**Case Study Roadmap**

A bike-sharing program swiftly navigates the path to success.

* **Ask**
  + **Guiding Questions:**
    - **What is the problem I’m trying to solve?**
      * In what ways do annual members and casual members (those with single ride or full-day passes) utilize our service differently?
      * How can we encourage casual members to transition into becoming annual members?
    - **How can my insights drive business decisions?**
      * Upon delivering well-supported and compelling data insights along with professional data visualization, I anticipate securing approval for my recommendation from the Cyclistic executive team. Following this approval, the Cyclistic marketing team is poised to implement fresh marketing strategies specifically tailored to target casual members, with the goal of enticing them to transition into becoming annual members.
  + **Key Tasks:**
    - **Identify the business task.**
      * Analyze and delineate the distinct patterns of Cyclistic bike utilization between annual members and casual riders.
    - **Consider the key stakeholders.**
      * Furnish the Cyclistic executive team with detailed data insights and impactful visualizations, offering strategic recommendations aligned with organizational goals. This task is crucial for informing key decision-makers and steering Cyclistic toward success.
  + **Deliverable:**
    - Clear statement of business task.
      * Conduct a comprehensive analysis to distinguish and outline the unique patterns of Cyclistic bike utilization among annual members and casual riders.
* **Prepare**
  + **Guiding Questions:**
    - **Where is your data located?**
      * The data is hosted on a public index accessible for public viewing and downloading; however, direct editing or uploading to the index is restricted.
    - **How is the data organized?**
      * The data is structured to incorporate essential attributes crucial for addressing the business task, encompassing ride details such as time, date, locations, and bike type. Each ride is represented with comprehensive and detailed information.
    - **Are there issues with bias or credibility in this data? Does your data ROCCC?**
      * Certainly, the data exhibits reliability, as it is accurate, complete, and free from bias. The originality is inherent, given that it is automatically gathered by the company through user interactions with the application. The data is comprehensive, sufficiently addressing our inquiries, and it's current as of 2023. Furthermore, it is cited, considering it originates from automatic collection during each ride by the company servers.
    - **How did you verify the data’s integrity?**
      * I ensured the integrity of the data through a multi-faceted approach. Firstly, the data originates from an internal source, collected within our company. Secondly, measures were taken to prevent bias, ensuring that it does not favor any particular group. Although some may contend that the representation of member types is not equal, it's important to note that the focus lies not in comparing their counts but rather in understanding the predominant behaviors and patterns among them.
    - **How does it help you answer your question?**
      * It aids me in discerning distinctions between annual and casual members by examining the locations where they predominantly initiate or conclude their rides, the times at which their rides commence or conclude, the duration of their rides, and the specific dates on which they initiate their rides.
    - **Are there any problems with the data?**
      * Regrettably, there are issues with the data. Approximately 22.02% of the records contain null values for either the start station, end station, or both, posing a significant challenge for handling the data effectively.
      * Additionally, a substantial number of rides commence and conclude within the same minute and at the same station, prompting the speculation that these instances may represent incomplete rides, possibly canceled ones.
      * Moreover, there is a higher representation of annual members compared to casual members in the data. While not ideal, this discrepancy in counts should not pose a major issue since our focus is on comparing behavioral patterns rather than numerical counts.
  + **Key Tasks:**
    - **Download data and store it appropriately.**
      * All necessary data has been downloaded and stored in a dedicated folder, encapsulated within the project directory.
    - **Identify how it’s organized.**
      * It's structured to incorporate ride ID and type, ride time and date, ride locations, and ride member type, all assigned with the appropriate data types.
    - **Sort and filter the data.**
      * Sorting the data may not currently provide significant value since it can be addressed during the process of querying the data.
    - **Determine the credibility of the data.**
      * There is no cause for distrust in the data, given that it is automatically gathered through customer interactions with the application, and its origin is internal, not reliant on human collection.
  + **Deliverable:**
    - **A description of all data sources used.**
      * It is available as an open source, allowing the public to access and download the data online. I specifically selected the data for the first quarter of the year 2023.
* **Process**
  + **Guiding Questions:**
    - **What tools are you choosing and why?**
      * PostgreSQL, I opted for SQL due to the file's size, which exceeded the capacity of spreadsheets at approximately 650 thousand records.
      * Tableau, For visualization.
    - **Have you ensured your data’s integrity?**
      * The data originates internally, and the table was established in the PostgreSQL server. Subsequently, the data was imported into the table, ensuring the correct data types, elimination of null values, adherence to the specified time period (1st quarter of 2023), and absence of duplicates or misspelled words.
    - **What steps have you taken to ensure that your data is clean?**
      * Removal of Null Values: A total of 141,133 records with null values, particularly in the start or end station, were deleted.
      * Attempted Grouping of Stations: An effort was made to create a new column (areas) grouping stations based on longitude and latitude. However, the accuracy was compromised, and some records had null values for longitude and latitude. Despite multiple attempts, this approach proved unsuccessful.
      * Current Record Count: The cleaned dataset, totaling 498,291 records, is deemed sufficient for analysis. The development team is advised to investigate and address data gaps, potentially leveraging longitude and latitude information.
      * Removal of Zero-Second Rides: Rides with a duration of 0 seconds, indicative of potential anomalies like bugs or immediate cancellations, were identified (13,421 records) and subsequently deleted, current total of records is 484,870.
    - **How can you verify that your data is clean and ready to analyze?**
      * Data cleanliness and analysis readiness are verified through a comprehensive checklist covering errors, null data, duplicates, and adherence to business logic.
    - **Have you documented your cleaning process so you can review and share those results?**
      * Certainly, all SQL code responsible for modifying the data is meticulously documented within the query script, accompanied by comments.
  + **Key Tasks:**
    - **Check the data for errors.**
      * Found 13,421 records with ride length less than 1 minute.
    - **Choose your tools.**
      * I chose SQL and Tableau, as the data was too heavy for spreadsheets when I tried.
    - **Transform the data so you can work with it effectively.**
      * Data was cleansed, validated and in ready-to-use format.
    - **Document the cleaning process.**
      * I wrote comments in the query file to explain each query used, and additionally, all visualizations created in Tableau have corresponding SQL documentation.
  + **Deliverable:**
    - **Documentation of any cleaning or manipulation of data.**
      * I provided comments within the query tool to explain each query utilized.