‘Cyclistic’ Project_notes

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company’s future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Additional details

- Bike company ‘Cyclistic’
- 5,824 fleet
- single-ride passes, full-day passes: **Casual members**
- annual memberships: **Cyclistic members**
- Profitability: **Cyclistic members> Casual members**

Business tasks

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

Business goal

- Maximizing Cyclistic members
- Converting **Casual members** into **Cyclistic members**

Requirements

- Understand how *annual members* and *casual riders* diff2er
- Why casual riders would buy a membership
- How digital media could affect their marketing tactics

Tool used:

- BigQuery
- Looker

Data Integrity, Consistency

- **Checking for rows with NULL values:**

```
SELECT *
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
WHERE
  ride_id IS NULL
  OR rideable_type IS NULL
  OR started_at IS NULL
  OR ended_at IS NULL
  OR start_station_name IS NULL
  OR start_station_id IS NULL
  OR end_station_name IS NULL
  OR end_station_id IS NULL
  OR start_lat IS NULL
  OR start_lng IS NULL
  OR end_lat IS NULL
  OR end_lng IS NULL
  OR member_casual IS NULL
```

***RESULT: 1x stolen/abandoned bike shown on results = missing geographic end details*

- ***ride_id* integrity:**

```
SELECT counter
FROM
  (
    SELECT LENGTH(ride_id) AS counter
    FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
  ) AS summary – inner query
WHERE summary.counter!=16
```

***RESULT: all ride_id have 16 characters*

- **Year consistency:**

```
SELECT
  DISTINCT EXTRACT(YEAR FROM started_at)
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
```

*** repeated also for 'ended_at'*

- ***start_station_name* integrity:**

```
SELECT DISTINCT start_station_name
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
```

- **Consistency between *start_station_name* and *start_station_id*:**

```
SELECT
  t1.start_station_id
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020` AS t1
LEFT JOIN
  (SELECT
    DISTINCT start_station_name,
    start_station_id
  FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`) AS t2
  ON t1.start_station_id=t2.start_station_id
  WHERE t2.start_station_id IS NULL -- check only for those values that are not in common
```

***repeated also for end_station_id*

- ***member_casual* integrity:**

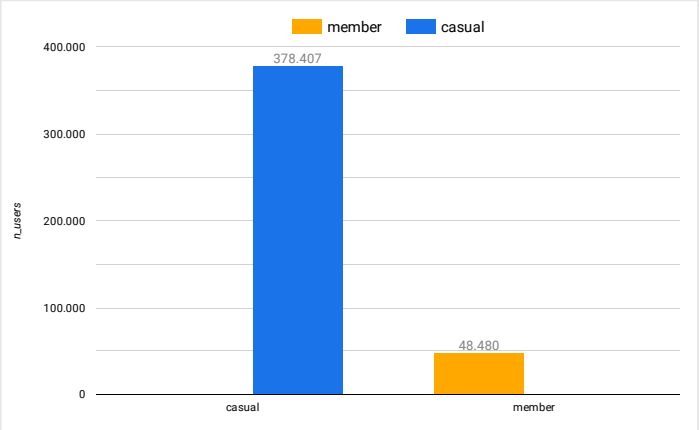
```
SELECT
  DISTINCT member_casual
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
```

Once verified the integrity and consistency of the data present in the dataset, it is now possible to continue with processing and analyzing it, in order to find answers to our business task.

Data Processing, Analysis (BigQuery, Looker)

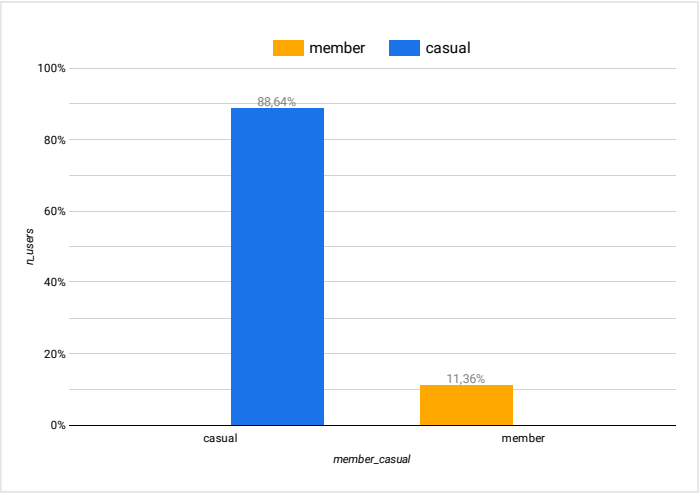
- **Number of users per member type:**

```
SELECT
  member_casual,
  COUNT (member_casual) AS n_users
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
GROUP BY member_casual
```



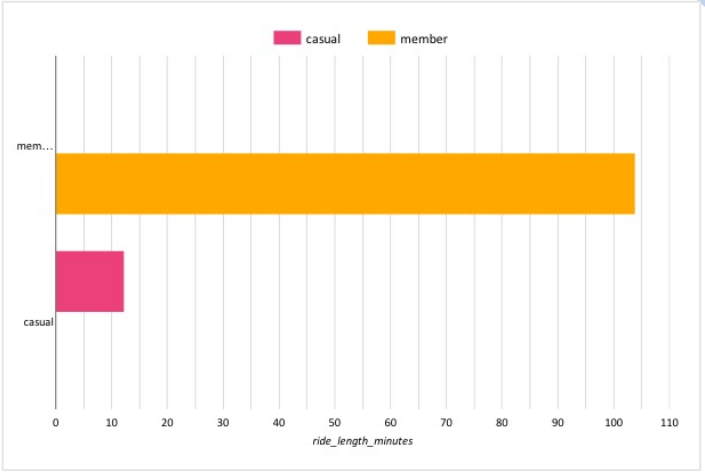
- **Percentage of users per member type:**

```
WITH counter AS
  (
  SELECT
    member_casual,
    COUNT (member_casual) AS n_users
  FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
  GROUP BY member_casual
)
SELECT
  member_casual,
  ROUND((counter.n_users/SUM(counter.n_users) OVER () * 100),2) AS perc
FROM counter
```



- **Average ride length (minutes) per member type:**

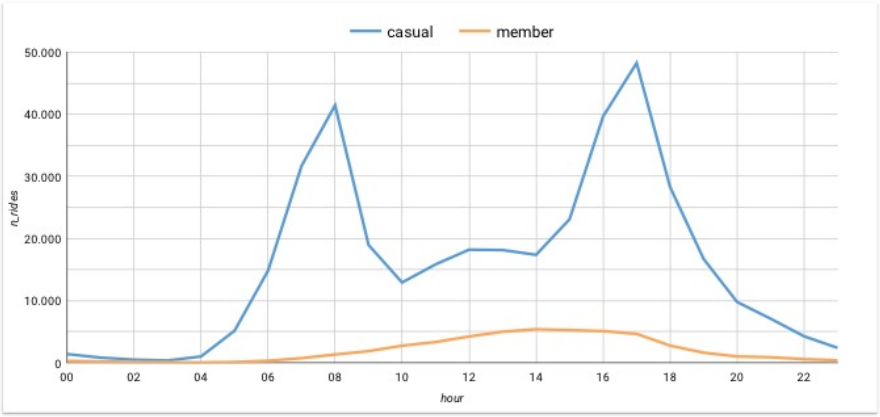
```
SELECT
  member_casual,
  ROUND(AVG(TIMESTAMP_DIFF(ended_at,started_at,MINUTE)),2) AS
ride_length_minutes -- avg time rounded up to 2nd decimal position
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
WHERE (TIMESTAMP_DIFF(ended_at,started_at,MINUTE))>0 -- excluding possible
negative results
GROUP BY member_casual
ORDER BY ride_length_minutes DESC
```



- **Popular ride times**

  Rides frequency per hour of the day, grouped by member type:

```
SELECT
  EXTRACT (HOUR FROM started_at) AS hour,
  COUNT (ride_id) AS casual_riders
FROM `portfolioproject-401814.Cyclistic.trips_q1_2020`
WHERE member_casual='casual'
GROUP BY hour
ORDER BY hour
```



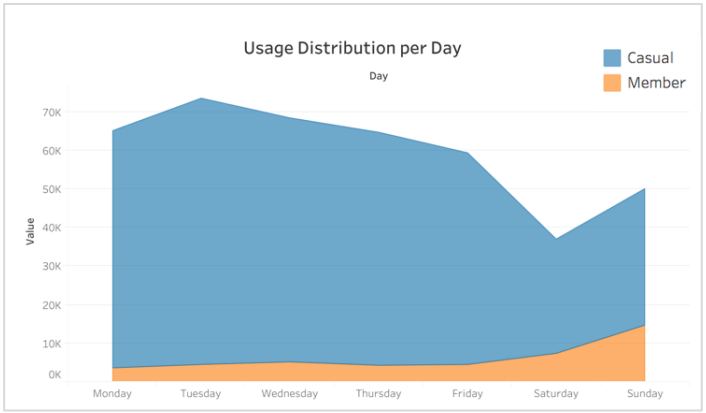
- **Popular ride days**

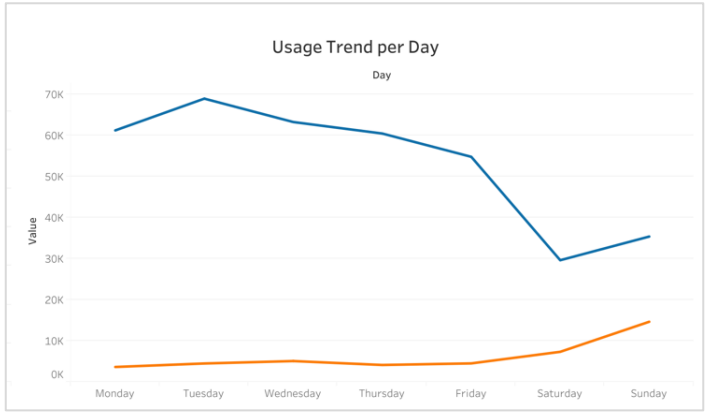
  Rides frequency per day of the week, grouped by member type:

```
SELECT
  FORMAT_DATE('%A',started_at) AS day,
  COUNTIF(member_casual='member') AS member,
  COUNTIF(member_casual='casual') AS casual
FROM
  (
    -- ride made on the single day AND ride length>0
    SELECT
      *
    FROM`portfolioproject-401814.Cyclistic.trips_q1_2020`
    WHERE DATE(started_at)=DATE(ended_at)
    AND
    (TIMESTAMP_DIFF(ended_at,started_at,MINUTE))>0
  )
GROUP BY day
ORDER BY
  CASE
    WHEN day='Monday' THEN 1
    WHEN day='Tuesday' THEN 2
    WHEN day='Wednesday' THEN 3
    WHEN day='Thursday' THEN 4
    WHEN day='Friday' THEN 5
    WHEN day='Saturday' THEN 6
    WHEN day='Sunday' THEN 7
  END
```

***the query above excludes the 1152 rows referred to rides lasted more than one day. Verified through:*

```
SELECT
  COUNT(started_at)
  FROM`portfolioproject-401814.Cyclistic.trips_q1_2020`
WHERE DATE(started_at)<>DATE(ended_at)
```





- **Top 5 used stations per member type:**

```
WITH StationCounter AS
(
  SELECT
      start_station_name,
      member_casual,
      COUNT(*) AS station_count,
      RANK() OVER (PARTITION BY member_casual
ORDER BY COUNT(*)DESC) AS count_rank
    FROM `portfolioproject-
401814.Cyclistic.trips_q1_2020`
    GROUP BY start_station_name, member_casual
)
SELECT
  start_station_name,
  member_casual,
  StationCounter.station_count
FROM StationCounter
WHERE StationCounter.count_rank <=5
```

member_casual	start_station_name	station_c...
casual	Canal St & Adams St	7.586
casual	Clinton St & Madison St	6.546
casual	Clinton St & Washington Blvd	5.823
casual	Kingsbury St & Kinzie St	4.491
casual	Columbus Dr & Randolph St	4.099
member	HQ QR	3.766
member	Lake Shore Dr & Monroe St	1.590
member	Streeter Dr & Grand Ave	1.530
member	Shedd Aquarium	998
member	Millennium Park	779