Google Capstone Project: Cyclistic Bike Company

Scenario

I am 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, the team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, the team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve my recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Ask

Three questions will guide the future marketing program:

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

2. Why would casual riders buy Cyclistic annual memberships?

3. How can Cyclistic use digital media to influence casual riders to become members?

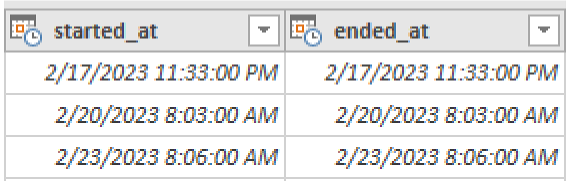
Prepare and Process

I used Cyclistic's historical trip data to analyze and identify trends. I did this by downloading the previous 12 months of Cyclistic trip data. Once the 12 Excel files were downloaded, I used power query in Excel to upload and combine each xlsx file. Since the files were very large, it was important to clean missing or incorrect data as much as possible. In power query, the sorting and filtering tools helped when identifying the holes in the data, since combing through the spreadsheet would have taken far too long given the 100k+ rows for each file. In Excel, once the data was combined and cleaned, I could make a connection and create pivot tables as shown below.

I also used R in this project since it is a better tool for statistical analysis, and I also wanted to see if there were trends that I could find in R that I didn't see in Excel. The prepare process was similar to Excel by assigning each file to a data frame, and combining those data frames into one. Then I used the Janitor RStudio Package to assist with cleaning missing or irrelevant data.

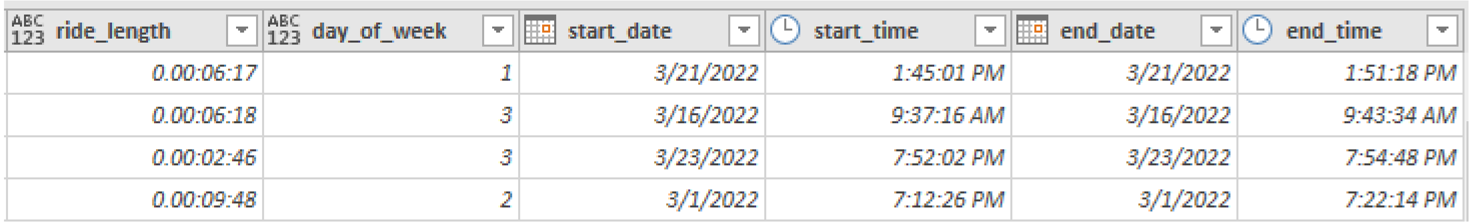
Analyze and Share

To start my analysis, I needed to create a ride length column, a day of the week column, and a start time/end time column. In power query, I separated the “started_at” and “ended_at” since the format combined the date and the time:



started_at	ended_at
2/17/2023 11:33:00 PM	2/17/2023 11:33:00 PM
2/20/2023 8:03:00 AM	2/20/2023 8:03:00 AM
2/23/2023 8:06:00 AM	2/23/2023 8:06:00 AM

I had to separate these columns to create an individual column for start time, end time, start date, and end date. I then used these new columns, along with a formula, to calculate the ride length column.

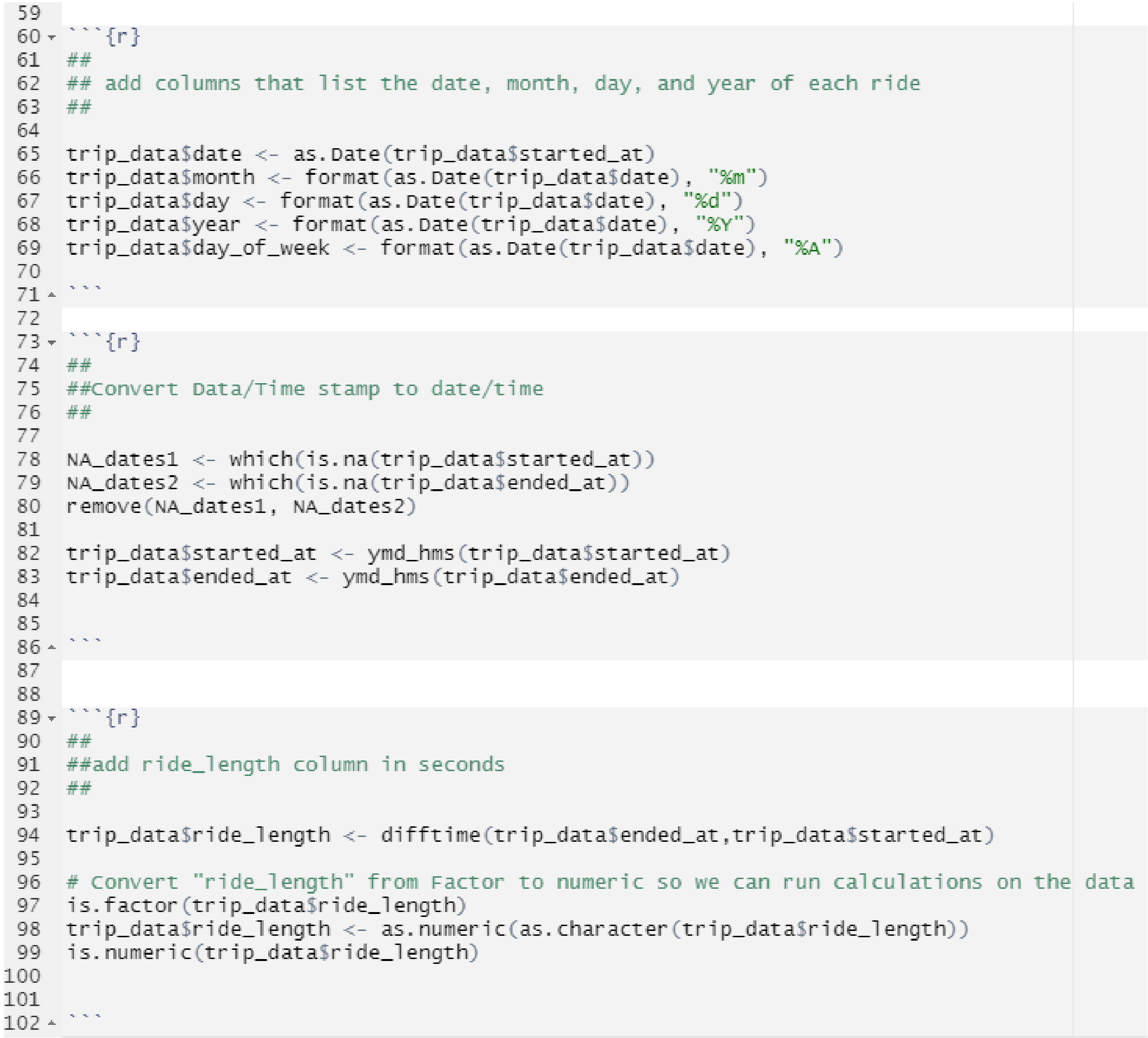


ride_length	day_of_week	start_date	start_time	end_date	end_time
0.00:06:17	1	3/21/2022	1:45:01 PM	3/21/2022	1:51:18 PM
0.00:06:18	3	3/16/2022	9:37:16 AM	3/16/2022	9:43:34 AM
0.00:02:46	3	3/23/2022	7:52:02 PM	3/23/2022	7:54:48 PM
0.00:09:48	2	3/1/2022	7:12:26 PM	3/1/2022	7:22:14 PM

For day of the week column, I used Date.DayOfWeek([started_at], 0)), which sets 0 as the first day of the week (Sunday) based on the start_date column.

Using the pivot tables I created, I chose to visualize the data using Pivot Charts that are shown below.

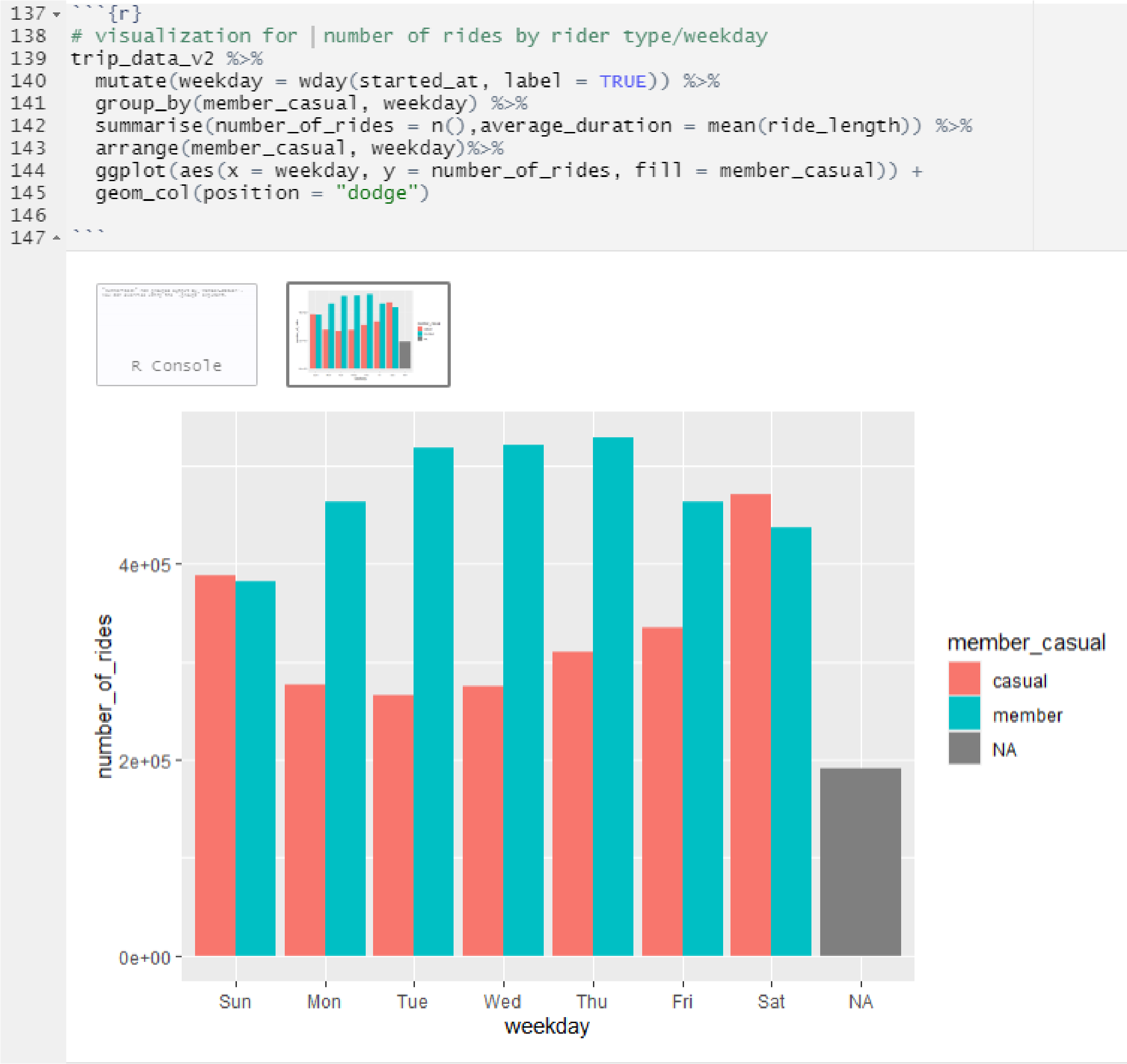
In RStudio, I chose to find and visualize the average duration of each ride, and the number of rider by rider type per weekday. To do this, I needed to separate the started at and ended at column again. The code for the that is below:



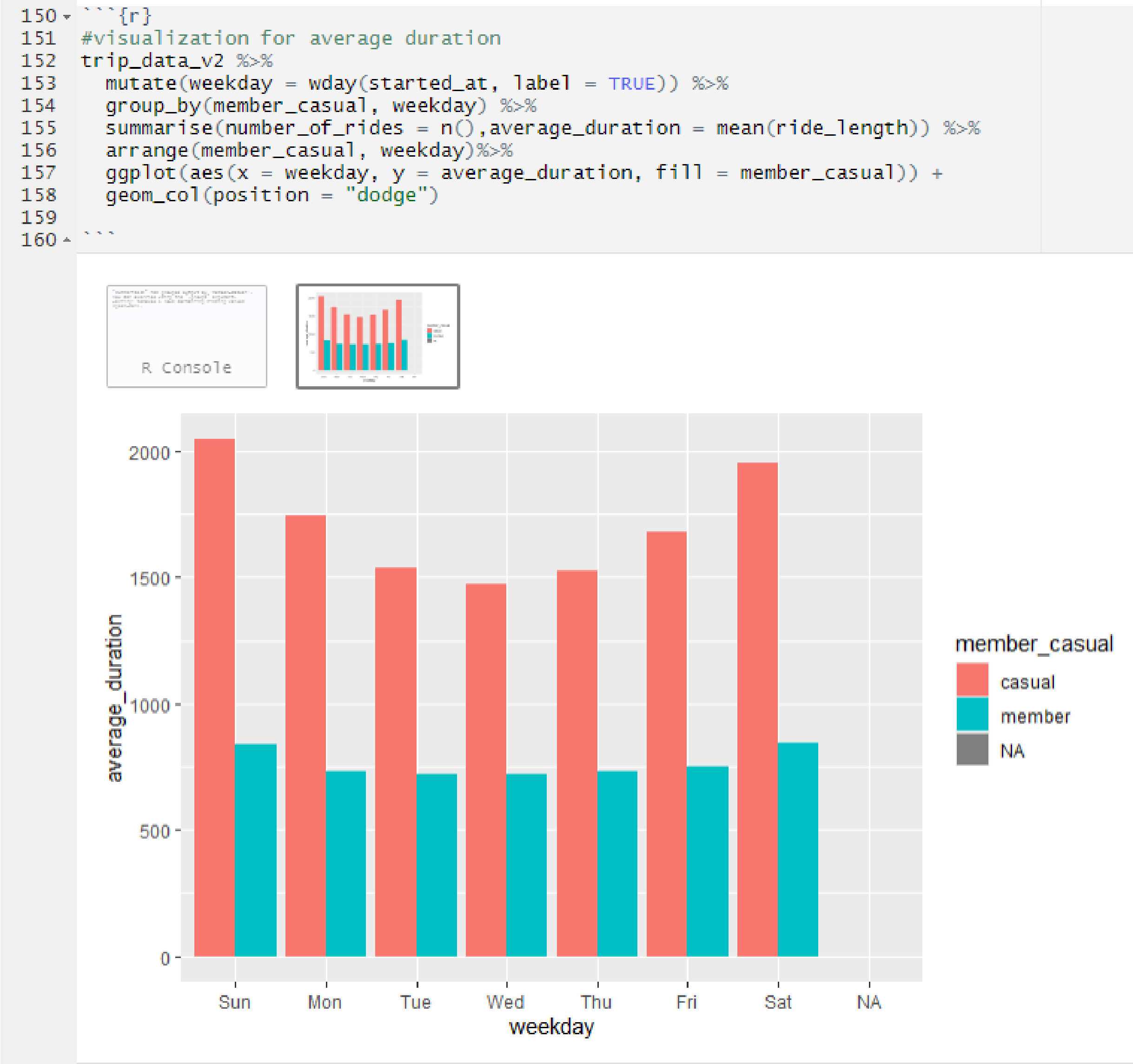
~~~
```{r}
##
## add columns that list the date, month, day, and year of each ride
##

trip_data$date <- as.Date(trip_data$started_at)
trip_data$month <- format(as.Date(trip_data$date), "%m")
trip_data$day <- format(as.Date(trip_data$date), "%d")
trip_data$year <- format(as.Date(trip_data$date), "%Y")
trip_data$day_of_week <- format(as.Date(trip_data$date), "%A")

~~~

```{r}
##
##Convert Data/Time stamp to date/time
##

NA_dates1 <- which(is.na(trip_data$started_at))
NA_dates2 <- which(is.na(trip_data$ended_at))
remove(NA_dates1, NA_dates2)

trip_data$started_at <- ymd_hms(trip_data$started_at)
trip_data$ended_at <- ymd_hms(trip_data$ended_at)


```

```{r}
##
##add ride_length column in seconds
##

trip_data$ride_length <- difftime(trip_data$ended_at,trip_data$started_at)

# Convert "ride_length" from Factor to numeric so we can run calculations on the data
is.factor(trip_data$ride_length)
trip_data$ride_length <- as.numeric(as.character(trip_data$ride_length))
is.numeric(trip_data$ride_length)


```
```
```

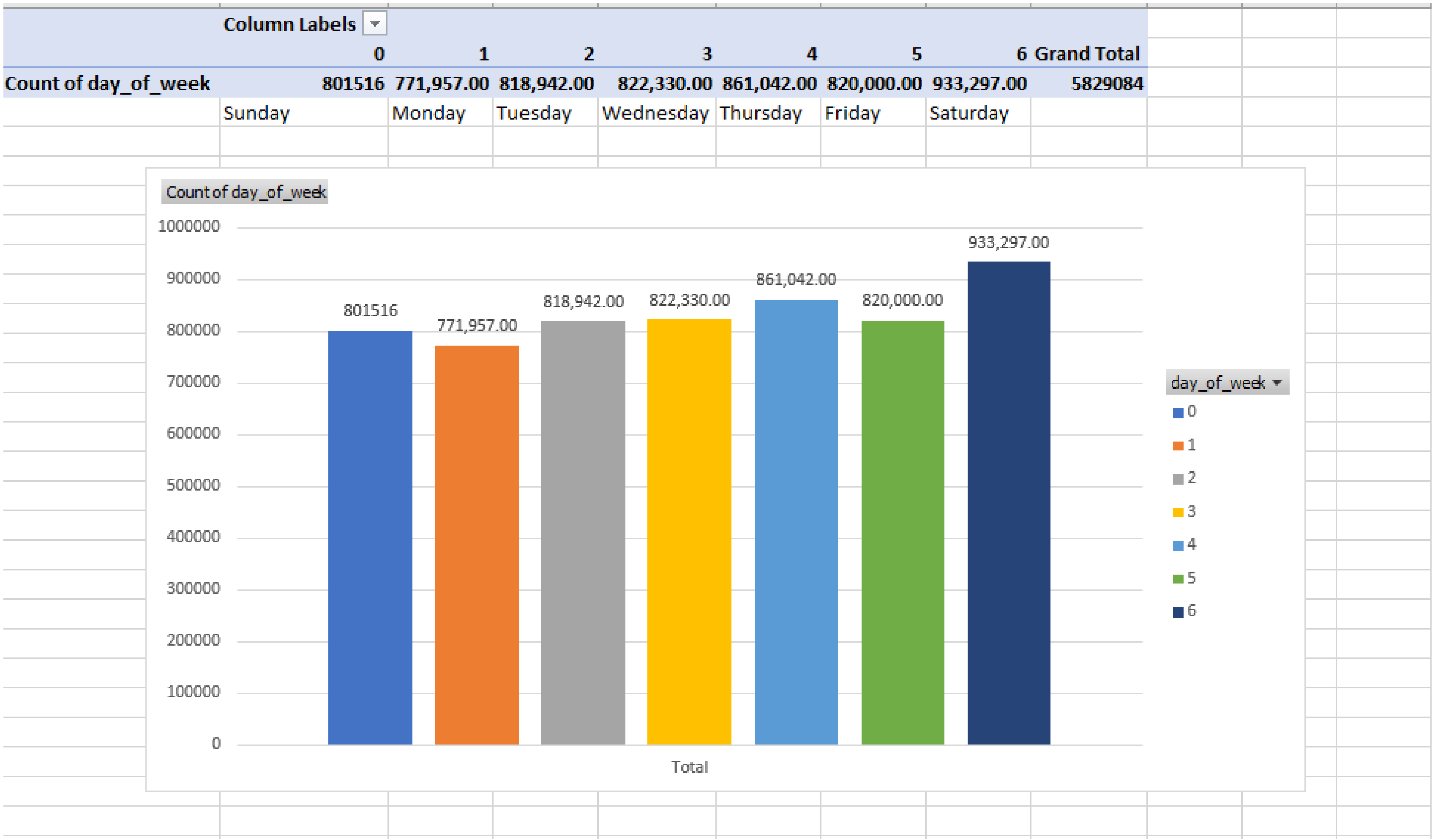


~~~
```{r}
# visualization for number of rides by rider type/weekday
trip_data_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(),average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)%>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")

```
~~~



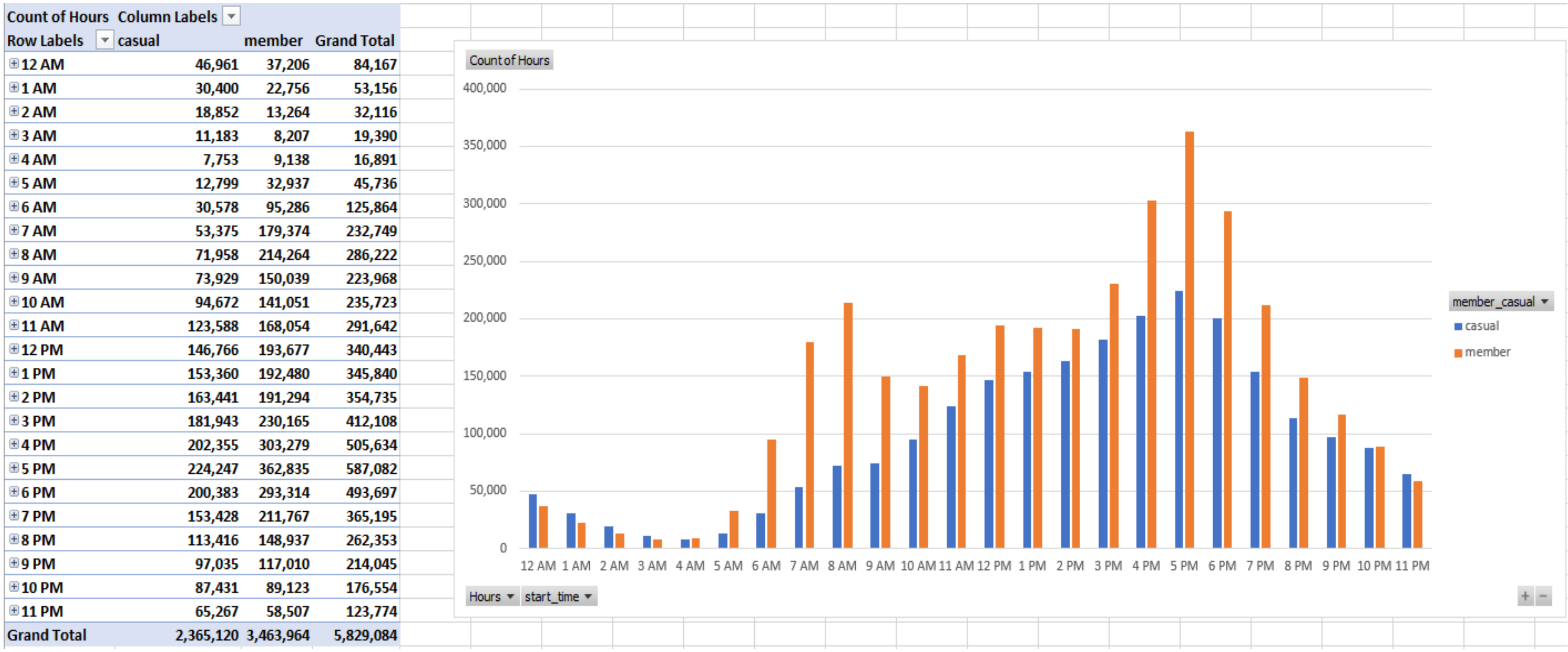
~~~
```{r}
#visualization for average duration
trip_data_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(),average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)%>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")

```
~~~

Count of Total Trips for each Day of the Week



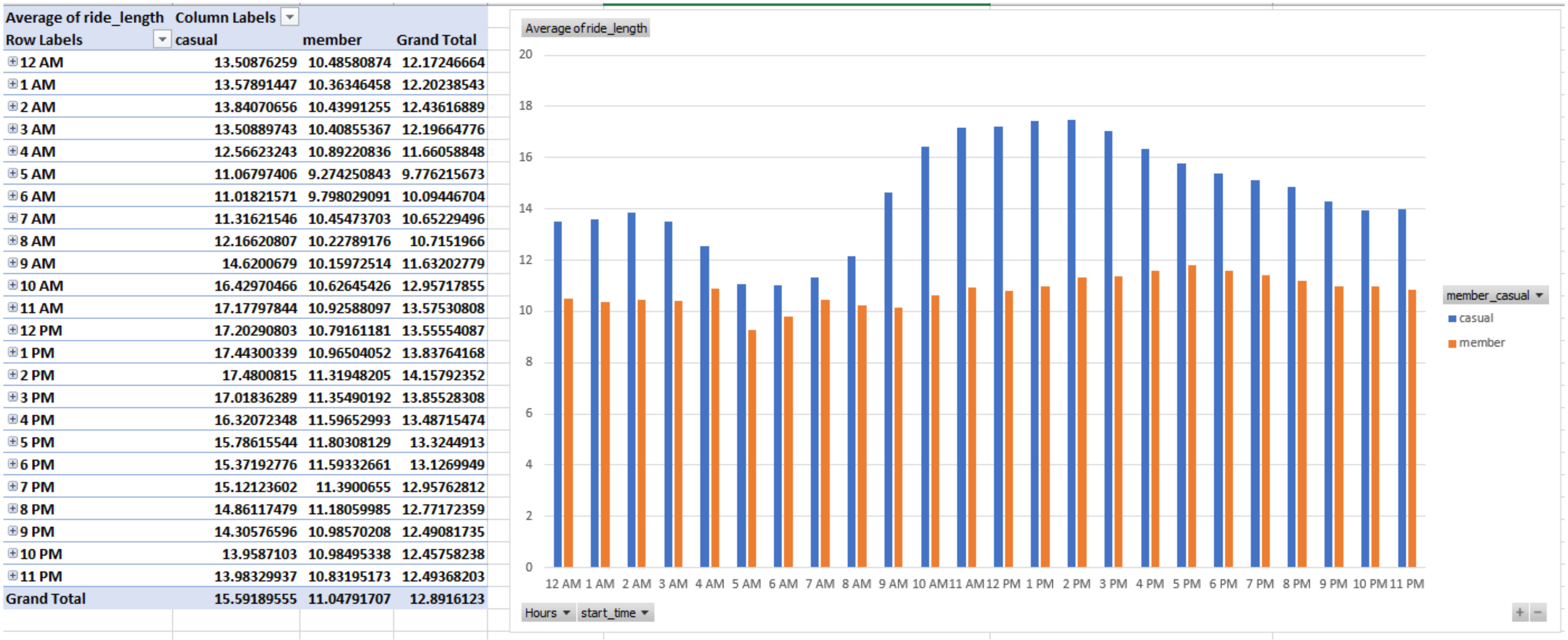
	Column Labels ▾							
	0	1	2	3	4	5	6	Grand Total
Count of day_of_week	801516	771,957.00	818,942.00	822,330.00	861,042.00	820,000.00	933,297.00	5829084
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	

Number of trips per hour



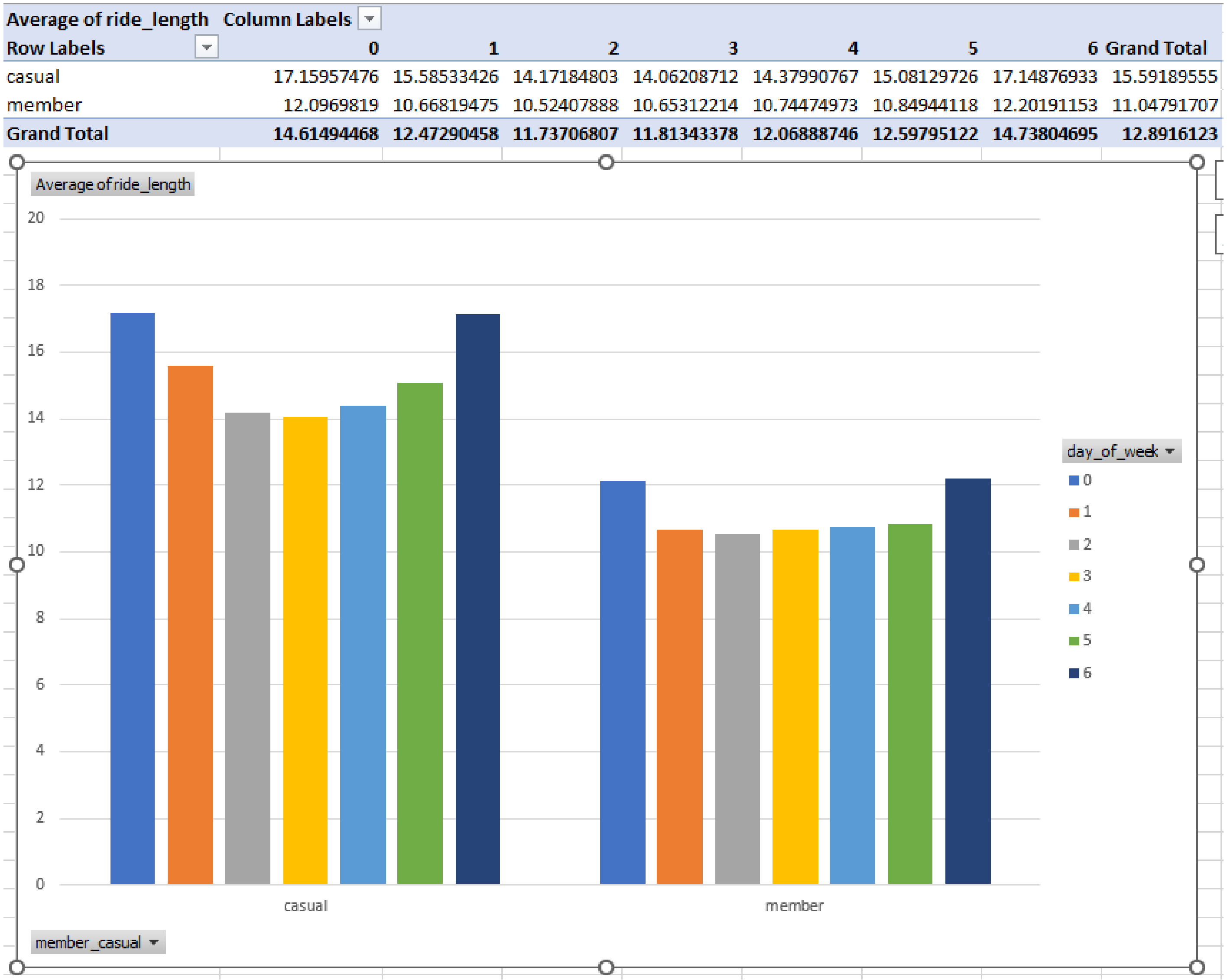
Count of Hours	Column Labels ▾		
Row Labels ▾	casual	member	Grand Total
⊞12 AM	46,961	37,206	84,167
⊞1 AM	30,400	22,756	53,156
⊞2 AM	18,852	13,264	32,116
⊞3 AM	11,183	8,207	19,390
⊞4 AM	7,753	9,138	16,891
⊞5 AM	12,799	32,937	45,736
⊞6 AM	30,578	95,286	125,864
⊞7 AM	53,375	179,374	232,749
⊞8 AM	71,958	214,264	286,222
⊞9 AM	73,929	150,039	223,968
⊞10 AM	94,672	141,051	235,723
⊞11 AM	123,588	168,054	291,642
⊞12 PM	146,766	193,677	340,443
⊞1 PM	153,360	192,480	345,840
⊞2 PM	163,441	191,294	354,735
⊞3 PM	181,943	230,165	412,108
⊞4 PM	202,355	303,279	505,634
⊞5 PM	224,247	362,835	587,082
⊞6 PM	200,383	293,314	493,697
⊞7 PM	153,428	211,767	365,195
⊞8 PM	113,416	148,937	262,353
⊞9 PM	97,035	117,010	214,045
⊞10 PM	87,431	89,123	176,554
⊞11 PM	65,267	58,507	123,774
Grand Total	2,365,120	3,463,964	5,829,084

Average Ride Length by Minutes: Member vs Casual Rider



Average of ride_length	Column Labels		
Row Labels	casual	member	Grand Total
⊞12 AM	13.50876259	10.48580874	12.17246664
⊞1 AM	13.57891447	10.36346458	12.20238543
⊞2 AM	13.84070656	10.43991255	12.43616889
⊞3 AM	13.50889743	10.40855367	12.19664776
⊞4 AM	12.56623243	10.89220836	11.66058848
⊞5 AM	11.06797406	9.274250843	9.776215673
⊞6 AM	11.01821571	9.798029091	10.09446704
⊞7 AM	11.31621546	10.45473703	10.65229496
⊞8 AM	12.16620807	10.22789176	10.7151966
⊞9 AM	14.6200679	10.15972514	11.63202779
⊞10 AM	16.42970466	10.62645426	12.95717855
⊞11 AM	17.17797844	10.92588097	13.57530808
⊞12 PM	17.20290803	10.79161181	13.55554087
⊞1 PM	17.44300339	10.96504052	13.83764168
⊞2 PM	17.4800815	11.31948205	14.15792352
⊞3 PM	17.01836289	11.35490192	13.85528308
⊞4 PM	16.32072348	11.59652993	13.48715474
⊞5 PM	15.78615544	11.80308129	13.3244913
⊞6 PM	15.37192776	11.59332661	13.1269949
⊞7 PM	15.12123602	11.3900655	12.95762812
⊞8 PM	14.86117479	11.18059985	12.77172359
⊞9 PM	14.30576596	10.98570208	12.49081735
⊞10 PM	13.9587103	10.98495338	12.45758238
⊞11 PM	13.98329937	10.83195173	12.49368203
Grand Total	15.59189555	11.04791707	12.8916123

Average Ride Length by Day of the Week: Member vs Casual Rider



Average of ride_length	Column Labels ▾							
Row Labels ▾	0	1	2	3	4	5	6	Grand Total
casual	17.15957476	15.58533426	14.17184803	14.06208712	14.37990767	15.08129726	17.14876933	15.59189555
member	12.0969819	10.66819475	10.52407888	10.65312214	10.74474973	10.84944118	12.20191153	11.04791707
Grand Total	**14.61494468**	**12.47290458**	**11.73706807**	**11.81343378**	**12.06888746**	**12.59795122**	**14.73804695**	**12.8916123**

Act

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

Based on my analysis, annual members tend to take rides more frequently, however, casual riders take significantly longer rides on average. In fact, they take longer rides every single day of the week, at every hour of the day.

Casual riders also take more rides on Saturdays and Sundays than annual members.

Annual members rides spike in the morning between 7am and 8am, and both casual riders and members rides spike in the afternoon between 3pm and 5pm. This indicates that members are probably using bikes to commute to work, while casual riders are using bikes for activities after work. Perhaps casual riders use bikes more for exercise or social gatherings, and members use rides for commuting more.

2. Why would casual riders buy Cyclistic annual memberships?

Casual riders would convert to annual memberships if they used Cyclistic bikes so often that they would save money buying an annual membership. Another reason for using Cyclistic bikes would be out of ease of access. Upping the number of bike stations could create more rides in general, which could turn to more memberships.

3. How can Cyclistic use digital media to influence casual riders to become members?

It would be interesting to see data on the amount of money some casual riders spend on an annual basis, and compare that to the price of an annual membership. When it comes to marketing, I think it would be beneficial to create advertisements that say the price of an annual membership is equal to, for example, 10 rides. Trying to lock in more memberships by making it clear to casual riders that take more than 10 rides a year, that they would save money through buying a membership.

In order to convert more casual riders to memberships, Cyclistic could consider charging more per mile for non members. Since casual riders tend to use Cyclistic bikes more often on the weekends than even members do, spending more money for advertisements on the weekends could be a good investment if they want to increase memberships.