Track 1 Bikes Case Study- January to April

N. White

January 25, 2022 - January 27, 2022

Cyclistic Bike Data

Main Question at Hand

How do annual members and casual riders use Cyclistic bikes differently?

The finance department has already determined that annual subscription riders are more profitable for the company that casual riders who use bikes for one ride at a time or have one day passes. Cyclistic hopes to grow the company and convert casual users to annual Cyclistic members.

We must first determine the differences in how the two users utilize Cyclistic bikes.

Accessing the data

Available online, 2021's data is in full as of January 2022 and separated by month. The bike data is accessed and used under the license detailed by Divvy Bikes. You can review the license here: Divvy license agreement. To review the data yourself, it is linked here: Divvy Bike Data. No step taken utilizing this information violates the agreement.

This data comes directly from the source of the bike-share company. It doesn't look to be skewed in any direction or to be built with a specific persuasive purpose in mind.

I downloaded the data and put it into Google Cloud and RStudio to do analysis in R and SQL, should I need to. Much of the data is large and over the limits to utilize Google Sheets or Excel without error messages or major slowing. In the ways I saved the data, it is not accessible while I work on this project so there are no concerns about unforeseen changes. There is no expectation to come across private data or personally identifiable information, so there is not an addtional need to increase security around the data.

Within R, the tables are saved under a 3 character shortened version of the month being analyzed. The year is not needed because all of the data is from the year 2021. To do the same, set up your R Workspace:

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

```
library(lubridate)
library(tinytex)
```

Set your project folder: setwd("/Users/redrosessweet/Documents/[Coding]/[Data Science]/Google Cert - Data Analytics/Capstone/Track 1 - Bikes")

```
setwd("/Users/redrosessweet/Documents/[Coding]/[Data Science]/Google Cert - Data Analytics/Capstone/Track 1 - Bikes")
```

Then save each csv file under it's 3 letter month:

I can now see the data and what is available, but let's double-check.

```
colnames(Jan)

##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

This data is rectangular, as to be expected, with 13 columns of data. There are few NULL values which may indicated the data has already had some cleaning. The values in each cell match with the type of data expected. Given the data is organized in a way that can be used, the next important part is to filter the information and sort it.

Already we can start to get some ideas of what information on behaviors will be observable. * How many trips are taken each month by which type of user * Which stations are most popular for annual members? Which for casual? * Which stations are most popular overall? * How long are trips for casual users? Members? * What is the average length of a bike trip by member type? * What time of day are the bikes most popular? * What are some more popular times of the day that members access bikes? Casual riders? * What type of bike is most popular by type of rider? * Compare the number of members over time

Processing the Data

I added a column that calulcates duration in seconds to each month..

Thanks Statology! Check out the code I learned at statology

I saw that there were many recorded trips that were under 10 seconds, and decided to remove rides 5 seconds or under across each month, filtering out staggeringly short trips where one would have barely been on the bike before returning it.

```
Jan_filtered <- Jan %>%
    select(ride_id, rideable_type, started_at, ended_at, member_casual, trip_duration1) %>%
    filter(trip_duration1 > 5)

Feb_filtered <- Feb %>%
```

```
    select(ride_id, rideable_type, started_at, ended_at, member_casual,
trip_duration2) %>%
    filter(trip_duration2 > 5)

Mar_filtered <- Mar %>%
    select(ride_id, rideable_type, started_at, ended_at, member_casual,
trip_duration3) %>%
    filter(trip_duration3 > 5)
```

In January, 96834 to 96619 observations. In February, 49622 to 49424 observations. In March, 228496 to 228107 observations. In April, 337230 to 336524 observations. In May, 531633 to 530441 observations. In June, 729595 to 728164 observations. in July, 822410 to 820755 observations. In August, 804352 to 802700 observations. In September, 756147 to 754623 observations. In October, 631226 to 630022 observations. In November, 359978 to 359165 observations. In December, 247540 to 247028 observations.

Looking over (Analyze)

January

In January we have 96,619 entries (after cleaning) representing 96,619 individually counted bikeshare rides.

Time to separate our members from casual riders.

```
Jan_members <- Jan_filtered %>%
    filter(member_casual == "member")
Jan_casual <- Jan_filtered %>%
     filter(member_casual == "casual")
```

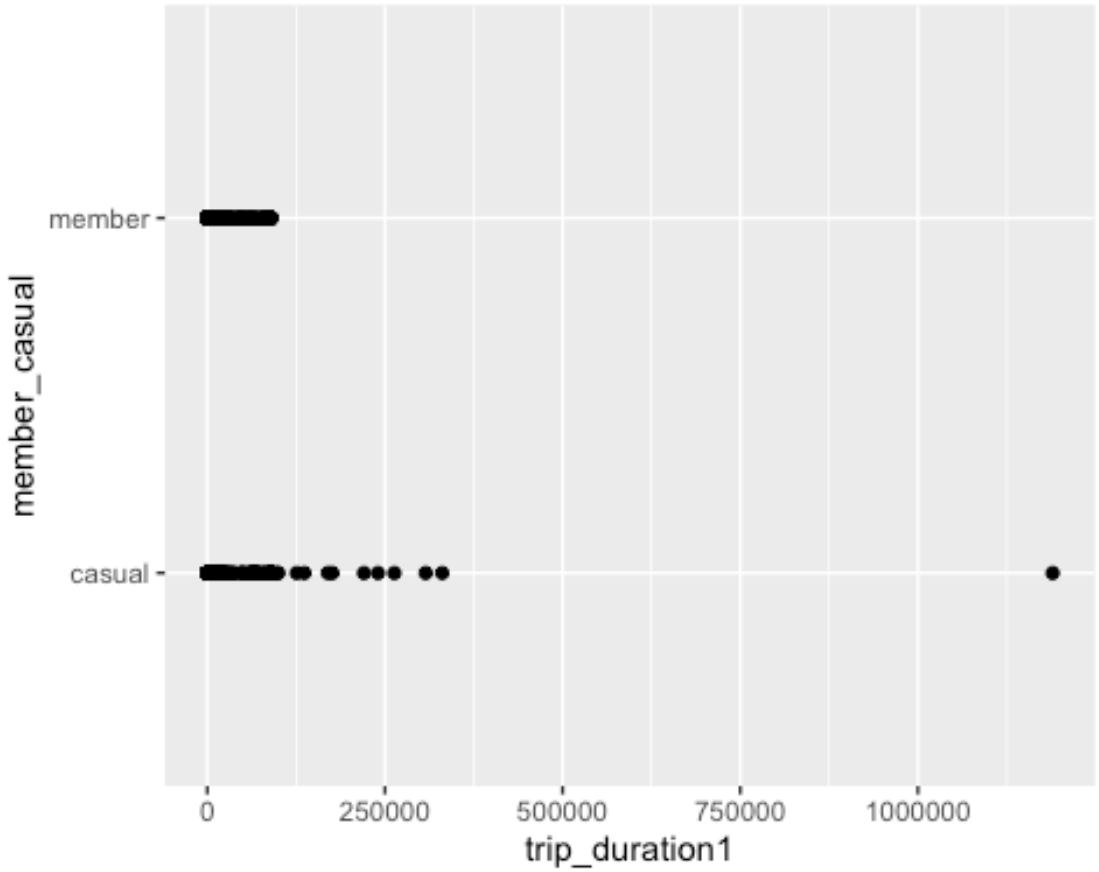
In January, there were 18091 rides from casual users and 78528 rides from annual members. Annual members had over 4 times as many trips than casual riders.

```
Jan_trip_avg <- (mean(Jan_filtered$trip_duration1))
Jan_m_trip_avg <- (mean(Jan_members$trip_duration1))
Jan_c_trip_avg <- (mean(Jan_casual$trip_duration1))
```

Trip length

The average trip length for all users in January was 918.18 seconds or about 15 minutes 18 seconds. For members, this was 774 seconds or about 12 minutes.

```
ggplot(data = Jan_filtered, aes(x = trip_duration1, y = member_casual)) +
    geom_point()
```



For casual riders the average trip length was 1543 seconds. This is where we find quite the outlier in the casual rides with one trip being 1189555 seconds or nearly 2 weeks. We could speculate as to why there was a ride so long like not knowing how to return the bike, but that is beyond the scope of what we are looking at. After removing the outlier, the average trip for casual users is 1477 seconds or 24 minutes and 37 seconds, about 16 minutes less. This doesn't warrant completely changing the data used, but is important information to have.

Max Trip and Min Trip

Before filtering for trips over 5 seconds, the shortest trip for casual riders is 1 second. For Members, the shortest trip was -625 seconds. There was also a -103 second trip, 4 trips of 0 seconds. Just like casual riders there are several trips of 1 second.

```
Jan_c_trip_max <- (max(Jan_casual$trip_duration1))
Jan_m_trip_max <- (max(Jan_members$trip_duration1))
Jan_c_trip_min <- (min(Jan_casual$trip_duration1))
Jan_m_trip_min <- (min(Jan_members$trip_duration1))
```

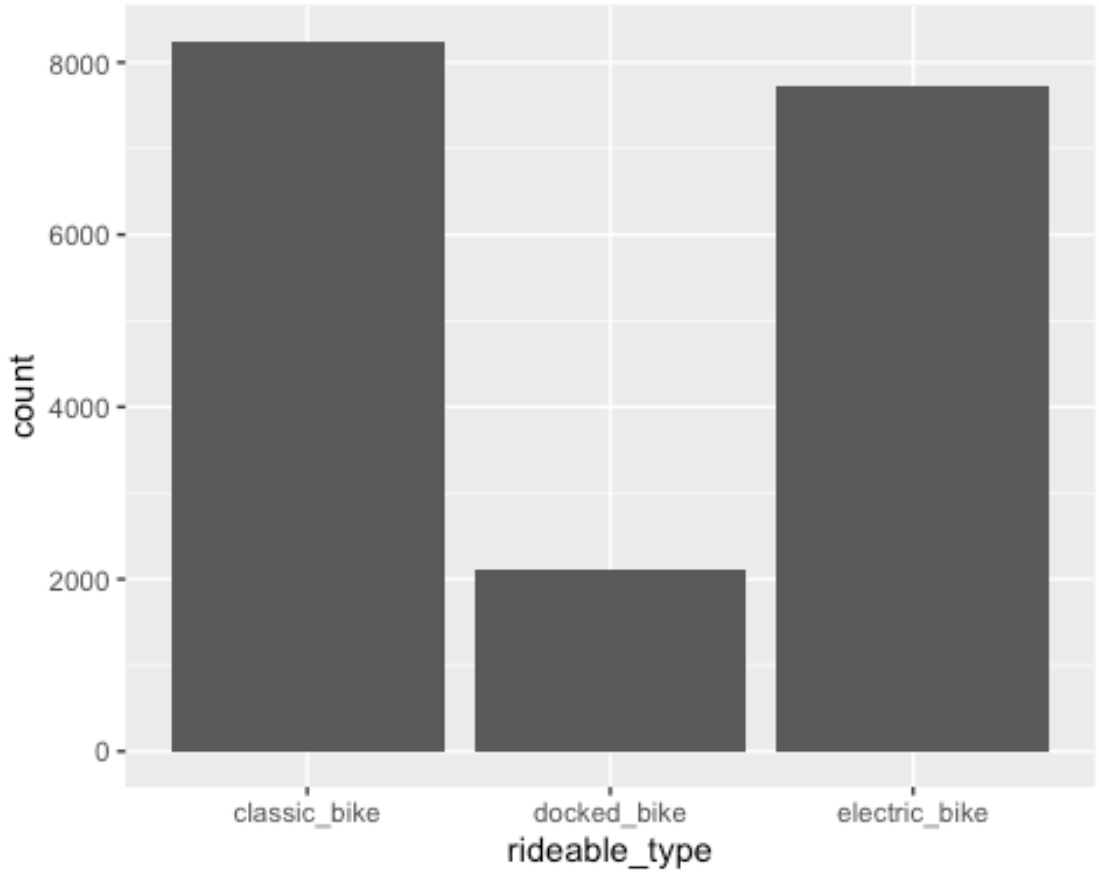
The longest trip for members is 89997 seconds - just over 1 day. The longest trip for casual riders is out outlier of 1189555 seconds, just under 2 weeks. The 2nd longest trip is 330578 seconds.

Most popular bike

There are 3 recorded bike types: electric, classic, and docked.

Casual Users

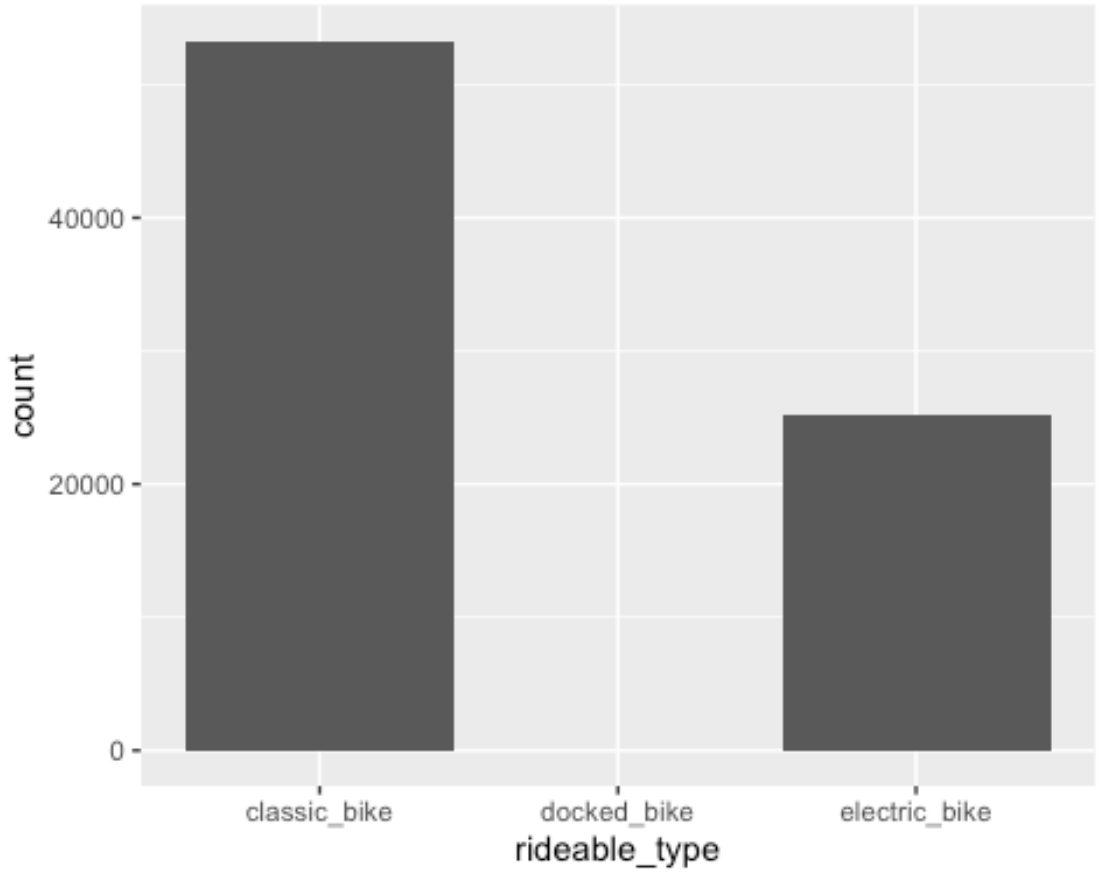
```
ggplot(data = Jan_casual, aes(x = rideable_type)) +
     geom_bar()
```



8254 rides use classic bikes. 7735 use electric bikes. 2101 trips with docked bikes.

Annual Members

```
ggplot(data = Jan_members, aes(x = rideable_type)) +
     geom_bar()
```



53307 trips were with the classic bikes. 25220 trips were with electric bikes. 1 trip was with a docked bike.

```
Jan_c_elec <- length(which(Jan_casual$rideable_type == "electric_bike"))
Jan_c_classic <- length(which(Jan_casual$rideable_type == "classic_bike"))
Jan_c_docked <- length(which(Jan_casual$rideable_type == "docked_bike"))

Jan_m_elec <- length(which(Jan_members$rideable_type == "electric_bike"))
Jan_m_classic <- length(which(Jan_members$rideable_type == "classic_bike"))
Jan_m_docked <- length(which(Jan_members$rideable_type == "docked_bike"))
```

February

In February we have 49424 entries after cleaning.

```
Feb_members <- Feb_filtered %>%
    filter(member_casual == "member")
Feb_casual <- Feb_filtered %>%
     filter(member_casual == "casual")
```

There are 10111 casual rides and 39313 rides by annual members - well over 3 times as much and nearly 4 times as much.

```
Feb_trip_avg <- (mean(Feb_filtered$trip_duration2))
Feb_m_trip_avg <- (mean(Feb_members$trip_duration2))
Feb_c_trip_avg <- (mean(Feb_casual$trip_duration2))
```

The average trip length for casual rides is 2968 seconds, 1086 seconds for members, and 1471 seconds overall. This is about 49 minutes on average for casual riders, 18 minutes for members, and 24 minutes on average for the month of February.

There are more very long trips with the two longest being nearly 3 weeks (2.99 and 2.97 weeks).

Max and Min

Before filtering, the shortest casual user's bike ride was 0 seconds. For members it was also 0 seconds. The longest bike ride for casual users was 1807754 seconds (around 2.99 weeks) and for members it was 89997seconds - once more just over a day.

```
Feb_c_trip_max <- (max(Feb_casual$trip_duration2))
Feb_m_trip_max <- (max(Feb_members$trip_duration2))
Feb_c_trip_min <- (min(Feb_casual$trip_duration2))
Feb_m_trip_min <- (min(Feb_members$trip_duration2))
```

Bike Type

```
Feb_c_elec <- length(which(Feb_casual$rideable_type == "electric_bike"))
Feb_c_classic <- length(which(Feb_casual$rideable_type == "classic_bike"))
Feb_c_docked <- length(which(Feb_casual$rideable_type == "docked_bike"))
```

For casual riders, 5685 trips were on classic bikes, 1268 on docked bikes, 3158 on electric bikes.

```
Feb_m_elec <- length(which(Feb_members$rideable_type == "electric_bike"))
Feb_m_classic <- length(which(Feb_members$rideable_type == "classic_bike"))
Feb_m_docked <- length(which(Feb_members$rideable_type == "docked_bike"))
```

For annual members, 29166 trips on classic bikes, 10147 trips on electric bikes. None were on docked bikes.

March

In March we have 228107 trips after cleaning.

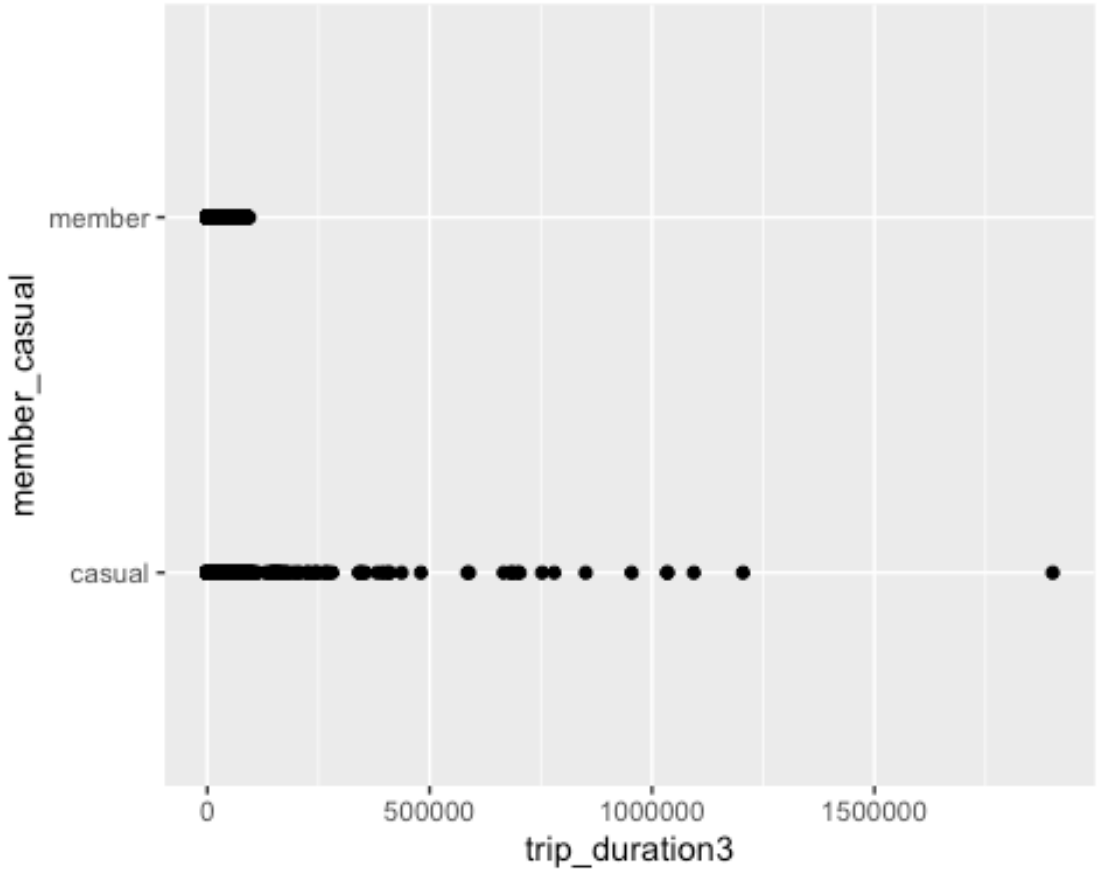
```
Mar_members <- Mar_filtered %>%
    filter(member_casual == "member")
Mar_casual <- Mar_filtered %>%
     filter(member_casual == "casual")
```

83958 casual rider trips. 144149 annual member trips. We are seeing that casual riders and annual members are taking trips at a much closer rate. Annual members had only 1.7 times as many trips than casual riders.

```
Mar_trip_avg <- (mean(Mar_filtered$trip_duration3))
Mar_m_trip_avg <- (mean(Mar_members$trip_duration3))
Mar_c_trip_avg <- (mean(Mar_casual$trip_duration3))
```

The average trip length in march was 1374 seconds, about 22 minutes. For casual riders, the average trip was 2291 seconds (38 minutes). For annual members, it was 840 seconds, only 14 minutes.

```
ggplot(data = Mar_filtered, aes(x = trip_duration3, y = member_casual)) +
    geom_point()
```



We see yet another very long trip, although there are more points that are very long this month than compared to the previous.

Max and Min

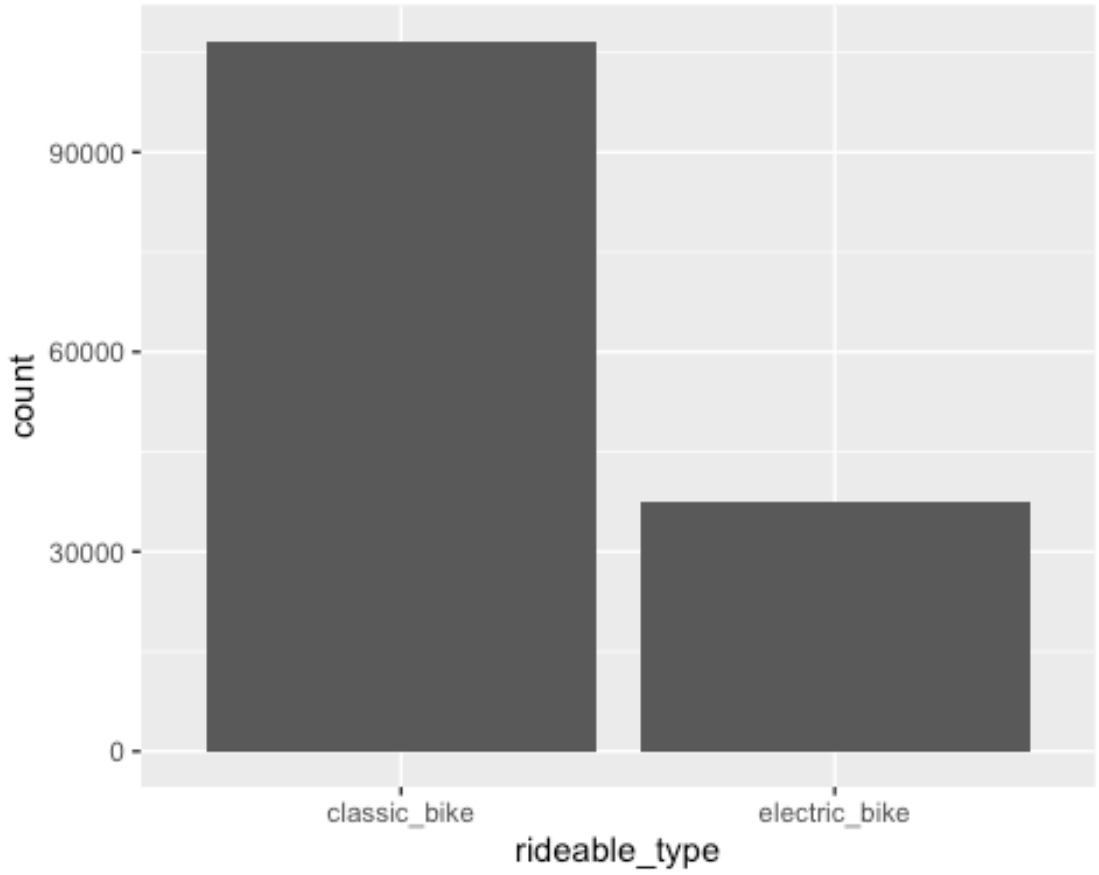
Before filtering, the shortest trip for members was -1 second, and 6 trips of 0 seconds. For casual riders the shortest trip was -1 seconds with 4 trips of 0 seconds afterwards. The longest bike ride for members was 93596 seconds or 1.08 days. For casual riders it was 1900899 seconds or over 3 weeks.

```
Mar_c_trip_max <- (max(Mar_casual$trip_duration3))
Mar_m_trip_max <- (max(Mar_members$trip_duration3))
```

```
Mar_c_trip_min <- (min(Mar_casual$trip_duration3))
Mar_m_trip_min <- (min(Mar_members$trip_duration3))
```

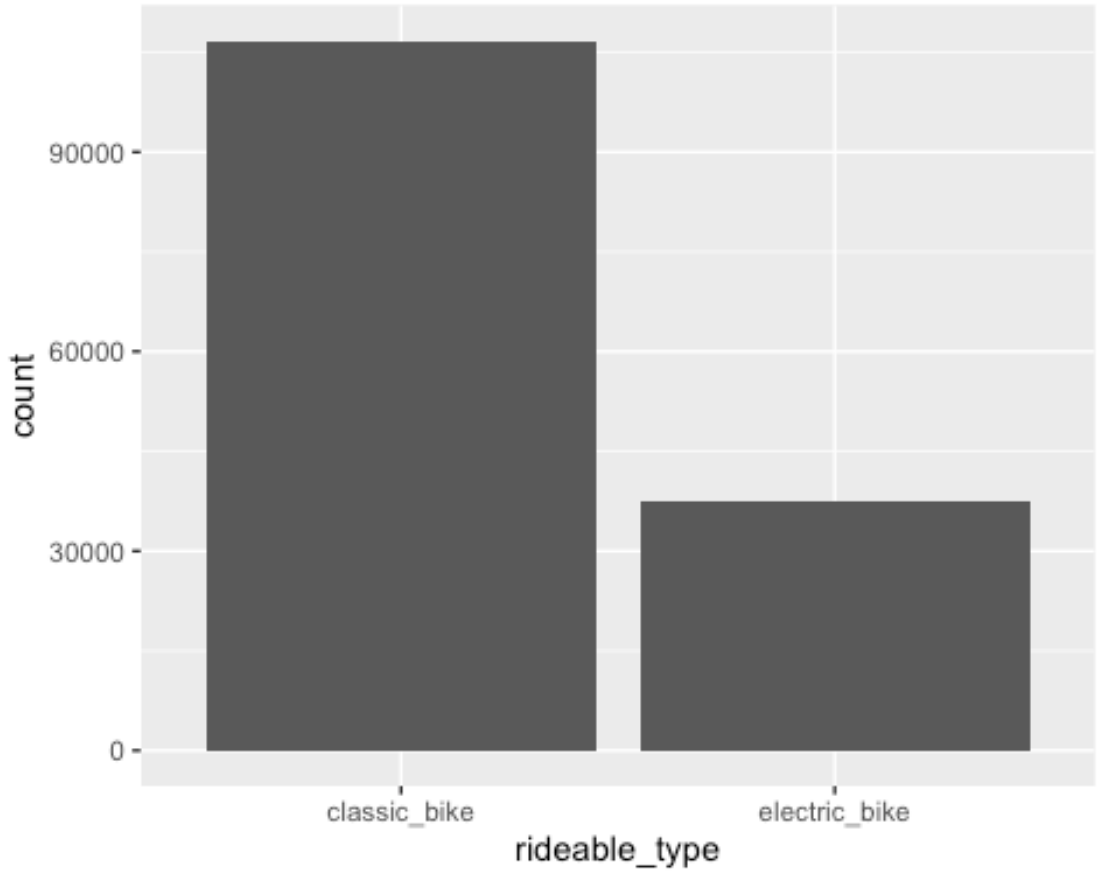
Bike Type

```
Mar_c_elec <- length(which(Mar_members$rideable_type == "electric_bike"))
Mar_c_classic <- length(which(Mar_members$rideable_type == "classic_bike"))
Mar_c_docked <- length(which(Mar_members$rideable_type == "docked_bike"))
ggplot(data = Mar_members, aes(x = rideable_type)) +
     geom_bar()
```



In March, 45492 trips by casual users were on classic bikes. 22821 trips were on electric bikes and 15645 trips were on docked bikes.

```
Mar_m_elec <- length(which(Mar_members$rideable_type == "electric_bike"))
Mar_m_classic <- length(which(Mar_members$rideable_type == "classic_bike"))
Mar_m_docked <- length(which(Mar_members$rideable_type == "docked_bike"))
ggplot(data = Mar_members, aes(x = rideable_type)) +
     geom_bar()
```



For Members, 106756 trips were on classic bikes, 37393 were on electric bikes, and 0 were on docked bikes. Consistently we see classic bikes being popular overall, but casual users are more likely to use docked bikes, despite this style becoming more popular overall.

April

```
Apr <- read_csv("202104-divvy-tripdata.csv")
Apr <- Apr %>%
  mutate(Apr, trip_duration4 = as.duration(ended_at - started_at))
Apr_filtered <- Apr %>%
    select(ride_id, rideable_type, started_at, ended_at, member_casual,
trip_duration4) %>%
    filter(trip_duration4 > 5)
```

In April we have, 336524 trips. 136433 trips by casual users, and 200091 trips by annual members. Annual members have 1.46 times as many trips as casual users.

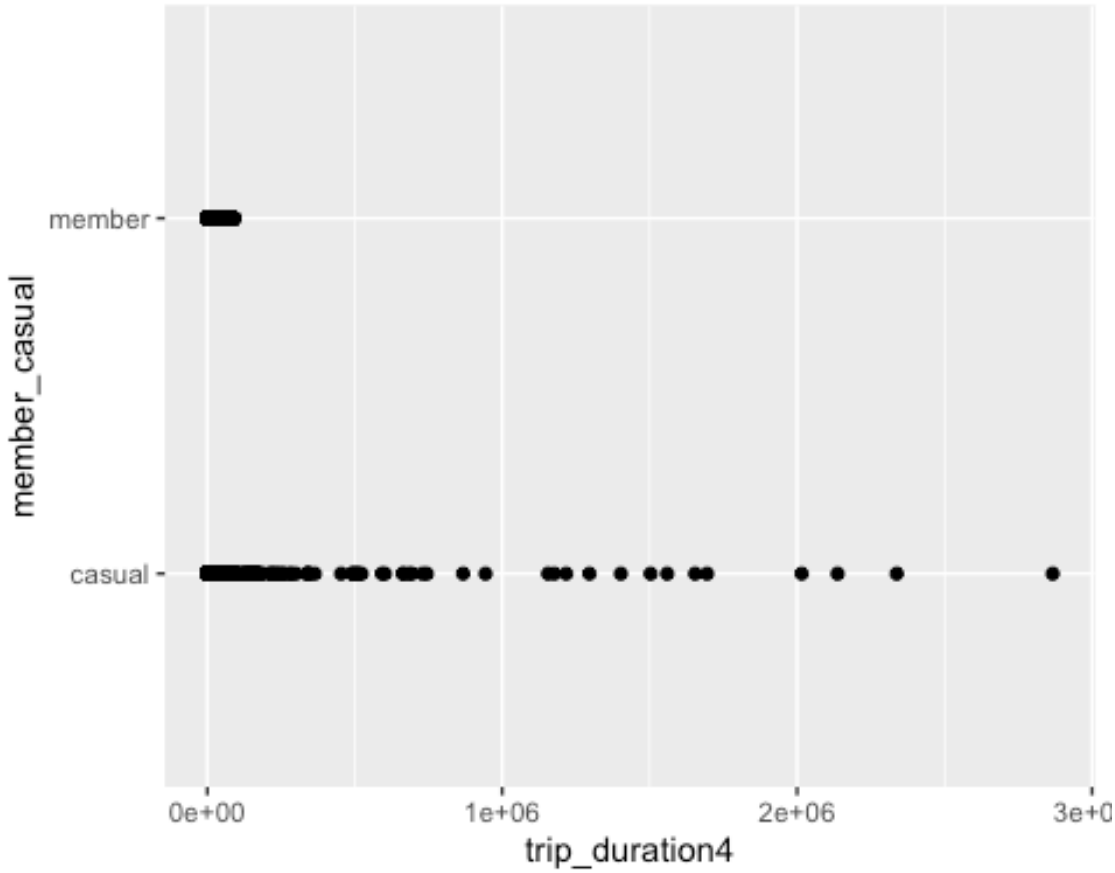
```
Apr_members <- Apr_filtered %>%
    filter(member_casual == "member")
Apr_casual <- Apr_filtered %>%
     filter(member_casual == "casual")
```

Trip Length

```
Apr_trip_avg <- (mean(Apr_filtered$trip_duration4))
Apr_m_trip_avg <- (mean(Apr_members$trip_duration4))
Apr_c_trip_avg <- (mean(Apr_casual$trip_duration4))
```

The average trip length for April was 1451 seconds, about 24 minutes. For members, the average trip length was 883 seconds, still around 14 minutes. For casual members we see the average trip is 2284 seconds, or about 38 minutes.

```
ggplot(data = Apr_filtered, aes(x = trip_duration4, y = member_casual)) +
    geom_point()
```



Max and Min

Before filtering, the shortest trip for annual members was -132 seconds. There were also trips of -10, -9, -7, and -2 seconds, and 22 trips of 0 seconds. For casual members, the shortest trip was of 0 seconds; there were 11 trips of 0 seconds for casual users.

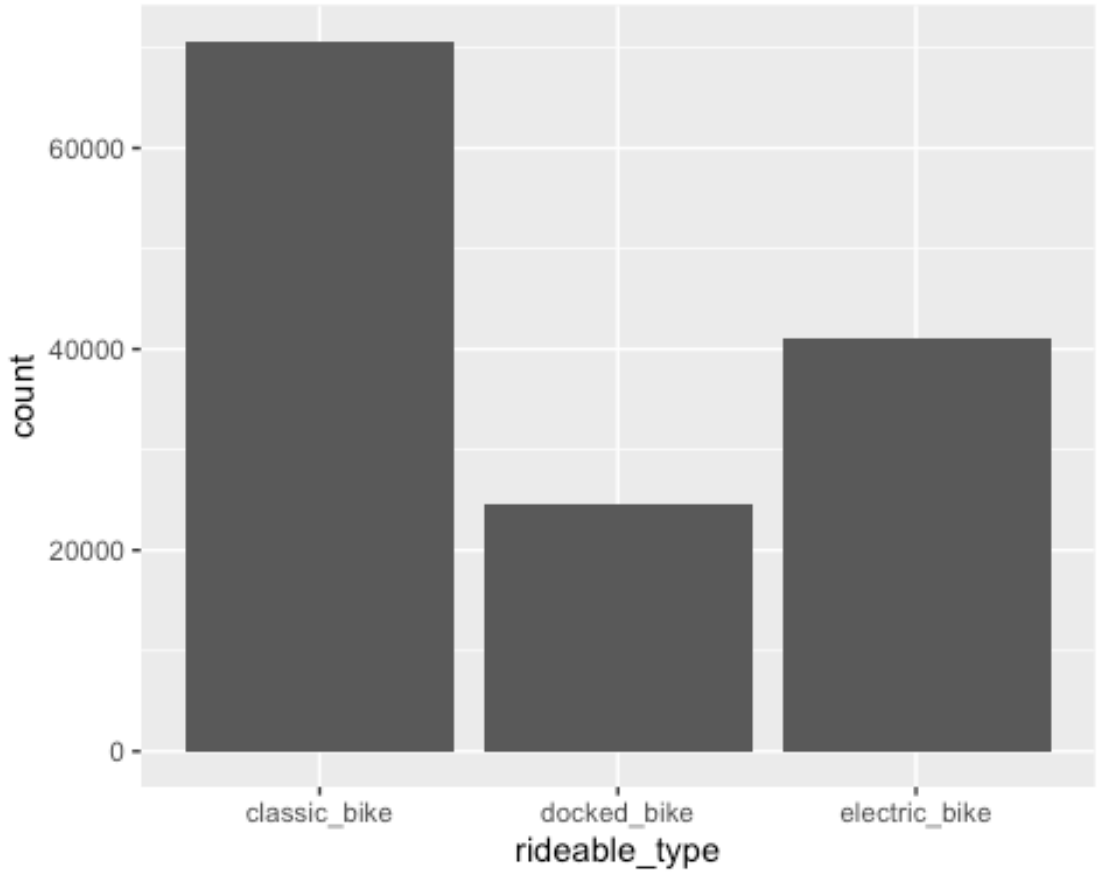
The longest trip for members was 89996 seconds, or just over a day. For casual riders, the longest trip was 2866602 seconds, or 4.74 weeks.

```
Apr_c_trip_max <- (max(Apr_casual$trip_duration4))
Apr_m_trip_max <- (max(Apr_members$trip_duration4))
```

```
Apr_c_trip_min <- (min(Apr_casual$trip_duration4))
Apr_m_trip_min <- (min(Apr_members$trip_duration4))
```

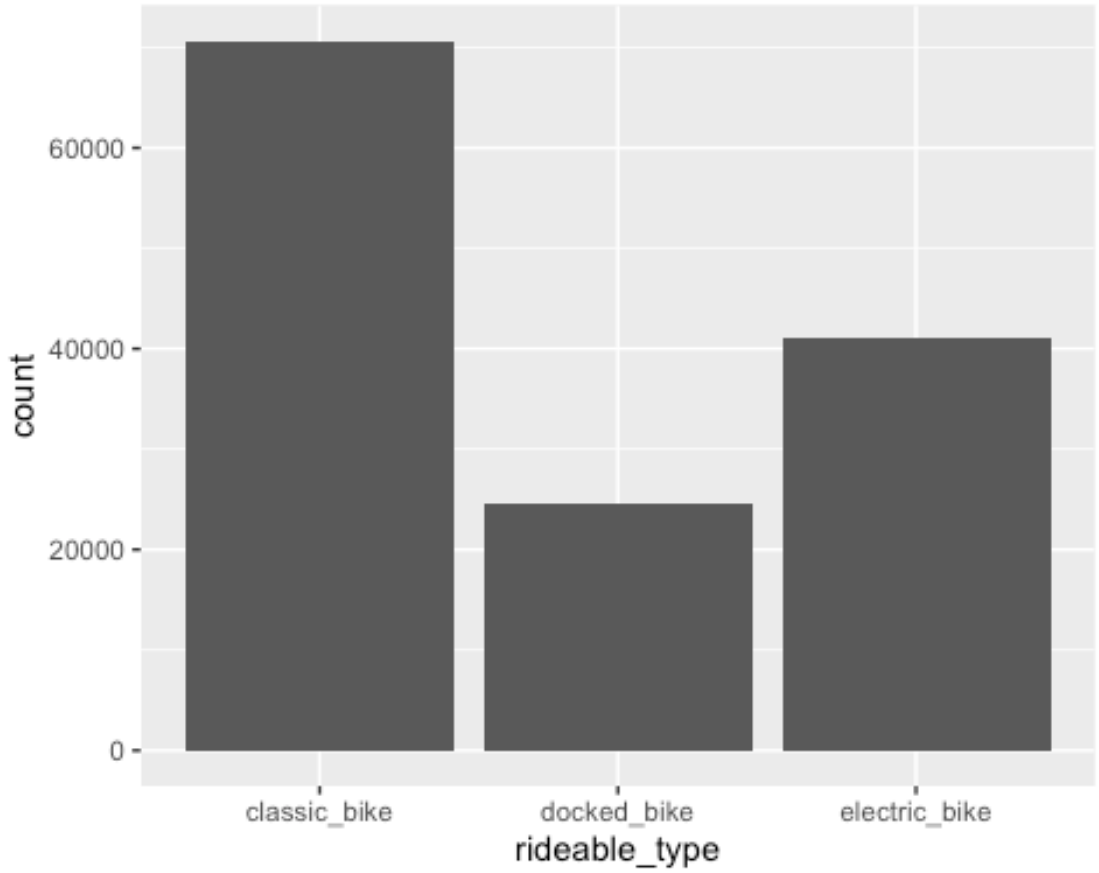
Bike Type

```
Apr_c_elec <- length(which(Apr_casual$rideable_type == "electric_bike"))
Apr_c_classic <- length(which(Apr_casual$rideable_type == "classic_bike"))
Apr_c_docked <- length(which(Apr_casual$rideable_type == "docked_bike"))
ggplot(data = Apr_casual, aes(x = rideable_type)) +
     geom_bar()
```



Casual riders took 70691 trips on classic bikes, 24693 trips on docked bikes, and 41049 trips on electric bikes.

```
Apr_m_elec <- length(which(Apr_members$rideable_type == "electric_bike"))
Apr_m_classic <- length(which(Apr_members$rideable_type == "classic_bike"))
Apr_m_docked <- length(which(Apr_members$rideable_type == "docked_bike"))
ggplot(data = Apr_casual, aes(x = rideable_type)) +
     geom_bar()
```



Annual members took 143444 trips on classic bikes, 56647 trips on electric bikes, and none on docked bikes.

These resources helped me along the way: * https://datacornering.com/filter-in-r/ * https://www.statology.org/remove-columns-in-r/ * https://www.statology.org/remove-rows-in-r/ * https://r4stats.com/examples/graphics-ggplot2/ * https://stackoverflow.com/questions/28195996/count-number-of-rows-matching-a-criteria