Cyclistic Case Study

Tools: Excel, SQL, Tableau

1. Introduction

The following article shows my process of completing the Cyclistic case study, a **capstone project** for the **Google Data Analytics Certification** on **Coursera**. I decided to use Excel, SQL, and Tableau to complete the case study, as Excel & SQL are the two most requested (and used) skills for Data Analysts—and thus those are the skills I wanted to hone through the completion of this project—and Tableau is my preferred data visualization tool. The format of this article will be in the structure recommended by the course, following the steps of **Ask, Prepare, Process, Analyze, Share, Act.**

Scenario

As per the course briefing:

> 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.

2. Ask

Business Task

The business task is to identify the differences between how **casual riders** and **annual members** utilize Cyclistic's service. The marketing team believes that annual members will be the key to the company's success, and thus we want to convert casual customers to members based on the differences in their usage of the service. We must present our findings with data-driven support and compelling visualizations in order to convey our solutions to Cyclistic executives.

3. Prepare

Data Source & Summary

Cyclistic's trip data is publicly accessible and is stored on the following page:

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

Accessibility & Usage

The datasets provided have a different name as Cyclistic is a fictional company for the sake of the case study. The **Cyclistic Trip Data** is collected by a real bike sharing company located in Chicago, and has been made available by **Motivate International**

Inc. under this **license.** The data has also been cleaned of any personally identifiable information in order to prevent any data-privacy issues.

Data Structure

As requested in the assignment, I only used data from the previous **12** months for analysis. I downloaded files in **CSV** format from February 2022 to January 2023.

Each file has a table with 13 columns and hundreds of thousands of observations. The columns are as follows:

- ride_id
- rideable_type
- started_at
- ended_at
- start_station_name
- start_station_id
- end_station_name
- end_station_id
- start_lat
- start_lng
- end_lat
- end_lng
- member_casual

The column names match in each month and are in the same order from left to right. Each row represents an observation with the details of a single ride.

Data Limitations & Integrity

The **Cyclistic Trip Data** is collected by a real bike sharing company so there are no issues with **bias** or **credibility**. The data is also **original**, **current**, and **cited**. However, this data is not **comprehensive** as it lacks some information. It does not include data about the use of options such as reclining bikes, hand tricycles and cargo bikes, which

is said to make up about 8% of total riders. Furthermore, data-privacy issues prevents me from using personally identifiable information. As such, I won't be able to determine if casual riders have purchased multiple single passes. Financial information such as the ticket fare of each Ride Id is not available either. If this information was available, I could have calculated whether or not casual riders would have benefitted from purchasing annual memberships. Overall, the data is **reliable** and **relevant**, and will suffice for the purpose of this assignment.

4. Process

Documentation of any cleaning or manipulation of data

Note: All SQL queries used for both processing and analysis can be found on Google Docs **here** or on my Github **here.**

1. Downloaded and saved files on PC in CSV format.
2. Added 'ride_length' and 'day_of_week' columns in Excel.
3. Sorted by 'ride_length' and deleted entries with negative values.
4. Uploaded files to Google Cloud, then imported the files into BigQuery to continue analysis using SQL.
5. Merged the 12 files into a single table.
6. The following steps were taken to check for data accuracy:
    a. Checked for any null values in member_casual.
    b. Checked for any records with duplicate ride_id.
    c. Checked for spaces around string values.
    d. Checked for misspellings in rideable_type and member_casual.
    e. Counted records with null values in either start_station_name or end_station_name.
7. I decided to keep records that had null values for start and end station names, as it was a large portion of the data set and the null values would not be impactful for most analysis and visualizations.

    a. For visualizations involving location, I removed these records.

8. Based on these observations, the following steps were taken to clean the data:
    a. Deleted any entries with null values for member_casual.
    b. Deleted any entries with duplicate ride_id.

5. Analyze

Exploring the data

Through a variety of SQL queries, I explored the nature of the data and discovered some data points:

▼ Member distribution:

- 59% of riders are annual members compared to 41% being casual riders.

▼ Type of bike preferred by member type:

- **Members preferred classic bikes more than casual riders:**
    - 51% of the time vs 41% of the time for casual riders. However, members were relatively even on classic vs electrical bikes.
- **Casual members highly preferred electric bikes:**
    - 59% of the time for casual riders vs 49% of the time for members.

▼ Average ride duration by member type:

- **Casual members have a longer average ride duration:**
    - 29 minutes vs 13 minutes for annual members.

▼ Distribution of rides by day of week and member type:

- **The majority of casual members rides were on the weekend.**
- Annual members' rides were evenly dispersed throughout the week, with the least rides being on the weekend.

▼ Ride duration by day of week and member type:

- For both members weekend rides were longer than weekday rides.

▼ Distribution of rides and average ride length by month:

- Average ride length increased in the summer months and was at its lowest in the colder months.

▼ Average ride length of round trips vs different starting/ending locations

- Average length of round trips was 26 minutes, which was longer than the length of one way trips at 16 minutes. This can be explained as round trips were likely for leisure, whereas one way trips were likely direct commutes for work used by annual members.

Aggregating information into key findings

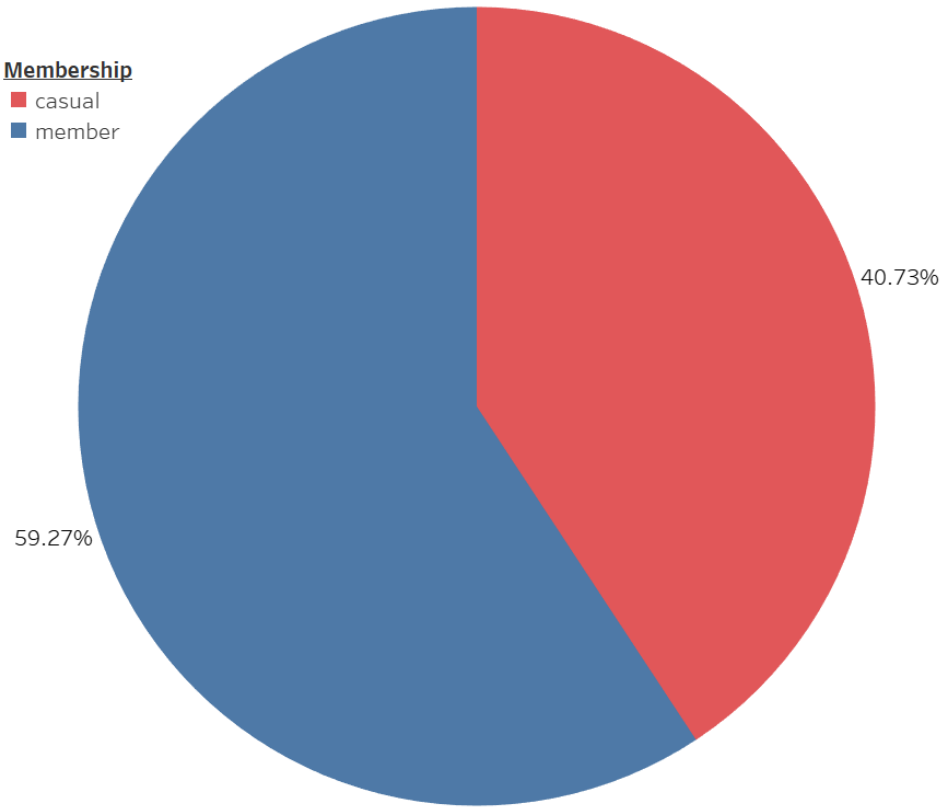
From these data points, we can see a key difference in the way that annual members and casual consumers are using the service. Annual members are using the service for direct commutes to and from work, supported by the higher frequency of trips during weekdays and the shorter ride length. On the contrary, casual riders are using the service for leisure rides, which mostly take place on weekends. So, in order to increase turnover to annual members, we will need to target and provide benefits for the way that casual consumers utilize the product.

6. Share

As stated in the course briefing, it’s important to provide compelling and clear visuals to share our findings in a succinct way. In order to accomplish this, I used Tableau as my data visualization tool of choice.

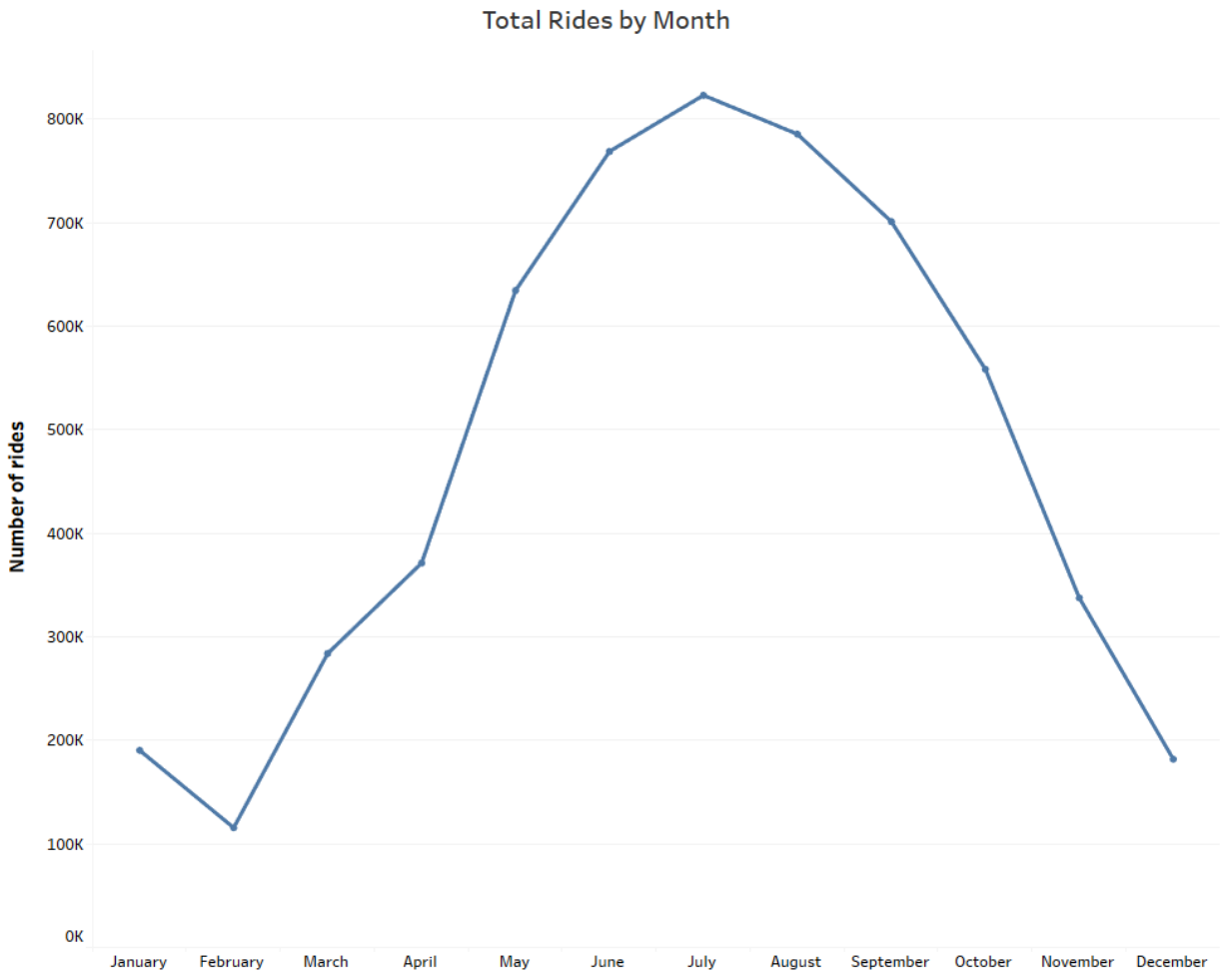
The following visualizations will highlight some key takeaways that will allow us to make informed decisions on how to properly target the needs of Cyclistic’s customers.

Member type distribution



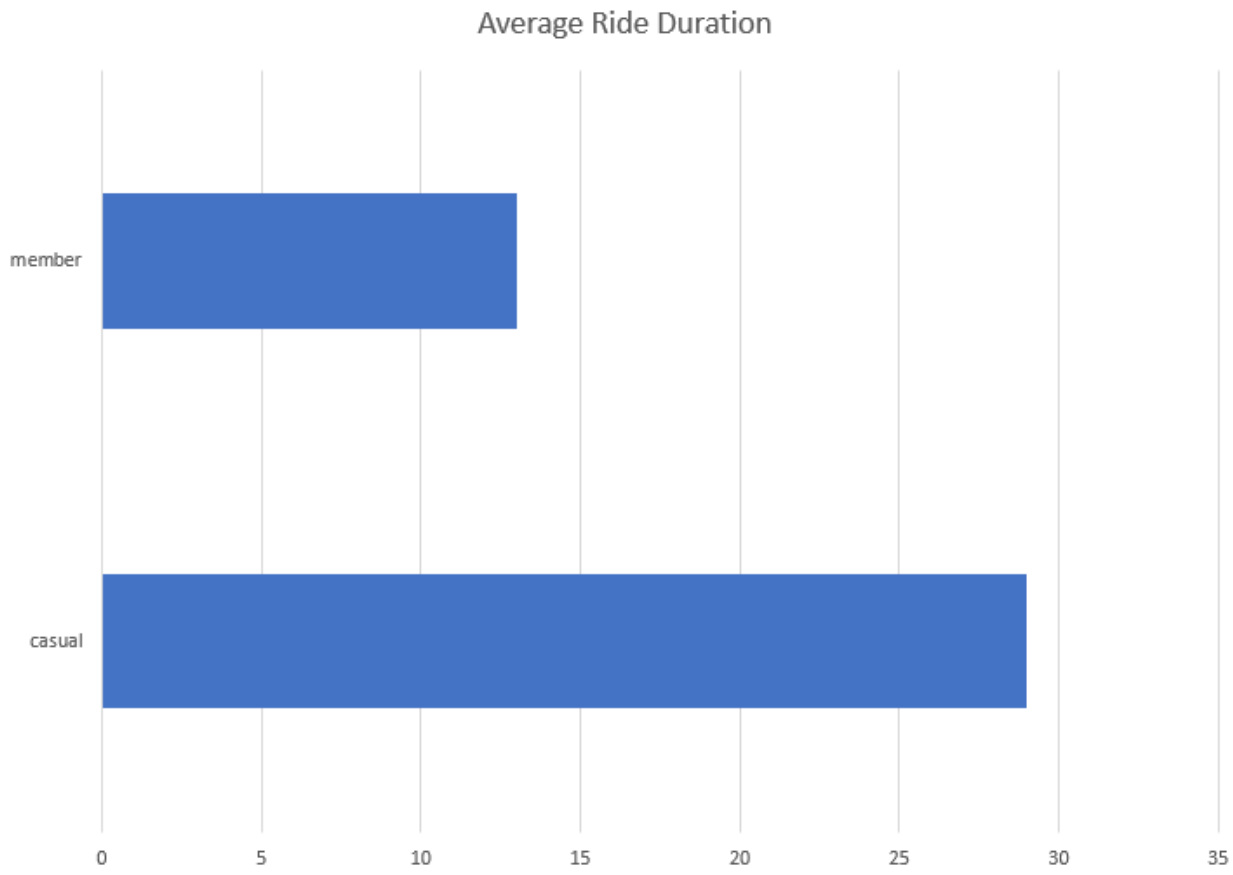
This pie chart shows that the majority of Cyclistic's customers are members. However, casual consumers still represent a large portion of customers, and converting them into annual members would increase revenue significantly. So, our original goal of increasing turnover from casual to annual members will be beneficial.

Monthly rides



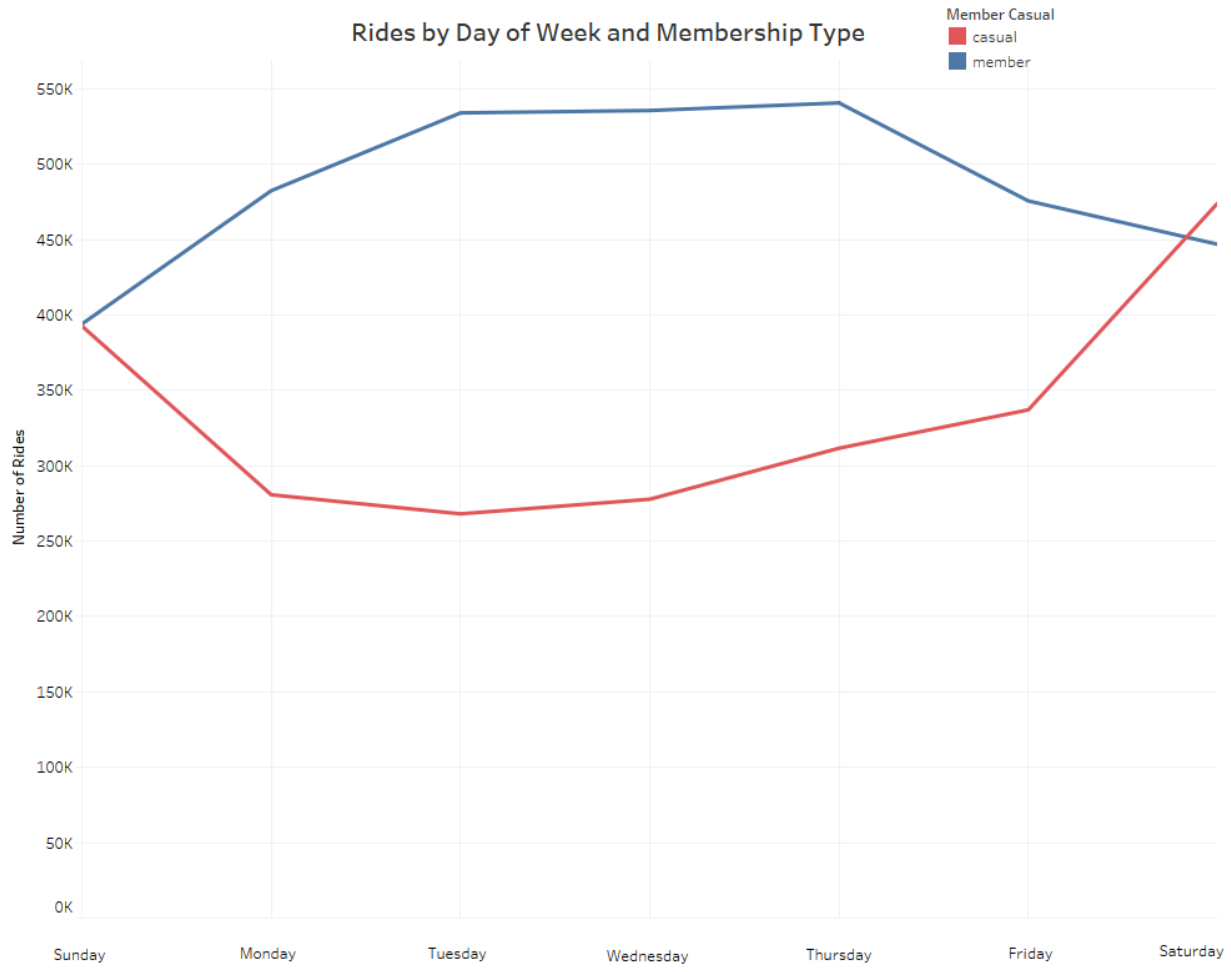
The data shows a **tendency for customers to ride more in the summer and warmer months.** This is an expected behavior, and is also not distinguishable between casual riders and annual members. However, it is still an important observation that we can take advantage of in our solutions. Furthermore, this displays that our data processing was done correctly as this is an expected behavior that correlates with our understanding of cyclists.

Average ride duration



We see that **casual customers have a longer average ride duration than annual members**. This can be explained as Cyclistic doesn't have time constraints on their bikes, and casual riders are often tourists or users who utilize the service to enjoy leisure rides and sightseeing. In comparison, members often are using their bikes with definitive plans—such as traversing to and from work.

Day of the week



The graph above further highlights the distinction in customer behavior based on membership type. We see that **casual customers have a higher frequency of trips on weekends**, whereas members have most of their rides during the weekday. This corroborates with our findings earlier, as members are using their rides for day to day tasks such as commuting to work, while casual customers use the service for leisure on weekends.

7. Act

Finding Summary

The results of the analysis show that casual consumers and annual members fundamentally use Cyclistic's service in different ways. Members are working locals of Chicago who use the bikes to commute daily to and from their workplaces. On the contrary, casual customers are tourists and leisure riders who use the bikes for sightseeing and with less intent. From these results, we can conclude that in order to convert casual customers to members, we need to incentivize and target the way that they use the service, rather than trying to convert them to a similar membership model that had attracted the current annual members.

Recommendations

For what actions we can take, lets first cater our approach towards the needs of our target audience, the casual users of Cyclistic. These users prefer to use Cyclistic bikes during the weekends, and their trip length is longer on average than annual members. First, under Cyclistic's current strategy ride length doesn't contribute to higher pricing. So, the marketing team should look into developing a system to charge customers based on ride length. This would also benefit current annual users who have lower average ride duration. Then, after implementing this system, Cyclistic could offer a couple annual packages that would target the needs of casual riders:

1. **An annual package with special offers for weekend users**
2. **An annual package with discounts for riders who take longer rides**
3. **A seasonal package for the summer**

Furthermore, the Cyclistic marketing team should design and target ad campaigns towards promoting these packages and reaching the casual audience. For example, they could create media campaigns that play on the seasonality of bike rides and encourage users to look into the seasonal package. Or they could have another campaign that emphasizes the enjoyment of weekend leisure rides and highlights the annual package for weekend users.