Marketing data analytics for bike sharing company using Rstudio and tableau

This article would be sharing how I approached this case study, and for my choice of weapon, I would be using RStudio and Tableau in this article. I will be publishing a separate article using BigQuery SQL instead.

Before we start, I will be using my understanding of the analysis process which is: Ask, Prepare, Process, Analyze, Share & Act.

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-sharing company in Chicago. Moreno (director of marketing) believes the company's future success depends on maximizing the number of annual memberships.

Our goal is to design marketing strategies aimed at converting casual riders into annual members. In order to do that, we need to understand how casual riders and annual members use Cyclistic bikes differently.

About the company

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic’s finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

These are the questions/business task that would guide the future of the marketing program:

1. To understand how annual members and casual riders use our Cyclistic bikes differently
2. Why would casual members upgrade to annual memberships
3. How can Cyclistic use digital media to influence casual riders to become members?

Prepare

We will be using Cyclistics historical trip data (here) from 2021 January till 2021 December (202101-divvy-tripdata.zip -> 202112-divvy-tripdata.zip). We will be extracting all of our files into a folder “Divvy_Monthlytripdata” to organize and provide context.

We will also be renaming the files to represent the data more clearly, it would also help with readability for other team members. Below is how I would do so:

(202101-divvy-tripdata.csv) -> (2021_01.csv)

(202102-divvy_tripdata.csv) -> (2021_02.csv)

and so on.

As a disclaimer, the data has been made available by Motivate International Inc. under this license

Process

In order to process the 5,595,063 total records, spreadsheets wouldn’t be able to handle the sheer amount of data. In this case, we would be using Rstudio instead.

Firstly, we would need to install & load the packages required for this process, which in this case will be: Tidyverse, Janitor & Lubridate.

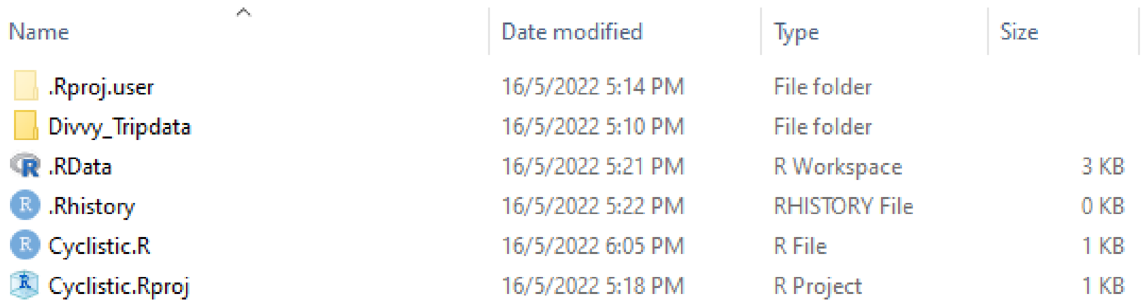
Disclaimer: Sentences followed after # are comments for the audience, not lines of codes

```
#Installing the packages

install.packages('tidyverse')

install.packages('janitor')

install.packages('lubridate')

#Loading the packages

library(tidyverse)

library(janitor)

library(lubridate)
```

Subsequently, we would need to import the csv’s into Rstudio, in which we would use read_csv. I would also like to add a new name for the imported csv’s to help with readability.

```
#Adding a name <- Importing the csv(file_location)

Jan2021 <- read_csv("Divvy_Tripdata/2021_01.csv")

Feb2021 <- read_csv("Divvy_Tripdata/2021_02.csv")

Mar2021 <- read_csv("Divvy_Tripdata/2021_03.csv")

Apr2021 <- read_csv("Divvy_Tripdata/2021_04.csv")

May2021 <- read_csv("Divvy_Tripdata/2021_05.csv")

Jun2021 <- read_csv("Divvy_Tripdata/2021_06.csv")

Jul2021 <- read_csv("Divvy_Tripdata/2021_07.csv")

Aug2021 <- read_csv("Divvy_Tripdata/2021_08.csv")

Sep2021 <- read_csv("Divvy_Tripdata/2021_09.csv")

Oct2021 <- read_csv("Divvy_Tripdata/2021_10.csv")

Nov2021 <- read_csv("Divvy_Tripdata/2021_11.csv")

Dec2021 <- read_csv("Divvy_Tripdata/2021_12.csv")
```

A disclaimer: My csv’s were located inside the folder “Divvy_Tripdata”, and my working directory is in “Google Case Study 1” to separate the file types as shown below:



Name	Date modified	Type	Size
.Rproj.user	16/5/2022 5:14 PM	File folder	
Divvy_Tripdata	16/5/2022 5:10 PM	File folder	
.RData	16/5/2022 5:21 PM	R Workspace	3 KB
.Rhistory	16/5/2022 5:22 PM	RHISTORY File	0 KB
Cyclistic.R	16/5/2022 6:05 PM	R File	1 KB
Cyclistic.Rproj	16/5/2022 5:18 PM	R Project	1 KB

The next step would be to merge all the csv’s(which we will now call dataset) into one table, however before doing that we would need to verify if there are any extra or missing columns.

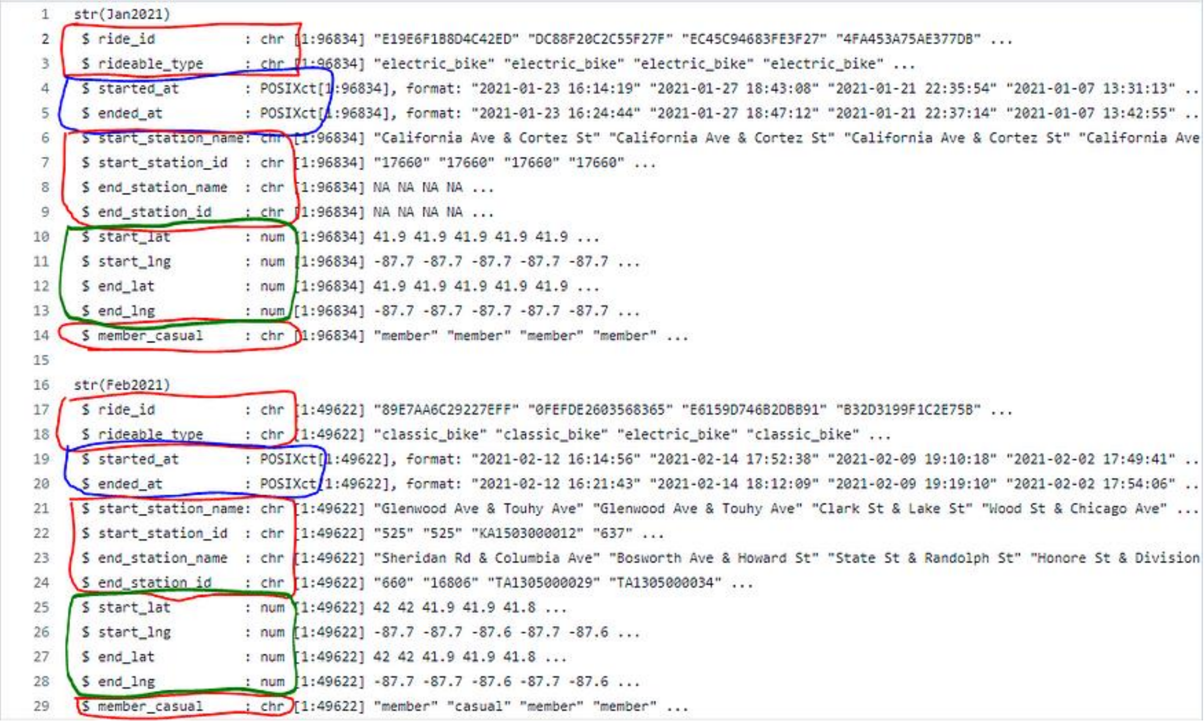
We would also need to check if there are any discrepancies with formatting as well (maybe one dataset’s ID column is formatted as INT and another dataset’s ID column is formatted as CHR). We would do this by using ‘str()’ to list the structure of our datasets.

```
#str(dataset_name)

str(Jan2021)

str(Feb2021)

str(Mar2021)

str(Apr2021)

str(May2021)

str(Jun2021)

str(Jul2021)

str(Aug2021)
str(Sep2021)

str(Oct2021)

str(Nov2021)

str(Dec2021)
```

We would get the following:

We will focus on what comes after : and before [

By right, we should be getting 10 more outputs of what we're seeing above, but to reduce the clutter I will not be including it.



```
str(Jan2021)
 $ ride_id           : chr [1:96834] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...
 $ rideable_type     : chr [1:96834] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
 $ started_at        : POSIXct[1:96834], format: "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 22:35:54" "2021-01-07 13:31:13" ..
 $ ended_at          : POSIXct[1:96834], format: "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-01-07 13:42:55" ..
 $ start_station_name: chr [1:96834] "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" "California Ave
 $ start_station_id  : chr [1:96834] "17660" "17660" "17660" "17660" ...
 $ end_station_name  : chr [1:96834] NA NA NA NA ...
 $ end_station_id    : chr [1:96834] NA NA NA NA ...
 $ start_lat         : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ end_lat           : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr [1:96834] "member" "member" "member" "member" ...

str(Feb2021)
 $ ride_id           : chr [1:49622] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E75B" ...
 $ rideable_type     : chr [1:49622] "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
 $ started_at        : POSIXct[1:49622], format: "2021-02-12 16:14:56" "2021-02-14 17:52:38" "2021-02-09 19:10:18" "2021-02-02 17:49:41" ..
 $ ended_at          : POSIXct[1:49622], format: "2021-02-12 16:21:43" "2021-02-14 18:12:09" "2021-02-09 19:19:10" "2021-02-02 17:54:06" ..
 $ start_station_name: chr [1:49622] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St" "Wood St & Chicago Ave" ...
 $ start_station_id  : chr [1:49622] "525" "525" "KA1503000012" "637" ...
 $ end_station_name  : chr [1:49622] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Randolph St" "Honore St & Division
 $ end_station_id    : chr [1:49622] "660" "16806" "TA1305000029" "TA1305000034" ...
 $ start_lat         : num [1:49622] 42 42 41.9 41.9 41.8 ...
 $ start_lng         : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ end_lat           : num [1:49622] 42 42 41.9 41.9 41.8 ...
 $ end_lng           : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ member_casual     : chr [1:49622] "member" "casual" "member" "member" ...
```

All 12 datasets should have identical column names & formatting types. We also should be checking if there are any wrongly formatted data types, IE:

- Columns that contain characters are formatted as num/int
- Columns that contain numbers/decimals but are formatted as chr
- Columns that contain dates/time but is formatted as num/int

However, based on what we’re looking at, nothing needs to be transformed.

Now we’re ready to merge all datasets into one table(which will be called a dataframe), which I will be naming merged_df, and to do that we would be using bind_rows.

After that, we will also be cleaning our column names to remove spaces, parentheses, camelCase, and so on. This will also automatically separate words by capital case. IE:

- This.Is An.Example -> this_is_an_example
- This is An example -> this_is_an_example

```
#Creating new dataset name <- binding rows(all_your_datasets)

merged_df <- bind_rows(Jan2021, Feb2021, Mar2021, Apr2021, May2021, Jun2021,

Jul2021, Aug2021, Sep2021, Oct2021, Nov2021, Dec2021

#Cleaning & removing any spaces, parentheses, etc.

merged_df <- clean_names(merged_df)
```

We should also remove any empty columns and rows in our

dataframe, we can do so by using remove_empty().

```
#removing_empty(dataset_name, by leaving c() empty, it selects rows & columns)

remove_empty(merged_df, which = c())
```

After a brief view of the data, with our current information, we would only be able to aggregate at ride-level, AKA popular stations, types of bikes used, percentages of casual members to annual members, etc.

However, before we proceed, it is of utmost importance to revisit our business task/the problem that we are trying to solve in an attempt to refocus ourselves as we may get too carried over the little details and lose track of what we’re supposed to do in the first place.

After going through our business task, we would need to add columns that would list:

- Day of the week — By using wday()

more info regarding wday here

- Start hour — By using format(as.POSIXct))

more info regarding time_formats here

- Month — By using format(as.Date))

more info regarding date_formats here & datatypes here

- Trip Duration — By using difftime()

$ is used to extract/access/select a specific dataframe column

~~Analyze~~

After processing and cleaning our data, it is essential to clean one last time, as our added columns may have errors within them.

In this situation, we would need to remove any rows which have trip_durations of 0 seconds or less. However, in this case, we would just create a new dataframe that does not contain trip durations of <0 seconds using ‘!’

```
#In laymans term, '!' means is not equals to
```

```
cleaned_df <- merged_df[!(merged_df$trip_duration<=0),]
```

For those that don't understand, we’re essential creating a new dataframe called cleaned_df from merged_df that does not have trip_durations of 0 seconds or less

If you’re using a separate visualization tool such as Tableau or PowerBI, we need to export our dataframe using write.csv:

```
write.csv(cleaned_df, file ='Cyclistic_df.csv')
```

Analyze & Share (RStudio)

As a side note, I will be using ggplot in RStudio for this part of the article. I will be including another section in which I used Tableau instead as well.

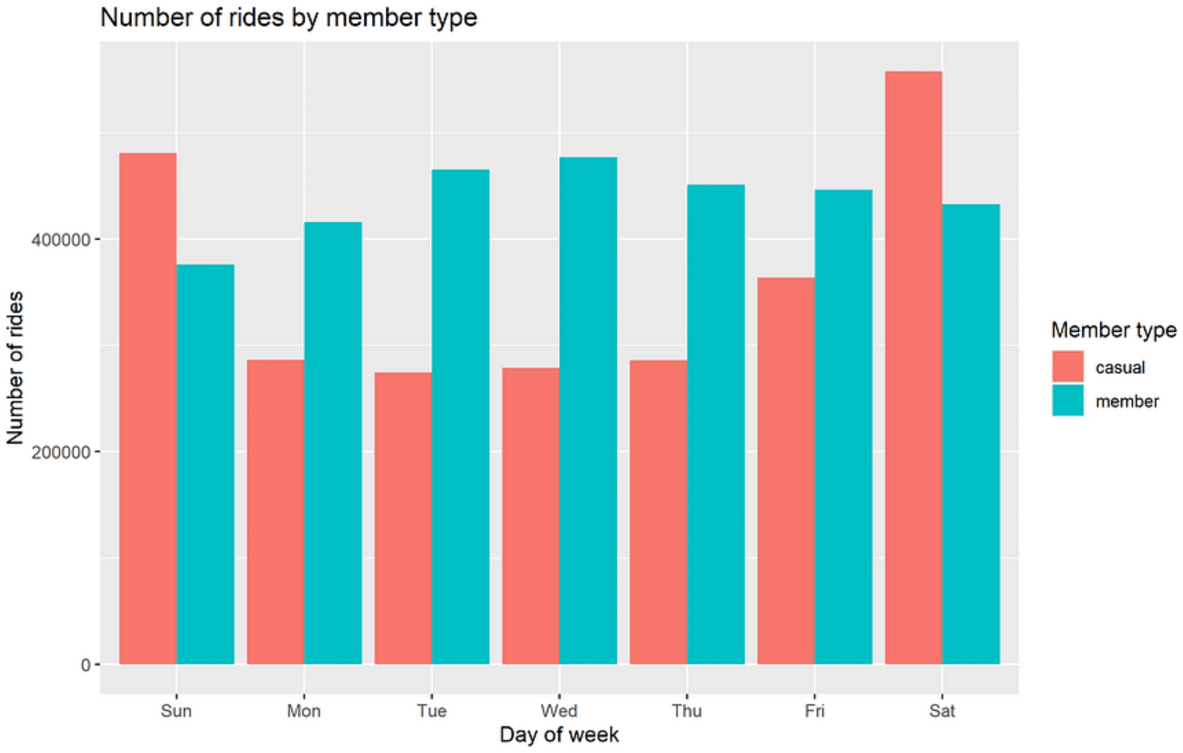
Now it’s time to analyze the data and look for key information that we can perform analysis on, and afterward, plot/visualize it!

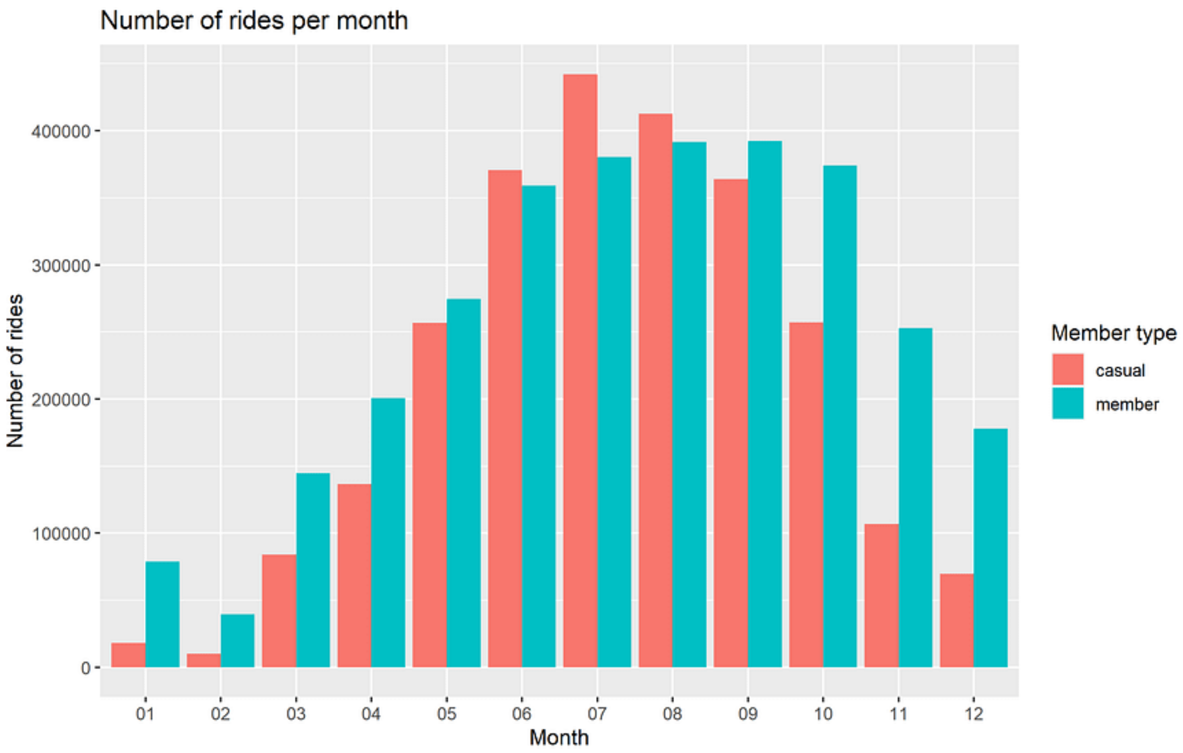
As mentioned just a moment ago, it is imperative to always remind yourself of the business task at hand during this stage. In order to answer our first business question, it would be beneficial to plot a few of our observations revolving around:

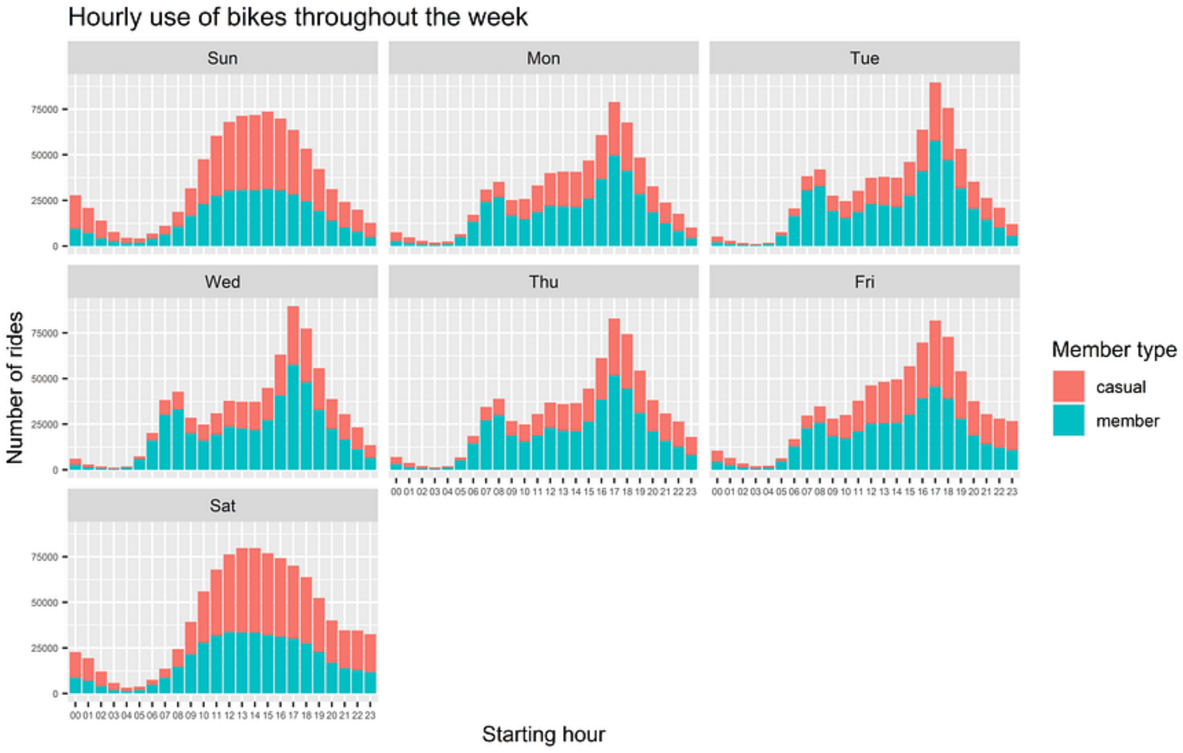
1. How do casual and members use their bikes differently throughout the week
2. Peak hours of bike usage between casual and annual members
3. Bike usage throughout the year
4. The average trip duration between casual and annual members
5. Most popular stations among casual and annual members

In order to carry out our plotting, we will be using ggplot() in RStudio. I will not be commenting/explaining the basic lines of codes for ggplot.

without scipen, the value of ‘Number of rides’ would be displayed as 2e+05, 4e+05, and so on.




Number of rides per month
Number of rides
400000
300000
200000
100000
0
01
02
03
04
05
06
07
08
09
10
11
12
Month
Member type
casual
member


Hourly use of bikes throughout the week
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Number of rides
75000
50000
25000
0
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Starting hour
Member type
casual
member

Due to weird errors with RStudio not being able to locate my objects, I was not able to produce a chart for the average trip duration between casual and annual members.

Instead, I will be carrying out descriptive analysis by using aggregate().

- Our first column would be the row ID
- The second column would be member type
- The third column would be the day of the week
- The fourth column would be the mean/average trip duration

Based on a brief view of our output text, we can clearly see that casual members almost spend double the amount of time using the bikes as compared to annual members!

In order to find out the most popular stations, we need to carry out descriptive analysis again. To do this, we need to filter out by member type and sort in descending order the most frequently visited stations!

As a side note, I will not be creating any more dataframe's to reduce the strain on my system.

To do so, I will be using the following code to find out the most popular stations for our annual members:

Press the highlighted word for more information about count & filter

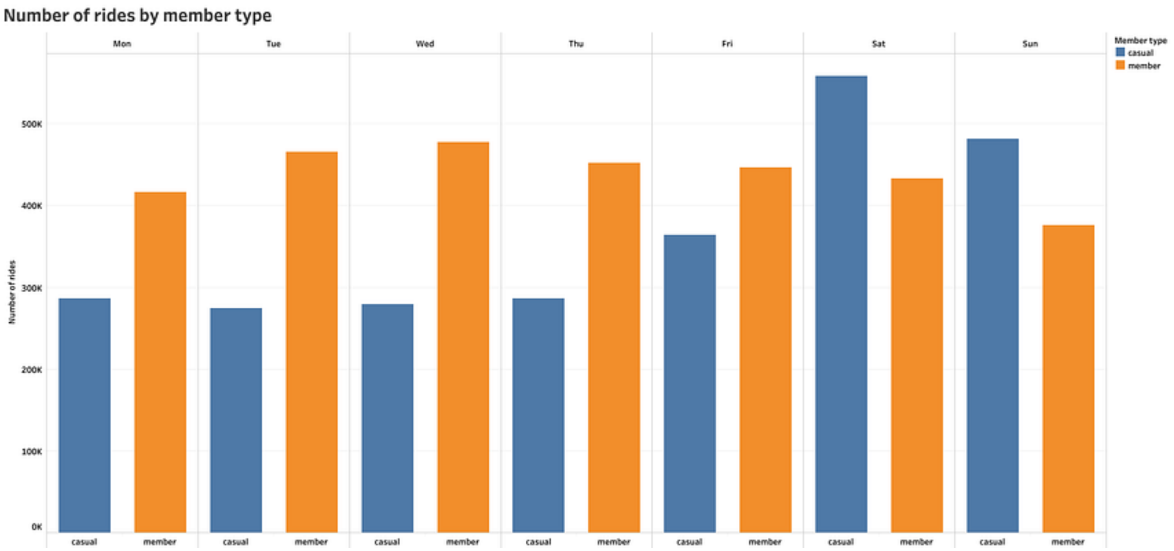
The output would be:

The first 2 output is for annual members and the last 2 is for casual members

With a quick read, we can tell which starting/ending stations are the most popular among casual and annual members!

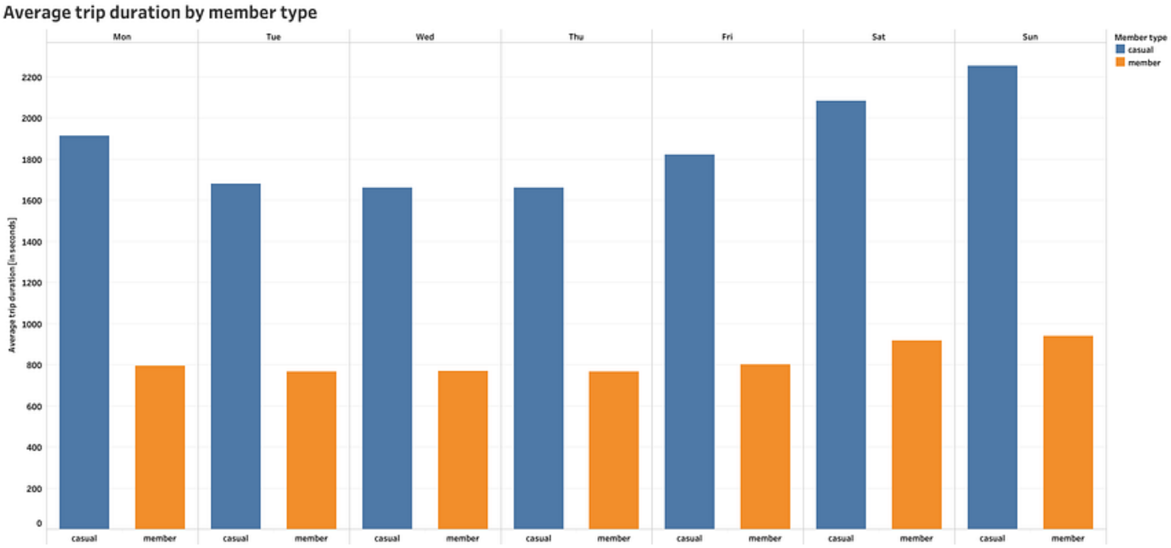
Analyze & Share (Tableau)

By using a visualization tool such as Tableau, we can create more intricate, well-designed visuals to assist in the Share phase. Here are my findings.



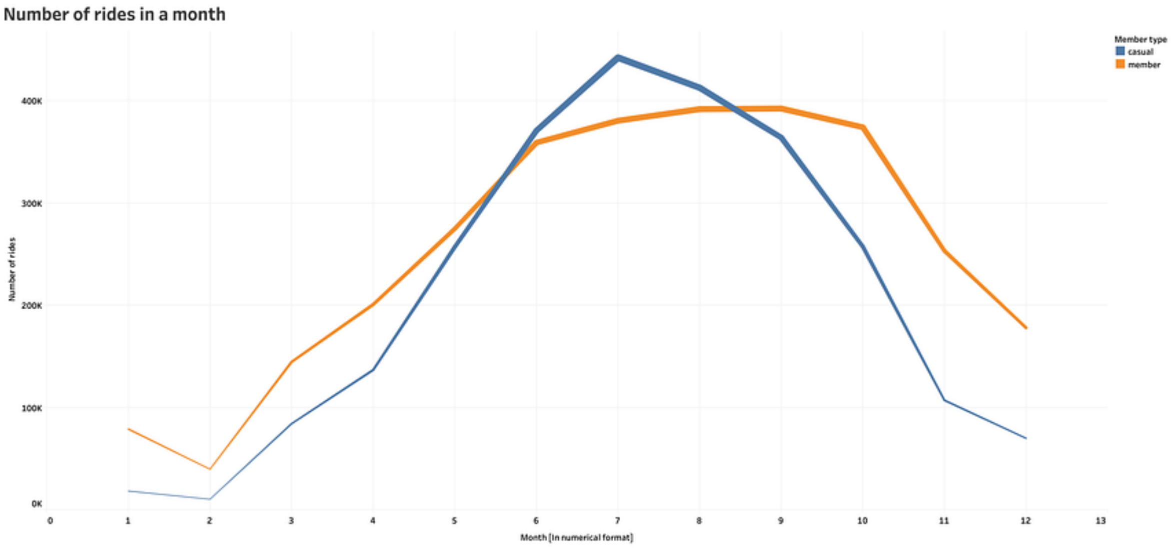
Here, we can find that overall ridership for annual members is fairly stable across the week. Such would indicate that there is a possibility that annual members are using the bikes as their main mode of transportation.

On the other hand, ridership for casual members is fairly low on the weekdays but starts to ramp up on Fridays and eventually peaks on Saturdays.



Here, we see that casual members have significantly longer average trip durations than annual members, nearly double in fact. This could signify that casual members are mainly using the bikes for leisure and or possibly sports activities which would.

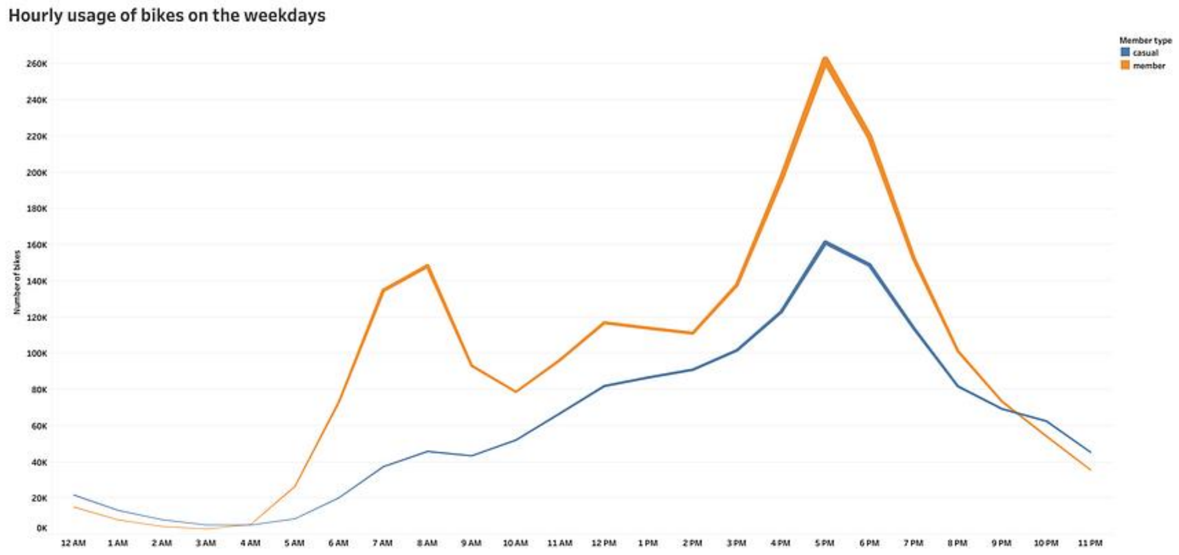
Whereas annual members would very likely use the bikes to commute from their living quarters to their offices and vice versa.



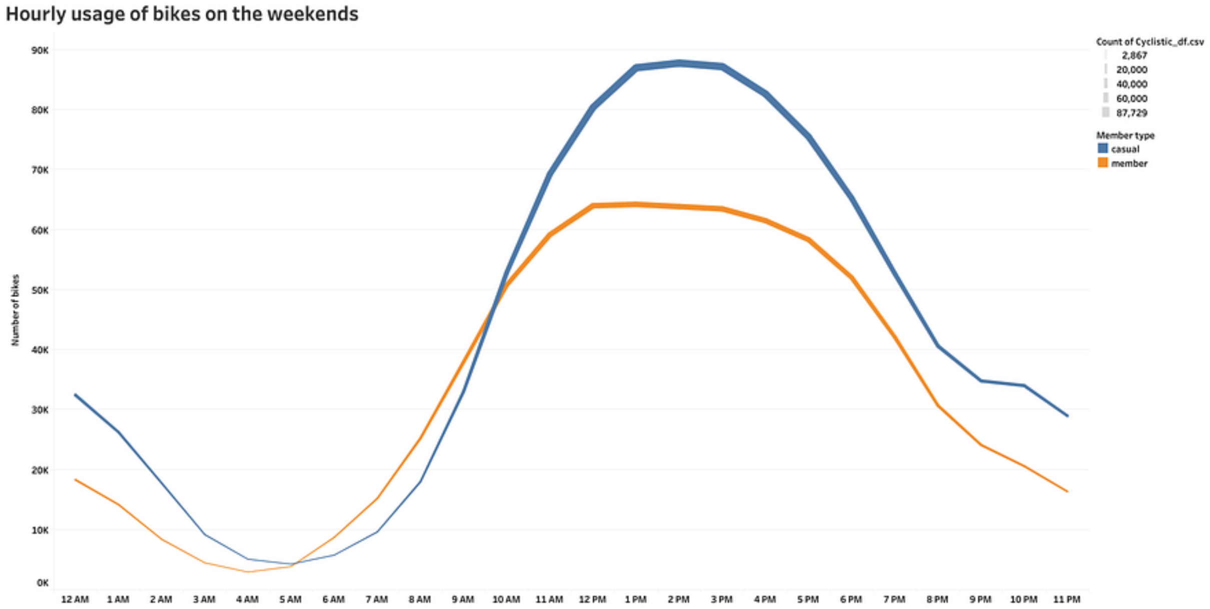
Numer of rides in a month throughout the period of a year

As we can see, ridership starts to freefall during the later months of the year. This is possible due to the change of seasons as usually in the months of October, the temperature starts to drop and the possibility of snow entails after.

Interestingly, casual member ridership peaks in the month of July. Knowing that schools in Chicago end around the middle of July, and resumes at end of August, we could hypothesize that the majority of the said users are not high school students.



Hourly bike usage on the weekdays

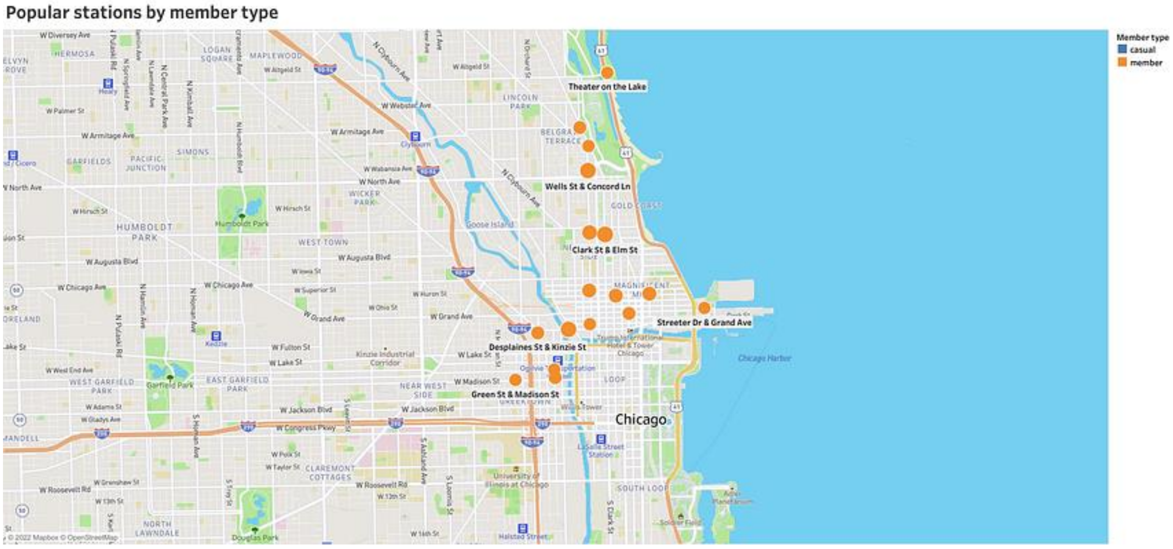


Hourly bike usage on the weekends

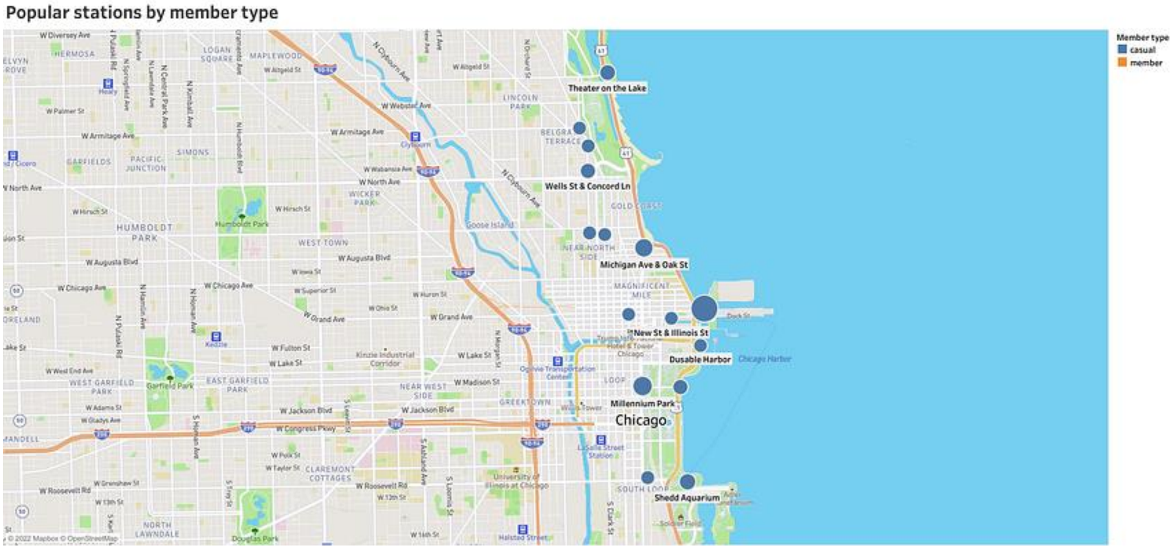
Based on the weekday graph, it would further reinforce my previous hypothesis whereby annual members are working adults, as we can see from:

- 7 am-8 am: A rally in usage, which could indicate when they’ve begun commuting to work
- 12 pm: An increase in usage, which would indicate lunch hour
- 5 pm: A peak in usage, which again falls in line with the office off-hours

The weekends, on the other hand, see a dramatic increase in ridership for the casual members starting from 10 am and peaks in the afternoon.



Popular stations among annual members (16k records and above only)

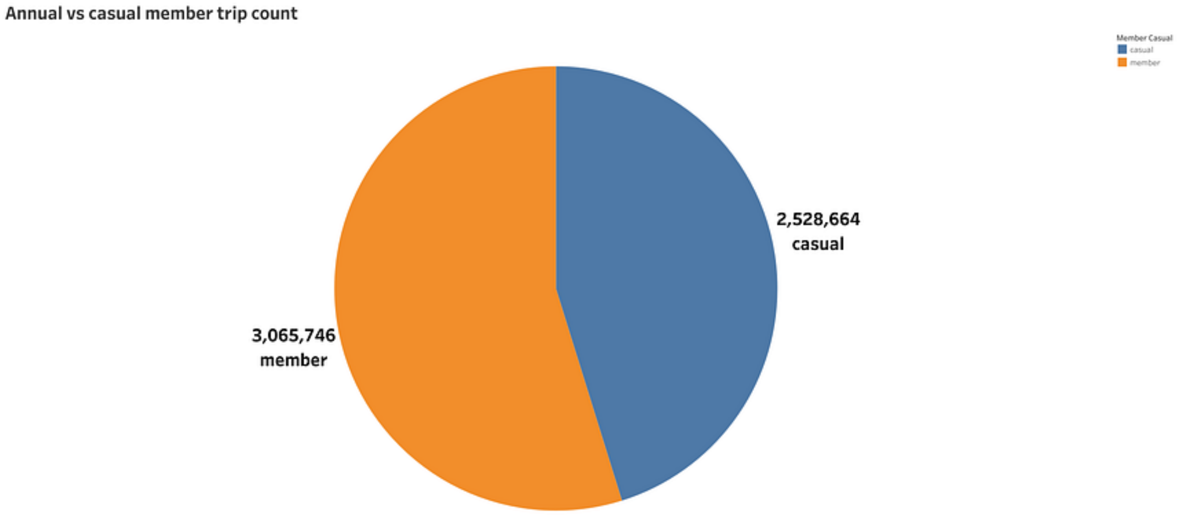


Popular stations among casual members (16k records and above only)

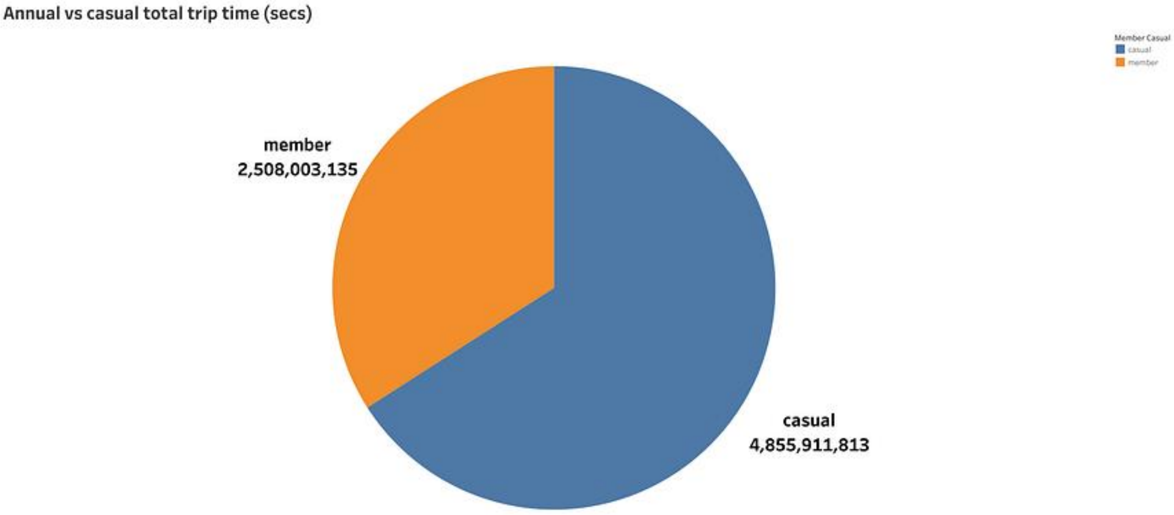
As shown in the plots above, we see that the frequently visited stations among the annual members are evenly spread and densely located nearby offices and working spaces.

Whereas the popular stations among casual members are located closer to the coastline, which could indicate possible sightseeing and or leisure activities carried out by the said members.

Before we end, I would like to confirm the trip counts, and trip durations by each member type and include any possible outliers as well.



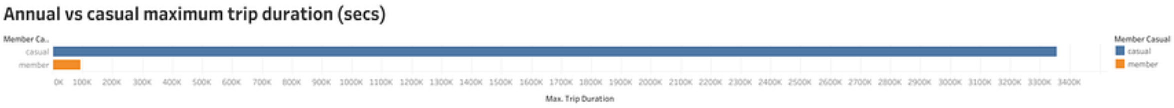
Total trips made by member type



total trip duration by member type

As shockingly shown, even though annual members made the most trips, a fair bit more than casual members, the total time spent on the bikes themselves is nearly double that of annual members.

This would further strengthen the previous hypothesis that casual members are using bikes for sightseeing/leisure purposes.



As we can see, for some reason, there is a casual user which managed to clock in a ridiculous 3,356,649 seconds on his bike. That’s equivalent to approximately 932 hours on a bike.

Act

Based on my findings after my analysis and along to conclude my observations, I would like to share my hypothesis on this matter.

1. I strongly believe that the casual members are mainly composed of tourists and or families who wish to spend their trips and or weekends sightseeing as well as carrying out leisure activities.
2. There’s a strong inclination to believe that annual members are mainly compromised of working adults which use our services as their means of transportation.
3. There’s a possibility that an exploit exists within the single ride pass which would allow irresponsible users to basically own a bike for themselves while only paying a fee for one-time use.

Now, to answer Moreno’s and her team’s request, which was: To design marketing strategies to convert casual riders into annual members. I would suggest the following:

1. We can clearly see a peak in casual riders on a few occasions: On the weekends as well as in the months of June, July & August. we should prioritize marketing on the said occasions.
2. As a follow up to the previous suggestions, we should advertise promotions on the previous point whereby current casual members would be able to upgrade to annual members at a discount.
3. I would suggest strategically enforcing location-based advertisements (featured on Instagram & Facebook) to target the popular stations among the casual members.

Let's move along, into now recommendations that I
would suggest to encourage casual members to upgrade to annual ones.

1. Increase the pricing of single-day & full-day passes. By strategically pricing it higher, it would appeal to upgrade to an annual membership. A great example is described here.
2. Charge/Impose additional fees for non-annual members based on trip duration. A great start would be to impose an additional 10% of your membership fee every 10 minutes after hitting the daily quota.

Additional remarks:

- It would be great if I had a dataset that did not format single-ride passes and full-day passes into ‘casual’ in the member types as we would be able to more intricately analyze the data regarding trip durations and make more specific recommendations
- Unique user IDs would allow me to count how many times an individual has used the service which would allow for more intricate and strategic promotions