Cyclistic Capstone Report

Google Data Analytics

Max Jewell

Ask.......... 2
Business Task.......... 2
Prepare.......... 3
Divvy_Trips_2019_Q1.......... 4
Divvy_Trips_2020_Q1.......... 4
Process.......... 6
Analyse.......... 7
Act.......... 9
Recommendation 1.......... 9
Recommendation 2.......... 9
Recommendation 3.......... 10

Ask

The problem is identifying how annual members and casual riders use Cyclistic bikes differently, in order to uncover behavioral patterns and preferences that can inform marketing strategies aimed at converting more casual riders into annual members.

By analyzing historical bike usage data, we can uncover key differences in ride patterns (e.g., duration, time of day, day of week, location) between casual riders and annual members. These insights can help the marketing team design targeted campaigns, improve messaging, and tailor promotions to appeal to casual riders' habits and motivations—ultimately encouraging them to become annual members.

Business Task

To analyze and compare the usage behavior of annual members and casual riders using Cyclistic bike-share data, in order to provide data-driven insights that will support the development of targeted marketing strategies aimed at converting casual riders into annual members.

Prepare

The dataset is provided in a **Google Sheets document** shared by the Cyclistic marketing analytics team. This document contains historical trip data for the Cyclistic bike-share program, including information on both **casual riders** and **annual members**.The data is given in tabular format, so each row represents an individual bike ride. Columns in these datasets have similar and relevant information, however not identical. For this reason not all data will be used due to compatibility and necessity.

Following the data ruling of ROCCC, the data is reliable as it has come from the company's internal system, meaning no external company has been able to alter any of the raw data presented to me. Original as it is first party data. Comprehensive - it gives lots of information on the users behaviours when using bikes, which will enable me to answer the business task for my stakeholders. It includes data such as; duration of ride, where the user has started their journey and where they have ended their journey and most importantly their membership type. Current - the datasets provided are current as they are dated quarter 1 of 2019 and quarter 1 of 2020, current and ideal for direct comparison as they fall within the same timeframes. Cited - the datasets are internal, and coming directly from where the raw data has been collected.

The datasets given do not contain any information that could identify a person exactly, such as government names, personal addresses or IP addresses. Meaning the data is anonymised. Since Cyclistic is a fictional company for educational purposes there are no legal restrictions on using or sharing the data for this project.

To ensure the dataset was accurate and reliable for analysis, several verification steps were undertaken:

1. **Column Naming and Consistency**
   Each column name was reviewed to confirm it followed a consistent naming convention, was spelled correctly, and accurately described the data it contained. This step ensured clarity and prevented confusion during later stages of analysis.

2. **Identification of Null or Invalid Values**
   The dataset was checked for null values, missing entries, or data types that were inconsistent with the column’s intended purpose (e.g., text in numeric fields). Detecting and addressing these anomalies helped maintain the accuracy of the results.

3. **Duplicate Record Detection**
   Duplicate rows were identified and removed, as their presence could skew aggregate metrics and lead to misleading conclusions.

4. **Format and Logical Validity Checks**
   The data was examined for correct formatting, particularly for time-related fields. Decimal values in time columns were corrected to avoid misinterpretation. Logical checks were also applied — for example, confirming that ride end times always occurred after start times and that all recorded ride durations were valid.

The datasets provided key information — including ride quantity, trip duration, start and end locations, and membership type — that directly supports the business task. Each of these elements contributes valuable insight, enabling a comprehensive analysis of how annual members and casual riders use Cyclistic bikes differently.

During the initial review of the raw datasets, several issues were identified that could lead to invalid or inaccurate insights if left unaddressed. These are outlined below:

Divvy_Trips_2019_Q1

1. **Inconsistent Naming Conventions**
   - Column headers lack consistent naming formats.

2. **Trip Duration Formatting**
   - The *duration* field is not formatted as a time value, making it less intuitive to interpret.

3. **Missing Values**
   - Significant null values exist in the `gender` column (19,713 blanks) and the `birthyear` column (18,025 blanks), identified using the `COUNTIF` function.

4. **Date and Time Fields**
   - The `start_time` and `end_time` columns contain both date and timestamp in a single field. Separating these into distinct columns will allow more flexible time-based analysis.

5. **Extraneous Rows**
   - Two empty rows are present at the end of the dataset.

Divvy_Trips_2020_Q1

1. **Unnecessary Columns**
   - The `rideable_type` column is redundant since all Cyclistic bikes are docked bikes.

2. **Identifier Inconsistency**

    - This dataset uses `ride_id` to identify rides, whereas the 2019 dataset uses `trip_id`.

3. **Location Data Differences**
    - Start and end latitude/longitude coordinates are provided alongside station names. While potentially useful, these fields are not present in the 2019 dataset, limiting direct comparability.

4. **Missing Duration Field**
    - Unlike the 2019 dataset, there is no pre-calculated trip duration column. This metric will need to be derived from the `started_at` and `ended_at` columns.

5. **Date and Time Fields**
    - Both `started_at` and `ended_at` combine date and time in a single column. Separating these will facilitate more granular analysis.

6. **Membership Type Naming Differences**
    - Subscription type values differ between datasets:
        - 2019: “Subscriber” and “Customer”

        - 2020: “Member” and “Casual”
          These need to be standardised for accurate comparisons.

7. **Extraneous Rows**
    - Two empty rows are present at the end of the dataset.

Overall, while the data presents some inconsistencies and formatting issues these can be addressed through cleaning and transformation. The datasets provide the core information necessary to evaluate differences between casual and member riders, supporting actionable business insights.

Process

For this project, **Google Sheets** was chosen as the primary tool for both data cleaning and visualization. Its intuitive interface and robust filtering and formula functions enabled efficient identification and correction of data quality issues such as null values, duplicates, and inconsistent formatting. Additionally, Google Sheets' charting capabilities allowed for the creation of clear and effective visualizations to support the analysis.

Using a single platform streamlined the workflow, making it easier to maintain data integrity throughout the cleaning and analysis process while producing presentation-ready visuals for stakeholders.

To ensure the integrity of the data used for analysis, I followed a rigorous cleaning protocol addressing completeness, relevance, and consistency:

- **Handling Missing Values:**
  Significant null values were present in the `gender` and `birthyear` columns of the 2019 dataset. Rather than removing rows with missing values—which would have resulted in losing over 10% of the data—I opted to remove these columns entirely. This decision was made because these fields were not only incomplete but also unavailable in the 2020 dataset, making them irrelevant for a comparative analysis of annual and casual riders.

- **Removing Irrelevant Data:**
  Columns that did not contribute to answering the business question, such as `gender` and `birthyear`, were excluded to focus the analysis on meaningful, comparable variables.

- **Standardizing Naming Conventions:**
  To enhance readability and reduce the risk of error, I standardized column headers across datasets. For example, columns indicating the start and end stations of rides were labeled consistently between the 2019 and 2020 datasets, facilitating seamless merging and analysis.

- **Documentation and Transparency:**
  All modifications—from raw data to cleaned datasets—are documented comprehensively in the Data Cleaning Log. This record supports transparency, maintains data integrity, and enables reproducibility for future reviews or extensions of the project.

A full record of all changes from the raw datasets to the cleaned versions are available in the Data Cleaning Log for reference. Links to the modified datasheets are accessible through the data cleaning log.

Analyse

How do annual members and casual riders use Cyclistic bikes differently?

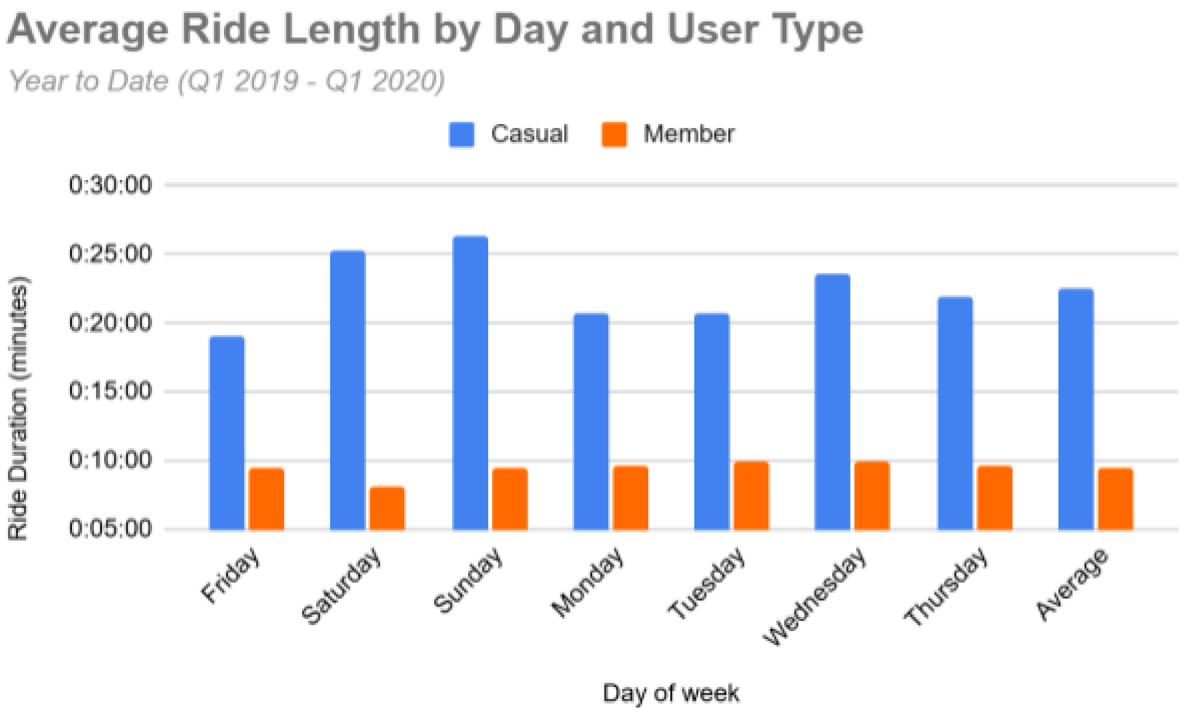

Figure 1

The first chart (*Figure 1*) shows the **average ride duration by day of the week** for each user type. A distinct difference is immediately noticeable: **casual riders tend to have significantly longer rides**, especially on weekends. For instance, the longest average ride duration for casual users occurs on Sundays (00:26:17), while for annual members, the highest is on Wednesdays (00:09:55).

Casual users also show **greater variation in ride duration** across the week, with a difference of 7 minutes and 12 seconds between the longest and shortest days. In contrast, annual members are much more consistent, with only a 1 minute 44 second difference — again suggesting more routine use. These patterns imply that **annual members are likely using bikes for commuting**, while **casual users may be using them for leisure or recreational purposes.**

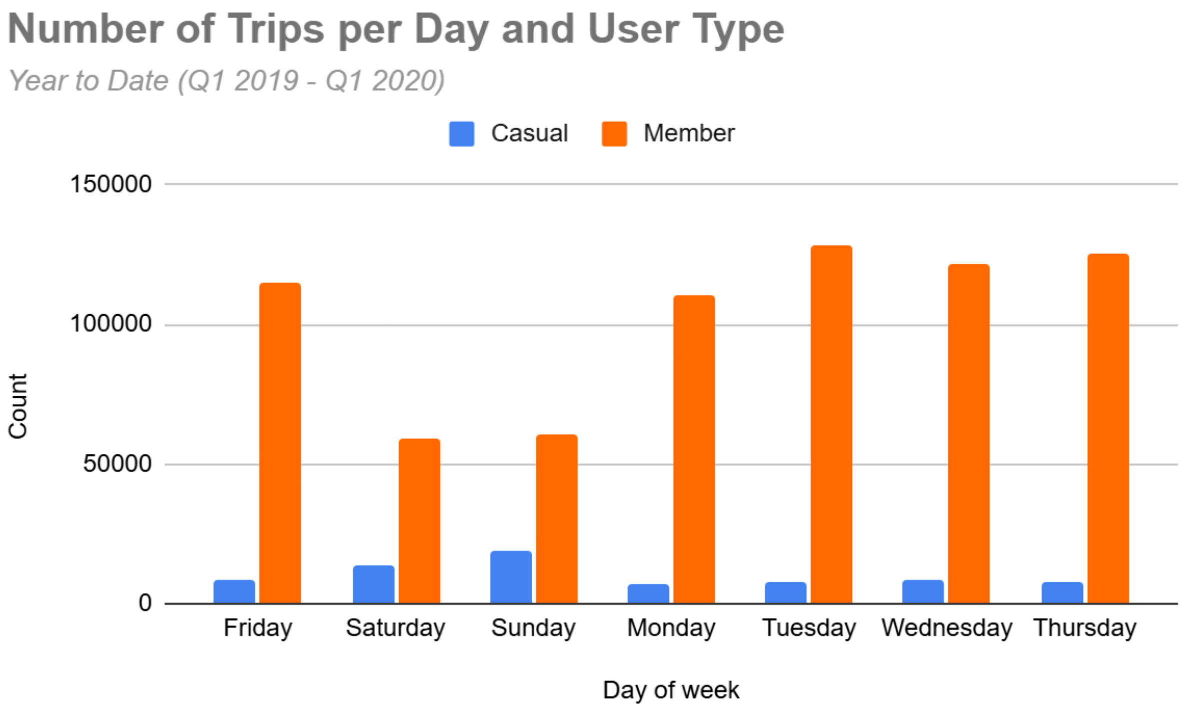

Figure 2

The second chart (*Figure 2*) explores the **number of trips per day of the week by user type.** While this visualization highlights weekly usage trends well, it's important to acknowledge a limitation in the dataset: there is a significant class imbalance, with approximately **400,000 records for annual members and only around 80,000 for casual users**.

As a result, **raw trip counts are not directly comparable** between groups. However, the shape of each group's usage across the week remains meaningful. Notably, annual members show peak usage during weekdays — particularly Monday through Friday — which aligns with commuting patterns. Meanwhile, casual users peak on Saturdays and Sundays, supporting the earlier conclusion that they are more likely using the service for leisure.

While absolute numbers differ due to the imbalance, the **relative patterns** over the days of the week provide valuable insights. In future analyses, this imbalance could be addressed through **normalization techniques** or **sampling** for even more direct comparisons.

Act

In summary, Cyclistic’s annual members and casual riders show clear behavioral differences. Casual riders take **longer trips, especially on weekends**, suggesting a leisure-focused usage pattern. Annual members take **shorter, more consistent rides throughout the workweek**, likely indicating routine commuting behavior. These insights — along with the recognition of data imbalance — provide valuable direction for targeted marketing and user engagement strategies in the next phase.

The business can apply these insights by tailoring services and features to the distinct behaviors of casual and annual riders. For example, Cyclistic could optimize bike availability based on peak usage times (e.g., more bikes in leisure areas on weekends), develop user-type-specific pricing plans (commuter vs. leisure), and create targeted in-app experiences. These strategies would help increase user satisfaction, encourage more frequent usage, and support casual-to-member conversion goals.

Recommendation 1

Weekend Pass Promotion to Encourage Casual Rider Conversion

Analysis indicates that casual riders demonstrate peak activity on weekends and tend to have longer trip durations, suggesting a predominantly leisure-oriented usage pattern. Introduce targeted weekend pass offers, bundled ride packages, or loyalty incentives for casual users who ride on Saturdays and Sundays. These initiatives aim to encourage repeat usage among casual riders and increase the likelihood of converting them into annual members.

Recommendation 2

Promote Memberships to Commuters Through Workplace Partnerships

Annual members exhibit consistent usage throughout the workweek, indicating that many are using the service for commuting purposes. Therefore, establishing partnerships with local employers and public transit providers to position Cyclistic membership as a cost-effective and sustainable commuting option is beneficial. This could include corporate membership discounts, workplace cycling incentive schemes, or integrated transit-bike pass offerings to encourage new annual memberships.

Recommendation 3

Optimise Bike Distribution to Align with Demand Patterns

Usage trends reveal that annual members primarily ride during weekdays, while casual riders are most active on weekends. This indicates differing demand patterns between user groups. To increase use of Cyclistic we could adjust bike redistribution schedules and station stocking levels to ensure optimal availability based on these patterns. For example, increase bike availability at commuter hubs and transit-adjacent stations during weekday mornings and evenings, while reallocating more bikes to parks, recreational areas, and tourist locations on weekends. This operational adjustment would enhance service reliability and user satisfaction for both rider types.

This report has identified clear behavioural differences between Cyclistic's annual members and casual riders, and proposed actionable strategies to address these patterns. By implementing the recommendations outlined, Cyclistic can optimise operations, enhance customer satisfaction, and drive membership growth. Ongoing monitoring and data collection will ensure that strategies remain effective and responsive to changing rider behaviours.