# Project Report

## ASK Phase:

A clear statement of the business task: My task is to identify the riding behaviour of annual member and casual member. I have to find out difference of riding behaviour between these two groups. My findings will help marketing team to understand the future goal. Marketing team will understand should they encourage casual member to convert to annual member or not?

## Data Prepare:

I will use Cyclistic's historical trip data to analyze and identify trends. Previous 12 months of Cyclistic trip data of 2021 can be found - https://divvy-tripdata.s3.amazonaws.com/index.html. This data is available by Motivate International Inc. This is public data that I can use to explore how different customer types are using Cyclistic bikes. There are a lot of data in above location but I will work with only year of 2021 data. After opening data by Microsoft excel, I can see that data need to be cleaned.

## Data Process:

1. After downloading I have unzipped the files.
2. I have created a folder on my desktop to house the files. My location is : ***C:\Coursera\Capstone\2021***.
3. I have named the data of different months in different names. For example: Trip_Data_Jan.csv. Only month name has been changed , other format is same.
4. I have chosed Microsoft excel for data process. Some months data are so big for processing in google sheets. Data cleaning and processing steps by Microsoft Excel are below:
    a. First, I have centered, aligned middle and wraped text of each data.
    b. Then I have clicked data tab and clicked Remove Duplicates. I have checked is there any null cells or not. I have found maximum null cells at station name and id columns. My main concentration is starting date and ride duration.
    c. I have inserted five new columns:
        - STARTED_AT_DATE [= INT(STARTED_AT)>FORMAT CELLS>CUSTOM>DD-MMMM-YY],
        - STARTED_AT_WEEKDAY [=TEXT (STARTED_AT_DATE,"DDDD")],
        - STARTED_AT_YEAR[=YEAR(STARTED_AT_DATE)],

- STARTED_AT_MONTH[=MONTH(STARTED_AT_DATE)] AND
- RIDE_DURATION [= STARTED_AT-ENDED_AT>FORMAT CELLS>CUSTOM>HH:MM:SS].

5. Value of new created columns are not integer type. To analyse the data, we need the data in integer type. So beside each new created column I have created a new extra column and implemented [=value(column value)] formula to get the integer type data. After converting each data in integer type, I have copied the whole column and paste only the values to original column. So no formula cell remained in sheets.

## Data Analysis in Python:

### Importing data from CSV into Python:

```
# import each month data
jan = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_Jan.csv")
Feb = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_Feb.csv")
Mar = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_Mar.csv")
April = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_April.csv")
May = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_May.csv")
June = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_June.csv")
July = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_July.csv")
August = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_August.csv")
September = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_September.csv")
October = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_October.csv")
November = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_November.csv")
December = pd.read_csv("C:\\Coursera\\Capstone\\2021\\Trip_Data_December.csv")
```

### Merging all the data frame into one data frame:

```
# Aggregating all months data
All_Months_Data = [jan, Feb, Mar, April, May, June, July, August, September, October, November, December]
```

### Converting ride duration column into time format:

```
# declaring ride duration as a time. Column shall be in hh:mm:ss format
all_data_df["ride_duration"] = pd.to_timedelta(all_data_df["ride_duration"])
```

# Project Report

## Finding total ride duration in weekdays of member and casual riders:

```
# segregating summation of ride duration of different starting weekdays' data depending on member and casual
member_rides_duration = all_data_df[all_data_df["member_casual"] == "member"].groupby('started_at_weekday')[
    'ride_duration'].sum()
casual_rides_duration = all_data_df[all_data_df["member_casual"] == "casual"].groupby("started_at_weekday")[
    'ride_duration'].sum()
```

## Plotting graph of total ride duration data:

```
# plotting figure for comparison depending on ride duration(started_at_weekday)
plt.plot(member_rides_duration, color ='g', label ='member')
plt.plot(casual_rides_duration, color ='r', label ='casual')
plt.title('weekday vs rides duration')
plt.xlabel('started_at_weekday')
plt.ylabel('ride duration')
plt.legend()
plt.show()
```

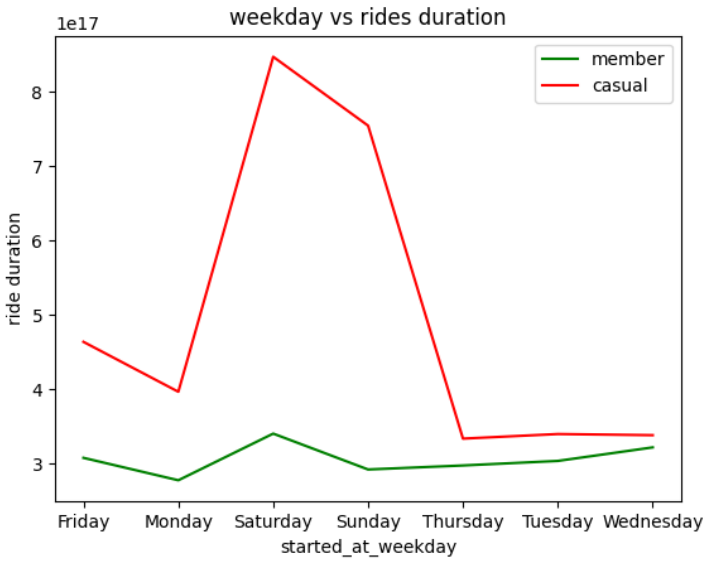


*Figure 1: weekday vs rides duration*

## Observation:

We see that total ride duration for casual users are more than member users. For both casual and member users most bike used day is Saturday. But the peak for casual users is more than members.

# Project Report

## Counting total number of rides in weekdays of member and casual riders:

```
# counting total ride id of different starting weekdays' data of member and casual
member_rides_id = all_data_df[all_data_df["member_casual"] == "member"].groupby('started_at_weekday')[
    'ride_id'].count()
casual_rides_id = all_data_df[all_data_df["member_casual"] == "casual"].groupby("started_at_weekday")[
    'ride_id'].count()

# plotting figure for comparison depending on ride duration(started_at_weekday)
plt.plot(member_rides_id, color='g', label='member')
plt.plot(casual_rides_id, color='r', label='casual')
plt.title('weekday vs total ride ID')
plt.xlabel('started_at_weekday')
plt.ylabel('total ride ID')
plt.legend()
plt.show()
```

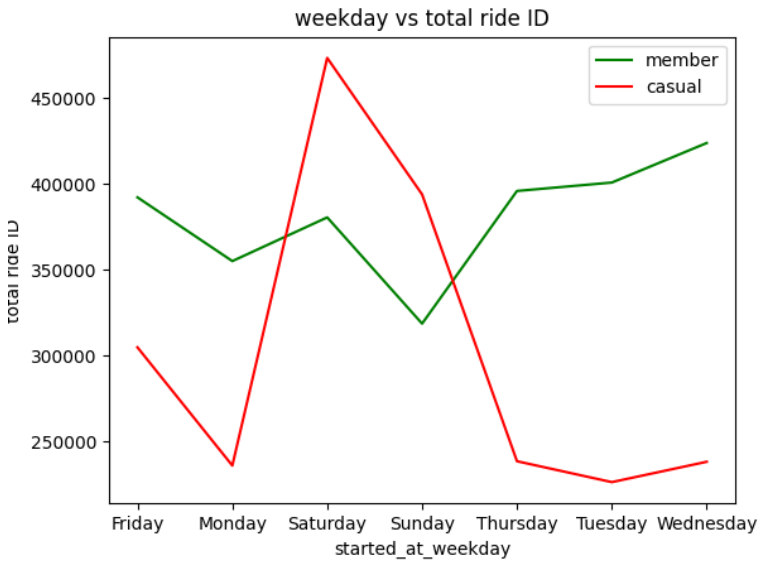


*Figure 2: weekday vs total ride id*

## Observation:

From above figure, we see that without Saturday and Sunday, member users are using bikes more frequently than member users. Again, bike using behaviour of member users is more irregular than casual users. It is changing more frequently day by day.

# Project Report

## Finding mean ride duration of member and casual riders in different weekdays:

```
# finding mean of ride duration of different starting weekdays' data of member and casual
member_rides_duration_mean = all_data_df[all_data_df["member_casual"] == "member"].groupby('started_at_weekday')[
    'ride_duration'].mean()
casual_rides_duration_mean = all_data_df[all_data_df["member_casual"] == "casual"].groupby("started_at_weekday")[
    'ride_duration'].mean()
plt.plot(member_rides_duration_mean, color='g', label='member')
plt.plot(casual_rides_duration_mean, color='r', label='casual')
plt.title('weekday vs mean rides duration')
plt.xlabel('started_at_weekday')
plt.ylabel('mean ride duration')
plt.legend()
plt.show()
```

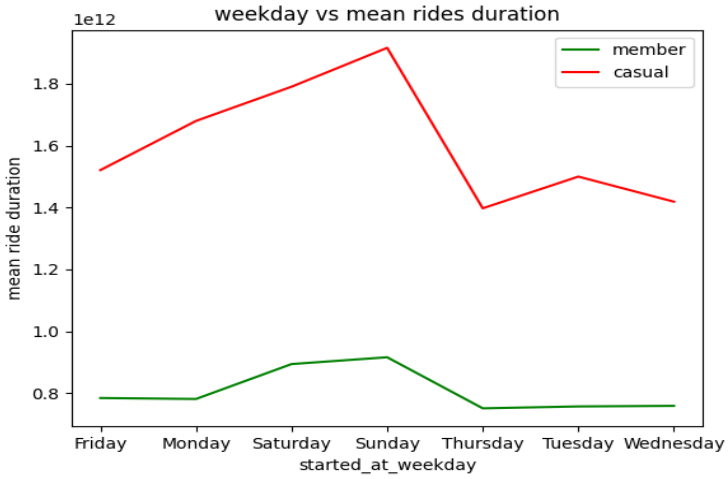


*Figure 3: Mean ride duration*

### Observation:

We see that mean ride duration is more for casual riders than member users. Thursday to Wednesday, mean ride duration for member users is same but for casual users, it keeps ups and downs.

## Finding max ride duration of member and casual riders in weekdays:

```
# finding max of ride duration of different starting weekdays' data of member and casual
member_rides_duration_max = all_data_df[all_data_df["member_casual"] == "member"].groupby('started_at_weekday')[
    'ride_duration'].max()
casual_rides_duration_max = all_data_df[all_data_df["member_casual"] == "casual"].groupby("started_at_weekday")[
    'ride_duration'].max()

# plotting figure
plt.plot(member_rides_duration_max, color='g', label='member')
plt.plot(casual_rides_duration_max, color='r', label='casual')
plt.title('weekday vs max rides duration')
plt.xlabel('started_at_weekday')
plt.ylabel('max ride duration')
plt.legend()
plt.show()
```

# Project Report

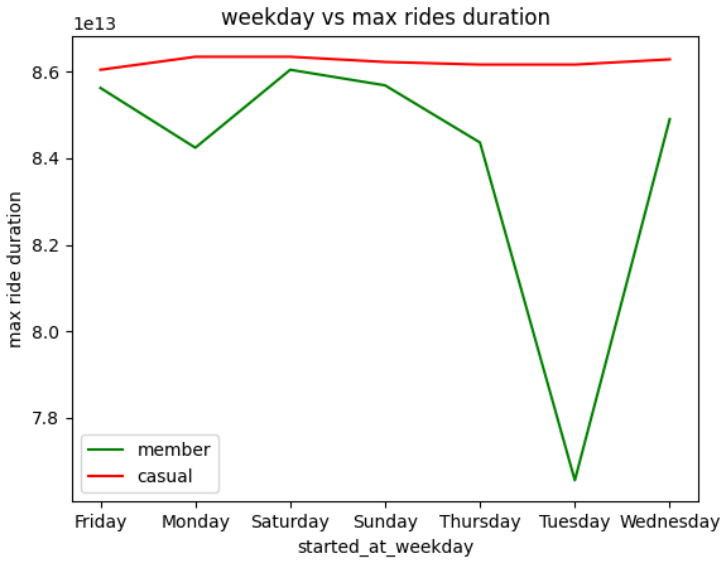


*Figure 4: Weekday vs max rides duration*

## Observation:

Maximum rides duration is same for all weekdays for casual riders, but it is very inconsistent for member users. Most max ride duration occurs for member riders on Saturday and minimum occurs on Tuesday.

## Finding modes of rides for member and casual rider:

```
# finding modes of ride of different starting weekdays' of member and casual
member_rides_duration_mode = all_data_df[all_data_df["member_casual"] == "member"]["started_at_weekday"].mode()
casual_rides_duration_mode = all_data_df[all_data_df["member_casual"] == "casual"]["started_at_weekday"].mode()

print(member_rides_duration_mode)
print(casual_rides_duration_mode)
```

## Result:

```
0    Wednesday
Name: started_at_weekday, dtype: object
0    Saturday
Name: started_at_weekday, dtype: object

Process finished with exit code 0
```

## Observation:

Member riders use bikes mostly on Wednesday and Casual riders on Saturday.

# Project Report

## Finding total number of users:

```
# finding total number of users
member_total_numbers = all_data_df[all_data_df["member_casual"] == "member"]['member_casual'].count()
casual_total_numbers = all_data_df[all_data_df["member_casual"] == "casual"]['member_casual'].count()

plt.bar('member', member_rides_duration_mode)
plt.bar('casual', casual_rides_duration_mode)
plt.title('No. of Total riders')
plt.show()
```

## Figure 5:

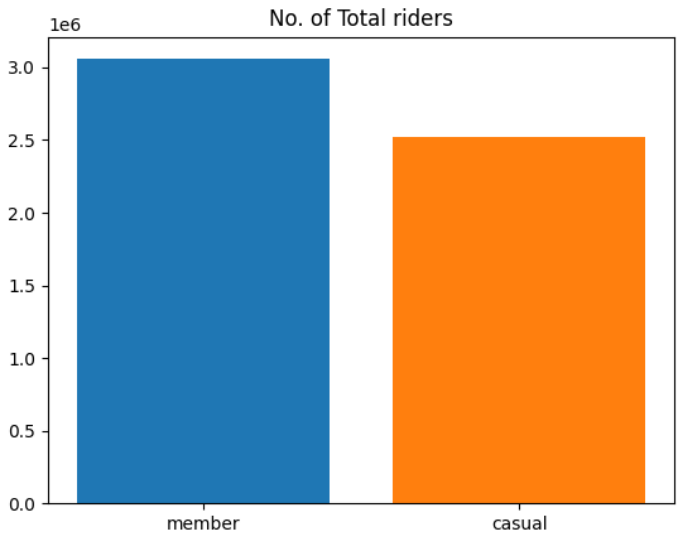


*Figure 5: No. of Riders*

## Observation:

Most bike users are member riders.

## Users' percentage:

```
#drawing pie chart
member_total_numbers = all_data_df[all_data_df["member_casual"] == "member"]['member_casual'].count()
casual_total_numbers = all_data_df[all_data_df["member_casual"] == "casual"]['member_casual'].count()
total_users_number=[member_total_numbers, casual_total_numbers]
member_type=['member', 'casual']


plt.pie(total_users_number, labels=member_type, autopct='%1.1f%%')
plt.title('total users')
plt.show()
```

# Project Report

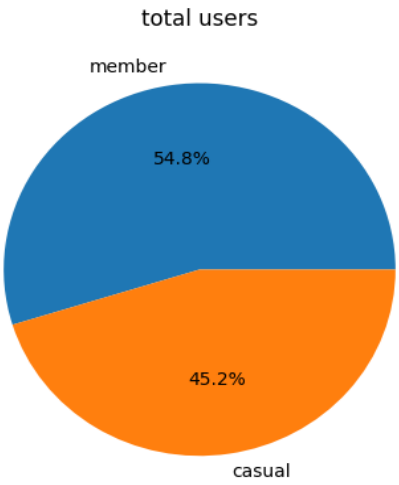


*Figure 6: % of total users.*

## Observation:

More than 50 percent users are Member users.

## Monthly bike riding users:

```
#Monthly data
plt.figure(figsize=(8,6))
sns.countplot(x="member_casual",hue="started_at_month", data=all_data_df)
plt.show()
```

# Project Report

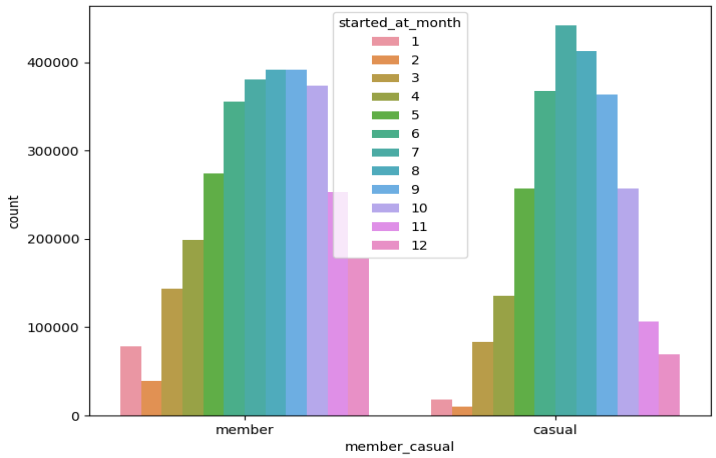


*Figure 7: Monthly bike users*

## Observation:

Bike usage of August and September is same for members. But  for casual riders, ride occurs more frequently on August.

## Rideable Bike:

```
#rideable_bike
plt.figure(figsize=(8,6))
sns.countplot(x="member_casual", hue="rideable_type", data=all_data_df)
plt.show()
```

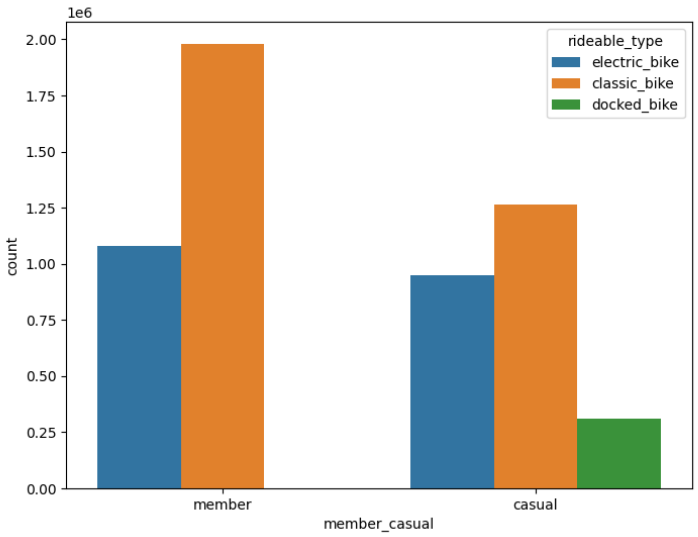


*Figure 8: Rideable Bike*

## Observation:

Most used bike for members are classic bike but they do not used docked bike at all.

# Project Report

## Recommendations:

1. We see that total ride duration for casual users are more than member users. For both casual and member users most bike used day is Saturday. But the peak for casual users is more than members. Member users can be encouraged by giving incentives after any number of rides.
2. We observe that without Saturday and Sunday, member users are using bikes more frequently than member users. Again, bike using behaviour of member users is more irregular than casual users. It is changing more frequently day by day. So in weekends, discounts can be provided to member riders.
3. We observe that mean ride duration is more for casual riders than member users. Thursday to Wednesday, mean ride duration for member users is same but for casual users, it keeps ups and downs. Bonus points can be provided to member users for encouraging them.
4. Maximum rides duration is same for all weekdays for casual riders, but it is very inconsistent for member users. Most max ride duration occurs for member riders on Saturday and minimum occurs on Tuesday. For Tuesday, discounts and reimbursement can be provided for member users.
5. Bike usage of August and September is same for members. But for casual riders, ride occurs more frequently in August. In the month of August, discounts and reimbursement can be provided for member users.
6. Most used bike for members is classic bike but they do not used docked bike at all. New initiatives can be taken to increase docked bike usage for member riders. Then casual riders will also be encouraged to get membership.