**CYCLISTIC BIKE SHARE DOCUMENTATION**

**ASK**

My team lead has asked me to analyze the datasets to find any differences between how the members and casuals use the bikes.

**PREPARE**

* Download the datasets from the bucket and validate the ROCCC (Reliability, Originality, Comprehensive, Cited, Current)
* Save them as XLSX files
* N.B a lot of empty cells in the station id columns, ride id lengths are inconsistent

**PROCESS**

* I copied the dataset into a new sheet which I renamed “Worksheet” since that’s where I did the cleaning.
* I hid columns E to L since they were not relevant to the cleaning.
* I created an “id\_length” column to observe the length of the ids and deleted any rows which didn’t fall within the 16 characters limit since 16 had the most rows by far, which suggested that was the right number of characters for the cell.
* I created a column to verify that the “started\_at” wasn’t greater than the “ended\_at” with a formula (=IF(C2>D2,”D”,””), I deleted all rows which returned a “D”.
* I created a “ride\_length” column to determine how long each ride lasted (D2-C2).
* I created “weekday” columns for both the start and end dates to see what days of week the rides began and ended.

**ANALYZE**

I proceeded to query the now clean data in Microsoft SQL Server Studio in search of insight that could help the organization make better decisions. See the queries below.

**SHARE**

**ACT**

**202204**

Copy the dataset into another sheet and rename it Worksheet.

Delete the station\_id & long/lat columns since I don’t plan to do any geographical analysis

Create a column to check the length of the ride\_id and filter does which are less or more than 16. Delete the rows for those the filtered data since 16 seems to be the right number of characters for the id. Delete the column afterwards.

Separate the date and time for the start & end of the trips for ease of analysis.

Create a column to find out what weekdays the trips started and ended.

Create a column to find out the length of the ride and then create another column to convert it into seconds by multiplying it by 24\*60\*60

Filtered the start & end station names for any blank cells and delete their rows, filtered the ride\_length column for any discrepancy and removed them.