Cyclistic Bike-share Analysis Capstone Project using spreadsheets, SQL, and Tableau

In this case study, I analyzed historical data from a Chicago-based bike-share company to identify trends in how their customers use bikes differently. The main tools I use are spreadsheets, SQL, and Tableau.

- Tableau for data visualizations
- Pavan-portfolio
- Cyclistic bike-share slide presentation

A more in-depth breakdown of the case study scenario is included below, followed by my full report.

Scenario

Cyclistic is a bike-share company based in Chicago with two types of customers. Customers who purchase single-ride or full-day passes are known as **casual riders**, while those who purchase annual memberships are known as **members**. Cyclistic's financial analysts have concluded that annual members are much more profitable than casual riders. The marketing director believes the company's future success depends on maximizing the number of annual memberships.

The marketing analytics team wants to understand how casual riders and annual members use Cyclistic bikes differently. The team will design a new marketing strategy from these insights to convert casual riders into annual members. The primary stakeholders for this project include Cyclistic's director of marketing and the Cyclistic executive team. The Cyclistic marketing analytics team are secondary stakeholders.

Defining the problem (Ask phase)

The main problem for the marketing and marketing analytics team director is this: Design marketing strategies to convert Cyclistic's casual riders into annual members. Three questions will guide this future marketing program. For the scope of this project, I will analyze the first question;

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

We will first get a broad sense of specific patterns occurring in the two groups by looking at the data. Understanding the differences will provide more accurate customer profiles for each group. These insights will help the marketing analytics team design high-quality targeted marketing for converting casual riders into members. For the Cyclistic executive team, these insights will help Cyclistic maximize the number of annual members and increase the company's future growth.

Business task

Analyze historical bike trip data to identify trends in how annual members and casual riders use Cyclistic bikes differently.

Data Sources (Prepare phase)

We will be using Cyclistic's historical bike trip data from the last 12 months, which is publicly available.

- Cyclistic-bikeshare-data

The data is structured and organized in rows(records) and columns(fields). Each record represents one trip, and each trip has a unique field that identifies it: ride_id.
In terms of bias and credibility, the data sources we are using ROCCC:

- **Reliable and original:** this is public data that contains accurate, complete, and unbiased info on Cyclistic's historical bike trips. It can be used to explore how different customer types are using Cyclistic bikes.
- **Comprehensive and current:** these sources contain all the data needed to understand how members and casual riders use Cyclistic bikes. The data is from the past two years. It is current and relevant to the task at hand. This is important because the usefulness of data decreases as time passes.
- **Cited:** these sources are publicly available data provided by Cyclistic company. Governmental agency data and vetted public data are typically good sources of data.

Data cleaning and manipulation (Process phase)

Google Spreadsheets: initial data cleaning and manipulation

Our next step is to make sure the data is stored appropriately and prepared for analysis. After downloading all 12 zip files and unzipping them, I housed the files in a temporary folder on my desktop. I also created subfolders for CSV files and the XLS files so that I have a copy of the original data. Then, I uploaded each CSV file into google sheets.
For each CSV file, I did the following:

- Changed the format of the started_at and ended_at columns
    - Formatted as custom DATETIME
    - Format > Numbers > DATETIME

- Created a column called ride_length
    - Calculated the length of each ride by subtracting the column started_at from the column ended_at (example: =D2-C2)
    - Formatted as Duration
    - Format > Number > Duration

- Created a column called ride_date
    - Calculated the date of each ride started using the DATE command (example: =DATE(YEAR(C2), MONTH(C2), DAY(C2)))
    - Format > Number > date

- Created a column called start_time
    - Calculated the start time of each ride using the started_at column
    - Formatted as TIME
    - Format > Number > Time

- Created a column called end_time
    - Calculated the end time of each ride using the ended_at column
    - Formatted as TIME
    - Format > Number > Time

- Created a column called day_of_week
    - Calculated the day of the week that each ride started using the WEEKDAY command (example: =WEEKDAY(C2, 1))
    - Formatted as a Number with no decimals
    - Format > Number > Number
    - Note: 1 = Sunday, and 7= Saturday

BigQuery: Further data cleaning and manipulation via SQL

Since these datasets are so large, it makes sense to move our analysis to a tool that is better suited for handling large datasets. I chose to use SQL via BigQuery.

In order to continue processing the data in BigQuery, I created a bucket in Google Cloud Storage to upload all 12 files. I then created a project in BigQuery and uploaded these files as datasets. I have provided my initial cleaning and transformation SQL queries here for reference:

Data cleaning using SQL:

```
1)  -- Data cleaning: ride_length
    -- Replace hashtags with '0:00:00' in ride_length field
    UPDATE
      `smiling-landing-379505.bike_share_analysis.M-2021-02_divvy_tripdata`
    SET
      ride_length = '0:00:00'
    WHERE
      ride_length =
    '##########################################################################
    ##########################################################################
```

```
##############################################################################
##################'
```

Repeated this step for all 12 tables.

2) Converting ride_length from TIME data type to STRING data type

```
-- Updating data type for ride length part 1
-- Convert from time to string
SELECT
  ride_id,
  rideable_type,
  started_at,
  ended_at,
  CAST(ride_length AS STRING) AS ride_length,
  day_of_week,
  start_station_name,
  start_station_id,
  end_station_name,
  end_station_id,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  member_casual
FROM
  `smiling-landing-379505.bike_share_analysis.M-2021-01_divvy_tripdata`
Saved the result tables after running the query.
```

- I also deleted the previous tables with the ride_length column as the time data type.

3)Convert the ride_column data type from string to interval.

```
-- Updating data type for ride length part 2
-- Convert from string to interval
SELECT
  ride_id,
  rideable_type,
  started_at,
  ended_at,
  CAST(ride_length AS INTERVAL) AS ride_length,
  day_of_week,
  start_station_name,
  start_station_id,
```

```
  end_station_name,
  end_station_id,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  member_casual
FROM
  `smiling-landing-379505.bike_share_analysis.M-2022-12_divvy-tripdata`
```

Create quarterly tables

In order to perform analysis by season, let's combine these tables. We'll create Q1, Q2, Q3, and Q4 tables for analysis. We'll have two Q1 tables-one for 2021 and one for 2022.

```
4) -- We'll create five tables for analysis:
 -- Table 1) 2021_Q1 -> 2,3
 -- Table 2) 2021_Q2 -> 4,5
 -- Table 3) 2021_Q3 -> 11,12
 -- Table 4) 2022_Q1 -> 1,2,3,4
 -- Table 5) 2022_Q2 -> 11,12
--create 2021_Q1 and then repeat for the remaining four tables
SELECT
  ride_id,
  rideable_type,
  started_at,
  ended_at,
  ride_length,
  day_of_week,
  start_station_name,
  start_station_id,
  end_station_name,
  end_station_id,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  member_casual
FROM
  `smiling-landing-379505.bike_share_analysis.M-2021-02-divvy-tripdata`
UNION DISTINCT
```

```sql
SELECT
  ride_id,
  rideable_type,
  started_at,
  ended_at,
  ride_length,
  day_of_week,
  start_station_name,
  start_station_id,
  end_station_name,
  end_station_id,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  member_casual
FROM
  `smiling-landing-379505.bike_share_analysis.M-2021-03-divvy-tripdata`
```

5)
```sql
-- Update the format for 'day_of_week' from float to string
-- Start with 2021_Q1_02-03 and repeat for other remaining tables
-- Update 'day_of_week' format with CAST()
SELECT
  ride_id,
  rideable_type,
  started_at,
  ended_at,
  ride_length,
  CAST(day_of_week AS STRING) AS day_of_week,
  start_station_name,
  start_station_id,
  end_station_name,
  end_station_id,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  member_casual
FROM
  `smiling-landing-379505.bike_share_analysis.2022-Q5_11-12`
```

6)
```sql
-- Update 'day_of_week' values in Q1, Q2, Q3, Q4
```

```
-- Start with 2021-Q1 and repeat for other four tables
-- Update 'day_of_week' values with CASE WHEN
UPDATE
  `smiling-landing-379505.bike_share_analysis.2022-Q2-11-12`
SET
  day_of_week =
    CASE
      WHEN day_of_week = '1' THEN 'Sunday'
      WHEN day_of_week = '2' THEN 'Monday'
      WHEN day_of_week = '3' THEN 'Tuesday'
      WHEN day_of_week = '4' THEN 'Wendesday'
      WHEN day_of_week = '5' THEN 'Thursday'
      WHEN day_of_week = '6' THEN 'Friday'
      WHEN day_of_week = '7' THEN 'Saturday'
    END
WHERE
  day_of_week IN ('1', '2', '3', '4', '5',: '6', '7')
```

- Repeated for all remaining tables.

Analysis: Exploring the data for finding relations and trends.

2021_Q1 - quarterly data exploration

We'll select a few columns from 2021-Q1-02-03 to preview in a temporary table. This will help give us an idea of potential trends and relationships to explore further:

1)

```
-- Select columns from Q1 data to preview
SELECT
  ride_id,
  started_at,
  ended_at,
  ride_length,
  day_of_week,
  start_station_name,
```

```
  end_station_name,
  member_casual
FROM
  `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
ORDER BY
  ride_id DESC
```

- This gave a total of 278118 trips.

Total trips

- We'll create total columns for overall, annual members, and casual riders. We'll also calculate percentages of the overall total for both types.

2)

```
      -- Total Trips: Members vs Casual
-- Looking at overall, annual member and casual rider totals
SELECT
  TotalTrips,
  TotalMemberTrips,
  TotalCasualTrips,
  ROUND(TotalMemberTrips/TotalTrips,2) * 100 AS MemberPercentage,
  ROUND(TotalCasualTrips/TotalTrips,2) * 100 AS CasualPercentage
FROM
  (
  SELECT
    COUNT(ride_id) AS Totaltrips,
    COUNTIF(member_casual = 'member') AS TotalMemberTrips,
    COUNTIF(member_casual = 'casual') AS TotalCasualTrips,
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  )
```

- Out of 278118 total trips of 2021-Q1, 66% are from annual members and the remaining 34% are from casual members.

Average ride lengths

- How does the average ride_length differ for these groups?

3)

```
    -- Average Ride Lengths: Members vs Casual
-- Looking at overall, member and casual average ride lengths
SELECT
  (
  SELECT
```

```
    AVG(ride_length)
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  ) AS AvgRideLength_Overall,

  (
  SELECT
    AVG(ride_length)
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  WHERE
    member_casual = 'member'
  ) AS AvgRideLength_Member,

  (
  SELECT
    AVG(ride_length)
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  WHERE
    member_casual = 'casual'
  ) AS AvgRideLength_Casual
```

- The overall average ride length is around 23 minutes, the member average ride length is around 15 minutes, and the casual average ride length is around 39 minutes.

Max ride lengths

- We'll look at the maximum values for ride_length to see if anything extreme is influencing the casual rider average:

4)
```
-- Looking at max ride lengths to check for outliers
SELECT
  member_casual,
  MAX(ride_length) AS ride_length_MAX
FROM
  `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
GROUP BY
  member_casual
ORDER BY
```

```
  ride_length_MAX DESC
LIMIT
  2
```

- There is a hint that the 23 minutes more average ride length for casual riders than that of members is due to longer ride lengths of casual members like 528 hours or 22 days of ride length for casual riders.

5)
```
-- Looking at top 100 longest trips for casual riders
SELECT
  member_casual,
  ride_length
FROM
  `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
WHERE
  member_casual = 'casual'
ORDER BY
  ride_length DESC
LIMIT
100
```

This result proves that more than one casual rider's ride length is influencing the average ride length of a casual member.

Median ride lengths

- Since there are more than a few outliers impacting the average, we're going to use the median instead of the average. The median will be more accurate for our analysis:

6)
```
-- Looking at median ride lengths
SELECT
  DISTINCT median_ride_length,
  member_casual
FROM
  (
  SELECT
    ride_id,
    member_casual,
    ride_length,
    PERCENTILE_DISC(ride_length, 0.5) OVER(PARTITION BY member_casual) AS
median_ride_length
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  )
```

```
ORDER BY
  median_ride_length DESC
```

- The median ride length for casual and annual members is 18 and 10 minutes respectively.

Here the difference is not much like 23 minutes in average ride length.
The busiest day for rides

- Let's see which day has the most rides for annual members and casual riders:

7)
```
-- Looking at which days have the highest number of rides
SELECT
  member_casual,
  day_of_week AS ModeDayOfWeek # Top number of day_of_week
FROM
  (
  SELECT
    DISTINCT member_casual, day_of_week, ROW_NUMBER() OVER (PARTITION BY member_casual
ORDER BY COUNT(day_of_week) DESC) AS rn
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  GROUP BY
    member_casual, day_of_week
  )
WHERE
rn = 1

ORDER BY
  member_casual DESC
```

- Saturday is the most popular day for both casual and annual members.

Median ride length per day
Let's look at the median ride lengths per day for both annual members and casual riders. Since Saturday is the most popular overall, do we think it will also have the highest median ride length?

8)
```
-- Looking at median ride lengths per day for annual members
SELECT
  DISTINCT median_ride_length,
  member_casual,
  day_of_week
```

```
FROM
  (
  SELECT
    ride_id,
    member_casual,
    day_of_week,
    ride_length,
    PERCENTILE_DISC(ride_length,0.5) OVER(PARTITION BY day_of_week)median_ride_length
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  WHERE
    member_casual = 'member'
  )

ORDER BY
  median_ride_length DESC
```

9)

```
-- Looking at median ride lengths per day for casual members
SELECT
  DISTINCT median_ride_length,
  member_casual,
  day_of_week
FROM
  (
  SELECT
    ride_id,
    member_casual,
    day_of_week,
    ride_length,
    PERCENTILE_DISC(ride_length,0.5 IGNORE NULLS) OVER(PARTITION BY day_of_week) AS
median_ride_length
  FROM
    `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
  WHERE
    member_casual = 'casual'
  )

ORDER BY
  median_ride_length DESC
```

- It is interesting to find that the median ride length for casual riders on the top five days (SUN, SAT, MON, TUE, WED) is nearly double the amount for annual members on their top five days (SAT, SUN, MON, TUE, WED).

Total rides per day

- Let's look at the total rides per day. We'll create columns for the overall total, annual members, and casual riders:

10)
```sql
-- Looking at total number of trips per day
SELECT
  day_of_week,
  COUNT(DISTINCT ride_id) AS TotalTrips,
  SUM(CASE WHEN member_casual = 'member' THEN 1 ELSE 0 END) AS MemberTrips,
  SUM(CASE WHEN member_casual = 'casual' THEN 1 ELSE 0 END) AS CasualTrips
FROM
  `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
 GROUP BY
  day_of_week
ORDER BY
  TotalTrips DESC
```

- Calculated the total trips, member trips, and casual trips per day.

Start stations

- Next, we'll look at the most popular start stations for trips. We'll again include columns for the overall, annual member, and casual rider totals per start station:

11)
```sql
-- Start stations: member vs casual
-- Looking at start station counts
SELECT
  DISTINCT start_station_name,
  SUM(CASE WHEN ride_id = ride_id AND start_station_name = start_station_name THEN 1
ELSE 0 END) AS Total,
  SUM(CASE WHEN start_station_name = start_station_nameAND member_casual = 'member'
THEN 1 ELSE 0 END) AS Member,
  SUM(CASE WHEN start_station_name = start_station_name AND member_casual = 'casual'
THEN 1 ELSE 0 END) AS Casual

FROM
  `smiling-landing-379505.bike_share_analysis.2021-Q1-02-03`
GROUP BY
```

```
  start_station_name
ORDER BY
  Total DESC
```

Here the interesting trend is that different riders favored different stations, it is not a regular trend. We can get the top 10 start stations used by casual and annual riders by using Casual DESC and Member DESC commands.

- We found that there is only one start station that is present in both the top 10 tables i.e Clark St & Elm St. which ranked #1 for annual members and #10 for casual members.
- The casual riders seem to prefer stations near the water like Lake Shore Dr & Monroe St and Streeter Dr & Grand Ave.
- While annual members favored the stations like Dearborn St & Erie St and Kingsbury St & Kinzie St.

Initial Hypothesis: Casual riders favored stations near water which are tourist places because they want to use the bikes for weekend entertainment. Annual members favored stations near downtown, retail areas because they are using bikes for their work and shopping trips.

Data visualizations

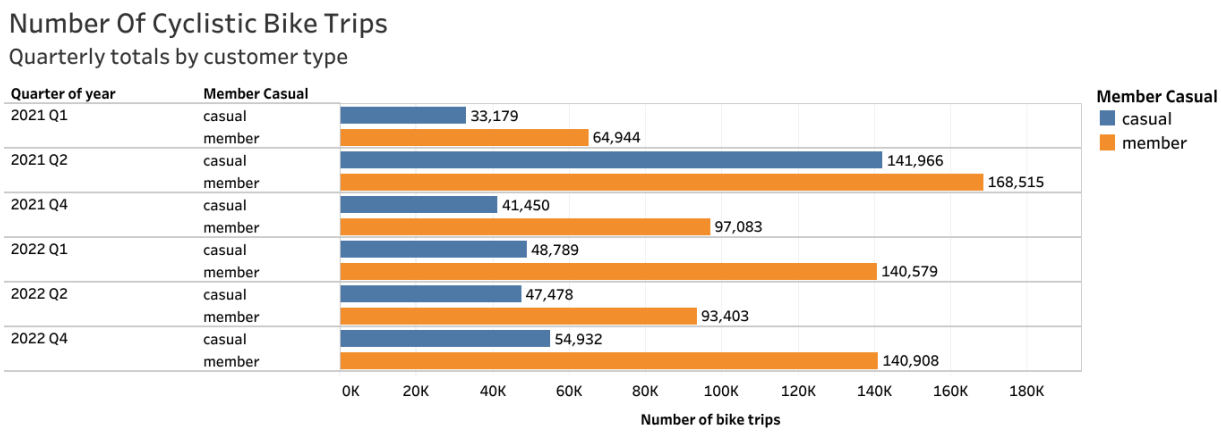
In order to analyze all twelve months together, we'll combine the five quarterly tables into one table. For a summary and overall visualization of the full-year analysis, please visit my Tableau account. I will also highlight some of the interesting trends and relationships I discovered below.

- Tableau: Data visualizations

Seasonal trends

The busiest time of year for overall bike trips is Q2- April, May. This makes sense because these months are mainly springtime. Bike riding is better suited for warmer weather, which is also why we see a major drop-off in total rides during the winter months of Q1- February, and March.

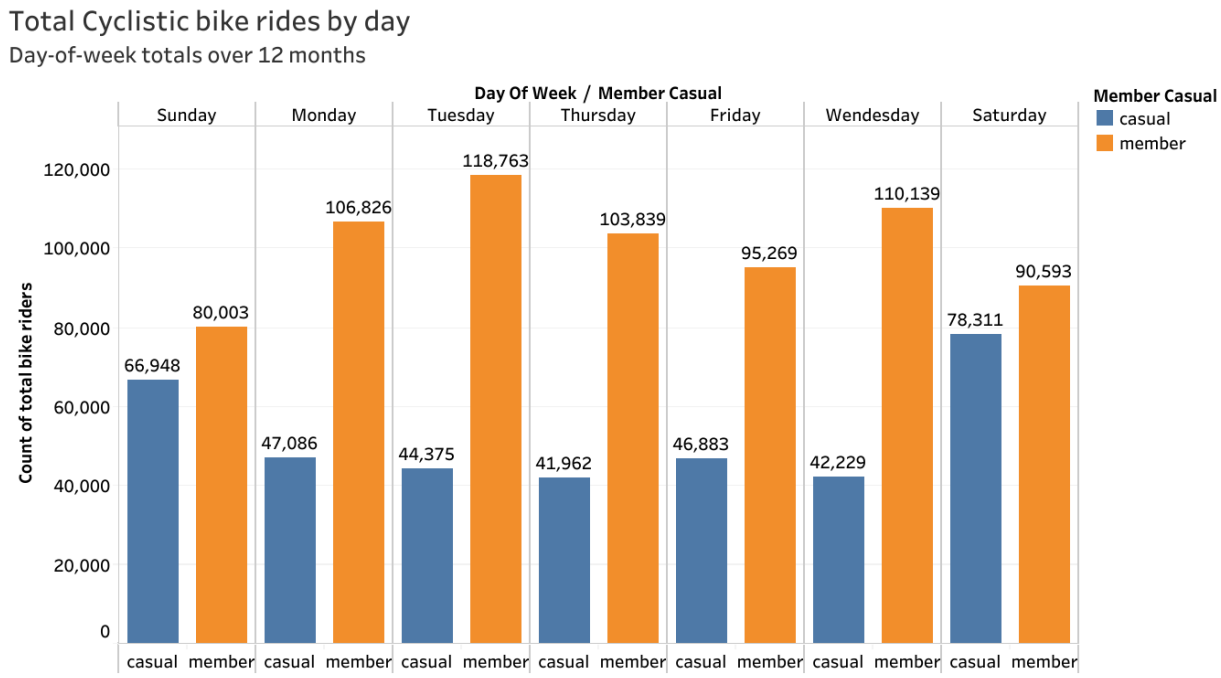
Annual members outnumbered casual riders in every quarter. Interestingly, the annual members nearly doubled the casual ridership in Q1, and Q3 while with a very low difference in Q2.

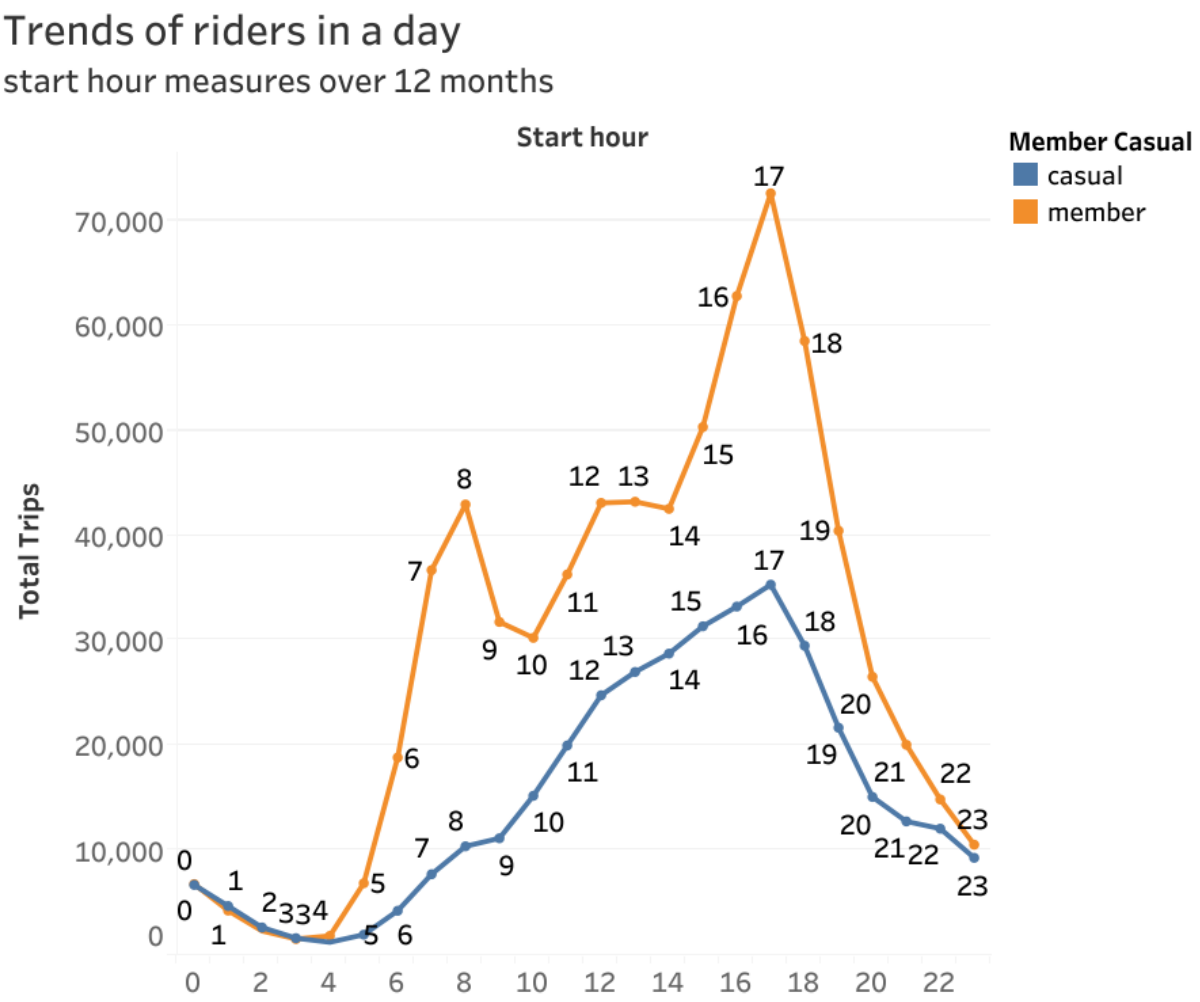


Day of week

Which days of the week have the highest number of rides for casual riders vs annual members? Let's look at the mode for each quarter and for the full year:

Casual raiders were extremely consistent, with Saturday revealing itself as their preferred day of the week for each quarter and across the full year. Meanwhile, the annual members looked for a favor in the middle of the week for their bike use. The most popular day for them across the full year was Wednesday.





Conclusion

Stakeholder presentation

I have provided links below for my stakeholder presentation, which includes the following:

- A summary of my analysis
- Supporting visualizations and key findings
- Three recommendations based on my analysis.

Presentation: Cyclistic-Bikeshare-analysis