Google Analytics - Capstone Project 1

Author: Victoria Fernández
Date: 2023/08/1

Case study: how does a bike-share navigate speedy success?

The Google Analytics Certificate includes this final project as a way to exercise everything learned in their program. They facilitate two projects with their datasets and problematics to solve; this one being the first of them. The dataset used is public and made available by Motivate International Inc., but the names of the people and company are fictional

To the project itself:

Cyclistic is a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships, so they want to design marketing strategies aimed at converting casual riders into annual members.

They want to understand how annual members and casual riders differ, why casual riders buy a membership and how digital media could affect their marketing tactics.

Questions to answer:

* How do annual members and casual riders use Cyclistic bikes differently?
* Why would casual riders buy Cyclistic annual memberships?
* How can Cyclistic use digital media to influence casual riders to become members?

Datasets and cleaning process

We have access to data going back as far as 2013, but for our analysis, we will be using datasets from the last year, ranging from June 2022 to June 2023. We make sure to name every file properly and file them neatly in appropriate labeled folders:There is one for each month of the past year.

For this process we will be using RStudio since the files are too heavy for Spreadsheets to handle and because R will help us with cleaning and statistics.

Setting our environment

First, we load the packages we will use throughout the process:

```
library(tidyverse) #helps wrangle data
library(lubridate) #helps wrangle date attributes
library(ggplot2) #helps visualize data
librarr(readr) #saves .csv datasets
```

Then we will set our work directory and upload every .csv file and store them in an object each.

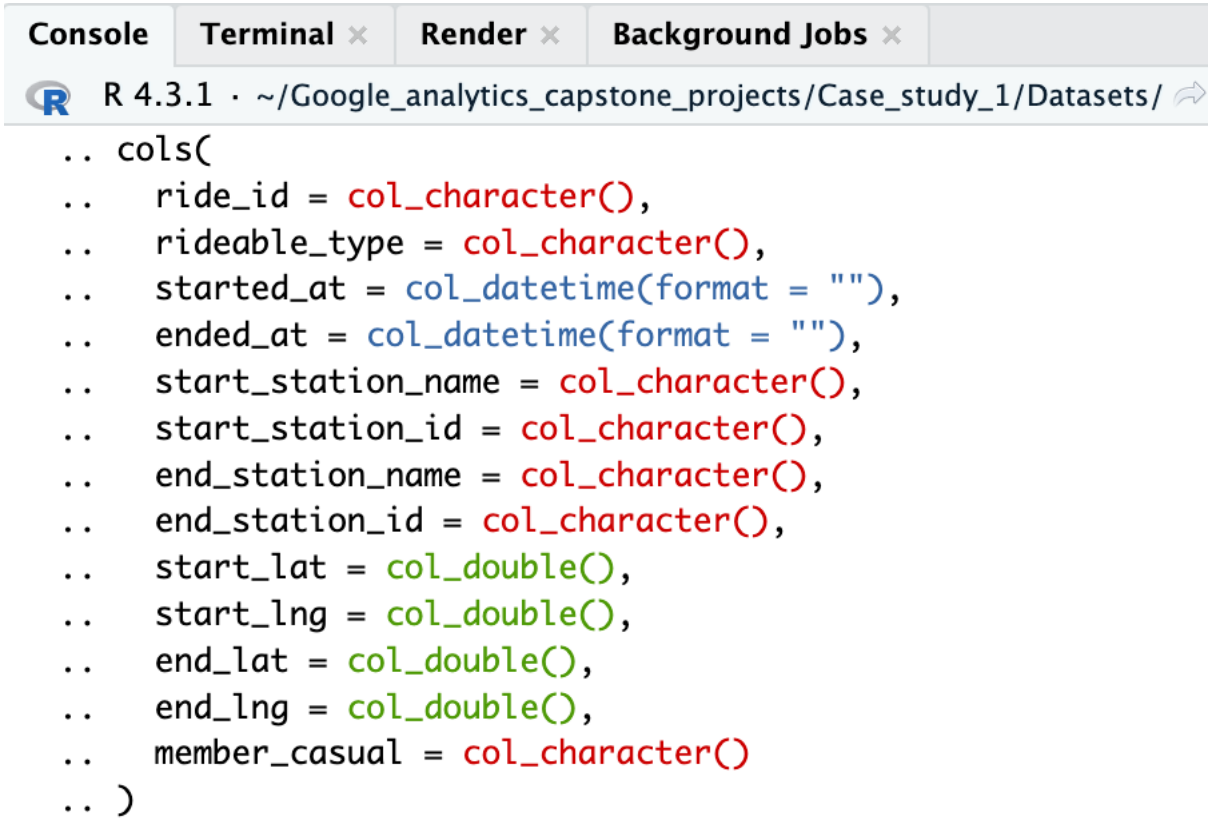
```
getwd()
setwd("my directory")
ds_2022_06 <- read_csv("202206-divvy-tripdata.csv") #repeat with every file
```

Verifying and cleaning data

Once all our data sets are loaded, we will get a view of each: we want to get to know the structure of this data and a visual of how it is organized.

```
str(ds_2022_06)
as_tibble(ds_2022_06) #this is a heavy dataset, so View() might not be the best
```

We pass every dataset through this functions to verify that all columns are named the same: this is important since we want to merge this tables to create a unified dataset.



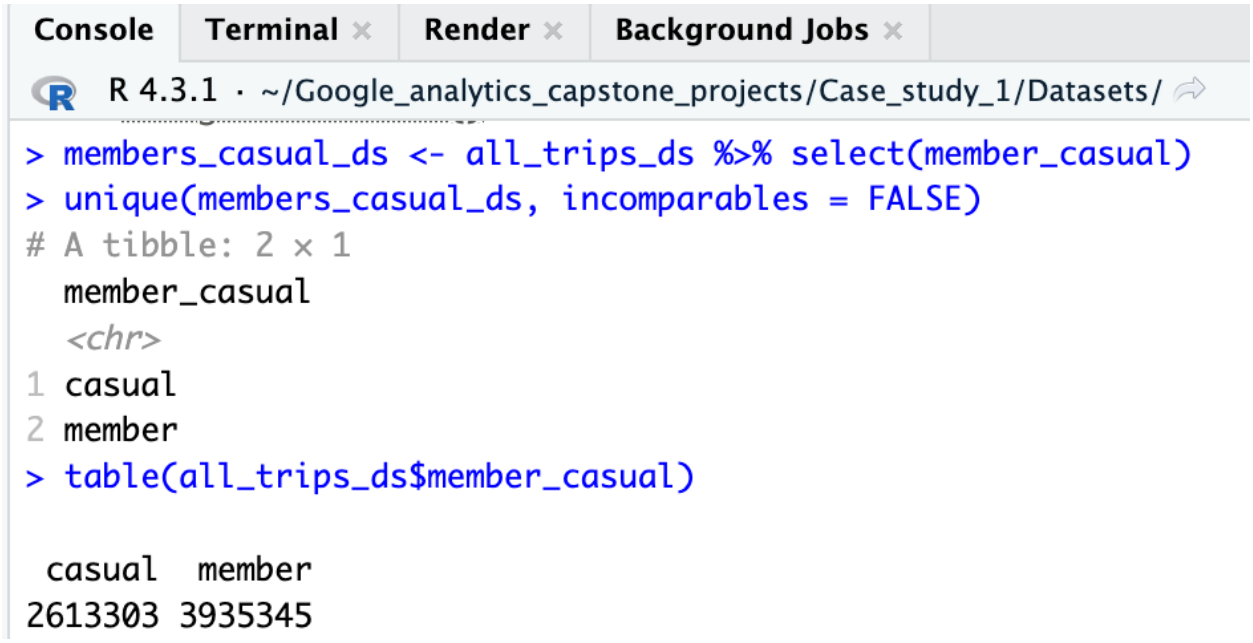

```
#Combining into a single dataset
all_trips_ds <- bind_rows(ds_2022_06,ds_2022_07,ds_2022_08,ds_2022_09,ds_2022_10,
                          ds_2022_11,ds_2022_12,ds_2023_01,ds_2023_02,ds_2023_03,
                          ds_2023_04,ds_2023_05,ds_2023_06)
write_csv(all_trips_ds, file="all_trips_ds.csv") #saving a backup
```

We can check the data set quickly and ask for a summary:

```
str(all_trips_ds)
summary(all_trips_ds)#This is ouonsole output
```

We are most interested in the member_casual column since it describes if the trip was made by a casual rider or a membership rider. Since having other labels or if the input is misspelled can skew our observations, lets make sure that they are right, and if so, how many of them are by category:

```
#Making sure we only have two kinds of labels in the member_casual
columnmembers_casual_ds <- all_trips_ds %>% select(member_casual)
unique(members_casual_ds, incomparables = FALSE)
##counting trips made by these categories.
table(all_trips_ds$member_casual)
```



Aggregating data

We also want to be able to do some calculations with the “started_at” column, so we need to aggregate this data in a way we can use it. We will create new columns for month, day, year and day of the week.

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

We also want a “ride_length” column: this is the result of subtracting the “started_at” column from the “ended_at” column:

```
all_trips_ds$ride_length <- difftime(all_trips_ds$ended_at,all_trips_ds$started_at,
                                     units="mins")
# Convert "ride_length"  to numeric so we can run calculations on the dataall_trips_ds$ride_length <- as.numeric(as.character(all_trips_ds$r
```

Some values are negative. Ideally, we could create a new dataset dropping all values bellow 0, but that creates a strange amount of NA values. We will deal with negative values adding filters when it comes to it.

Analyzing data

Descriptive Analysis

Now that we have a final clean dataset to work with we can start getting some indicators on the “ride_length” values:

```
mean(all_trips_ds$ride_length>0) #straight average
median(all_trips_ds$ride_length > 0) #standar deviation
median(all_trips_ds$ride_length>0) #midpoint number
max(all_trips_ds$ride_length) #longest ride
min(all_trips_ds$ride_length>0) #shortest ride
##summary gives the same kind of results, but taking the negative values
##into account.
```

The results are:

- Mean: 0,999 min
- Standard deviation: 0,0098
- Median: 1 min
- Max: 41387.25 min
- Min: 0 min

It seems most trips are pretty short.

Now we want to compare casual riders with annual members using our ride_length variable.

```
aggregate(all_trips_ds$ride_length>0 ~ all_trips_ds$member_casual, FUN = mean)
aggregate(all_trips_ds$ride_length>0 ~ all_trips_ds$member_casual, FUN = sd)
aggregate(all_trips_ds$ride_length>0~ all_trips_ds$member_casual, FUN = median)
aggregate(all_trips_ds$ride_length ~ all_trips_ds$member_casual, FUN = max)
aggregate(all_trips_ds$ride_length>0 ~ all_trips_ds$member_casual, FUN = min)
```

```
> aggregate(all_trips_ds$ride_length>0 ~ all_trips_ds$member_casual, FUN = mean)
  all_trips_ds$member_casual all_trips_ds$ride_length > 0
1                     casual                    0.9998886
2                     member                    0.9999113
> aggregate(all_trips_ds$ride_length>0 ~ all_trips_ds$member_casual, FUN = sd)
  all_trips_ds$member_casual all_trips_ds$ride_length > 0
1                     casual                  0.010551823
2                     member                  0.009416773
> aggregate(all_trips_ds$ride_length~ all_trips_ds$member_casual, FUN = median)
  all_trips_ds$member_casual all_trips_ds$ride_length
1                     casual                 12.31667
2                     member                  8.70000
> aggregate(all_trips_ds$ride_length ~ all_trips_ds$member_casual, FUN = max)
  all_trips_ds$member_casual all_trips_ds$ride_length
1                     casual                41387.250
2                     member                 1559.667
> aggregate(all_trips_ds$ride_length>0 ~ all_trips_ds$member_casual, FUN = min)
  all_trips_ds$member_casual all_trips_ds$ride_length > 0
1                     casual                           0
2                     member                           0
```

We can see that casual riders have a slight tendency to longer trips (because of the median, and because they have bigger trips with similar dispersion), which might indicate that casual riders only take rides when the distances are long.

Lets analyze the days of the week too:

```
all_trips_ds$day_of_week <-ordered(all_trips_ds$day_of_week, levels=c("Sunday",
            "Monday", "Tuesday", "Wednesday", "Thursday", "Friday","Saturday"))
aggregate(all_trips_ds$ride_length ~ all_trips_ds$member_casual +
          all_trips_ds$day_of_week, FUN = mean)
```

	all_trips_ds$member_casual	all_trips_ds$day_of_week	all_trips_ds$ride_length
1	casual	Sunday	33.30530
2	member	Sunday	13.89266
3	casual	Monday	27.72467
4	member	Monday	11.91477
5	casual	Tuesday	25.15340
6	member	Tuesday	12.00073
7	casual	Wednesday	24.11926
8	member	Wednesday	11.94681
9	casual	Thursday	24.38924
10	member	Thursday	12.09829
11	casual	Friday	27.70881
12	member	Friday	12.45835
13	casual	Saturday	32.34183
14	member	Saturday	14.07524

Once again, casual riders are the one taking the longest rides, and the difference is the greatest on Sundays.

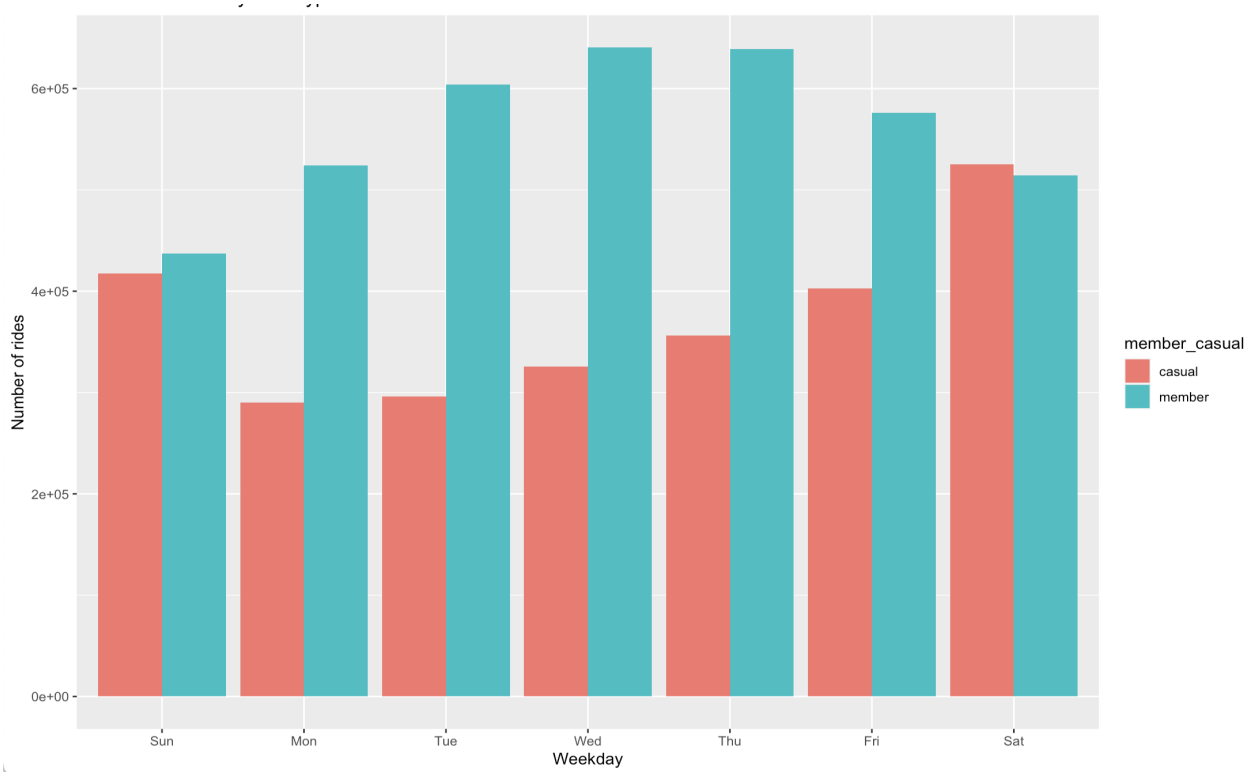
```
# analyze ridership data by type and weekday
all_trips_ds %>%
mutate(weekday = wday(started_at, label = TRUE)) %>% #creates weekday field using wday()
group_by(member_casual, weekday) %>% #groups by usertype and weekday
summarise(number_of_rides = n() #calculates the number of rides and average duration
,average_duration = mean(ride_length)) %>%  # calculates the average duration
arrange(member_casual, weekday) # sorts
```

	member_casual	weekday	number_of_rides	average_duration
	<chr>	<ord>	<int>	<dbl>
1	casual	Sun	417263	33.3
2	casual	Mon	290093	27.7
3	casual	Tue	295933	25.2
4	casual	Wed	325650	24.1
5	casual	Thu	356260	24.4
6	casual	Fri	402980	27.7
7	casual	Sat	525124	32.3
8	member	Sun	437001	13.9
9	member	Mon	524415	11.9
10	member	Tue	604024	12.0
11	member	Wed	640425	11.9
12	member	Thu	639181	12.1
13	member	Fri	576204	12.5
14	member	Sat	514095	14.1

We can see that , even though the annual members take more trips, this trips are shorter in average.

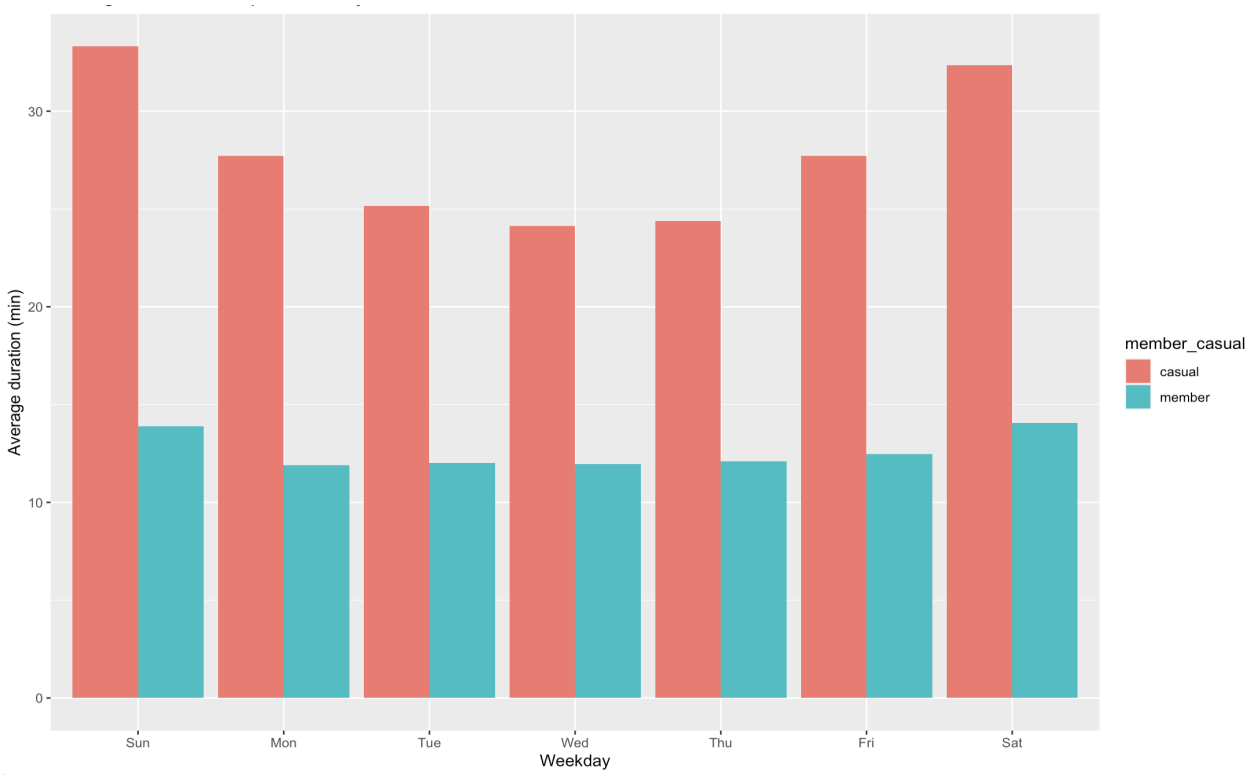
Results

Number of Rides by rider type



In this graphic we can appreciate how Wednesday is the most busy days for those who are annual members, but the weekends are the days the casual riders use the service the most.

Average ride duration per weekday



Thanks to this graphic we can easily see how casual riders take longer trips, specially on the weekends.

Conclusions

Lets go back to our questions:

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

Thanks to our analysis we can take two main takeaways in regards of riders differences, and those are that casual riders use this service for longer rides, specially on the weekends, while annual members take shorter rides mostly during the week. This may indicate the casual riders commute to work with other kinds of transportation and take rides for leisure on the weekends, or only when the distance is too long and the expense is justifiable.

On the other hand, those with annual subscriptions use the bikes more intensely and also for shorter trips: if they are already paying, why not use the bike?

Casual riders can become annual members by realizing that commuting by bike can be more rewarding than other services: it serves as exercise, is eco-friendly, and maybe the price of longer trips adds up faster and becomes more expensive than an annual subscription.

Cyclist could take this takeaways to create advertising tailored to this groups focused on this very topics: how to save money by buying an annual subscription and how commuting by bike can be beneficial for your body and the planet.