Scenario

Cyclistic is a bike-share company offering the bike rental services to customers to address their mobility needs. Since 2016, Cyclistic has made 5,824 bicycles available for customers to travel within a network of 692 stations across Chicago. Until now, Cyclistic's marketing team has been focusing on attracting more customers with the flexibility of its pricing plans: single-ride passes, full-day passes and annual memberships. Customers who purchase single-ride passes and full-day passes are referred to as casual riders, while those who subscribe to the annual memberships are referred to as annual members. According to Cyclistic's finance analyst, annual members are found more profitable than the casual riders. This finding has been a very good inspiration for the director of marketing, Lily Moreno to drive more marketing effort towards maximizing the number of annual members for company's better financial achievement in the future, and she has also seen the opportunity to convert the casual riders into annual members in order to accomplish this goal.

In this case study, we will go through a comprehensive data analysis process to help Cyclistic in achieving the goal. This data analysis process will be conducted in a step-by-step manner in six stages as below:

1) **Ask** : A clear statement of the business task
2) **Prepare** : A description of all data sources used
3) **Process** : Documentation of any cleaning or manipulation of data
4) **Analyze** : A summary of the analysis
5) **Share** : Supporting visualization and key findings
6) **Act** : Recommendations based on the analysis

Ask

A clear statement of the business task

As mentioned above, the director of marketing, Lily Moreno aims to convert the casual riders into annual members as she believes that is the key for company's future success. The marketing analytics team is responsible to collect, organize and analyze the historical trip data to discover useful insights and make recommendations for the executive team to launch effective marketing program to meet the objective. However, the marketing team has to overcome some challenges before they can successfully launch the new marketing program to fulfil the task, and these include understanding the differences of annual members and casual riders in using the bikes, considering why the casual riders would buy a membership and how to leverage digital media to influence the casual riders to become members. In this case study, we will focus on exploring the differences of behavior between annual members and casual riders.

Prepare

A description of all data sources used

Link to download : https://divvy-tripdata.s3.amazonaws.com/index.html

Source: Coursera Google Data Analytics Course

License : Motivate International Inc. https://ride.divvybikes.com/data-license-agreement

The historical trip data can be downloaded from the link above and we note that customer's personally identifiable information is not accessible due to privacy and security issues. This has put some limitations to our analysis such as we are not able to determine if casual riders live in Cyclistic service area or if they have purchased multiple single passes. In spite of that, we are still able to proceed because the customer type has been recorded for each trip, which is the most important attribute for us to study the differences between casual riders and annual members. Here we have adopted historical data from year 2021 for this analysis which comes in 12 separate CSV files for each month and each file consists of the variables as below:

- ride_id : The ID that represent each trip
- rideable_type : The bike type – 'classic_bike', 'electric_bike' or 'docked_bike'
- started_at : The date and time when the trip started
- ended_at : The date and time when the trip ended
- start_station_name : The name of the start station
- start_station_id : The ID of the start station
- end_station_name : The name of the end station
- end_station_id : The ID of the end station
- start_lat : The latitude of the start station
- start_lng : The longitude of the start station
- end_lat : The latitude of the end station
- end_lng : The longitude of the end station
- member_casual : The customer type - 'member' or 'casual'

Process

Documentation of any cleaning or manipulation of data

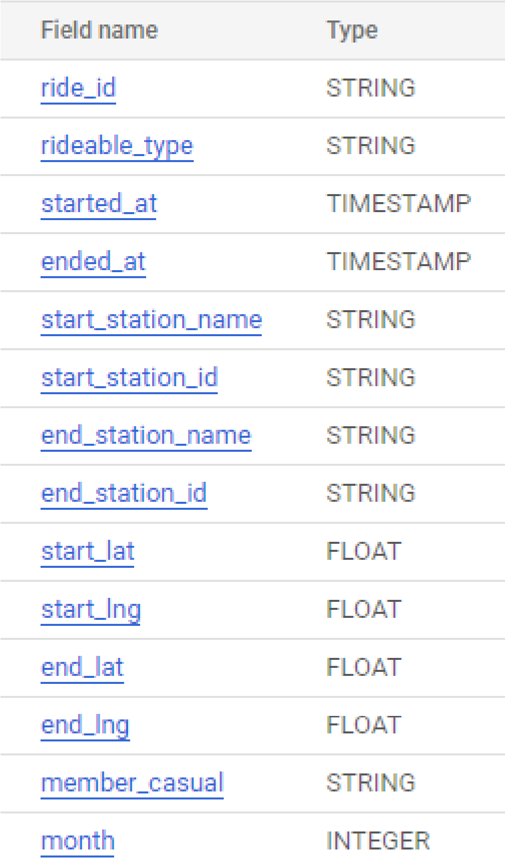
In this part, we will be cleaning the data to get rid of any errors that may potentially affect the results of the analysis by using Bigquery, which is a very useful SQL programming tool that can handle large amount of data. The detailed steps of the data cleaning process are stated below :

1) Upload all CSV files into the Bigquery database
   Dataset ID : Cyclistic
   Table Name : 202101_tripdata ~ 202112_tripdata (Total 12 CSV files for each month)

2) Merge all CSV files into one single table for easy processing of all data in one place. Here we have created the 'month' column to label each file so that we can identify the source of the data from the merged table.
   Merged Table Name : 2021_tripdata

```
SELECT *, 1 AS month
FROM Cyclistic.202101_tripdata
UNION ALL
SELECT *, 2 AS month
FROM Cyclistic.202102_tripdata
UNION ALL
SELECT *, 3 AS month
FROM Cyclistic.202103_tripdata
UNION ALL
SELECT *, 4 AS month
FROM Cyclistic.202104_tripdata
UNION ALL
SELECT *, 5 AS month
FROM Cyclistic.202105_tripdata
UNION ALL
SELECT *, 6 AS month
FROM Cyclistic.202106_tripdata
UNION ALL
SELECT *, 7 AS month
FROM Cyclistic.202107_tripdata
UNION ALL
SELECT *, 8 AS month
FROM Cyclistic.202108_tripdata
UNION ALL
SELECT *, 9 AS month
FROM Cyclistic.202109_tripdata
UNION ALL
SELECT *, 10 AS month
FROM Cyclistic.202110_tripdata
UNION ALL
SELECT *, 11 AS month
FROM Cyclistic.202111_tripdata
UNION ALL
SELECT *, 12 AS month
FROM Cyclistic.202112_tripdata
```

3) Check data format. When we upload the CSV file into the database, the system has automatically detected the data type. The below schema table shows that each column has been turned into the correct format.



Field name	Type
ride_id	STRING
rideable_type	STRING
started_at	TIMESTAMP
ended_at	TIMESTAMP
start_station_name	STRING
start_station_id	STRING
end_station_name	STRING
end_station_id	STRING
start_lat	FLOAT
start_lng	FLOAT
end_lat	FLOAT
end_lng	FLOAT
member_casual	STRING
month	INTEGER

4) Check data input

Double check if all input for column 'started_at' are belong to year 2021 to avoid wrong data input.

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE EXTRACT(YEAR FROM started_at) <> 2021
```

Result : 0

Double check if all input for column 'started_at' are corresponded to their respective month to avoid wrong data input.

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE EXTRACT(MONTH FROM started_at) <> month
```

Result : 0

5) Check duplicate

Each row should represent a single trip with a unique ID. The below codes verified that there is no repeated ride_id in this dataset.

```
SELECT ride_id, COUNT(ride_id)
FROM Cyclistic.2021_tripdata
GROUP BY ride_id
HAVING COUNT(ride_id) > 1
```

Result : There is no data to display

6) Check NULL value

We have found quite some null values for the station columns which should not be blank in this dataset, but we will just ignore them since we are not looking into the geographical factor in this case study.

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE ride_id IS NULL ;
```

Result : 0

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE rideable_type IS NULL ;
```

Result : 0

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE
started_at IS NULL OR
ended_at IS NULL ;
```

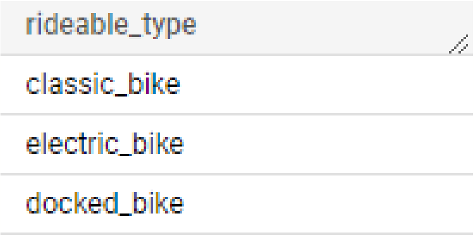
Result : 0

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE member_casual IS NULL ;
```

Result : 0

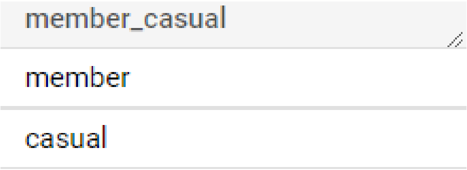
7) Check spelling

```
SELECT DISTINCT(rideable_type)
FROM Cyclistic.2021_tripdata
```



rideable_type
classic_bike
electric_bike
docked_bike

```
SELECT DISTINCT(member_casual)
FROM Cyclistic.2021_tripdata
```



member_casual
member
casual

8) Data validation

```
SELECT COUNT(*)
FROM Cyclistic.2021_tripdata
WHERE TIMESTAMP_DIFF(ended_at, started_at, SECOND) < 0
```

```
Result : 147
```

We found that there are 147 number of rows having the ended time earlier than the start time which does not make sense, but we will just exclude these rows from this analysis.

9) Additional columns

Based on the dataset given, we have generated two additional attributes to help in this analysis process :

a) day_of_week

```
SELECT EXTRACT(DAYOFWEEK FROM started_at) AS day_of_week
FROM Cyclistic.2021_tripdata
```

b) ride_length

The ride_length is calculated by subtracting the column ‘started_at’ from the column ‘ended_at’. We have turned the ride length into seconds because it is much easier to be used for calculation.

```
SELECT TIMESTAMP_DIFF(ended_at, started_at, SECOND) AS ride_length
FROM Cyclistic.2021_tripdata
```

Finally, we have wrapped up the codes above to get a fully cleaned dataset.

Table Name : cleaned_2021_tripdata

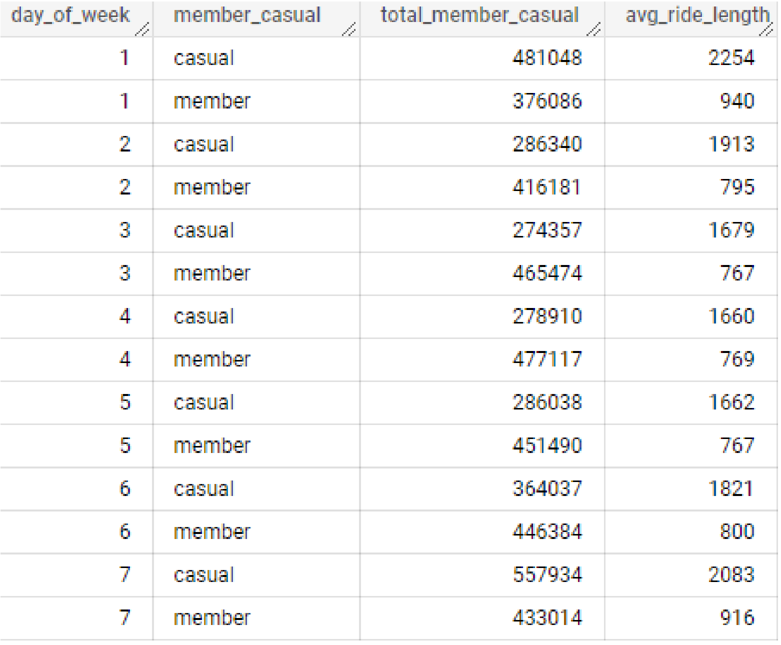
```
SELECT *,
EXTRACT(DAYOFWEEK FROM started_at) AS day_of_week,
TIMESTAMP_DIFF(ended_at, started_at, SECOND) AS ride_length
FROM Cyclistic.2021_tripdata
WHERE TIMESTAMP_DIFF(ended_at, started_at, SECOND) > 0
ORDER BY started_at
```

Analyze

A summary of the analysis

We have calculated the total count of member and casual, and average ride length for each day of week by writing the code below.

```
SELECT day_of_week, member_casual,
COUNT(*) AS total_member_casual,
CAST(AVG(ride_length) AS INT) AS avg_ride_length
FROM Cyclistic.cleaned_2021_tripdata
GROUP BY day_of_week, member_casual
ORDER BY day_of_week, member_casual
```



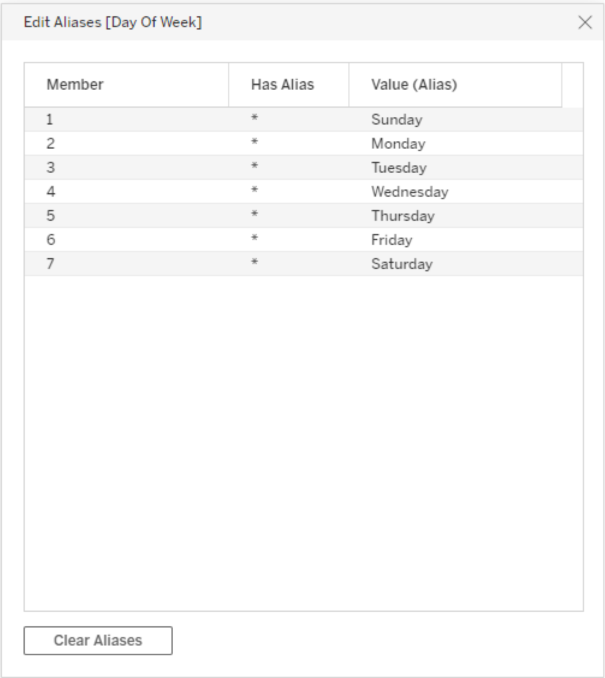
day_of_week	member_casual	total_member_casual	avg_ride_length
1	casual	481048	2254
1	member	376086	940
2	casual	286340	1913
2	member	416181	795
3	casual	274357	1679
3	member	465474	767
4	casual	278910	1660
4	member	477117	769
5	casual	286038	1662
5	member	451490	767
6	casual	364037	1821
6	member	446384	800
7	casual	557934	2083
7	member	433014	916

From the summary table above, we can find that on weekends – Saturday (7) and Sunday (1), the number of casual riders is greater than the member riders. While on the weekdays – Monday (2) to Friday (6), the number of member riders is greater than the casual riders. Therefore, we can assume that most of the casual riders are using the bike as their leisure activities during the weekends, while the member riders are more likely using the bike as their commute to work during the weekdays. The table also shows that casual riders are spending more time on average in riding than the member riders throughout the weeks. This probably means that casual riders enjoy riding for a longer duration compared to member riders.

Share

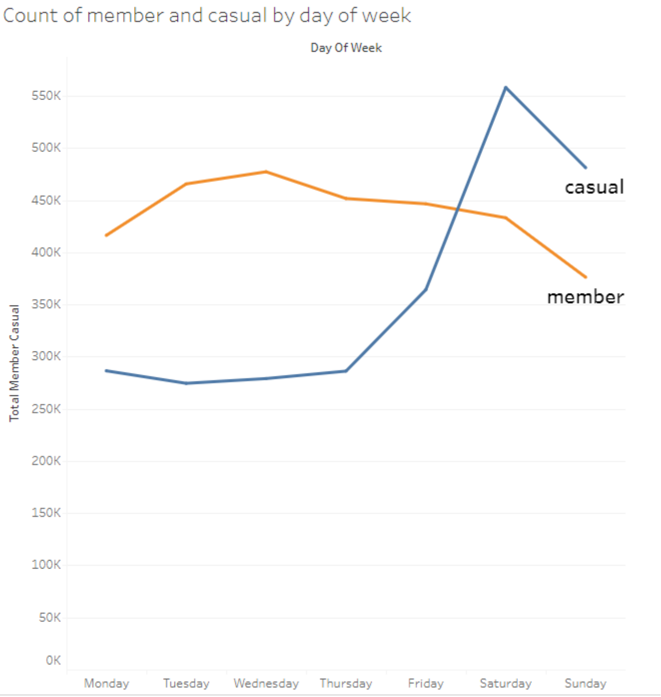
Supporting visualizations and key findings

With the summary table above, we can use Tableau to create visualization to showcase our findings. Before that, we have given aliases to the numbers in the column ‘day_of_week’ by converting them into the text format for better readability.



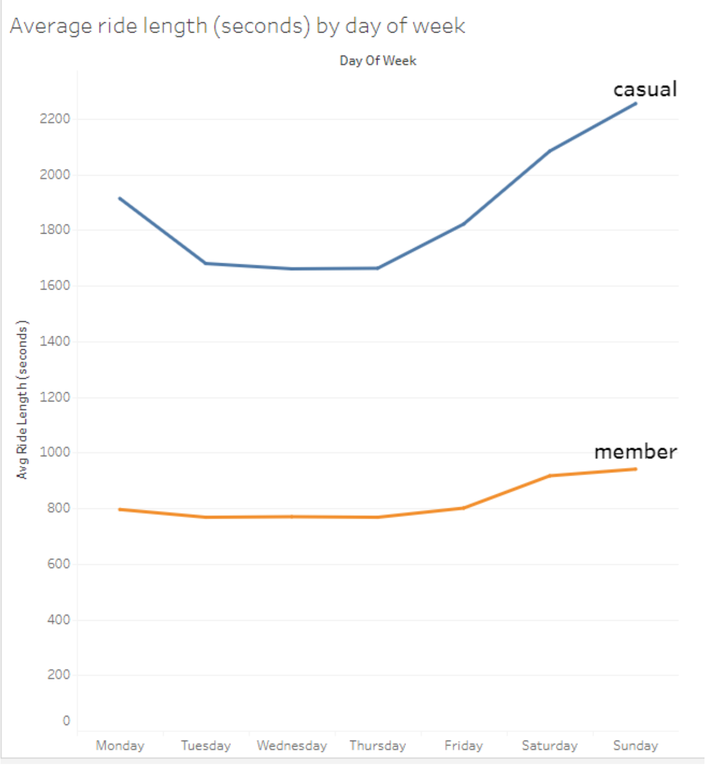
Member	Has Alias	Value (Alias)
1	*	Sunday
2	*	Monday
3	*	Tuesday
4	*	Wednesday
5	*	Thursday
6	*	Friday
7	*	Saturday

a) Count of member and casual by day of week



As per the line chart above, the casual riders (represented by blue line) are rather inactive during the weekdays, however the number of casual riders opposingly surpassed the member riders (represented by orange line) during the weekend.

b) Average ride length (seconds) by day of week



The line chart above showcases the average ride length in seconds for both member and casual riders. It can be observed that the average ride durations for casual riders are about twice as long as those of the member, with the discrepancy being even greater during the weekend. This highlights the fact that casual riders are more likely to use the bikes for leisure purposes, whereas member riders tend to use the bikes for their daily commute.

Act

Recommendations based on the analysis

In conclusion, the analysis of the trip data reveals that casual riders tend to use the bikes as their leisure activities and they are more active during the weekend compared to member riders. There are a few suggestions for the Cyclistic's marketing team to consider:

1) Offering weekend promotions to their members in order to attract the casual riders to join the membership.
2) Hosting a competition for members to reward them based on their ride length. This initiative not only encourages the existing members to travel more with bikes, but also potentially entices some casual riders who always spend a great amount of time in riding to become a member to stand a chance for winning exciting prizes.
3) Hiring some skilled cyclists to become mentors towards their members in the purpose of improving their riding skills and promoting their health levels. Casual riders will likely be drawn to the opportunity to learn from the expert to advance their skills and improve their overall fitness, making the membership more appealing to them.