Case Study 1: How Does a Bike-Share Navigate Speedy Success?

Michele Jean-Louis

Last updated: October 12, 2023

Google Data Analytics Professional Certificate

Cyclistic is a fictional bike-share company in Chicago. The company wants to maximalize the number of annual memberships and therefore wants to understand how do annual members and casual riders use Cyclistic bikes differently?

STEP 1: ASK

Business Task (What is the problem you are trying to solve?)

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

Key Stakeholders

Lily Moren, director of marketing and manager

Cyclistic executive teams

How can your insights drive business decisions?

My insights will help the marketing team design marketing strategies aimed at converting casual riders into annual members.

STEP 2: PREPARE

- Where is your data located? The data has been made available by Motivate International Inc.
- How is the data organized? The data is organized by trip by quarter up to 2020 and then we get monthly data files up to July 2023. For each trip, we have the start and end time, the bike ID, the bike type, the starting location and ending location (with latitude and longitude) and the usertype (casual or member).
- Are there issues with bias or credibility in this data? Does your data ROCCC? There doesn't seem to be any issues with bias or credibility in this data. The data is reliable, original, comprehensive, current and cited.
- How are you addressing licensing, privacy, security, and accessibility? The data was made available under the divvybikes.com data license agreement.

● How did you verify the data's integrity? I went to the source and verified that the source was credible.

● Are there any problems with the data? Some end times are before the start times. Some locations are missing.

We can filter the data by the customer vs subscriber and analyze the difference in number of trips, trip durations, genders and ages.

STEP 3: PROCESS

● What tools are you choosing and why? I am choosing to work in R since it is a very large dataset, but I will do some data manipulation in Excel and use Tableau for visualizations.

Data cleaning & wrangling

Microsoft Excel

1. First made a copy of the original data before wrangling.
2. Changed the format of started_at and ended_at columns to Custom: yyyy-mm-dd h:mm:ss
3. Created a column named trip_duration by substracting started_at from ended_at (=D2-C2) and then changed the format to hh:mm:ss.
4. Checked for the largest and smallest value to make sure there are no negative values and no values greater than 24 hours. There are 0 values for trip duration where the start time equals the end time. There are also some negative values, the end start is before the start time, this could be a user error. Since we have so many values in each monthly dataset (over 700,000), we will remove those entries.

Data File	Original number of entries	Negative trip_duration entries removed
July 2023	767,650	30
June 2023	719,618	7
May 2023	604,827	10
April 2023	426,590	4
March 2023	258,678	0
February 2023	190,545	1
January 2023	190,301	
December 2022	181,806	
November 2022	337,735	41
October 2022	558,685	4
September 2022	701,339	9
August 2022	785,932	11

Trips that span over multiple days.

5. Created a date column (=DATE(YEAR(C2),MONTH(C2),DAY(C2)))
6. Created a month column and converted it to number (=VALUE(MONTH(O2))

7. Created a quarter column (=IFS(OR(P2=1,P2=2,P2=3),"Q1",OR(P2=4,P2=5,P2=6),"Q2",OR(P2=7,P2=8,P2=9),"Q3",OR(P2=10, P2=11,P2=12),"Q4"))
8. Created a day column for the day of the week (=WEEKDAY(O2,1). 1 refers to Sunday and 7 to Saturday.

Moving to RStudio and converting script to RMarkdown

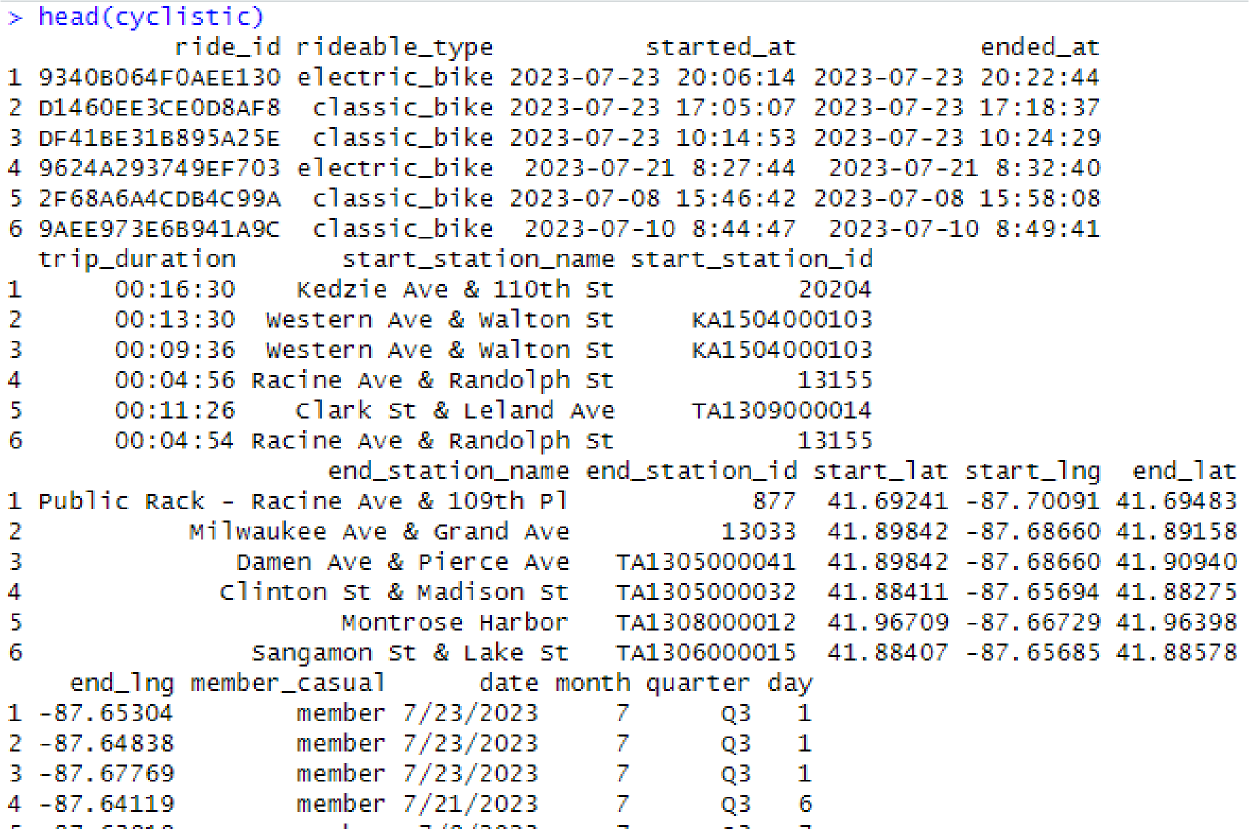
```
# First load and install the necessary packages

install.packages("tidyverse")

library(tidyverse)

library(janitor)

library(ggmap)

library(geosphere)

library(lubridate)


# Import the data

jul23 <- read.csv("202307-divvy-tripdata.csv")

jun23 <- read.csv("202306-divvy-tripdata.csv")

may23 <- read.csv("202305-divvy-tripdata.csv")

apr23 <- read.csv("202304-divvy-tripdata.csv")

mar23 <- read.csv("202303-divvy-tripdata.csv")

feb23 <- read.csv("202302-divvy-tripdata.csv")

jan23 <- read.csv("202301-divvy-tripdata.csv")

dec22 <- read.csv("202212-divvy-tripdata.csv")

nov22 <- read.csv("202211-divvy-tripdata.csv")

oct22 <- read.csv("202210-divvy-tripdata.csv")

sep22 <- read.csv("202209-divvy-publictripdata.csv")

aug22 <- read.csv("202208-divvy-tripdata.csv")


# Combining data frames with same column name
```

# We will only take the most 3 recent datasets as the data is too big

cyclistic_3months <- rbind(jul23, jun23, may23)

cyclistic <- rbind(jul23, jun23, may23, apr23, mar23, feb23, jan23, dec22, nov22, oct22, sep22, aug22)

head(cyclistic)

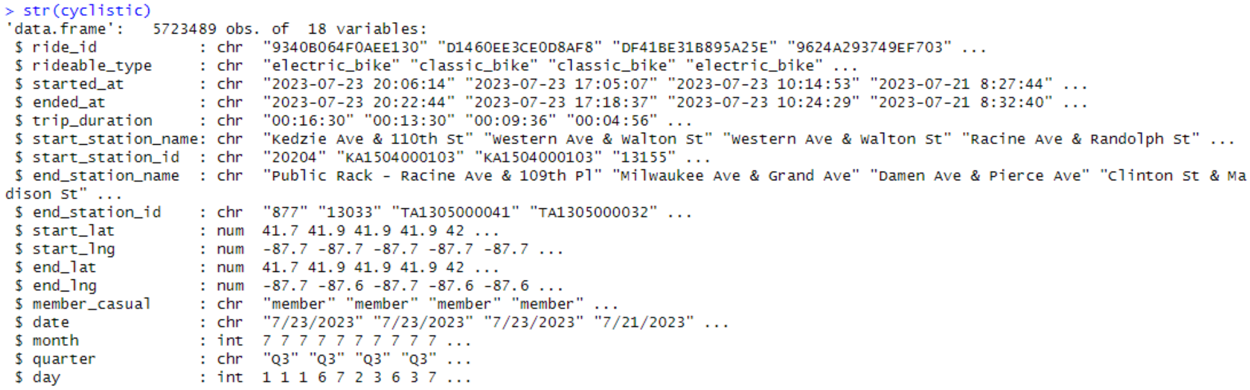


```
> head(cyclistic)
           ride_id rideable_type          started_at            ended_at
1 9340B064F0AEE130 electric_bike 2023-07-23 20:06:14 2023-07-23 20:22:44
2 D1460EE3CE0D8AF8  classic_bike 2023-07-23 17:05:07 2023-07-23 17:18:37
3 DF41BE31B895A25E  classic_bike 2023-07-23 10:14:53 2023-07-23 10:24:29
4 9624A293749EF703 electric_bike  2023-07-21 8:27:44  2023-07-21 8:32:40
5 2F68A6A4CDB4C99A  classic_bike 2023-07-08 15:46:42 2023-07-08 15:58:08
6 9AEE973E6B941A9C  classic_bike  2023-07-10 8:44:47  2023-07-10 8:49:41
  trip_duration       start_station_name start_station_id
1      00:16:30    Kedzie Ave & 110th St            20204
2      00:13:30  Western Ave & Walton St     KA1504000103
3      00:09:36  Western Ave & Walton St     KA1504000103
4      00:04:56 Racine Ave & Randolph St            13155
5      00:11:26    Clark St & Leland Ave     TA1309000014
6      00:04:54 Racine Ave & Randolph St            13155
                    end_station_name end_station_id start_lat start_lng  end_lat
1 Public Rack - Racine Ave & 109th Pl            877  41.69241 -87.70091 41.69483
2           Milwaukee Ave & Grand Ave          13033  41.89842 -87.68660 41.89158
3              Damen Ave & Pierce Ave   TA1305000041  41.89842 -87.68660 41.90940
4             Clinton St & Madison St   TA1305000032  41.88411 -87.65694 41.88275
5                     Montrose Harbor   TA1308000012  41.96709 -87.66729 41.96398
6               Sangamon St & Lake St   TA1306000015  41.88407 -87.65685 41.88578
    end_lng member_casual      date month quarter day
1 -87.65304        member 7/23/2023     7      Q3   1
2 -87.64838        member 7/23/2023     7      Q3   1
3 -87.67769        member 7/23/2023     7      Q3   1
4 -87.64119        member 7/21/2023     7      Q3   6
```

# We now have 5,723,489 observations with 18 variables. We can now start our analysis.

# Let's look at the structure of the data

str(cyclistic)



```
> str(cyclistic)
'data.frame':	5723489 obs. of  18 variables:
 $ ride_id           : chr  "9340B064F0AEE130" "D1460EE3CE0D8AF8" "DF41BE31B895A25E" "9624A293749EF703" ...
 $ rideable_type     : chr  "electric_bike" "classic_bike" "classic_bike" "electric_bike" ...
 $ started_at        : chr  "2023-07-23 20:06:14" "2023-07-23 17:05:07" "2023-07-23 10:14:53" "2023-07-21 8:27:44" ...
 $ ended_at          : chr  "2023-07-23 20:22:44" "2023-07-23 17:18:37" "2023-07-23 10:24:29" "2023-07-21 8:32:40" ...
 $ trip_duration     : chr  "00:16:30" "00:13:30" "00:09:36" "00:04:56" ...
 $ start_station_name: chr  "Kedzie Ave & 110th St" "Western Ave & Walton St" "Western Ave & Walton St" "Racine Ave & Randolph St" ...
 $ start_station_id  : chr  "20204" "KA1504000103" "KA1504000103" "13155" ...
 $ end_station_name  : chr  "Public Rack - Racine Ave & 109th Pl" "Milwaukee Ave & Grand Ave" "Damen Ave & Pierce Ave" "Clinton St & Ma
dison St" ...
 $ end_station_id    : chr  "877" "13033" "TA1305000041" "TA1305000032" ...
 $ start_lat         : num  41.7 41.9 41.9 41.9 42 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.7 41.9 41.9 41.9 42 ...
 $ end_lng           : num  -87.7 -87.6 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
 $ date              : chr  "7/23/2023" "7/23/2023" "7/23/2023" "7/21/2023" ...
 $ month             : int  7 7 7 7 7 7 7 7 7 7 ...
 $ quarter           : chr  "Q3" "Q3" "Q3" "Q3" ...
 $ day               : int  1 1 1 6 7 2 3 6 3 7 ...
```

# We can see that the structure of started_at, ended_at & trip_duration columns is character. We need to change is to the appropriate type.

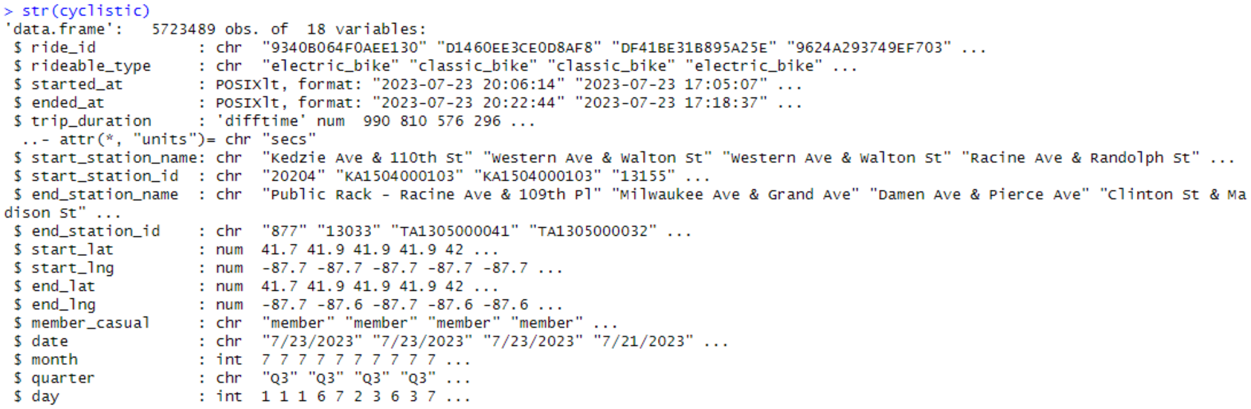
cyclistic$started_at <- strptime(cyclistic$started_at, "%Y-%m-%d %H:%M:%S")

cyclistic$ended_at <- strptime(cyclistic$ended_at, "%Y-%m-%d %H:%M:%S")

cyclistic$trip_duration <- difftime(cyclistic$ended_at, cyclistic$started_at)

# Let's look at the structure of the data again

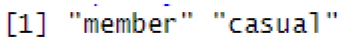
str(cyclistic)



```
> str(cyclistic)
'data.frame':   5723489 obs. of  18 variables:
 $ ride_id           : chr  "9340B064F0AEE130" "D1460EE3CE0D8AF8" "DF41BE31B895A25E" "9624A293749EF703" ...
 $ rideable_type     : chr  "electric_bike" "classic_bike" "classic_bike" "electric_bike" ...
 $ started_at        : POSIXlt, format: "2023-07-23 20:06:14" "2023-07-23 17:05:07" ...
 $ ended_at          : POSIXlt, format: "2023-07-23 20:22:44" "2023-07-23 17:18:37" ...
 $ trip_duration     : 'difftime' num  990 810 576 296 ...
  ..- attr(*, "units")= chr "secs"
 $ start_station_name: chr  "Kedzie Ave & 110th St" "Western Ave & Walton St" "Western Ave & Walton St" "Racine Ave & Randolph St" ...
 $ start_station_id  : chr  "20204" "KA1504000103" "KA1504000103" "13155" ...
 $ end_station_name  : chr  "Public Rack - Racine Ave & 109th Pl" "Milwaukee Ave & Grand Ave" "Damen Ave & Pierce Ave" "Clinton St & Ma
dison St" ...
 $ end_station_id    : chr  "877" "13033" "TA1305000041" "TA1305000032" ...
 $ start_lat         : num  41.7 41.9 41.9 41.9 42 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.7 41.9 41.9 41.9 42 ...
 $ end_lng           : num  -87.7 -87.6 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
 $ date              : chr  "7/23/2023" "7/23/2023" "7/23/2023" "7/21/2023" ...
 $ month             : int  7 7 7 7 7 7 7 7 7 7 ...
 $ quarter           : chr  "Q3" "Q3" "Q3" "Q3" ...
 $ day               : int  1 1 1 6 7 2 3 6 3 7 ...
```

# Type of members vs casual

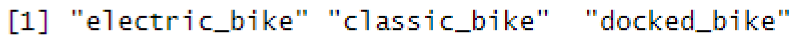
unique(cyclistic$member_casual)



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

# Type of ride type

unique(cyclistic$rideable_type)

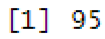


```
[1] "electric_bike" "classic_bike"  "docked_bike"
```

# There are 3 different types of ride in our dataset: electric_bike, classic_bike and docked_bike.

# Checking for duplicates, there are 95 duplicates

sum(duplicated(cyclistic$ride_id))



```
[1] 95
```

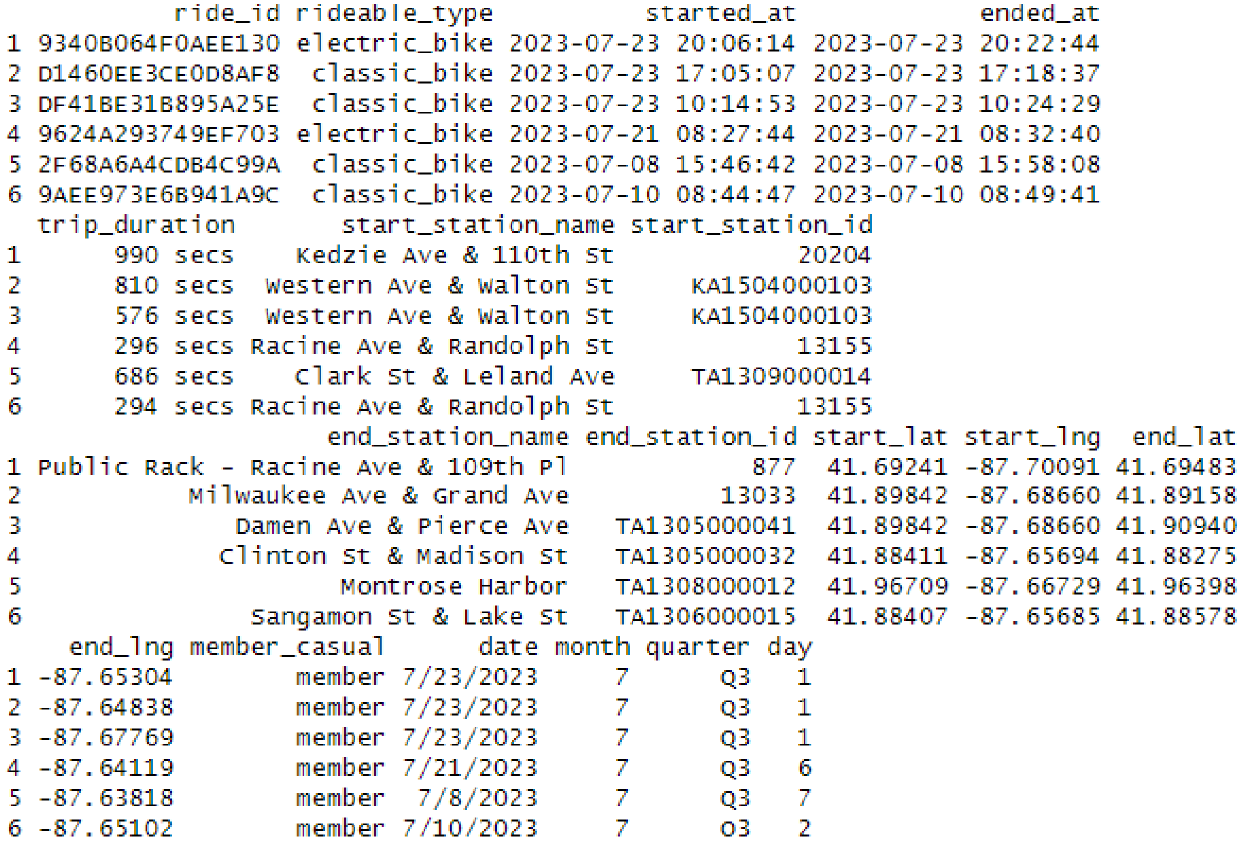
# Let's check again for trip_duration that is negative

cyclistic <- cyclistic%>%

  distinct(ride_id, .keep_all = TRUE)%>%

```
  filter(trip_duration > 0)
```

```
head(cyclistic)
```



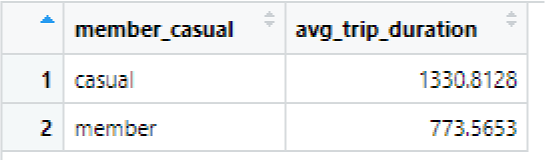
```
           ride_id rideable_type          started_at            ended_at
1 9340B064F0AEE130 electric_bike 2023-07-23 20:06:14 2023-07-23 20:22:44
2 D1460EE3CE0D8AF8  classic_bike 2023-07-23 17:05:07 2023-07-23 17:18:37
3 DF41BE31B895A25E  classic_bike 2023-07-23 10:14:53 2023-07-23 10:24:29
4 9624A293749EF703 electric_bike 2023-07-21 08:27:44 2023-07-21 08:32:40
5 2F68A6A4CDB4C99A  classic_bike 2023-07-08 15:46:42 2023-07-08 15:58:08
6 9AEE973E6B941A9C  classic_bike 2023-07-10 08:44:47 2023-07-10 08:49:41
  trip_duration       start_station_name start_station_id
1      990 secs    Kedzie Ave & 110th St            20204
2      810 secs  Western Ave & Walton St     KA1504000103
3      576 secs  Western Ave & Walton St     KA1504000103
4      296 secs Racine Ave & Randolph St            13155
5      686 secs    Clark St & Leland Ave     TA1309000014
6      294 secs Racine Ave & Randolph St            13155
                    end_station_name end_station_id start_lat start_lng  end_lat
1 Public Rack - Racine Ave & 109th Pl            877  41.69241 -87.70091 41.69483
2           Milwaukee Ave & Grand Ave          13033  41.89842 -87.68660 41.89158
3              Damen Ave & Pierce Ave   TA1305000041  41.89842 -87.68660 41.90940
4             Clinton St & Madison St   TA1305000032  41.88411 -87.65694 41.88275
5                     Montrose Harbor   TA1308000012  41.96709 -87.66729 41.96398
6               Sangamon St & Lake St   TA1306000015  41.88407 -87.65685 41.88578
    end_lng member_casual      date month quarter day
1 -87.65304        member 7/23/2023     7      Q3   1
2 -87.64838        member 7/23/2023     7      Q3   1
3 -87.67769        member 7/23/2023     7      Q3   1
4 -87.64119        member 7/21/2023     7      Q3   6
5 -87.63818        member  7/8/2023     7      Q3   7
6 -87.65102        member 7/10/2023     7      Q3   2
```

```
# Now let's remove those entries with no station info

cycli <- cyclistic%>%

  drop_na()

nrow(cycli)
[1] 5516602
```

```
# We now have (5,516,602) rows. This is the data we will use to make analysis in Tableau.

write.csv(cycli, "cyclistic.csv")

write.table(cycli, "cyclistic.txt")
```

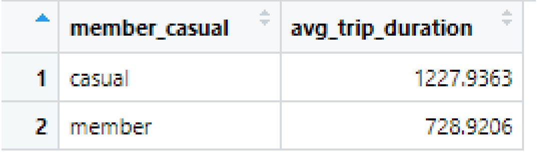
```
# Let's look at the number of rides per customer type.
```

```
ride_per_customer <- cycli%>%

  group_by(member_casual)%>%

  summarize(n=n())%>%

  mutate(percentage = n*100/sum(n))

view(ride_per_customer)
```



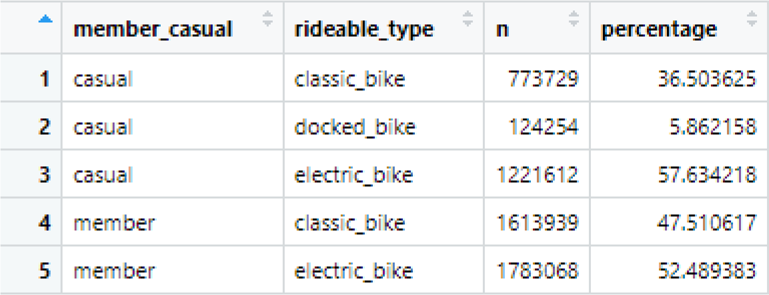
	member_casual	avg_trip_duration
1	casual	1330.8128
2	member	773.5653

```
ggplot(data = cycli, mapping= aes(x= member_casual, fill= member_casual)) + geom_bar() +
labs(title="No of Rides per Customer Type")


# Trip duration per customer type

cycli$trip_duration <- as.numeric(cycli$trip_duration)

trip_duration_per_customer <- cycli%>%

  group_by(member_casual)%>%

  summarize(avg_trip_duration = mean(trip_duration))

view(trip_duration_per_customer)
```



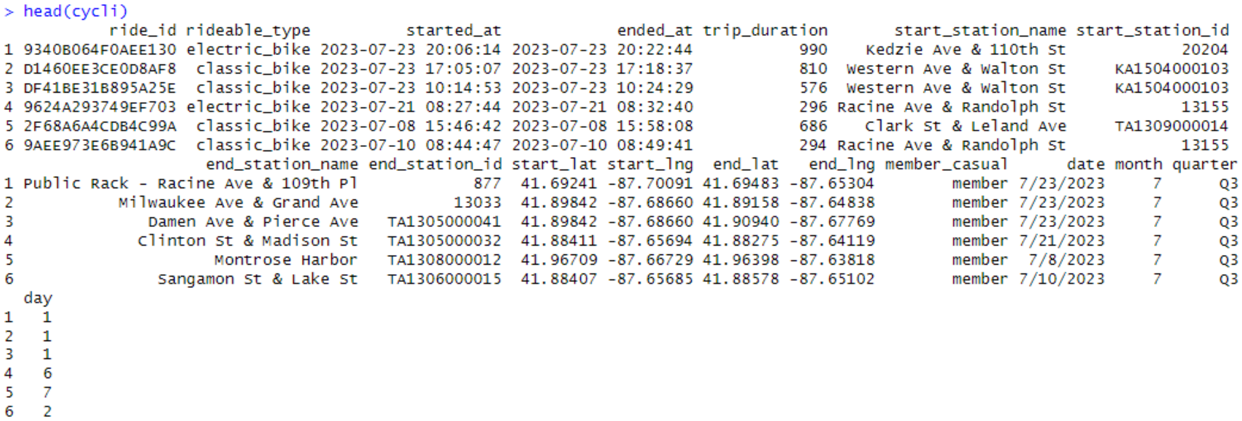
	member_casual	avg_trip_duration
1	casual	1227.9363
2	member	728.9206

```
# Ride Type per customer type

ridetype_per_customer <- cycli%>%

  group_by(member_casual, rideable_type)%>%

  summarize(n=n())%>%

  mutate(percentage = n*100/sum(n))

view(ridetype_per_customer)
```



	member_casual	rideable_type	n	percentage
1	casual	classic_bike	773729	36.503625
2	casual	docked_bike	124254	5.862158
3	casual	electric_bike	1221612	57.634218
4	member	classic_bike	1613939	47.510617
5	member	electric_bike	1783068	52.489383

head(cycli)



```
> head(cycli)
           ride_id rideable_type          started_at            ended_at trip_duration       start_station_name start_station_id
1 9340B064F0AEE130 electric_bike 2023-07-23 20:06:14 2023-07-23 20:22:44           990    Kedzie Ave & 110th St            20204
2 D1460EE3CE0D8AF8  classic_bike 2023-07-23 17:05:07 2023-07-23 17:18:37           810  Western Ave & Walton St     KA1504000103
3 DF41BE31B895A25E  classic_bike 2023-07-23 10:14:53 2023-07-23 10:24:29           576  Western Ave & Walton St     KA1504000103
4 9624A293749EF703 electric_bike 2023-07-21 08:27:44 2023-07-21 08:32:40           296 Racine Ave & Randolph St            13155
5 2F68A6A4CDB4C99A  classic_bike 2023-07-08 15:46:42 2023-07-08 15:58:08           686    Clark St & Leland Ave     TA1309000014
6 9AEE973E6B941A9C  classic_bike 2023-07-10 08:44:47 2023-07-10 08:49:41           294 Racine Ave & Randolph St            13155
                     end_station_name end_station_id start_lat start_lng  end_lat   end_lng member_casual      date month quarter
1 Public Rack - Racine Ave & 109th Pl            877  41.69241 -87.70091 41.69483 -87.65304        member 7/23/2023     7      Q3
2           Milwaukee Ave & Grand Ave          13033  41.89842 -87.68660 41.89158 -87.64838        member 7/23/2023     7      Q3
3              Damen Ave & Pierce Ave   TA1305000041  41.89842 -87.68660 41.90940 -87.67769        member 7/23/2023     7      Q3
4             Clinton St & Madison St   TA1305000032  41.88411 -87.65694 41.88275 -87.64119        member 7/21/2023     7      Q3
5                     Montrose Harbor   TA1308000012  41.96709 -87.66729 41.96398 -87.63818        member  7/8/2023     7      Q3
6               Sangamon St & Lake St   TA1306000015  41.88407 -87.65685 41.88578 -87.65102        member 7/10/2023     7      Q3
  day
1   1
2   1
3   1
4   6
5   7
6   2
```

View(cycli)

TABLEAU

In Tableau, I created a new calculated field (trip duration in minutes) for easier visualization.

I created different plots in order to quickly view and understand the data.

Report

You will produce a report with the following deliverables:

1. A clear statement of the business task

2. A description of all data sources used

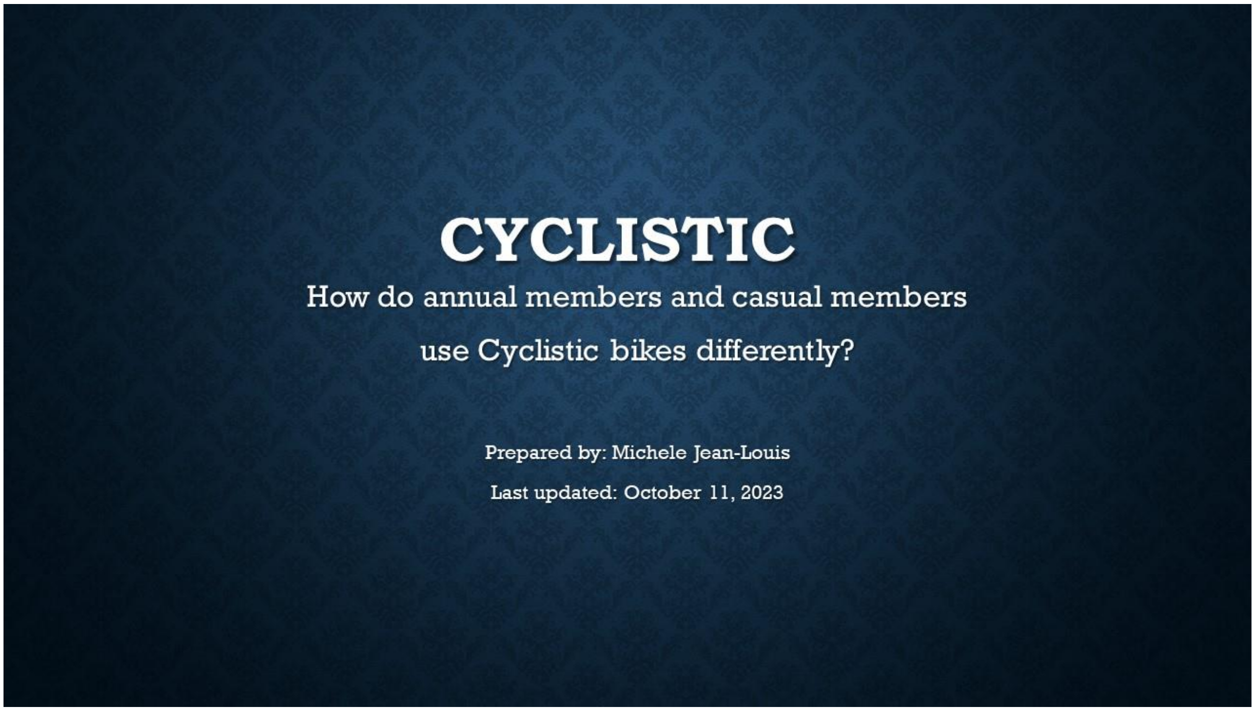
3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

6. Your top three recommendations based on your analysis

Here is the report:



CYCLISTIC

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

Prepared by: Michele Jean-Louis

Last updated: October 11, 2023

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

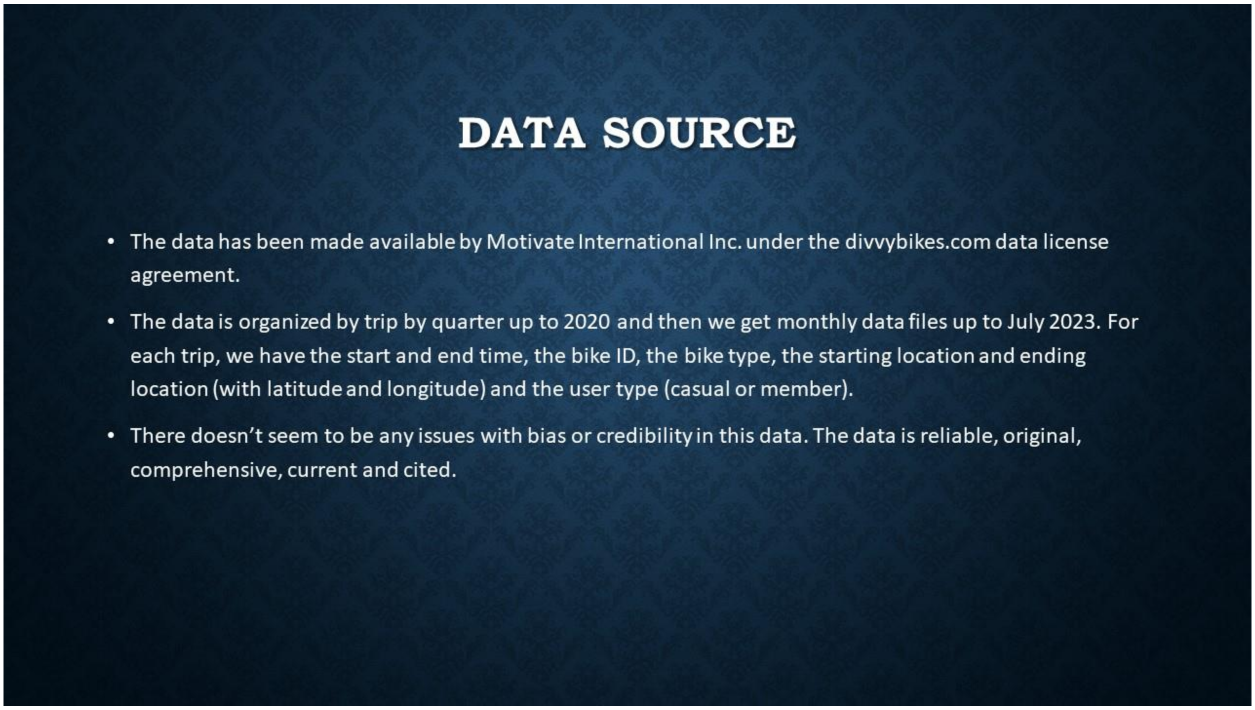
BUSINESS TASK

Cyclistic is a fictional bike-share company in Chicago. The company wants to maximalize the number of annual memberships and therefore wants to understand how do annual members and casual riders use Cyclistic bikes differently?

My insights will help the marketing team design marketing strategies aimed at converting casual riders into annual members.

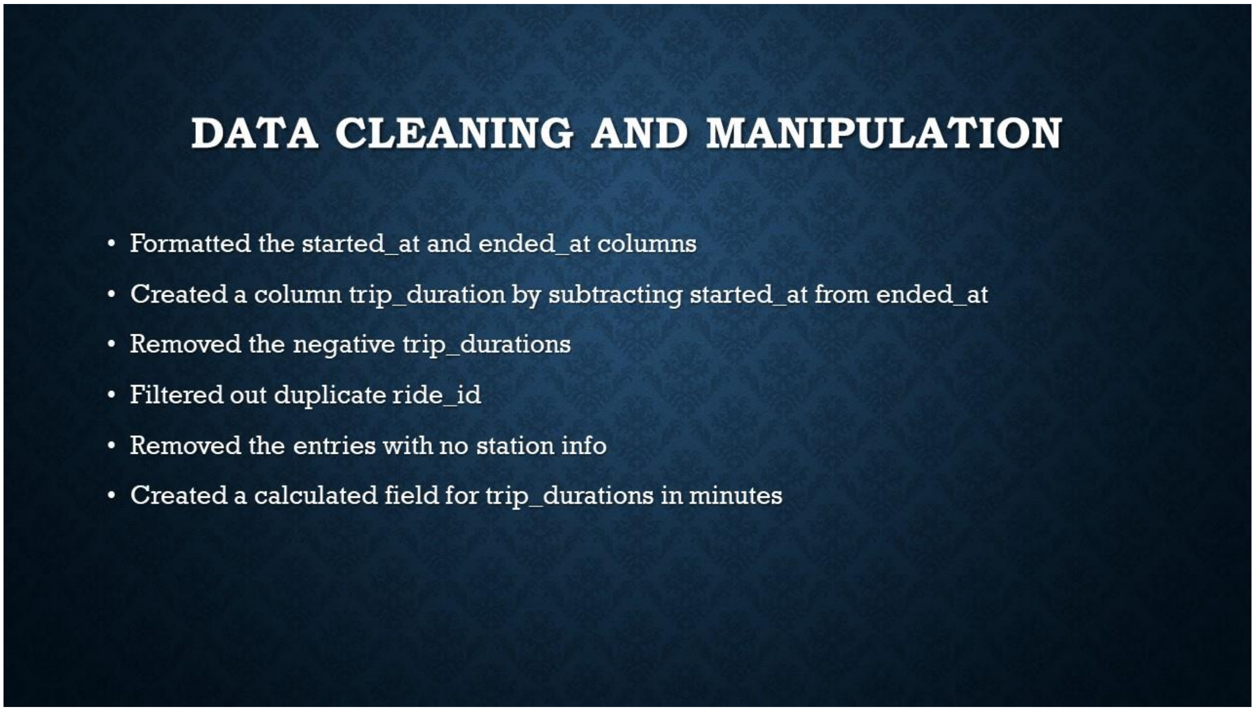
Key Stakeholders

- Lily Moren, director of marketing and manager
- Cyclistic executive team



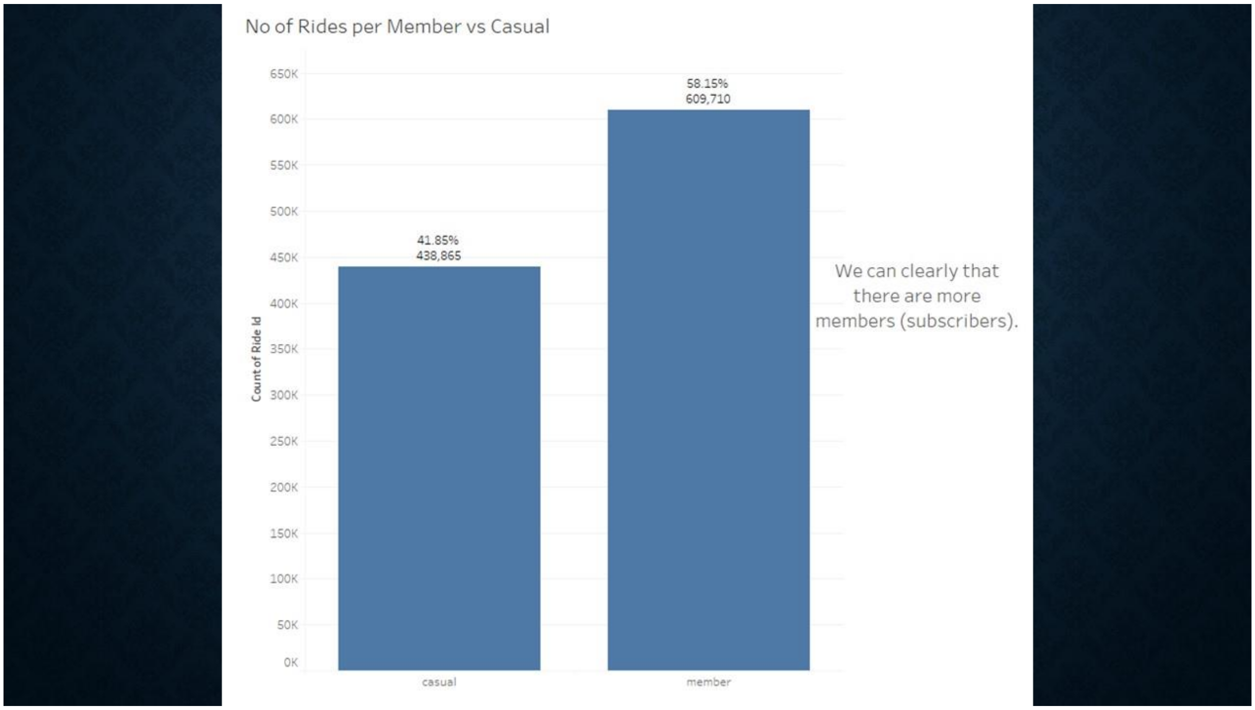
DATA SOURCE

- The data has been made available by Motivate International Inc. under the divvybikes.com data license agreement.
- The data is organized by trip by quarter up to 2020 and then we get monthly data files up to July 2023. For each trip, we have the start and end time, the bike ID, the bike type, the starting location and ending location (with latitude and longitude) and the user type (casual or member).
- There doesn't seem to be any issues with bias or credibility in this data. The data is reliable, original, comprehensive, current and cited.

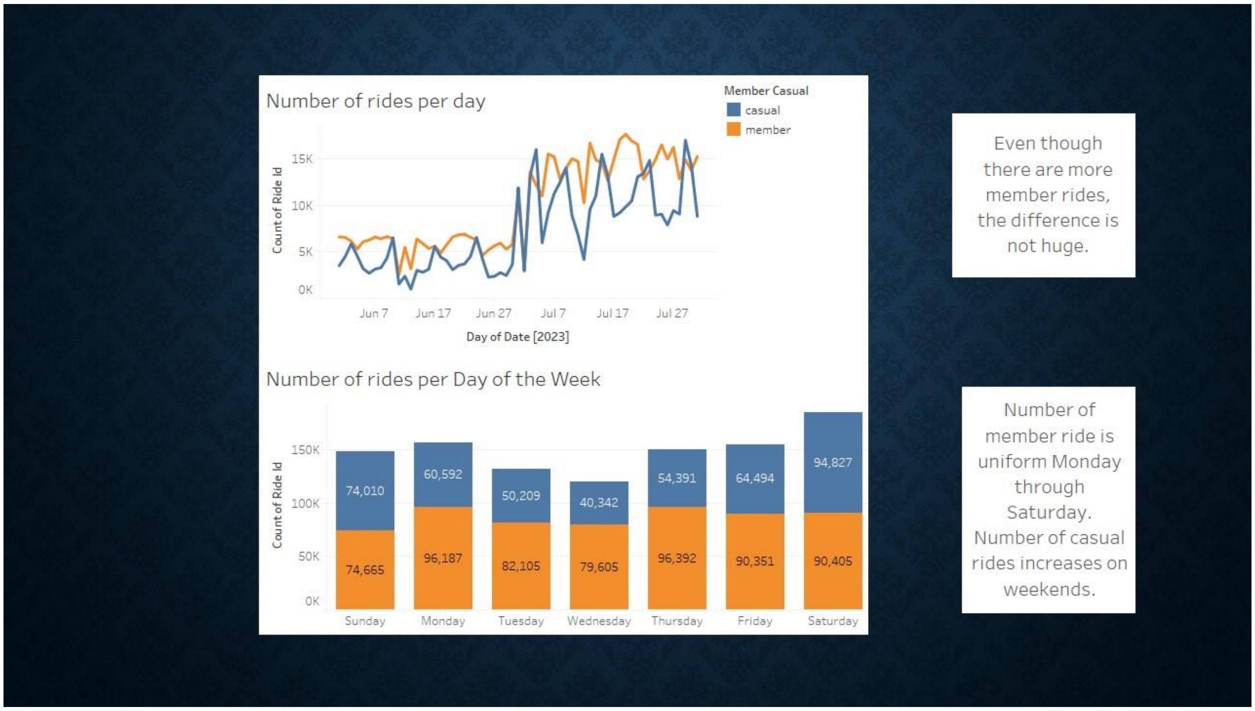


DATA CLEANING AND MANIPULATION

- Formatted the started_at and ended_at columns
- Created a column trip_duration by subtracting started_at from ended_at
- Removed the negative trip_durations
- Filtered out duplicate ride_id
- Removed the entries with no station info
- Created a calculated field for trip_durations in minutes

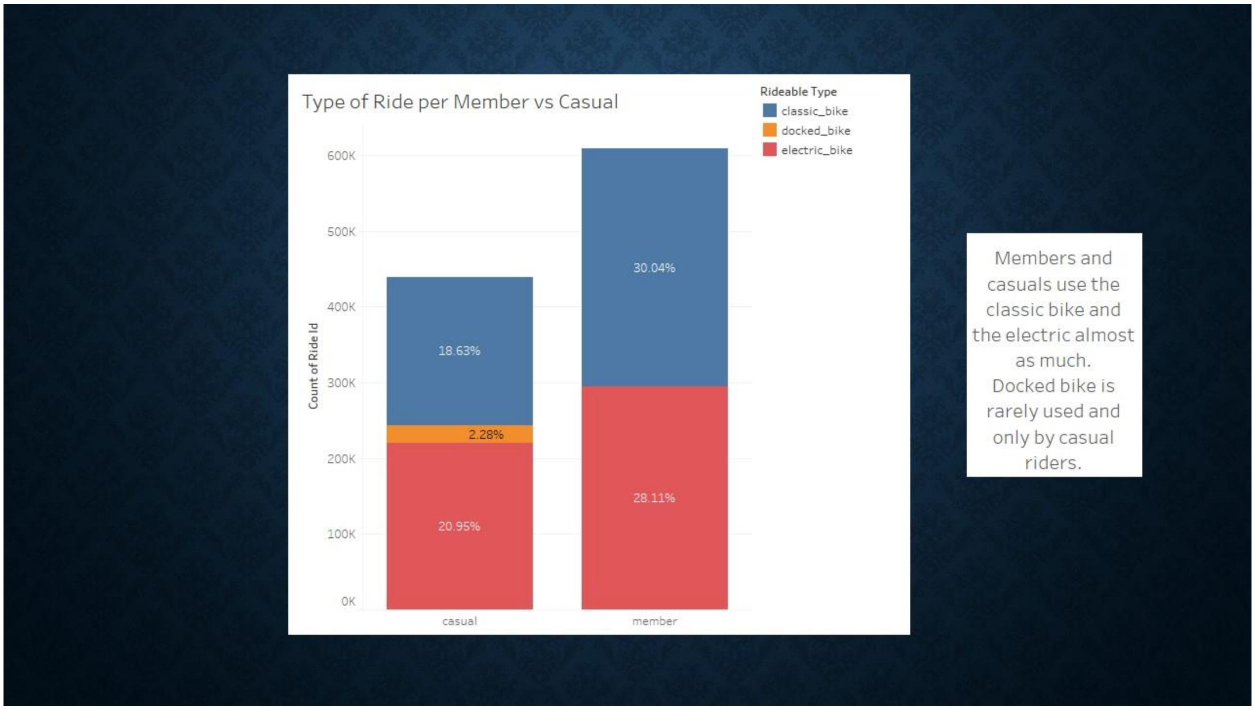


We can clearly that there are more members (subscribers).

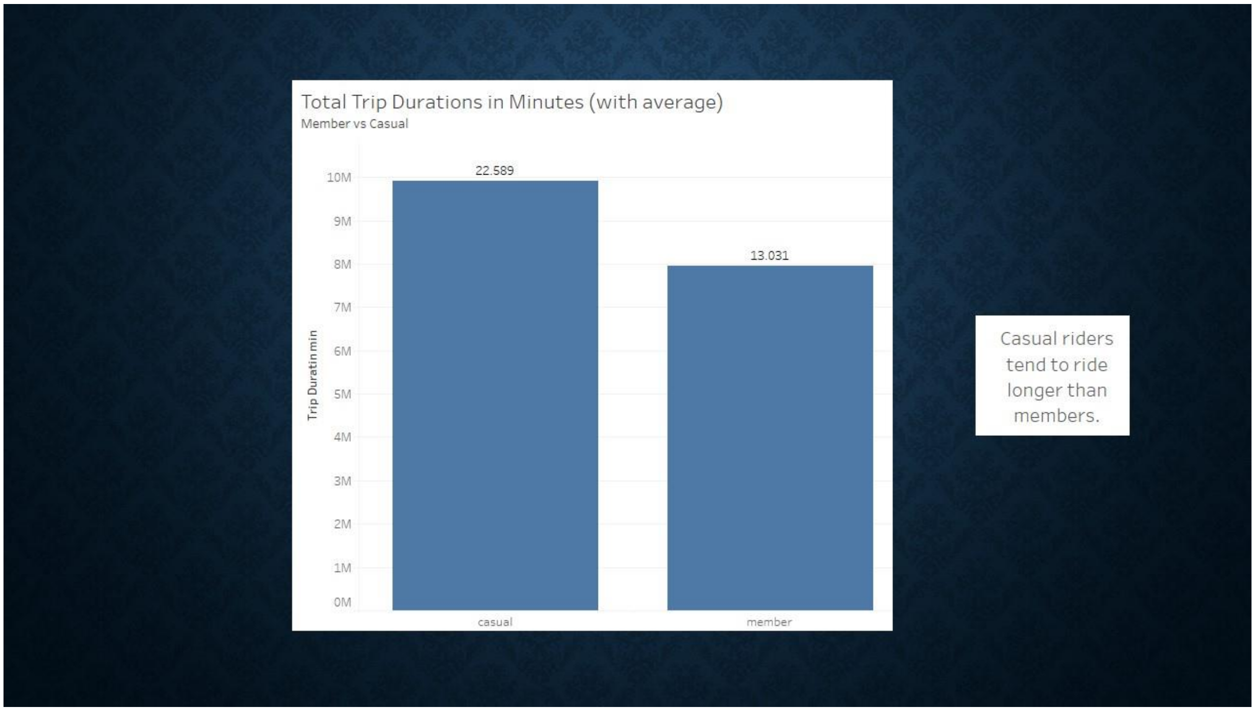


Even though there are more member rides, the difference is not huge.

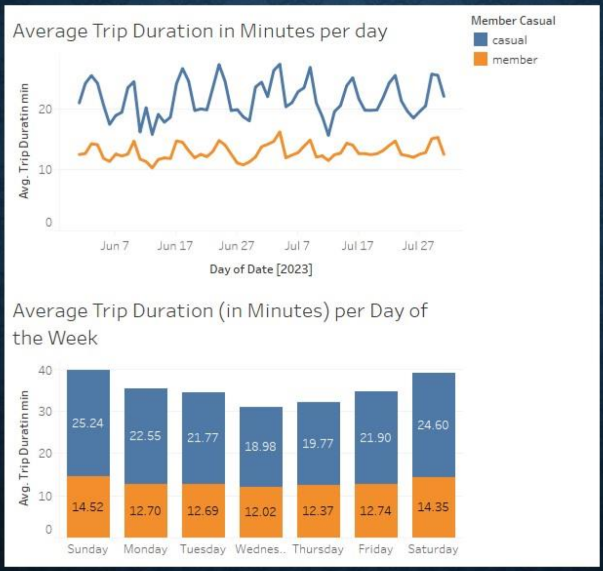
Number of member ride is uniform Monday through Saturday. Number of casual rides increases on weekends.



Members and casuals use the classic bike and the electric almost as much. Docked bike is rarely used and only by casual riders.



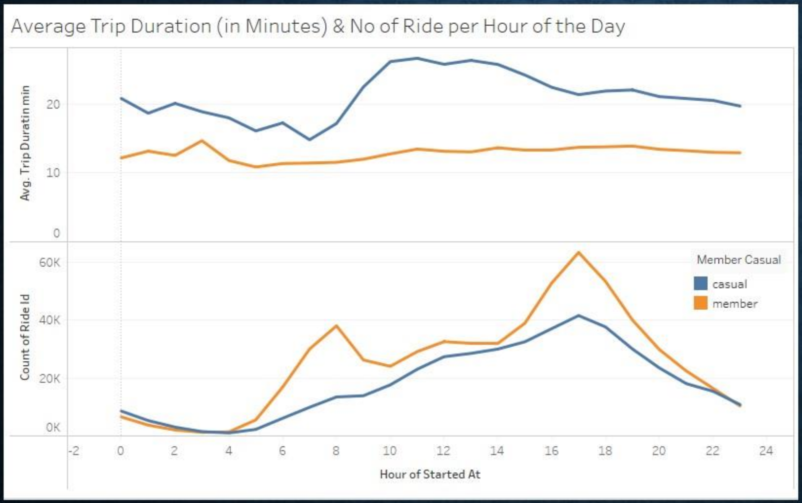
Casual riders tend to ride longer than members.



Average Trip Duration in Minutes per day

Average Trip Duration (in Minutes) per Day of the Week

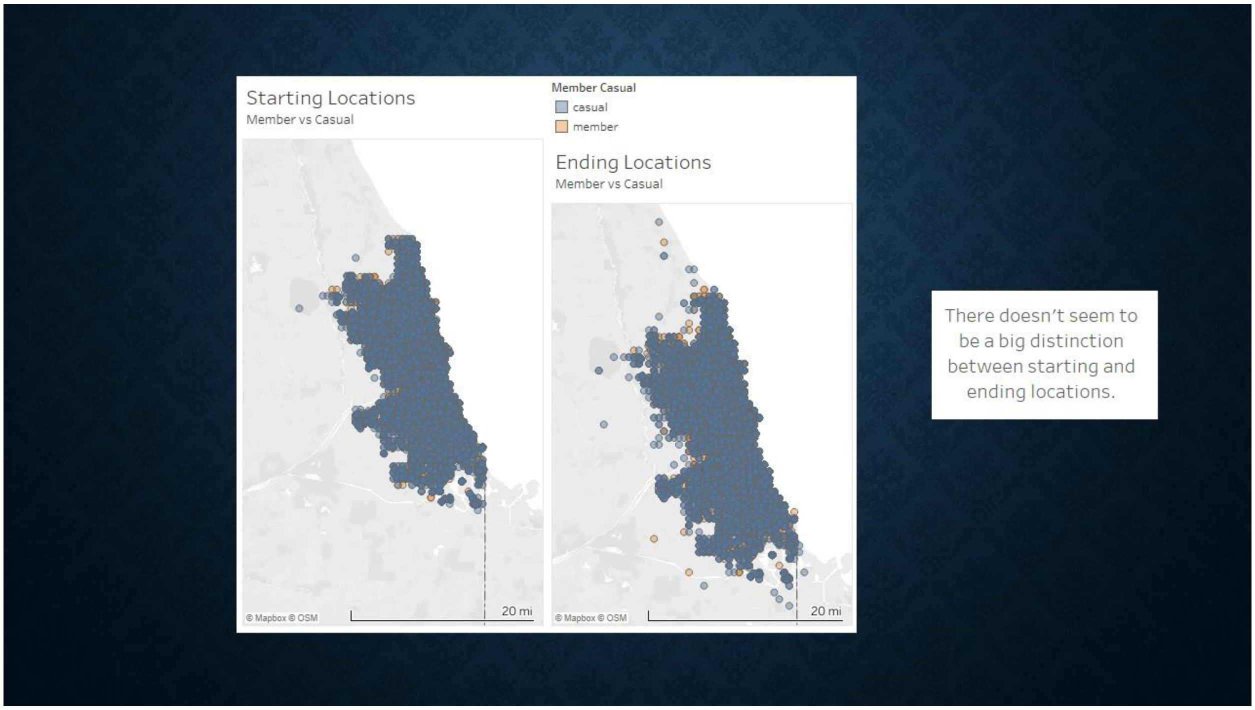
Whether by date or day of the week, casual rides duration are longer than member rides duration.



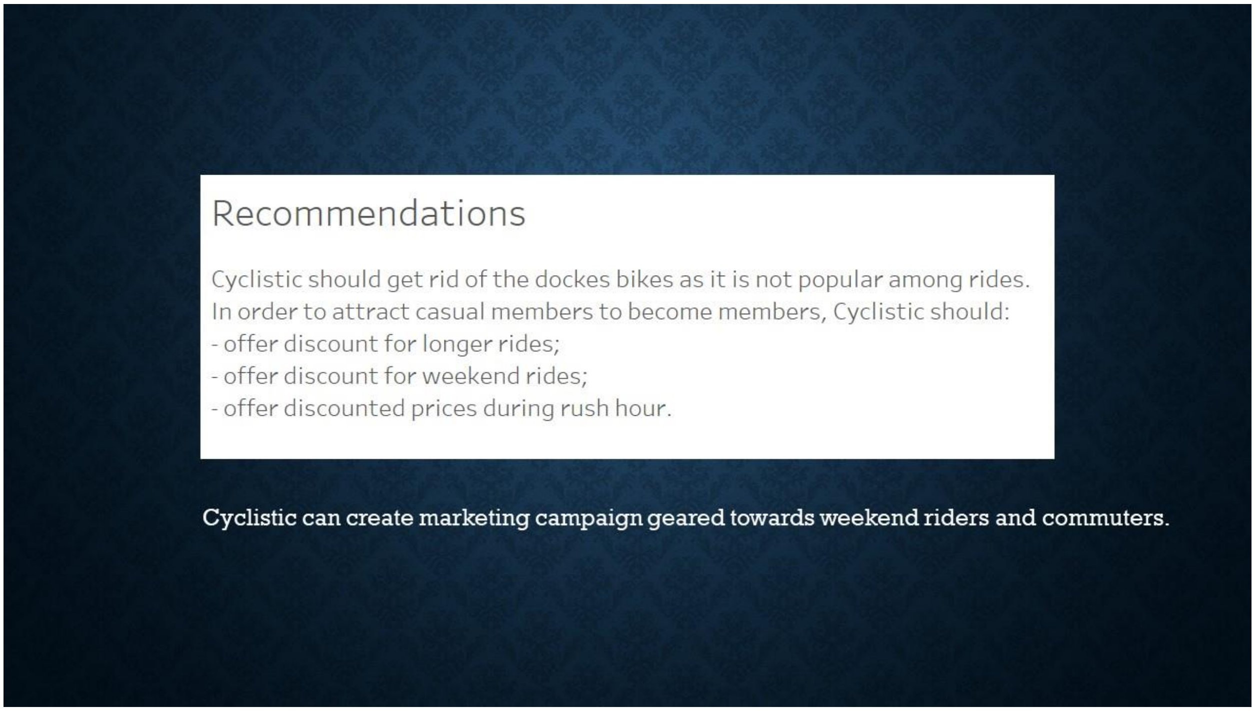
Average Trip Duration (in Minutes) & No of Ride per Hour of the Day

- The average trip duration is uniform for members throughout the data, but for casual riders the length of trips increase between 9AM and 5 PM.

- There is a peak in the number of rides for members at 8AM and 5PM which would indicate commute to and from work.
- For casual riders, the number of rides increases linearly from 5AM to 5PM and then decreases.



There doesn't seem to be a big distinction between starting and ending locations.



Recommendations

Cyclistic should get rid of the dockes bikes as it is not popular among rides.
In order to attract casual members to become members, Cyclistic should:
- offer discount for longer rides;
- offer discount for weekend rides;
- offer discounted prices during rush hour.

Cyclistic can create marketing campaign geared towards weekend riders and commuters.

REFERENCES

- Tableau Dashboard: https://public.tableau.com/app/profile/michele.jean.louis/viz/Cyclistic_16961906044630/Cyclistic?publish=yes
- GihHub: https://github.com/michelejl/Cyclistic
- Linkedin: https://www.linkedin.com/in/michelejeanlouis

Links

- Tableau Dashboard: https://public.tableau.com/app/profile/michele.jean.louis/viz/Cyclistic_16961906044630/Cyclistic?publish=yes
- GihHub: https://github.com/michelejl/Cyclistic
- Linkedin: https://www.linkedin.com/in/michelejeanlouis