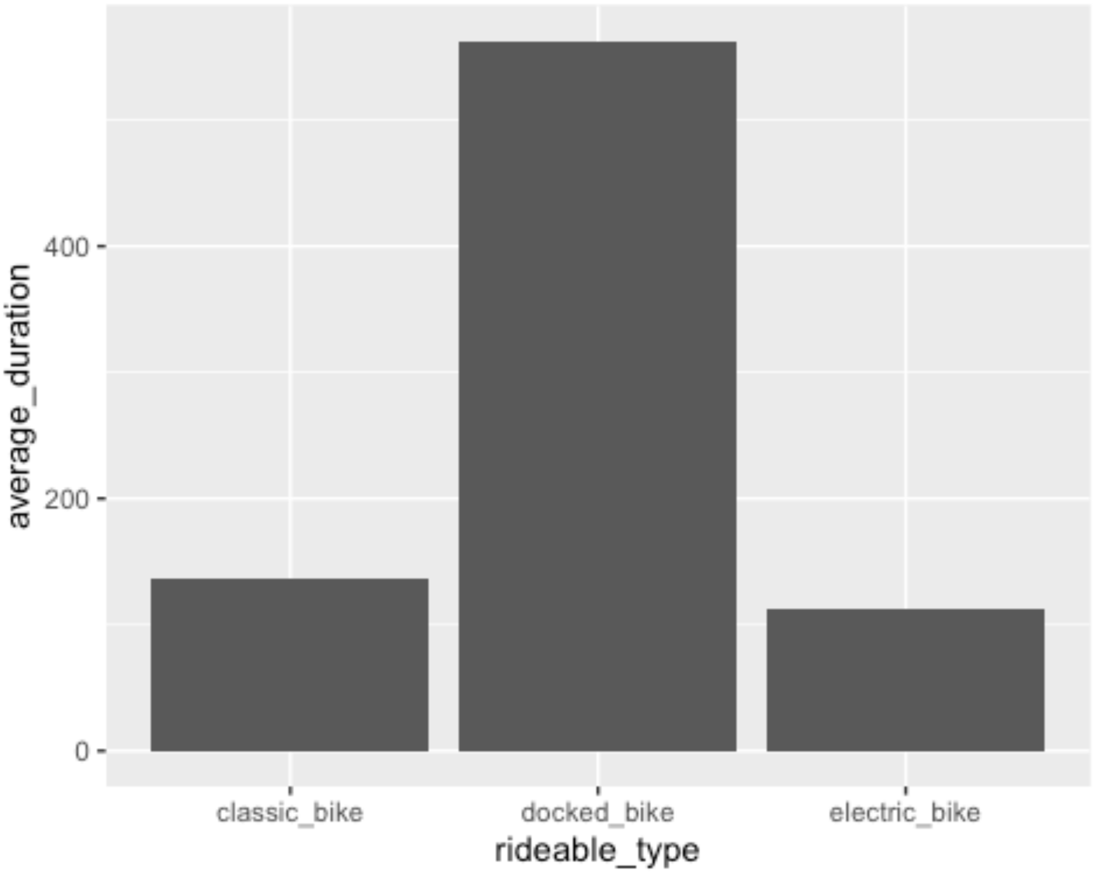


(7) I compared average riding time on each day of the week.

Overall, casual riders used almost 2 times longer riding time than member riders, and the longest riding time concentrated on Sunday.

```
# Average riding length(duration) on different rideable types
trip2021_v3 %>%
  group_by(rideable_type, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  ggplot(aes(x = rideable_type, y = average_duration)) +
  geom_col()
```

```
## `summarise()` has grouped output by 'rideable_type'. You can override usin
g the
## `.groups` argument.
```



(8) I compared average riding time on different rideable types.

Docked_bike had the longest duration of riding. But since I found that only casual riders used docked bike and casual riders used longer than member riders. Then classic bike has the second longest riding length.

4. Summary

- What did I discover from the data?

Member riders and casual riders act differently in weekdays. Casual riders have different riding length from member riders. Only casual riders use docked bikes.

- Trends from the data?

More member riders than casual riders. More riders on the weekends and casual riders take bigger proportion. The number of member users in the weekdays approximately stay the same. Member riders prefer classic bike or electric bikes. Casual riders have longer riding time than member riders. The volume of users is affected by weather and temperature. Small number of riders use bikes in cold weather, and big number of riders use bikes in summertime.

- How this relates to business problem?

The purpose of using the bikes between members and casuals could be different. The member riders use that to going to work with the fixed schedule during the weekday, while the casual riders use that on weekend for recreation and it is not scheduled.

- **To encourage casual riders to become members**:

(1) Set a cap of riding time for casual users.
(2) Raise the price for non-member use during the weekend and summertime.
(3) Set price promotion during the weekday for casual users.
(4) Introduce 2 types of members, one is regular members, the new one is members could enjoy priority and unlimited riding length during the peak-hour.
(5) Raise the price for docked bike or set a lower price for members.