Cyclisitic Case Study
Data Preparation and Cleaning
February 2024

Data Preparation

- Cyclistic's historical bike trip data from 2023 was used for the analysis
- Data for 2023 was available monthly, and all 12 months were used
- The UNION ALL function in SQL was used to combine all 12 tables for 2023 into one table. This query can be found in Attachment 1. After combining the 12 datasets, over 5 million records were stored in a new Bigquery table for cleaning and analysis
- The data used was credible as it came directly from consumer usage, and it was the most recent full-year data that was available
- Data was examined before usage to ensure completeness and richness of information. There were columns with NULL values, but the ones used for the analysis were largely complete
- Bigquery was used for cleaning and analysis, and Excel was used for visualisation
- Three new columns were added in Bigquery: day_of_the_week, month and ride_length
- The query for ride_length, day_of_the_week and month columns can be found in Attachment 2

Data Cleaning

- In cleaning the data, MAX, MIN, LEN, and IS NULL were some of the functions used. These were used to check for nulls in the columns used in the analysis and outliers in the calculation of ride_length. Using the MIN function, negative figures were discovered in

the ride_length column, which meant the ended_at times were earlier than the started_at times, indicating errors in the data entry. The WHERE function was then used to identify the number of negative values which exist (11,734 were found). These were removed, and results were stored in a new Bigquery table

- Cleaning was done on the ride_id column as 510 entries had ride_ids less than 16 characters. It is assumed that ride_ids should be 16 characters long. The query to identify the ride_ids with less than 16 characters can be found in Attachment 8. These were removed using the WHERE clause. Results were stored in a new Bigquery table
- Forty-five entries were associated with the year 2024 and were found in the December 2023 table. It is assumed that the riders activated the bikes late on New Year's Eve and returned them the next morning, which is New Year's Day. These remained in the dataset
- Rideable type and member_casual columns were clean and free from errors