Cyclistic BikeShare Data Analysis

Author: Nguyen Duong Khai

Date: 1/3/2022

Introduction

Cyclistic is a bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. Most riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

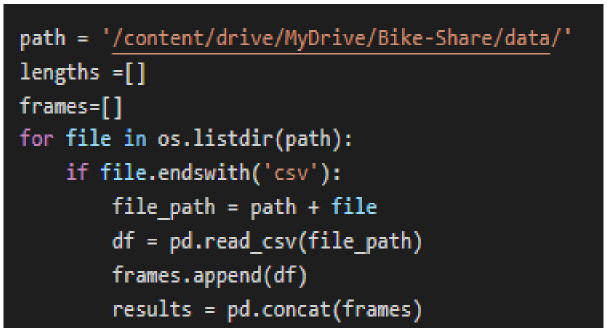
Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing 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.

Business Task

Working as a data analyst, I will explore what is different between casual and member users so that casual users become a member rider.

Prepare Data

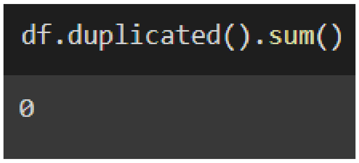
For this project, I used 12 datasets named "divvy-tripdata" from Feb-2021 to Jan-2022. Click the link to download and use these datasets. I am using Pandas to support me during this project. After downloading enough files, I merged 12 datasets into a csv named "2020-trip-data".



```
path = '/content/drive/MyDrive/Bike-Share/data/'
lengths =[]
frames=[]
for file in os.listdir(path):
    if file.endswith('csv'):
        file_path = path + file
        df = pd.read_csv(file_path)
        frames.append(df)
        results = pd.concat(frames)
```

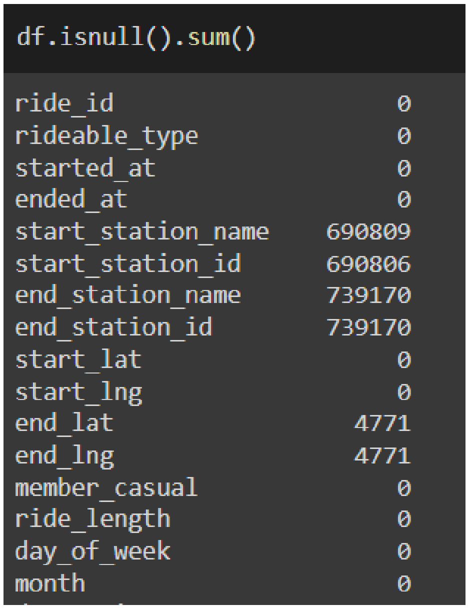
Processing

Before analyzing the data, I make sure the data clean. To perform this, I start check my data whether it has duplicate values and null values.



```
df.duplicated().sum()

0
```

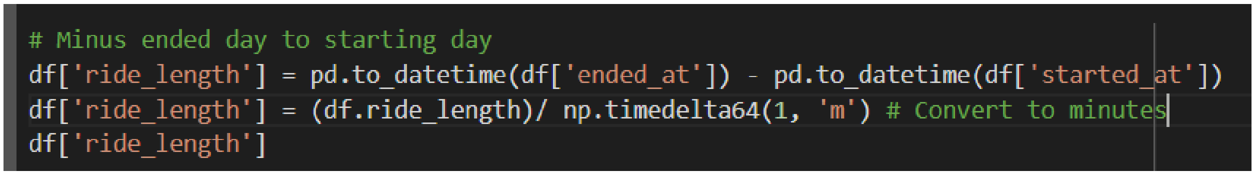


```
df.isnull().sum()

ride_id                    0
rideable_type              0
started_at                 0
ended_at                   0
start_station_name    690809
start_station_id      690806
end_station_name      739170
end_station_id        739170
start_lat                  0
start_lng                  0
end_lat                 4771
end_lng                 4771
member_casual              0
ride_length                0
day_of_week                0
month                      0
```

While the duplicated values do not exsit, there are many values null at start_station_name, start_station_id, end_station_name, end_station_id, emd_lat and end_lng columns. I realize all these columns relate to the location but it does not affect too much to my analysis processing, I remain and have no change for these.

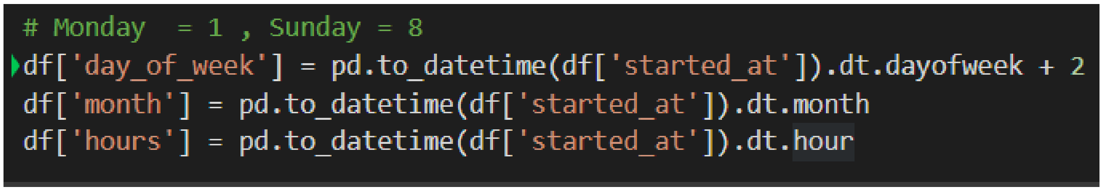
So far, I could not see anything meaningful, so I decide to extract some information from dataset by a making a new column through calculating the ride length which is subtraction between ended_at and started_at columns.



```
# Minus ended day to starting day
df['ride_length'] = pd.to_datetime(df['ended_at']) - pd.to_datetime(df['started_at'])
df['ride_length'] = (df.ride_length)/ np.timedelta64(1, 'm') # Convert to minutes
df['ride_length']
```

However, one noted point is the ride_length must be greater than 0, so, I drop values is negative and what I have 5595063 rows compared 5593819 rows. It means 147 rows is dropped.

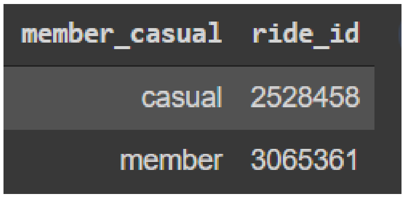
In addition, I continue to extract days of hours, week, and month from started_at columns. I want to Monday is 2 instead of 0 like default and Sunday = 8 ,so I plus 2 to the code below

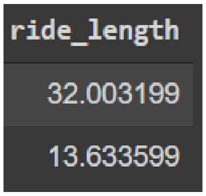


```
# Monday  = 1 , Sunday = 8
df['day_of_week'] = pd.to_datetime(df['started_at']).dt.dayofweek + 2
df['month'] = pd.to_datetime(df['started_at']).dt.month
df['hours'] = pd.to_datetime(df['started_at']).dt.hour
```

Analyze

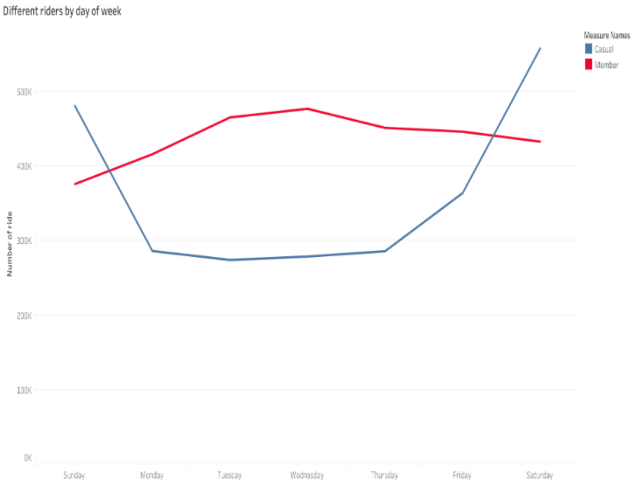
First, to discover the different between member and casual, I want to know how much exactly the number of each user and the average ride time of all members. As we can see, the causal has less than the member 500 thousand users, the average ride length is 2.5 times compared to members. Each ride, casuals ride more than 30 minutes and members ride less than 15 mins.

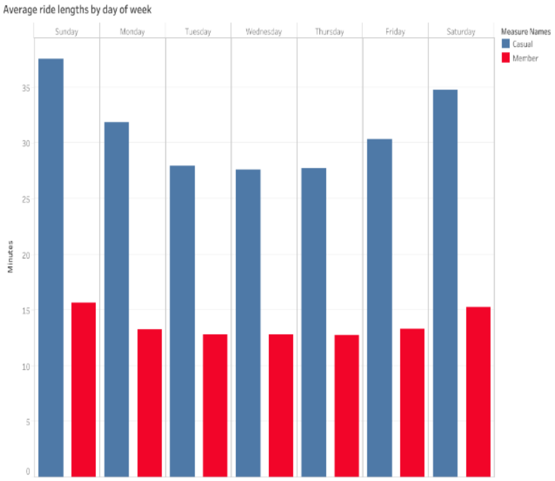




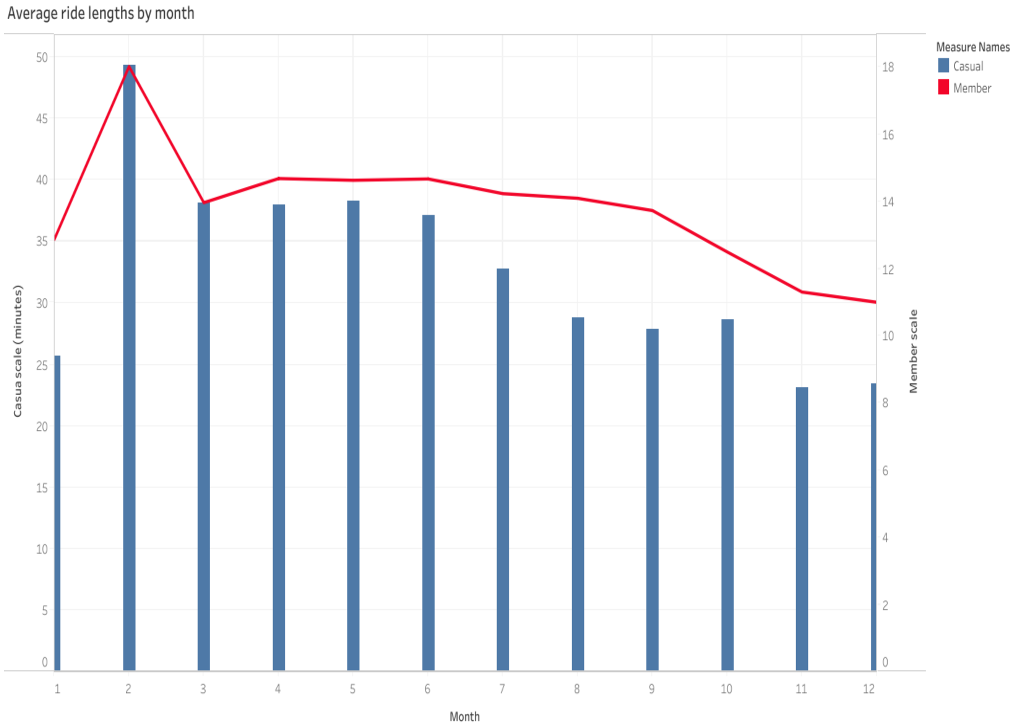
member_casual	ride_id	ride_length
casual	2528458	32.003199
member	3065361	13.633599

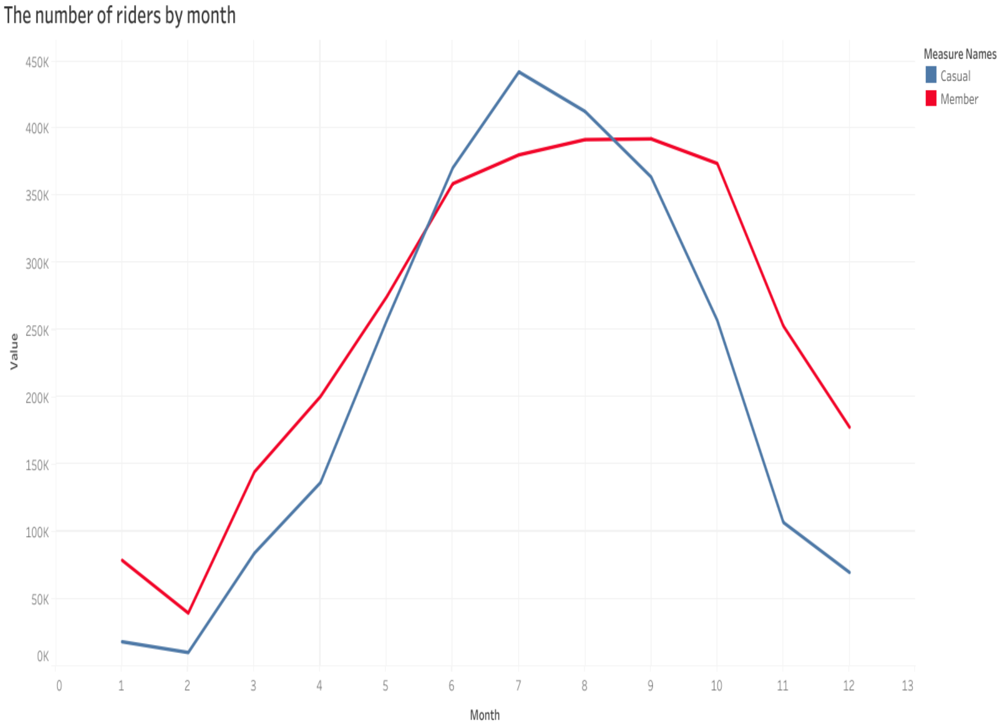
Every week, on Sunday and Saturday, the casual has the number of rides nearly 500,000. That's why the average ride lengths of causal are higher than weekday. Although, on Tuesday and Wednesday, the member has a ride compared to the rest of days, the average ride lengths remain unchanged days of weeks



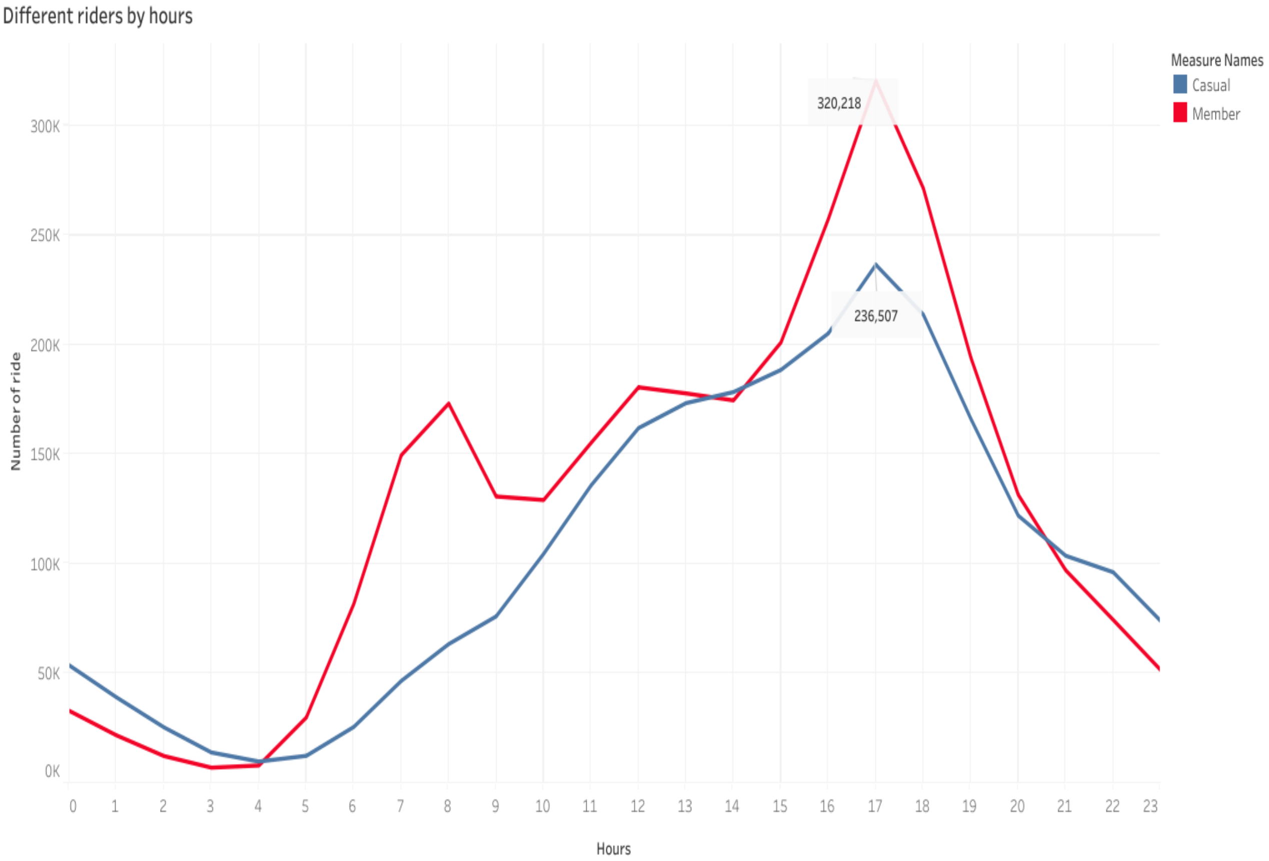


Although the average ride length by months of all members reaches a peak on February, the highest number of ride of casual is June and November is a member





To deeply analyze, I want to know the number of rides per hours of all members. As we can see the picture below, from 16 to 18, all members using bicycle more than the rest hours. It might be explained that this time is people getting off work and return home.



Conclusion

After summary the crucial information, and visualize the data through a chart, I realize there are some points different between casual and member:

- ✓ The average ride lengths of casual is 2.5 times compared to member
- ✓ Casual often has a higher number of rides on Saturday and Sunday
- ✓ Casual has a peak number of rides in July while member has a peak on November

Here, some suggestion:

- ✓ During peak periods, provide members priority access.
- ✓ Offering free ride minutes for every minute beyond 30 minutes of usage
- ✓ Use billboards/posters around the top 20 most popular stations for casual users to promote yearly membership fees