Key to Speedy Success for Cyclistic

Statement of Business Task:
How annual members and casual riders differ?
Data Source:
A publicly available data set, “divvy-tripdata”. I used one year, year 2021, of trip data in my analysis. One year data is sufficient to get actionable insights from the data.
Documentation:
This detailed report, an SQL file with actual queries and a Tableau “Data Story” is available will be ready by the end of this project.

Detailed Analysis Process:

After reading the requirements, I determined three questions.

- Design marketing strategies aimed at converting casual riders into annual members.
- To better understand how annual members and casual riders differ.
- Why casual riders would buy a membership, and how digital media could affect their marketing tactics.

Out of these 3 questions, second one is relevant to data analytics team. “***To better understand how annual members and casual riders differ.***”

For this task, I needed at least 1 year of data, so that I can analyze it with different angles. So, I downloaded data related to year 2021. Before moving forward, it was important to check the completeness and accuracy of the data. I used BigQuery as well as PostgreSQL as tools for preparing, cleaning and analyzing data. I used Tableau to visualize my analysis.

Data Combining

- So, as a first step, I uploaded all 12 CSV files in BigQuery.
- Now this was the time to use my SQL knowledge to prepare, clean and analyze data using SQL queries.
- I checked the data types for all the columns in all 12 tables in BigQuery just to make sure that data types are aligned, because I had to combine these tables into one.
- After making sure data types are identical for all the columns in all tables, I created one separate table by combining all the tables with a final row count of **6,733,219**.
- I created a copy of the combined table and performed data cleaning steps.
- I faced a roadblock here, Sandbox account didn't allow me to delete rows with NULL values, so I had to figure out the way to upload all data to PostgreSQL.
- I successfully created a table with the correct data types in Pg Admin and used the copy command to copy all 12 CSV files to the table.
- Here again, I created a copy of that table and performed cleaning tasks on that copied table. The original table is still in its original form.

Data Cleaning

- After creating a backup table, I started cleaning data. First, I deleted all rows containing NULL values in all columns. It was important to delete these Nulls for the sake of correct analysis and to avoid bias.
- After deleting NULL values, I got 4,588,302 rows of clean data.
- After exploring data, another round of cleaning was conducted. I found some discrepancy where station_name and station_id was same. Total 10 rows with this inconsistent data were removed.
- After further exploration, I found that there were some rows in the data where the difference between end_time and start_time was in negative. It was important to eradicate these rows. So I wrote some queries to further clean the data.
  Note: I further explored this data, and found there were thousands of rows of data where the difference between end time and start time was a couple of seconds and start station name and end station name were the same, I was convinced to delete these rows, but I kept them anyway.

Verification

- As duplicate rows could impact the result of analysis, so I verified data for the duplicated rows in the table. Fortunately, there were no duplicate records.

Getting to know the Data

- Before analysis, it was important to know the actual data. So I wrote some queries to get the knowledge of data.
- Total Unique Rows: 4,588,094
- Rideable types of bikes: "classic_bike", "docked_bike", "electric_bike"
- Start Station Name Count: 839
- End Station Name Count: 837
- Types of Riders: "casual", "member"

Preparing Data for Analysis

As the primary question to answer from this data is "***To better understand how annual members and casual riders differ.***" so I decided to break this question in to smaller pieces. To do this detailed analysis, to show the difference between type of riders, I needed to manipulate this data table. So, I decided to add some columns in the table instead of writing queries again and again, I decided to permanently add frequently used columns.

Data Analysis

To better answer the business question, I broke down the question in smaller steps to show how different users use different types of bikes and when they frequently use it.

1. Number of rides booked (%) per user category during 2021.
2. Number of rides booked per month per user category during 2021.
3. Which day of the week is most popular among different category riders?

4. Which month of the year is the most busy by different category riders?
5. What is the average ride time for different category riders, by year?
6. What is the average ride time for different category riders by month?
7. Top 1,000 longest rides belong to which type of riders?
8. Top 5 stations as a starting point for “member” category riders?
9. Top 5 stations as a starting point for “casual” category riders?
10. What time of the day, member and casual riders start their ride at?
11. Which bike type is most popular among different riders?
12. Bike use by different category riders during working week & Weekend?

Conclusion

After a detailed analysis, keeping in mind the core business task, I have come to the conclusion that:

- We can safely say that the members book more rides as compared to casual riders.
- Percentage of booking rides by members, is more than casual riders, except for July and August. In July and August, booking of casual rides exceed member rides.
- Casual riders book more rides on Saturday and Sunday as compared to members. Members book more rides from Monday to Friday compared to casual riders.
- The month of July is the most popular month for casual riders, whereas July, August & September are most popular among member riders.
- Ride average time of casual riders is more than double of average ride time of members.
- The monthly average ride time of casual riders is always way more than the average monthly ride time of members throughout the year.
- Top 1,000+ longest bike rides belong to casual riders.
- Top 5 start stations of members are very different from the top 5 start stations of “casual” riders.
- Majority of the members start their rides at 6PM. But there is also a spike from 7AM to 9AM. Peak hour for casual riders is also 6PM.
- Classic bike type is most popular among both type of riders. Casual riders use other two types of bikes as well. Members do not use “Docked” bikes.
- Members book more rides than casual riders during the working week, i.e. Monday - Friday, and casual riders book more rides over the weekend as compared to members.

Recommendations

- From the data, we know the most popular days of the week for casual riders are Saturday and Sunday, we should run a marketing campaign one or two days before the weekend highlighting member benefits.
- June-Sep are most popular among casual riders. We should focus these months of the year targeting casual riders.
- We know top starting points by casual riders, we can place billboards type marketing near these stations.
- Longest rides are booked by casual riders, we can run a campaign by highlighting cost and benefits of member subscription.

Data Limitations

- Financial data is not available. I don't know how much revenue comes from members and casual riders.
- Availability of financial data would help me better answer this business task, and it also would help write even better recommendations.