**How do Annual Members and Casual Riders Use “Cyclistic” Bikes Differently?**

Start: 11/30/2023 at 11:45 am

**Business Question:** How do casual riders use the service differently from annual members?

**Business objective:** Convert more casual riders into annual members.

**Data Source**

* 1st party data from the company accessed from [this link.](https://divvy-tripdata.s3.amazonaws.com/index.html)
* Data is publicly available under [this license.](https://divvybikes.com/data-license-agreement)
* The 2019 full year data will be used. All four quarter zip files containing the .csv files downloaded and unzipped locally.
* Personally identifiable data has already been removed from the datasets by the company.
* Observation: Customers (non-subscribers) seem more likely to skip entering their gender and birth years than subscribers.

**Data Cleaning/Manipulation**

* Tool: Excel and MySQL (see SQL script)
* Saved .csv files as .xlsx workbook copies
* For Q2 sheet, changed the column names to match those of the other 3 quarterly sheets. Also converted datetimes from codes to the conventional format in the other three sheets.
* Added two columns: One detailing the ride duration in hh:mm:ss and the other detailing the day of the week in which the ride started (1= Sunday and 7=Saturday)
* Added column detailing the quarter in case all quarterly sheets need to be combined
* Removed all formatting and reformatted dates, times, and weekdays
* All above steps repeated for each quarterly sheet
* Note: In the Q4 sheet, 13 rows appeared to have the start time and end time switched around (the time for the end time was earlier than that of the start time), so I corrected this.
* Note 2: The times and durations in the Q4 sheet appear to be rounded to the nearest minute.

**Data Analysis**

* Tool: Excel, MySQL (see SQL script)
* Formatted data range as table
* Created Summary overview of data. Summary overview revealed extreme outliers (e.g. ride durations exceeding thousands of hours each), skewing the average ride duration of casual riders (Avg was over an hour long!) vs Annual Members (in the 13-15 min range), ~5:1 ratio. As 98.8% of rides are under 1 hour, it was decided that the cutoff for outliers was any ride 1 hour or more. Many of those long duration outliers were most likely cases where bikes weren’t returned in a timely manner.
* Created duplicate sheet with outliers removed as well as a summary overview of this filtered data. The removal of outliers revealed that the average ride duration of casual riders was 22:54 vs 10:34 of annual members, roughly a 2:1 ratio.
* Eliminated all columns except those indicating the trip id, start and end dates, ride duration, quarter, and user type
* Created pivot charts and plot visualizations comparing casual riders to annual members with regards to a) total rides b) avg ride duration c) avg ride duration per day of the week.
* Analysis revealed that rides by user type were distributed differently throughout the week.
* Added column indicating if ride was on a weekday (M-F) or a weekend (Sat, Sun).
* All above steps repeated for each quarterly sheet.
* Combined all four datasets into a single FY 2019 sheet in SQL.
* Created and exported full years summary statistics tables using SQL.
* Note: The number of rows exceeds 1 million in some of these sheets, making the analysis phase very slow and frustrating at times. I was able to clean and analyze each quarter individually, however all attempts to combine the quarterly datasets into a single annual dataset in excel failed due to the large size. For the annual analysis, there are two options: a) copy and paste only the summary tables for each quarter into a new range/table in a new workbook. Or b) Just do everything in SQL then export to a BI tool such as Tableau.
* Update to note: WARNING- The dataset for Q3 is too large to load completely in Excel. 30% of the data is missing and so Q3 will have to be excluded from the Excel analysis. Q3 and annual cleaning and analysis will have to be done in SQL.

**Data Visualization**

* Tool: Excel, Tableau public
* Created pivot tables in Excel for Q1, Q2, and Q4
* Plotted visualizations of the data of these pivot tables
* Loaded full year summary statistics tables into Tableau
* Converted avg trip duration back to seconds in Tableau in order to be recognized by the software as a measure value.
* Created worksheets in Tableau with visualizations of number of trips and avg ride durations for both user types.
* Combined worksheets into several dashboards
* Combined dashboards into a story.

End: 12/8/2023 at 5 PM