Cyclistic Case Study Notes

The problems we are trying to solve for this case study:

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?

- The source for this project is historical data from Cyclistic, containing customer traveling and member information.

- The focus is a 12-month period to discover insights and potential trends to drive data-driven decision making.

- Customers' personal identifiable information is unavailable, meaning the analysis will not be able to identify if the customer is within the Cyclistic service area or has purchased multiple single passes.

- Due to thousands of missing data within station start/end columns, we cannot calculate most popular stations accurately.

- I removed multiple columns from dataset due to its irrelevancy of the business task. Although there may be potential insights to discover, it would be a distraction to our main focus.

- I ran into an issue where I could not calculate the minutes in my ride_length column in Power Query due to the format. I utilized generative ai to brainstorm solutions and I was able to solve my problem by transforming the data for it can be calculated. The formula used was "Duration.TotalMinutes([ended_at] - [started_at])"

- After creating a PowerQuery with relevant data, the total rows were over 5 million. Excel cannot handle such large numbers, so I copied the M Code then transferred it to Power BI to connect it with DAX studio. DAX will allow me to export all this data into a .csv file smoothly, which I can transfer to BigQuery to use SQL.

- I attempted to create visuals for the SQL finding with Power BI, but I ran into issues with how the data was formatted after saving from BigQuery. The months and other categories are not being recognized in the correct order it was extracted. Fixing this would require me to write code through Power BI's DAX query.

- Writing the code to organize it properly would likely take longer than using Excel pivot tables. Since I have already extracted the specific insights needed with BigQuery, there are much less rows stored in the files. This will allow me to utilize Excel without pushing its limits.

Data Processing Phase

Power Query, Power BI, DAX Studio

- Downloaded data from source and stored copies of original files in subfolder.

- Uploaded folder with csv files to Power Query for cleaning and transformation.

- Removal of columns because it is irrelevant to business tasks
  1. source.name
  2. ride_id
  3. start_station_name
  4. start_station_id
  5. end_station_name
  6. end_station_id
  7. start_lat
  8. start_lng
  9. end_lat
  10. end_lng

- Added columns to find additional insights
  1. **ride_length** to find minutes bike used (transformed data to be calculated for **minutes_by_group** column).
  2. **minutes_by_group** to know how long customers are likely to ride.
  3. **day_of_week** for insights of the most popular days.
  4. **month** for insights of the most popular months.
  5. **hour_of_day** for insights of the most popular start hour.

- Removed started_at, ended_at and ride_length columns because they are no longer necessary after creating our new columns.

- Transferred M Code of combined data and new columns to Power BI because it connects with DAX Studio, which is needed to convert large file into a csv.

Data Analysis Phase

BigQuery/SQL

- Upload csv file to big query to discover trends and aggregate customer usage.

- **Created and extracted SQL code with answers to many questions:**

- Comparing difference between members and casuals
- Distribution of ride types between customers
- Count of how many member and casual rides by minute groups
- Counting member/casual customers for each month
- What is the most popular day of the week
- What is the most popular day of the week customers
- What is the most popular day of the week for each customer type
- What is the most popular hour of day for customers
- What is the most popular hour of day for each customer type
- Most popular rideable type by month along with its ride count
- How long customers use each ride type within each minutes by group