**Project Description: Cyclistic Bike-Share Analysis**

**Introduction**

As a junior data analyst on the marketing analyst team at Cyclistic, a bike-share company in Chicago, I was tasked with understanding how casual riders and annual members use Cyclistic bikes differently. The goal is to design a marketing strategy to convert casual riders into annual members, as annual members are more profitable for the company.

**Phases of Data Analysis**

**Phase 1: Ask**

**Business Task:**

To identify the differences in usage patterns between annual members and casual riders to help design targeted marketing strategies for converting casual riders into annual members.

**Guiding Questions:**

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?

**Key Stakeholders**:

- Cyclistic Marketing Team

- Cyclistic Executive Team

**Phase 2: Prepare**

**Data Location and Organization**

1. **Where is your data located?**

I am going to use datasets that are in CSV files, specifically `Divvy\_Trips\_2020\_Q1.csv`, `Divvy\_Trips\_2019\_Q1.csv`, `divvy\_202004.csv`, and `divvy\_202104` for the analysis purposes.

**2. How is the data organized?**

The data is organized into rows and columns, with each row representing a single bike trip. The columns include details such as ride ID, bike type, start and end times, start and end station names, coordinates, user type (member or casual), and additional attributes in the 2019 dataset like gender and birth year.

**Data Quality and Preparation**

**3. Are there issues with bias or credibility in this data?** **Does your data ROCCC (Reliable, Original, Comprehensive, Current, Consistent)?**

**- Reliable:** The data appears to be reliable as it is sourced from a bike-sharing company and follows a structured format.

- **Original:** The data is original, as it is directly collected from the bike-share system.

- **Comprehensive**: The data covers a range of attributes needed for analysis, such as trip duration, start and end times, and user types.

- **Current**: The data includes records from Q1 of 2019 and 2020, also other datasets contain 202004 and 202104, which is relevant to the analysis period.

- **Consistent**: There may be some inconsistencies between the datasets (e.g., different column names). These need to be standardized during the data cleaning process.

**4. How are you addressing licensing, privacy, security, and accessibility?**

- **Licensing**: The data is publicly available and shared under the appropriate licenses from Motivate International Inc.

- **Privacy**: The data does not include personally identifiable information (PII) such as credit card numbers.

- **Security**: The data is securely stored and accessed in a controlled environment.

- **Accessibility**: The data is accessible in CSV format, which is widely supported by data analysis tools.

**5. How did you verify the data’s integrity?**

- Checked for missing or null values.

- Ensured that the date and time formats are consistent.

- Verified that trip durations are logical

For this verification process, I used SQL to show the problems of the data, which are missing value, column names inconsistencies, duplicate rows, invalid date, and negative ride length on the datasets Q1 of 2020 and 2019. I took action to correct them.

I used Excel Sheet to examine the divvy\_202004, and divvy\_202005 datasets, and correct duplicate rows, null values, and negative ride length.

**The SQL verification file is attached as clean.sql in the portfolio.**

**6. How does it help you answer your question?**

The data provides detailed trip information, which allows for the analysis of usage patterns between casual riders and annual members. This helps in understanding how each group uses the bike-share service differently, which is crucial for designing targeted marketing strategies.

**Phase 3: Data Processing**

**Data Cleaning Steps:**

- Check for and handle missing values.

- Standardize date formats.

- Create new columns for ride length and day of the week.

- Remove any irrelevant or duplicate entries.

For this verification process, I used SQL to show the problems of the data, which are missing value, column names inconsistencies, duplicate rows, invalid date, and negative ride length on the datasets Q1 of 2020 and 2019. I took action to correct them.

**You can find the detailed calculation under documentation for clean.sql.**

I used **Excel Sheet** to examine the divvy\_202004, and divvy\_202005 datasets, and correct duplicate rows, null values, and negative ride length.

**Documentation**:

/Data/

/Raw/

/Processed/

/SQL\_Files /

/Rcode /

/Tableau/

The data structure is as above. The documentation gives the raw datasets and processed datasets for further analysis.

In this phrase for the practice purpose, I used SQL and R programming did the same process twice and got the same results. Please see /Rcode for **RCode/Divvy\_Trips\_2019\_Q1\_process.R** and **RCode/2019\_2020\_Q1\_combine\_process.R**.

**Phase 4. Analyze**

**Data Analysis**:

I used R programming and SQL did the following analysis:

- Calculate descriptive statistics such as mean, median, and mode of ride length.

**This calculation is in descriptive analysis file: analysis.sql**

- Identify trends and patterns in usage by day of the week.

- Compare usage patterns between casual riders and annual members and visualizations.

I used R programming to visualize above tasks:

This is my Kaggle Notebook portfolio:

<https://www.kaggle.com/code/jake989/notebook4bcab63bc4>.

**Key Findings**:

* Members account for a higher proportion of total rides.
* This indicates that members use the service more frequently, likely due to their subscription.
* Members predominantly use bikes during commuting hours (morning and evening).
* Casual riders use bikes more evenly throughout the day, indicating leisure use.
* Members primarily use bikes on weekdays, suggesting commuting patterns.
* Casual riders use bikes more on weekends, indicating leisure activities.
* Members start rides near residential areas and transport hubs, indicating commuting use.
* Casual riders start rides at parks and tourist attractions, indicating recreational use.

**Phase 5. Share**

**Visualization Tools**:

- Tableau for creation of the visual dashboard.

Data source is from R programming combined files of **2019\_2020\_Q1\_combine\_process.R**

This is my Tableau portfolio for Cyclistic bike analysis.

<https://public.tableau.com/app/profile/jake.ma2236/vizzes>.

**Phase 6. Act**

**Recommendations**:

1. Targeted Promotions and Discounts at popular stations Offer Limited-Time Discounts: Provide discounts on annual memberships during peak riding seasons (e.g., summer). Offer special discounts for casual riders who have taken a certain number of rides in a month. Referral Programs: Implement a referral program where existing members can refer casual riders and receive rewards.
2. Personalized Communication Email Campaigns: Send personalized emails to casual riders highlighting their ride statistics and the benefits of becoming an annual member. Include testimonials from satisfied annual members. App Notifications: Use app notifications to remind frequent casual riders about the cost savings of an annual membership.
3. Enhanced User Experience Flexible Membership Plans for Peak hour usage. Introduce flexible membership plans such as quarterly or semi-annual memberships. Provide family or group membership options. Exclusive Benefits: Offer exclusive benefits to annual members such as priority access to new bikes, members-only events, and discounts at partner businesses.
4. Data-Driven Incentives Ride Credits: Provide ride credits to casual riders for every ride they take. These credits can be applied toward the purchase of an annual membership. Usage-Based Rewards: Implement a rewards program where casual riders earn points for each ride. Points can be redeemed for discounts on annual membership or other rewards.