Case Study 1: Selling Annual Bike Subscriptions

Garian Rice

10/7/2022

Introduction

Although I'm eager to use data of personal interest to answer my own questions about the world, like what's the key to happiness, I've decided to save these passions for next time and first follow two guided case studies offered in the Google Data Analytics certificate.

The advantage is that I'll improve familiarity with the tools and stages of the data analysis process: ask, prepare, process, analyze, share, and act, while not being bogged down by technical difficulties, and begin developing my professional portfolio.

Although these case studies have a specific question to answer and a roadmap with guided questions and objectives, all the details have been provided by me, Garian Rice. All the work you see is my own unless otherwise specified and reflects my own skills, abilities and knowledge.

The first scenario is as follows:

> You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Ask

Before we begin analyzing the data, there are five questions we should answer to consider the scope, audience and objectives of the project.

1. What's the problem?
- How can we maximize profit by converting casual riders into annual members.

2. What key factors are involved?
- How annual members and casual riders' bike usage differs.

3. Who is the audience?

- The executive team at Cyclistic who will approve of our proposed strategy.

4. What type of data is needed?

- Historical trip data from the past 12 months, so we can identify trends and ensure our analysis is up-to-date.

5. How will the data be obtained?

- Data is gathered automatically using geotracking services on each of the 5878 bicycle leased by the company Cyclistic. Historical data has been uploaded to Amazon Web Services, contained in CSV files exported from the company database that can be found here.

To summarize, our business task is to *develop a marketing strategy to convert annual members into casual members* and *present our solutions to the executive team.*

Prepare

This data is available for public use by Motivate International Inc (i.e. Cyclistic) according to this license. Sensitive personal and financial data has been omitted to respect privacy. Once again the data can be found here. It contains several ZIP files organized in different ways. There is quarterly data that could be merged together, but it's from 2020 so it's outdated. To get the most recent data from September 2022 and the 12 months before, I've stitched together 12 CSV files using `cat *.csv >combined.csv`in the Mac terminal. The result is saved as `combined.csv`in `data-copy` which is a back-up of the raw data.

Process

All data must be processed, or *cleaned* before analysis. I've chosen to use R for this because R handles all stages of the analysis process, including cleaning, visualization and this presentation. However SQL may be more efficient at processing such a large dataset. Some of the upcoming operations take a **long** time to load. Using Excel or Sheets would be tedious unless you had current quarterly data.

After importing the data, some columns such as `end_station_id` and `end_lat` are removed because they're irrelevant to the business task. Here's a quick look at the structure data:

```
# Importing and viewing data
tripdata <- read.csv("data-copy/combined.csv")

# Removing unnecessary columns
tripdata=subset(tripdata,select=-c(ride_id,start_station_id,end_station_id,start_lat,start_lng,end_lat,

# Viewing the structure of the data
str(tripdata)
```

```
## 'data.frame':    5883054 obs. of  6 variables:
##  $ rideable_type     : chr  "electric_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2022-04-06 17:42:48" "2022-04-24 19:23:07" "2022-04-20 19:29:08" "2022-0
##  $ ended_at          : chr  "2022-04-06 17:54:36" "2022-04-24 19:43:17" "2022-04-20 19:35:16" "2022-0
##  $ start_station_name: chr  "Paulina St & Howard St" "Wentworth Ave & Cermak Rd" "Halsted St & Polk S
##  $ end_station_name  : chr  "University Library (NU)" "Green St & Madison St" "Green St & Madison St
##  $ member_casual     : chr  "member" "member" "member" "casual" ...
```

Notice the `started_at` and `ended_at` columns are type `<chr>` we want to convert them to datetime format also known as POSIXlt.

```
# Convert start/end times to datetime (POSIXlt) format
tripdata$started_at <- strptime(tripdata$started_at,format="%Y-%m-%d %H:%M:%OS")
tripdata$ended_at <- strptime(tripdata$ended_at,format="%Y-%m-%d %H:%M:%OS")
```

This enables us to construct two new columns, one for the trip length in minutes and one for the weekday the trip started.

```
library(lubridate) # Dates and times made easy
```

```
# Creating a new column called ride length (in minutes)
tripdata$ride_length <- round(difftime(tripdata$ended_at,tripdata$started_at,units=c("mins")),digits=1)

# Convert to numeric
tripdata$ride_length <- as.numeric(as.character(tripdata$ride_length))

# Creating a column with days of the week trips started
tripdata$day_of_week <- weekdays(tripdata$started_at)
head(tripdata[c("ride_length","day_of_week")])
```

```
##   ride_length day_of_week
## 1        11.8   Wednesday
## 2        20.2      Sunday
## 3         6.1   Wednesday
## 4         9.4      Friday
## 5         5.7    Saturday
## 6         4.3    Thursday
```

Using the `unique()` function, we spot that there is an erroneous third category called "member_casual" with 11 observations. Since they're missing start and end times, we can safely remove them.

```
unique(tripdata$member_casual)
```

```
## [1] "member"        "casual"        "member_casual"
```

```
tripdata <- subset(tripdata, member_casual!="member_casual")
unique(tripdata$member_casual)
```

```
## [1] "member" "casual"
```

Now there are two categories, casual riders and members. Let's clean up any remaining missing values using `na.omit()` and save it to a backup dataframe.

```
# Remove NA
tripdata2 <- na.omit(tripdata)
```

Finally, let's reformat `days_of_the_week` to be a factor ordered from Sunday to Monday. This will make our visualizations more coherent.

```
tripdata2$day_of_week <- factor(tripdata2$day_of_week, levels= c("Sunday", "Monday",
    "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
```

Analyze

Now that the data has been processed into a single file, filtered for relevant rows, cleaned of missing/erroneous values and formatted correctly, it's time to start putting it to work. We'll compare members and casual users to analyze trends and relationships to help us convert casuals into members. First, let's conduct a five-number summary of `ride_length`.

```
summary(tripdata2$ride_length)
```

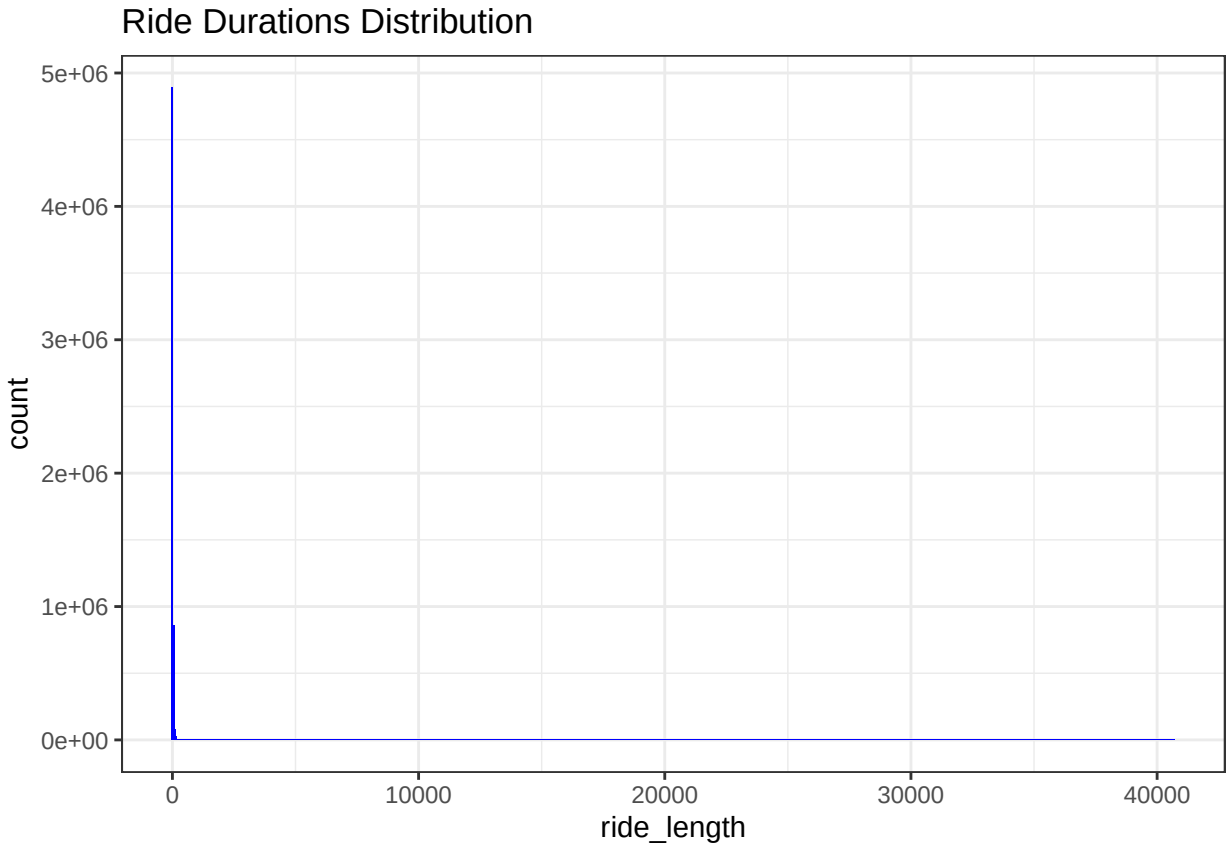
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  -137.40     6.00    10.70    19.75    19.30 40705.00
```

Uh-oh. The maximum trip duration is 40,705 minutes or 28 days and the minimum is negative. But I recall the project director saying the negatives can be removed since they represent inspections and tests. Here's a

histogram showing the new distribution of trip duration.

```
# Remove trips with durations less than 0
tripdata2 <- subset(tripdata2, ride_length>0)
```

```
library(ggplot2) # Data visualization package
# Inspect distribution of ride lengths
ggplot(tripdata2,aes(ride_length))+
  geom_histogram(binwidth=50,fill="blue")+
  labs(title="Ride Durations Distribution",face="bold")+
  theme_bw()
```



Yeah, that's probably the most skewed-left graph I've ever seen. Interestingly, even with the extreme outlier, the average trip duration is about `19.75` minutes. In fact, let's see if there's a difference between casuals and members.

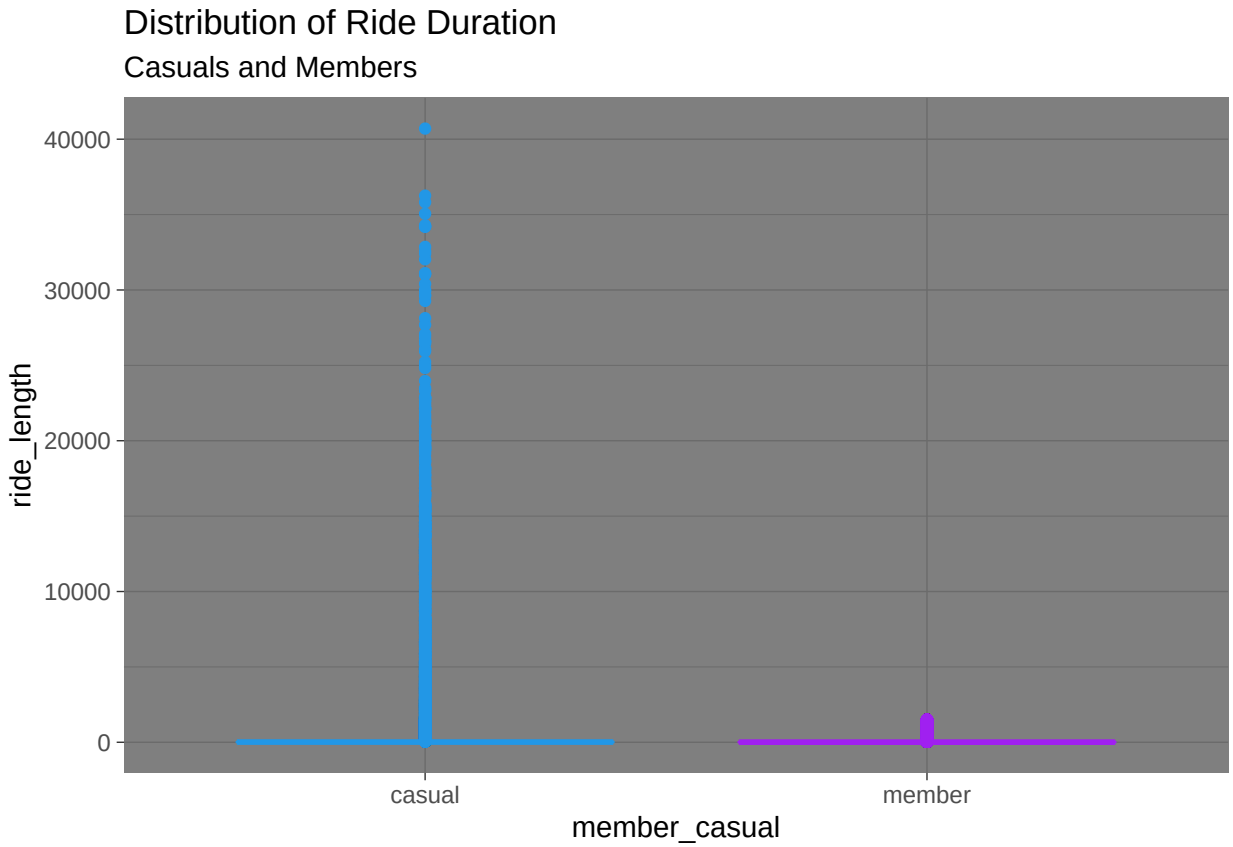
```
library(dplyr) # Pipe operator
```

```
tripdata2 %>%
  group_by(member_casual) %>%
  summarize(mean(ride_length),sd(ride_length))
```

```
## # A tibble: 2 x 3
##   member_casual `mean(ride_length)` `sd(ride_length)`
##   <chr>                       <dbl>             <dbl>
## 1 casual                       29.3             245.
## 2 member                       12.9              27.7
```

Wow! The standard deviation for casuals is 245 while that of members is only 28. Let's take a closer look:

```
ggplot(tripdata2,aes(member_casual,ride_length))+
         geom_boxplot(color=c("009900","purple"))+
  theme_dark()+
  labs(title="Distribution of Ride Duration",
       subtitle="Casuals and Members")
```

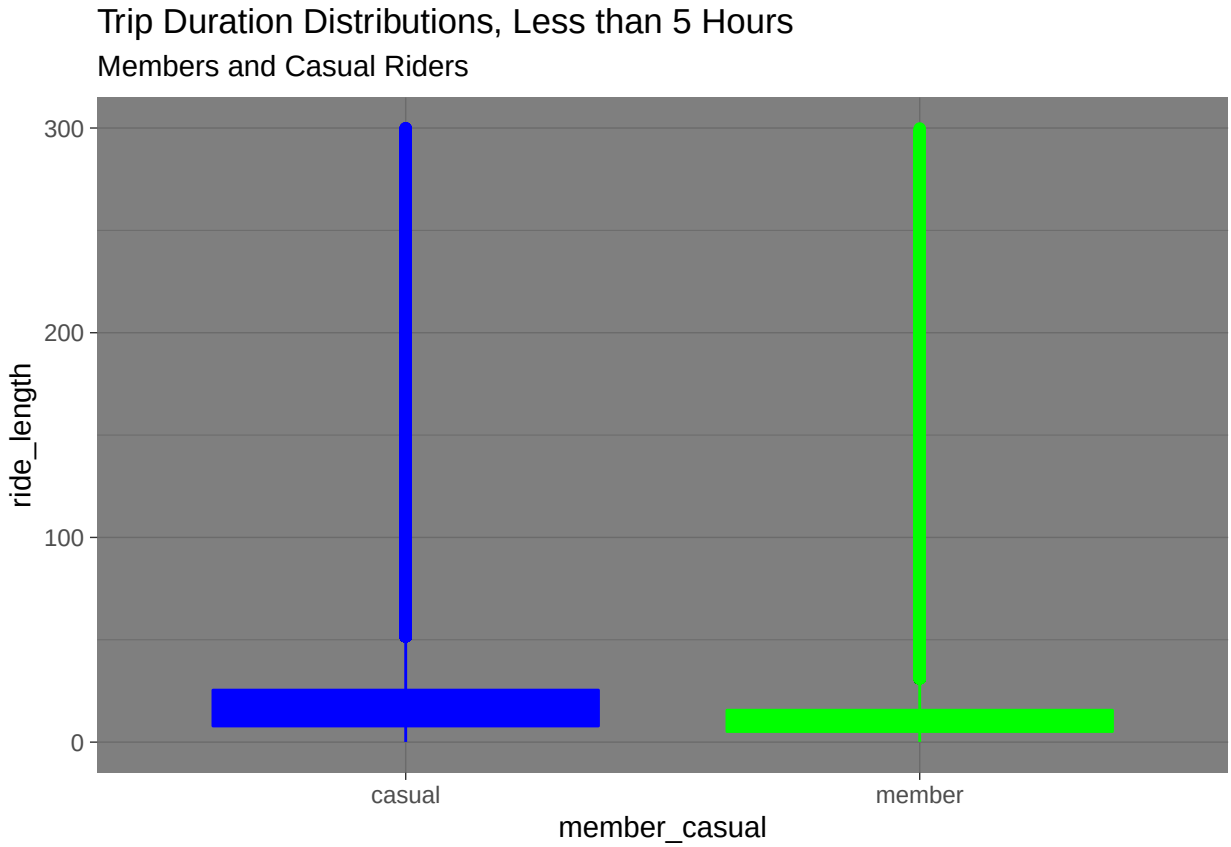


At this point, I seek out the advice of my colleagues. After meeting with the project manager, I learn that the bikes are never leased out for more than 5 hours, so any higher values are erroneous. We submitted these findings to the engineering department so they can review their data validation techniques and proceed as follows.

Let's take a look at the new distributions with trip lengths greater than 5 hours removed.

```
# Remove trips with durations greater than 300 minutes (5 hours)
tripdata2 <- subset(tripdata2, ride_length<300)
```

```
ggplot(tripdata2,aes(member_casual,ride_length))+
  geom_boxplot(fill=c("blue","green"),color=c("blue","green"))+
  theme_dark()+
  labs(title="Trip Duration Distributions, Less than 5 Hours",
       subtitle="Members and Casual Riders")
```



Both distributions are still skewed left. To get a better idea, let's summarize the average trip length and standard deviation grouped by members/casuals again.

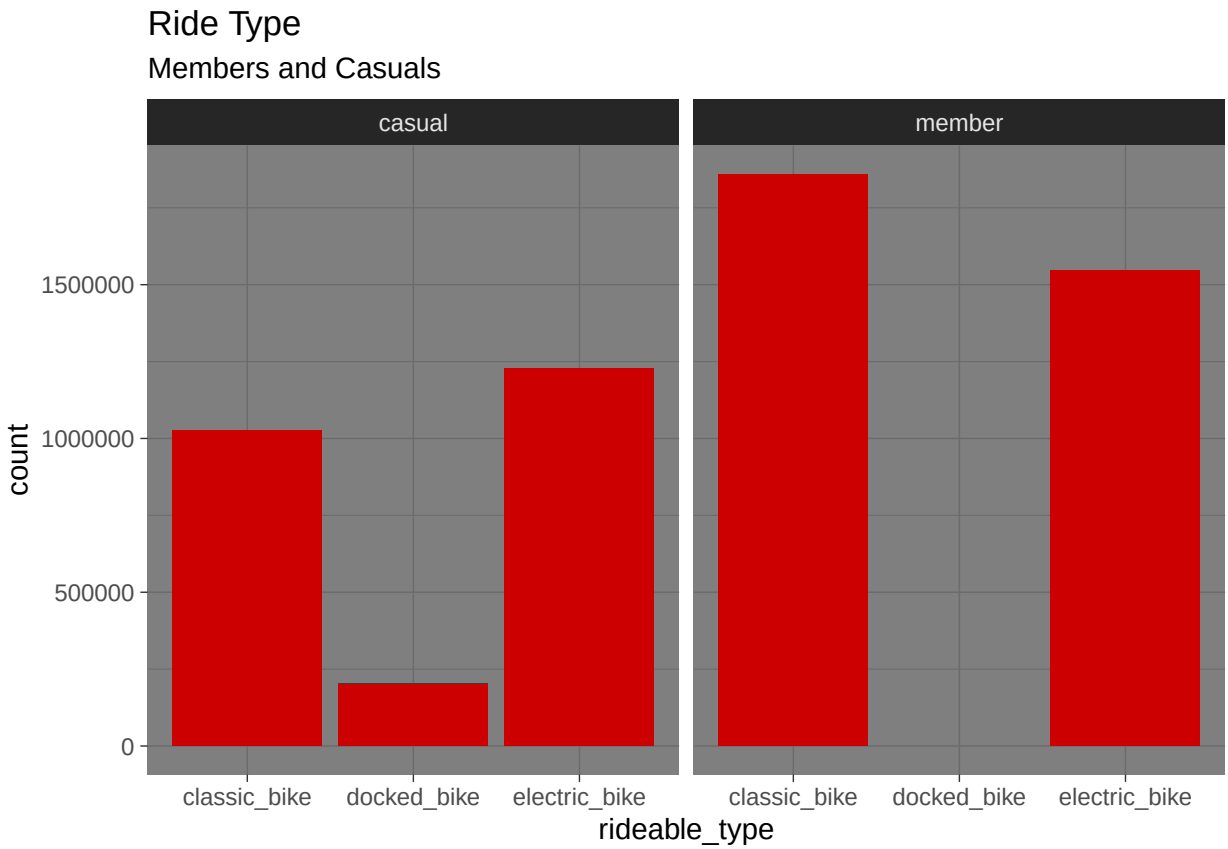
```
tripdata2 %>%
  group_by(member_casual) %>%
  summarize(mean(ride_length),sd(ride_length))
```

```
## # A tibble: 2 x 3
##   member_casual `mean(ride_length)` `sd(ride_length)`
##   <chr>                       <dbl>             <dbl>
## 1 casual                       21.8              25.5
## 2 member                       12.3              12.1
```

These findings show that there's probably not a significant difference between the average trip duration of casuals and members. One might wish to prove this by a two-sample hypothesis test but the skewness of the data could cause difficulties.
However, the skewness does provide an important insight. Consider the possibility that members tend to rent the bikes for work while casual riders tend to use them for longer, causla day-trips with friends and family. If this were the case, this would suggest we could convert more casuals into members by employing membership marketing strategies that appeal to customers who take longer trips.

Let's look for other differences. Let's see if casuals and members differ in the type of bike they prefer to use.

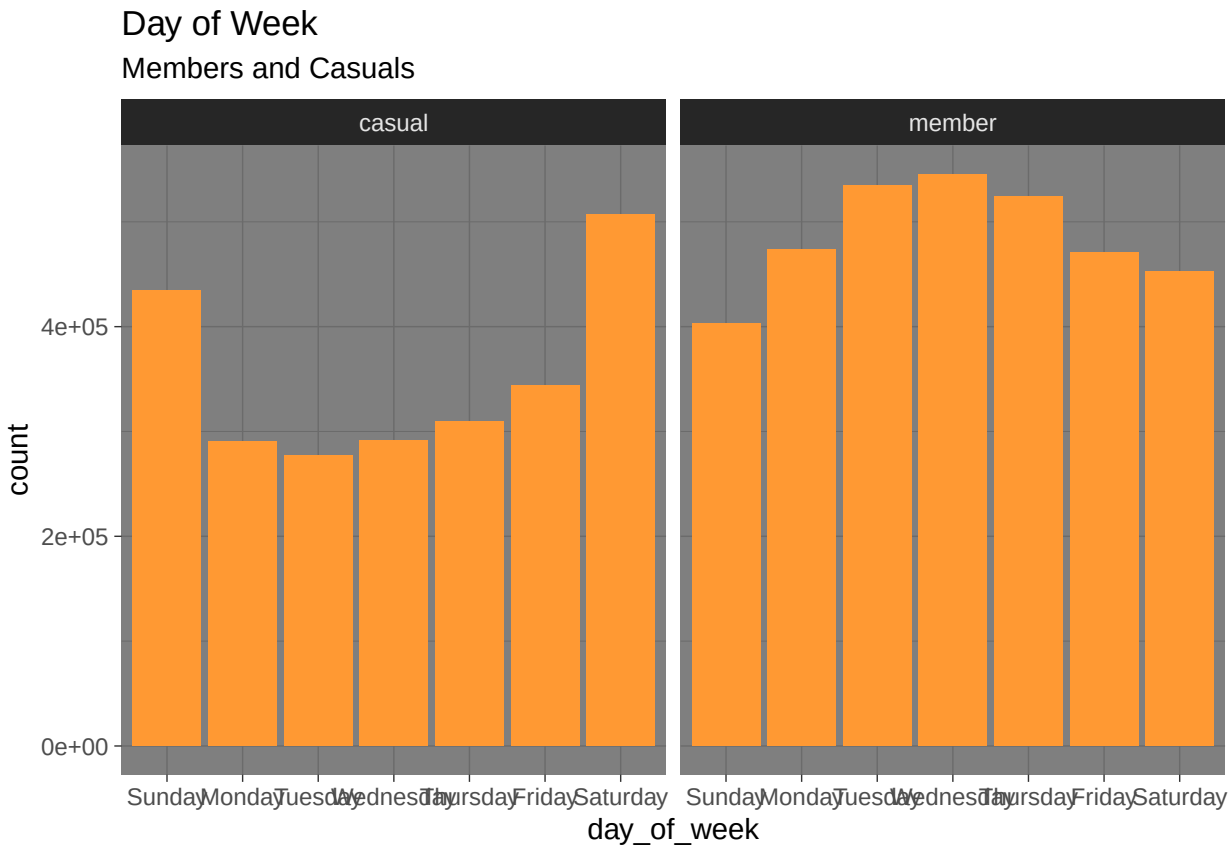
```
ggplot(tripdata2,aes(rideable_type))+
  geom_bar(fill="#CC0000")+
  labs(title="Ride Type",subtitle="Members and Casuals")+
  theme_dark()+
  facet_wrap(~member_casual)
```



Not much of a difference. Members seem to use the classic bike a bit more than the electric bike and don't use the docked bike at all. Casuals seem to use the docked bikes, and use electric bikes slightly more than classic bikes. Perhaps we can incentivize these options in marketing campaigns.

Finally, let's inspect what days of the week both customer groups tend to use.

```
ggplot(tripdata2,aes(day_of_week))+
  geom_bar(fill="#FF9933")+
  labs(title="Day of Week",subtitle="Members and Casuals")+
  theme_dark()+
  facet_wrap(~member_casual)
```



This graph reveals that casual users take bike trips more often on the weekend, Saturday and Sunday, while members' bike trips peak in the middle of the week on Wednesday. This suggests when we should try marketing membership passes to casuals. In the future we could also see whether the month would have an effect on a customer's willingness to buy a membership.

Summary

We conclude this analysis with a summmary of important findings:
1. We can't say casual users on average take longer trips than members, but there may be a niche market for longer trips due to a heavy presence of outliers (casuals with long trip times).
2. Some casual users use docked bikes, though members very rarely do. Moreover, casuals use electric bikes slightly more than classic bikes. Members use classical bikes the most.
3. Casual users take more trips on the weekends Saturday and Sunday, while members take more trips throughout the middle of the week.

Share

A dashboard summarizing these key findings has been created using Tableau and included in the main project directory with the name `capstone-viz.twb`.

Act

Thanks to this analysis, the data engineering team was able to identify a bug in its data validation techniques. Then the executive team reviewed our findings and used them to develop a new marketing strategy, dedicating more advertisements on Saturday and Sunday, with docked, classic and electric bike options, and offering special discounts for casual users who want to upgrade and take longer trips, maybe with their family or

friends.
This resulted in 10% increase in casual user conversion rate compared to last year.
In the future we'd like to learn more about what else compels casual riders to buy annual memberships, and how digital media can be used to help influence this decision.

Final Thoughts

I hope you enjoyed this case study. I tried to keep it short and simple so I'd have experience completing all steps of a project and a foundation to build upon, but overall it was more challenging than I expected. The dataframe is so large so I'd often have to do scratchwork on a smaller custom-made dataframe to see if I'd get the expected result before implementing it. Waiting for the code to compile was a difficulty caused by the way the data was processed.
Nevertheless, seeing this project through from beginning to end and codifying the different stages of the analysis has given me the confidence to complete more projects in the future. This took me about two days to complete, which is a typical deadline for a case study given by an employer.
Thanks for reading.