# **Metadata Report: Google Cyclistic Capstone Project - Analytical Process Documentation**

This report details the thought process, methodologies, and decisions made during the "Ask" and "Prepare" phases of the Cyclistic bike-share case study. It aims to provide transparent documentation of the analytical journey, demonstrating adherence to best practices in data analysis.

## **Ask Phase: Defining the Business Problem & Objectives**

The "Ask" phase is crucial for establishing a clear understanding of the project's purpose, identifying key stakeholders, and formulating precise questions that the analysis aims to answer. This ensures that the analytical efforts are directly aligned with business goals.

### **Business Task/Problem Statement**

Cyclistic, a bike-share company, faces a critical business challenge: while their service is widely used, their financial analysts have determined that **annual members are significantly more profitable than casual riders**. Despite this, the number of annual memberships has remained stagnant. The core business task is to **maximize the number of annual members**. This project specifically focuses on understanding and leveraging the existing engagement of casual riders to influence their conversion into annual members.

### **Key Stakeholders**

Successful analytical projects require clear communication and alignment with relevant stakeholders. For this project, the primary stakeholders include:

* **Lily Moreno:** As the Director of Marketing, Ms. Moreno is the direct manager and the key recipient of the analytical insights. She is responsible for developing and implementing marketing strategies, making the findings directly actionable for her department.
* **Cyclistic Executive Team:** The ultimate decision-makers who will review the proposed marketing strategies. Their approval is essential for allocating resources and launching new initiatives.
* **Cyclistic Marketing Analytics Team (Our Team):** Responsible for the entire data analysis pipeline, from data collection and cleaning to generating insights and providing actionable recommendations.

### **Guiding Questions for Analysis**

To address the business task effectively, the following guiding questions were formulated to direct the analytical exploration:

* How do annual members and casual riders use Cyclistic bikes differently? (This is the foundational analytical question, driving all subsequent investigations into usage patterns, duration, time, and location.)
* Why would casual riders choose to purchase Cyclistic annual memberships? (This question focuses on identifying the value proposition of membership from the casual rider's perspective, directly informed by usage differences.)
* How can Cyclistic use digital media to effectively influence casual riders to become members? (This question guides the development of specific, channel-oriented marketing recommendations.)

### **Key Deliverables (from the "Ask" phase)**

The outputs of the Ask phase include:

* A precisely defined business task.
* A clear identification of all relevant stakeholders.
* A set of well-defined guiding questions to steer the data analysis.

## **Prepare Phase: Data Collection & Credibility**

The "Prepare" phase focuses on identifying, collecting, and assessing the credibility of the data to ensure its suitability for the analysis. This critical step lays the groundwork for robust and reliable insights.

### **Data Source**

* **Provider:** Motivate International Inc., responsible for providing the public Divvy bike-share data.
* **Availability:** The dataset is publicly accessible through Cyclistic's website, reflecting a commitment to transparency and open data.
* **Timeframe:** The analysis utilizes monthly trip data spanning from **January 2025** to **June 2025**. This period provides a recent and comprehensive snapshot of user behavior.
* **File Format:** The data for each month is provided in individual **.csv** (Comma Separated Values) files.

### **Data Credibility (ROCCC Analysis)**

The credibility of the data was assessed using the ROCCC framework (Reliable, Original, Comprehensive, Current, Cited) to ensure its fitness for purpose:

* **Reliable:** The data is collected consistently across all months, maintaining a uniform structure and data types. It originates from Motivate International Inc., a well-established and trusted entity in the bike-share industry, lending significant reliability to the dataset.
* **Original:** The data represents raw trip logs directly generated by the Cyclistic bike-share system. This direct collection ensures its originality, as it has not undergone significant third-party aggregation or modification prior to acquisition.
* **Comprehensive:** The dataset includes essential details for each ride, such as unique ride IDs, bike types (rideable\_type), precise timestamps (started\_at, ended\_at), station information (start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id), geographical coordinates (start\_lat, start\_lng, end\_lat, end\_lng), and crucially, the user type (member\_casual). Covering a six-month period, it offers sufficient depth and breadth to observe trends and differentiate user behaviors.
* **Current:** The data set spans up to June 2025, ensuring that the analysis is based on recent user behavior and current operational patterns of Cyclistic. This timeliness is vital for developing relevant and actionable marketing strategies.
* **Cited:** The data source is clearly attributed to Motivate International Inc., affirming its provenance and allowing for traceability and verification.

### **Data Limitations & Considerations**

During the preparation phase, potential limitations and necessary considerations for the data were identified:

* **Anonymized Data:** The dataset is anonymized, meaning personal identifiable information (PII) is excluded. While beneficial for privacy, this prevents tracking individual user progression (e.g., a specific casual rider becoming a member) over time.
* **Data Volume:** The cumulative size of the monthly .csv files is substantial, necessitating the use of specialized tools capable of handling large datasets efficiently, rather than traditional spreadsheet software.
* **Potential for Outliers and Errors:** As with any large-scale operational data, the presence of outliers and data errors was anticipated. This includes instances of zero or negative ride durations, rides with missing station names, or incorrect geographical coordinates. These anomalies require dedicated processing steps to ensure data quality.
* **No External Data Integration:** The analysis is confined to the provided trip data. The absence of external datasets (e.g., demographic information, real-time weather data, historical marketing campaign performance data) limits the scope for broader contextual analysis or multi-variate modeling beyond internal trip patterns.

### **Data Storage**

All downloaded monthly .csv files were organized and stored locally within a structured directory system (e.g., a master project folder with subfolders for raw and processed data) to ensure easy access and maintain data integrity throughout the project lifecycle.

## **Process Phase: Data Cleaning & Transformation**

The "Process" phase involves meticulously cleaning, transforming, and preparing the raw data for analysis. This ensures data accuracy, consistency, and suitability for the analytical tools and methods. A robust processing pipeline is critical for generating reliable insights.

### **Tools Used & Rationale**

A combination of powerful tools was leveraged to handle the large dataset and perform the necessary cleaning and transformation:

* **Python (Pandas Library & geopy):** Utilized for initial data loading, inspection, concatenation of monthly files, and performing large-scale data manipulation and cleaning operations.
  + **Pandas:** The primary library for efficient DataFrame operations, including merging, type conversions, null handling, and filtering.
  + **geopy:** Specifically used for accurate geographical distance calculations (Haversine formula) between coordinates.
  + **Rationale:** Python's flexibility, extensive data manipulation capabilities with Pandas, and specialized libraries like geopy are ideal for robust and custom data processing tasks on large datasets.
* **SQL Server (via pyodbc and SQLAlchemy):** Employed as a robust intermediate data repository and for efficient storage of the cleaned and transformed dataset.
  + **Rationale:** SQL Server provides a scalable and reliable environment for storing the large processed dataset, making it easily accessible for subsequent analysis (e.g., direct connection from Power BI) and ensuring data integrity. pyodbc and SQLAlchemy facilitate seamless programmatic interaction between Python and SQL Server.
* **Microsoft Power BI Desktop (Power Query Editor & DAX):** The final tool in the process for further data modeling, creating advanced calculated columns and measures, and ultimately for visualization and analysis.
  + **Power Query Editor:** While initial heavy lifting was done in Python, Power BI's Power Query Editor could be used for minor final adjustments or verification if needed, and is the primary interface for loading the pre-processed data.
  + **DAX (Data Analysis Expressions):** Utilized for creating advanced calculated columns (e.g., time-based attributes, custom bins for durations and distances) and explicit measures (e.g., Total Rides, Average Trip Duration) that are essential for dynamic analysis and visualization within Power BI.
  + **Rationale:** Power BI's integrated environment streamlines the workflow from cleaned data to interactive dashboards, making it the ideal choice for the final analytical and sharing phases.

The combined approach of Python for heavy-duty initial processing, SQL for robust storage, and Power BI for final modeling and visualization offered the best balance of scalability, efficiency, and analytical flexibility for a project of this scope.

### **Step-by-Step Data Cleaning & Transformation**

The following detailed steps were executed to process the raw data and prepare it for the analysis phase, primarily using Python and then storing the result in SQL Server:

#### **1. Data Acquisition and Initial Consolidation:**

* The project began by identifying all .csv files within the Project\_Rawdata/ directory, specifically 202501-divvy-tripdata.csv through 202506-divvy-tripdata.csv.
* Each monthly .csv file was read into a Pandas DataFrame using pd.read\_csv().
* All individual DataFrames were then **concatenated** into a single, comprehensive DataFrame named combined\_df using pd.concat(all\_dfs, ignore\_index=True).
* The initial shape of the combined DataFrame was (2141425, 13), confirming all six months of data were successfully consolidated.

#### **2. Initial Missing Values & Data Type Overview:**

* An initial check for missing values using combined\_df.isnull().sum() revealed a significant number of nulls in station-related columns (start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id), and a small number in end\_lat and end\_lng.
  + start\_station\_name and start\_station\_id: ~20.06% missing.
  + end\_station\_name and end\_station\_id: ~20.88% missing.
  + end\_lat and end\_lng: ~0.10% missing.
* The ride\_id, rideable\_type, started\_at, ended\_at, and member\_casual columns were confirmed to have no missing values at this stage.
* Initial data types showed started\_at and ended\_at as object (string), and start\_lat, start\_lng, end\_lat, end\_lng as float64. Other textual columns were object.

#### **3. Handling Missing and Invalid Data (Early Stage):**

* A copy of the combined DataFrame (df\_cleaned) was created to ensure data integrity during cleaning operations.
* **Missing Coordinates:** Rows with null values in end\_lat or end\_lng were directly removed using dropna(subset=['end\_lat', 'end\_lng'], inplace=True). This was a deliberate choice to ensure accuracy for distance-based analysis, as the number of affected rows was minor (2187 rows out of 2.1M).
* **Missing Station Information:** For columns start\_station\_name, start\_station\_id, end\_station\_name, and end\_station\_id, null values were filled with the string 'UNKNOWN'. This approach was chosen to retain the ride data even if station information was missing, as dropping ~20% of rows would result in significant data loss.

#### **4. Data Type Conversion & Optimization:**

* started\_at and ended\_at columns were converted from object to datetime64[ns] using pd.to\_datetime(errors='coerce'). No new nulls were introduced by this conversion, indicating clean date strings.
* rideable\_type and member\_casual columns were converted to category dtype using astype('category') for memory efficiency and optimized operations on categorical data.
* Other object columns (ride\_id, start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id) were converted to string dtype for clearer representation and compatibility.

#### **5. Duplicate Handling:**

* A check for duplicate ride\_id values was performed using df\_cleaned.duplicated(subset=['ride\_id'], keep=False).
* No duplicate ride\_id values were found, confirming the uniqueness of each ride record based on its primary key.

#### **6. Outlier Handling & Feature Engineering - Trip Duration:**

* A new column, trip\_duration\_mins, was calculated by taking the difference between ended\_at and started\_at and converting it to total minutes: (df\_cleaned['ended\_at'] - df\_cleaned['started\_at']).dt.total\_seconds() / 60.
* **Outlier Removal for Duration:**
  + Rows with trip\_duration\_mins less than or equal to 0 were removed, as these represent invalid ride durations.
  + Rows where trip\_duration\_mins exceeded 1440 minutes (24 hours) were removed. This threshold was applied to filter out exceptionally long rides, which often indicate system errors (e.g., forgotten bikes, bikes not properly docked) and would skew average duration calculations.
  + This step removed 52106 rides, reducing the DataFrame shape from (2139238, 13) to (2087132, 14).
* The trip\_duration\_mins values were rounded to one decimal place for better readability using .round(1).

#### **7. Outlier Handling - Geographical Bounds:**

* To ensure rides were within the Chicago area and remove any erroneous far-flung GPS readings, geographical bounds were defined:
  + chicago\_lat\_min = 41.5, chicago\_lat\_max = 42.5
  + chicago\_lng\_min = -88.0, chicago\_lng\_max = -87.0
* Rides where either start\_lat/start\_lng or end\_lat/end\_lng fell outside these bounds were removed.
* This step removed only 2 rides, indicating the coordinates were largely within the expected range after initial null handling. The DataFrame shape became (2087130, 14).

#### **8. Outlier Handling - Categorical Consistency:**

* The unique values and their counts for rideable\_type and member\_casual columns were inspected using value\_counts().
* Confirmed that rideable\_type only contained 'electric\_bike' and 'classic\_bike', and member\_casual only contained 'member' and 'casual'. No further cleaning was required for these categorical columns.

#### **9. Feature Engineering - Distance Traveled:**

* A new column, distance\_traveled\_km, was calculated using the geopy.distance.geodesic function. This function computes the geodesic distance (shortest distance over the surface of an ellipsoid) between two points on the Earth's surface.
* A custom function calculate\_distance was applied row-wise to ensure robust calculation, handling potential NaN or 0.0 coordinates by returning pd.NA for invalid cases.
* The calculated distances were rounded to two decimal places for readability.
* All 2087130 rides successfully had their distances calculated, with 0 marked as NA after previous coordinate cleaning.
* A quick check confirmed that some rides had a 0.0 km distance, indicating start and end points at the exact same location.

#### **10. Intermediate Data Storage (Parquet):**

* The fully cleaned and transformed DataFrame (df\_cleaned) was saved to a Parquet file named cyclistic\_trips\_cleaned.parquet within the cleaned\_data/ folder.
* Parquet format was chosen for its efficiency in storing large tabular data, especially with mixed data types (including string and category dtypes from Pandas), and its optimized read/write performance for subsequent use in other tools like Power BI or further Python/SQL analysis.

#### **11. Data Ingestion to SQL Server:**

* The final cleaned DataFrame (df\_cleaned), loaded from the Parquet file, was then imported into a Microsoft SQL Server database.
* The connection was established using pyodbc and SQLAlchemy, configured for Windows Authentication (Trusted\_Connection=yes).
* The DataFrame was appended to a table named BikeTrips within the CYCLISTICBIKESHARE database using df\_cleaned.to\_sql(table\_name, con=engine, if\_exists='append', index=False, chunksize=1000).
* All 2087130 rows were successfully loaded into SQL Server, providing a robust and accessible central repository for the cleaned data.

### **C. Documentation of Process**

All data cleaning and transformation steps performed in Python are explicitly documented within the capstone\_cleaning.ipynb Jupyter Notebook. This includes code comments, print statements for intermediate checks (e.g., DataFrame shapes, null counts, unique values), and explicit rationale for outlier removal thresholds. The use of version control for the notebook/scripts ensures traceability of all changes. Power BI's Power Query Editor automatically records applied steps for any transformations performed within Power BI itself, while DAX formulas are self-contained within the Power BI data model.

## **Analyze Phase: Exploring Data & Identifying Trends**

The "Analyze" phase involved deep dives into the cleaned and prepared dataset to uncover patterns, differences, and key insights related to bike usage by casual riders and annual members. This phase directly addressed the guiding questions established during the "Ask" phase.

### **Methodology**

* **Descriptive Analysis:** Fundamental statistical measures (counts, averages, minimums, maximums) were used to summarize the characteristics of different ride attributes.
* **Comparative Analysis:** The core of the analysis involved direct comparisons between "member" and "casual" user types across various dimensions to highlight their distinct behaviors.
* **Visualization-Driven Exploration:** Interactive data visualizations were extensively used to identify trends, outliers, and relationships within the data, making complex patterns easily digestible.

### **Key Analytical Areas**

The following areas of analysis were systematically explored:

* **Overall Usage & Rider Distribution:** Examination of the total number of rides and the proportional split between annual members and casual riders. This provided a foundational understanding of the user base composition.
* **Ride Duration Analysis:** Comparison of average trip durations for both user types, complemented by a detailed analysis of ride duration distributions using specific time bins. This revealed significant differences in how long each group typically uses a bike.
* **Time-Based Usage Analysis (Temporal Patterns):** Investigation into how ride activity varies across different time granularities (e.g., hours of the day, days of the week, months of the year). This helped identify peak usage times and seasonal dependencies for each user segment.
* **Location-Based Analysis:** Identification of the most popular start and end stations for both members and casuals. This provided geographical insights into where each user group typically begins and ends their journeys.
* **Distance Analysis:** Calculation and comparison of average distances traveled by members and casuals, further detailed by distance bins. When combined with duration data, this offered insights into the typical speed or purpose of rides for each group.

### **Tools Used for Analysis**

* **Microsoft Power BI Desktop:** Served as the primary analytical tool. The cleaned data, prepared in Python and stored in SQL Server, was directly imported into Power BI. Power BI's capabilities were leveraged for:
  + Creating a robust data model.
  + Developing advanced DAX measures to calculate key metrics (e.g., total rides, average durations, percentages within bins).
  + Generating interactive visualizations (e.g., bar charts, line charts, pie charts) to represent findings effectively.
  + Allowing for iterative exploration and drilling down into specific data segments.

## **Share Phase: Communicating Insights**

The "Share" phase focuses on effectively communicating the analytical findings to stakeholders in a clear, concise, and actionable manner. The goal is to present complex data insights in an easily understandable format that supports decision-making.

### **Communication Strategy**

* **Target Audience:** Lily Moreno (Director of Marketing) and the Cyclistic Executive Team. The communication needed to be strategic, focusing on the business implications of the findings and directly addressing the casual-to-member conversion goal.
* **Emphasis on Visuals:** Given the complexity and volume of data, visual representations were prioritized to convey patterns and differences quickly and intuitively.
* **Concise Summaries:** Each key finding was accompanied by a brief, impactful summary explaining its significance to the business problem.

### **Key Deliverables for Sharing**

* **Interactive Power BI Dashboard/Report:** This served as the primary and most comprehensive deliverable. It allowed stakeholders to explore the data dynamically, filter by different dimensions, and drill down into details as needed. The dashboard included:
  + **Overview of User Distribution:** Highlighting the current member vs. casual split.
  + **Ride Duration Comparison:** Visuals illustrating average durations and the distribution of trips across various duration bins for both user types.
  + **Temporal Usage Trends:** Line charts showing daily, hourly, and monthly usage patterns, clearly differentiating member and casual behaviors.
  + **Top Station Analysis:** Bar charts displaying the most frequently used start and end stations, emphasizing the distinct geographical footprints of members and casuals.
  + **Distance Travelled Overview:** Visualizations summarizing trip distances and their distribution.
* **Summary Presentation/Report:** A concise document (like the final business report) to provide an executive summary of the findings, key conclusions, and actionable recommendations. This presentation would draw directly from the insights gathered in the Power BI dashboard, ensuring consistency and clarity.

### **Visualizations Utilized**

* **Clustered Bar Charts:** Effective for comparing quantitative measures (e.g., average duration, total rides) across different categories (member vs. casual).
* **Line Charts:** Ideal for showing trends over time (e.g., hourly, daily, monthly ride counts).
* **100% Stacked Column/Bar Charts:** Useful for illustrating the proportion of ride types within different bins (e.g., percentage of casual rides in longer duration bins).
* **Donut Charts:** For a quick overview of proportional distribution (e.g., overall member vs. casual split).

## **Act Phase: Developing Actionable Recommendations**

The "Act" phase is where the analytical insights are translated into concrete, data-driven recommendations. These recommendations directly address the business problem of converting casual riders into annual members, leveraging the identified behavioral differences.

### **Objective of Recommendations**

* To provide strategic, implementable marketing initiatives designed to appeal to casual riders.
* To clearly articulate the value proposition of annual membership based on casual rider usage patterns.
* To support Cyclistic's overarching goal of increasing annual member profitability.

### **Core Principles Guiding Recommendations**

* **Data-Driven:** All recommendations are directly derived from the distinct usage patterns identified in the "Analyze" phase.
* **Targeted:** Strategies are tailored to the specific motivations and behaviors of casual riders.
* **Actionable:** Recommendations are practical and suggest specific approaches or campaigns.

### **Categories of Recommendations**

Based on the comprehensive analysis of member and casual rider behavior, recommendations fall into several key categories:

* **Targeted Marketing Campaigns based on Temporal Patterns:**
  + Leveraging the **weekend and leisure-hour dominance** of casual riders.
  + Tailoring promotions around **seasonal peaks** for casuals (e.g., summer).
  + Developing specific messaging for casuals who *do* ride on weekdays, highlighting potential **commute-related benefits**.
* **Value Proposition Alignment for Different Ride Durations/Purposes:**
  + Creating membership tiers or offers that specifically appeal to casuals who take **longer, more leisurely rides**, emphasizing cost savings over multiple or extended trips.
  + Highlighting the **convenience and efficiency** of membership for shorter, regular trips that casuals might eventually transition to.
* **Location-Based Engagement Strategies:**
  + Developing partnerships or promotions around **popular tourist and recreational stations** frequently used by casual riders.
  + Targeting casuals with membership offers when they end rides at or near locations popular with members, suggesting a potential shift in usage purpose.
* **Loyalty and Incentive Programs:**
  + Considering programs that reward casual riders for accumulating certain ride durations or distances, providing a tangible incentive to explore membership benefits.
  + Offering trial memberships or introductory discounts to encourage casual riders to experience the "member" lifestyle without full commitment.
* **Operational Enhancements Supporting Conversion:**
  + Ensuring bike availability (especially electric bikes) at casual-heavy stations during peak leisure times.
  + Optimizing rebalancing efforts to align with observed usage patterns of both user types.

### **Measuring Success**

* Each recommendation would ideally be accompanied by suggested Key Performance Indicators (KPIs) to measure its effectiveness (e.g., increase in casual-to-member conversion rate, growth in annual memberships from targeted campaigns, reduction in cost per acquisition for new members).