Case Study 1 - Cyclistic - Attempt 2

Nika Levidze

2023-04-11

Tidyverse

We will use the Tidyverse packages to wrangle and analyze the data.

```
install.packages("tidyverse")
library(tidyverse)
install.packages("readr")
```

The Data

We imported the ride data for March 2022.

```
X202203 <- X202203_divvy_tripdata <- read_csv("202203-divvy-tripdata.csv")
```

For the ease of reference, we changed the name of the data frame.

Calculated Columns

To get better insights from this dataset, we'll create two new metrics based on the available ones.

1. Ride Length

We will add a new calculated row called ride_length which will be difference between the start time and the end time of the rides (indicated in started_at and ended_at columns respectively).

```
mutate(X202203_divvy_tripdata, ride_length = ended_at - started_at)

X202203_with_ride_length <- mutate(X202203_divvy_tripdata, ride_length = ended_at - started_at)
```

2. Day of the Week

We will add a new calculated row called day_of_week which will be calculated using the *wday* function based on the start time of the ride (indicated in the started_at column).

```
mutate(X202203_with_ride_length,day_of_week=wday(started_at))

X202203_with_ride_length_and_wday <- mutate(X202203_with_ride_length,day_of_week=wday(started_at))
```

3. Hour of the day

We will add a new calculated row called hour_of_day which will be calculated using the *hour* function based on the start time of the ride (indicated in the started_at column).

```
mutate(X202203_with_ride_length_and_wday, hour_of_day=hour(started_at))

X202203_with_ride_length_and_wday_and_hour <- mutate(X202203_with_ride_length_and_wday,
                                                     hour_of_day=hour(started_at))
```

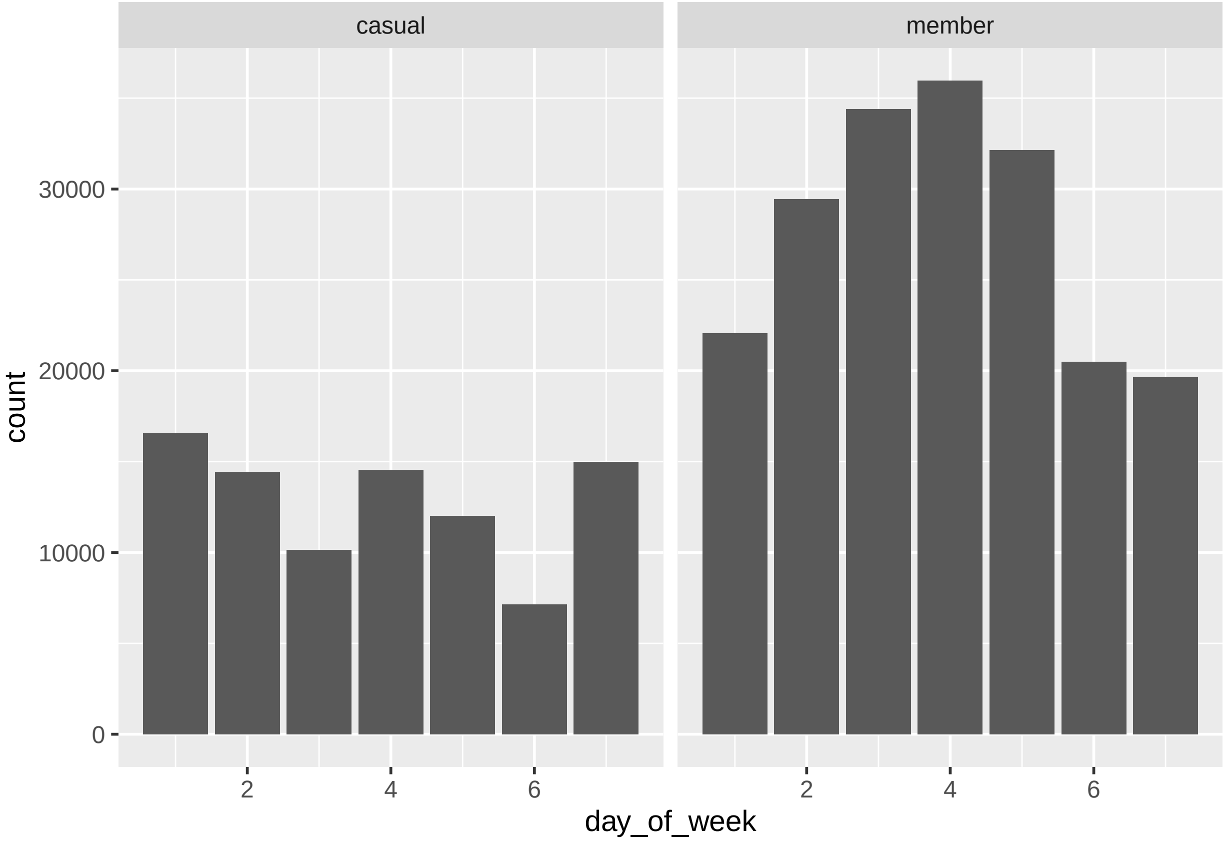
Analysis

Let's analyse the data and identify the behavioral differences between the Annual Members and the Casual users based on the following 2 KPIs.

1.Weekend Usage

It's important to identify the difference between the user types in terms of distribution of the rides throughout the days of the week. By comparing the usage during the weekends versus during the weekdays we can make informed assumptions about if the users use the bikes for their daily commute on the weekdays or for their leisure on the weekends.

```
ggplot(data=X202203_with_ride_length_and_wday)+
geom_bar(mapping=aes(x=day_of_week))+
facet_wrap(~member_casual)
```

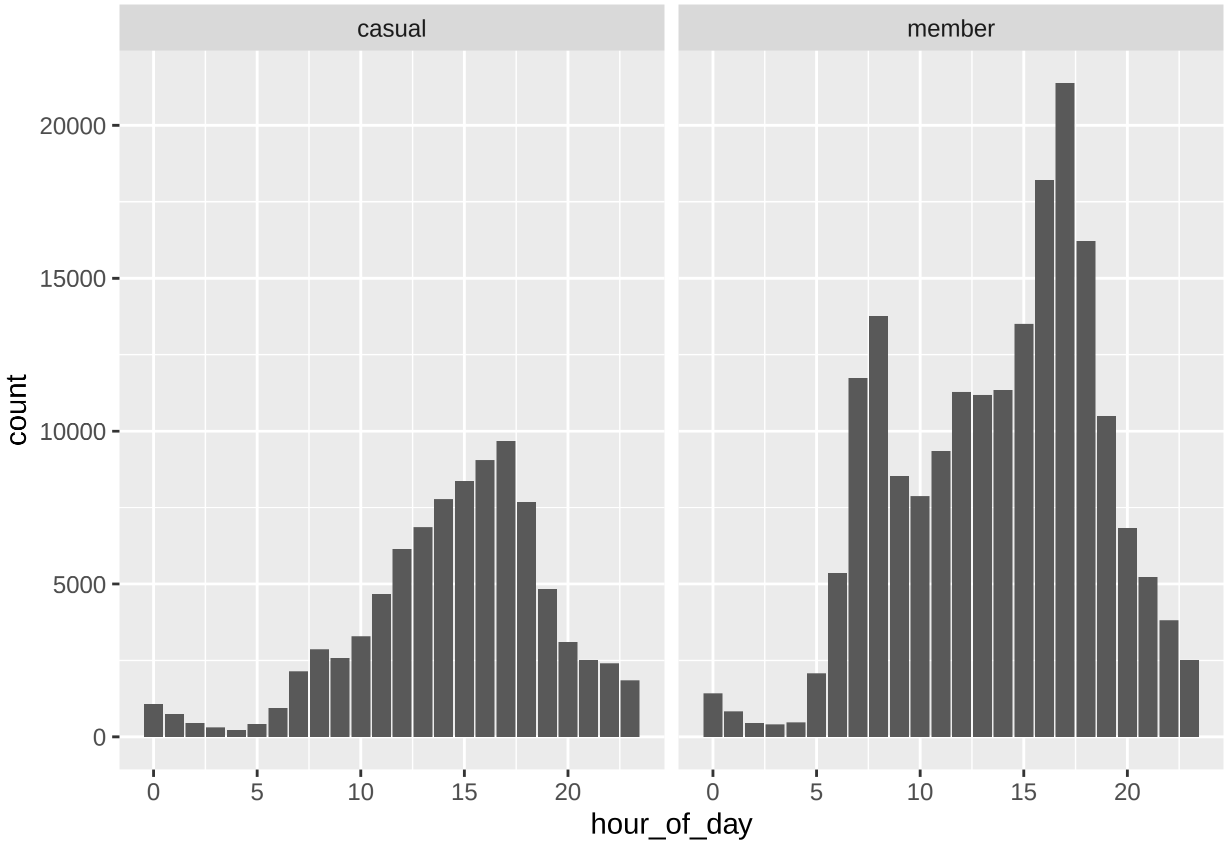


As expected, the annual members use the bikes more often on the weekdays and the casual riders use them more during the weekends.

2.Time of the day usage

Let's do the similar kind of analysis, but instead of the distribution throughout the weekdays, let's measure the distribution throughout the day.

```
ggplot(data=X202203_with_ride_length_and_wday_and_hour)+
geom_bar(mapping=aes(x=hour_of_day))+
facet_wrap(~member_casual)
```



As we can see, the pattern looks the same for both member types, except for one detail. For the annual members the usage surges from 6 AM to 9 AM. This is the time when most people commute to work. Yet another indication that the annual members use the bikes for their commute.

3.Proximity to Business Centers

To identify whether there's a pattern of . . . . or not and if these patterns can provide any useful insight, we will create a heat map, based on the provided coordinates of the start station locations.

To do this we will use Tableau. Here's a link to the Tableau dashboard.

As it shows on the heat map, the overall distribution of the rides throughout the city is somewhat similar between Members and Casuals. However, if we take a closer look at the highlights, we can see that among Members, the most popular locations are close to the downtown area which includes Universities. As for the, Casuals the most popular locations are closer to the areas with more leisure activities, such as the pier, parks and the theater.

Again, this visualization reinforces our claim that for the Annual Members, Citibike is more of a means of a day-to-day commute rather than a leisure activity. The opposite seems to be true for the Casual Riders.

4.Average monthly mileage

And finally, let's compare the average ride length between the two user types.

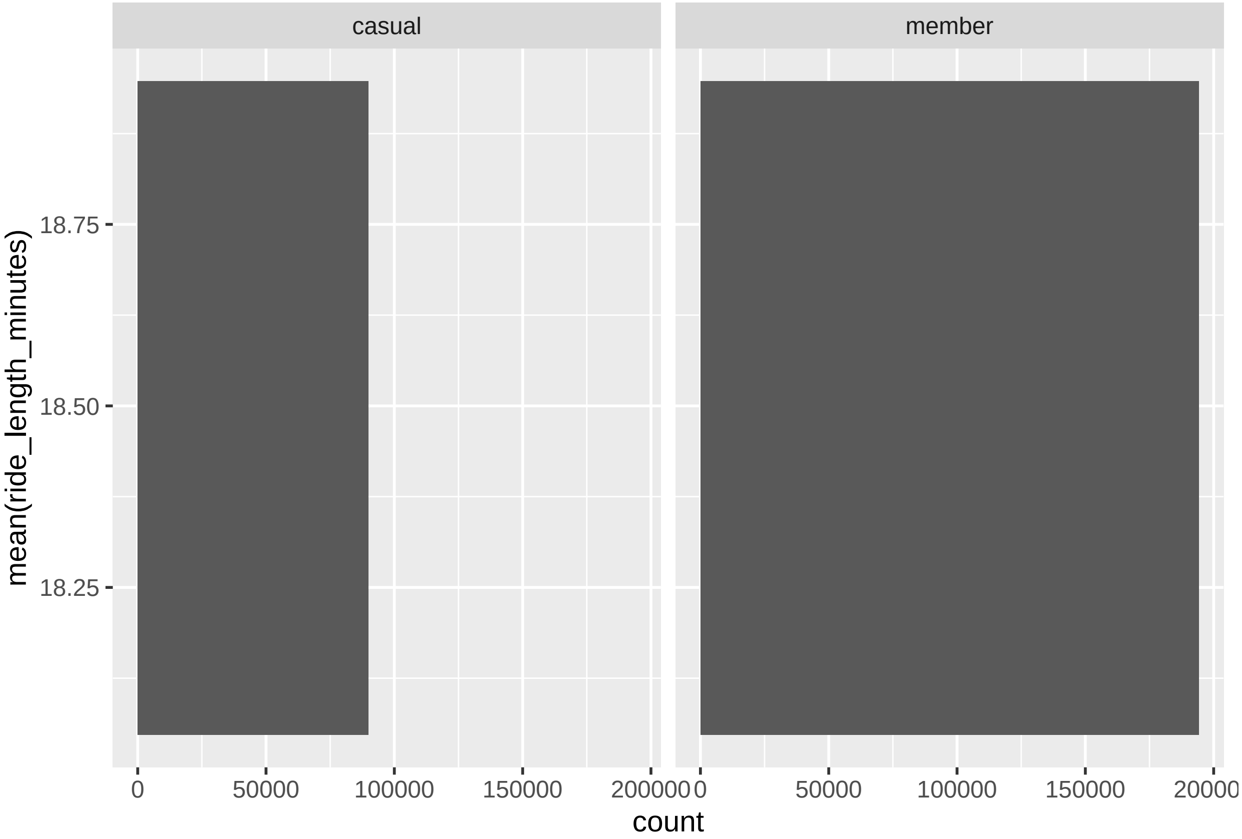
To do this, we're going to need a new calculated metric: ride_length_minutes.

```
mutate(X202203_with_ride_length_and_wday_and_hour, ride_length_minutes=ride_length/60)

X202203_with_ride_length_minutes_and_wday_and_hour <- mutate(X202203_with_ride_length_and_wday_and_hour
```

Let's create a visualization.

```
ggplot(data=X202203_with_ride_length_minutes_and_wday_and_hour)+
geom_bar(mapping=aes(y=mean(ride_length_minutes)))+
facet_wrap(~member_casual)
```



As we can see, the average ride length is the same for both user types. Since this can seem a bit odd, let's double check the validity by of this calculation with a secondary method.

First, let's calculate the average ride length for the Casual Riders:

```
print(mean(subset(X202203_with_ride_length_minutes_and_wday_and_hour,
                  member_casual="casual")$ride_length_minutes))
```

```
## Time difference of 18.49725 secs
```

Now let's do the same for the Annual members:

```
print(mean(subset(X202203_with_ride_length_minutes_and_wday_and_hour,
                  member_casual="member")$ride_length_minutes))
```

```
## Time difference of 18.49725 secs
```

It all checks out. There's no significant difference between the Members and Casuals in terms of the average ride length.