● Business Question

How do annual members and casual riders use Cyclistic bikes differently*?*

Case Study Roadmap - Ask

Guiding questions

● What is the problem you are trying to solve?

● How can your insights drive business decisions?

I must answer the following question: "How do annual members and casual riders use Cyclistic bikes differently?" This will help Cyclistic develop strategies to convert casual riders into annual members.

Although the pricing flexibility helps Cyclistic attract more customers, the marketing director believes that maximizing the number of annual members will be key to future growth.

Deliverable

● A clear statement of the business task

This case study consists in analyzing Cyclistic historical data, to find out patterns and insights that can help improve the annual membership numbers. The final report will be used by the director of marketing (Lily Moreno) and the Cyclist executive team to drive business decisions.

Case Study Roadmap - Prepare

Guiding questions

● Where is your data located?

It is in "https://divvy-tripdata.s3.amazonaws.com/index.html".

● How is the data organized?

The data is contained in .csv files, which are then inside a compressed .zip file. They will be imported into a database.

● Are there issues with bias or credibility in this data? Does your data ROCCC?

The data comes from the company itself, so it has credibility and adhere to the ROCCC principles. As it records both casual and annual members, it does not look biased towards a specific group.

● How are you addressing licensing, privacy, security, and accessibility?

The parent company granted non-exclusive, royalty-free, limited, perpetual license for any lawful purpose. For privacy reasons, we are not allowed to use any personally identifiable information, as so there should not be any problems with data security. Accessibility issues will be dealt in further steps.

● How did you verify the data's integrity?

The data was first uploaded to a database. Then, we checked the records using a wide array of SQL queries.

● How does it help you answer your question?

The data will provide valuable insights into the habits of different types of bicycle users.

● Are there any problems with the data?

We find out that some of the records do not have the gender or the birth year information (about 8% - more details below)

Deliverable

● A description of all data sources used

The dataset that will be used includes information about bicycle trips from the first and last quarters of 2019 (Q2 and Q3 could not be used due to size restrictions). They were imported into a table in a database, containing the following information:

- Start and end time of the trip (MM/DD/YYYY HH:MM format)

- ID of the bicycle used

- Trip duration (in seconds)

- IDs and names of the start and end stations

- User type ("Subscriber", for annual users, and "Customer", for casual)

- Gender and birth year (about 8% of records do not have this information)

Case Study Roadmap - Process

Guiding questions

● What tools are you choosing and why?

Database (Postgres), as the data is relatively big and SQL is the language that I am most proficient at. Initially, BigQuery was the database tool chosen, but since there are restrictions in a free user account, I decided to switch to Postgres, which is open source and I can use locally in my computer.

● Have you ensured your data's integrity?

The following verifications were performed using SQL queries:

- end time always comes after start time: 13 records had an end time that were less than start time. In all of them, the trip duration remains consistent if we add an extra hour to the end time. For example, one record has a start time of 1:43, end time of 1:09 and duration of 1594 seconds (or 26.5 minutes). If we add one hour to the end time (2:09), the difference between 2:09 and 1:43 is 26 minutes, which is consistent with the trip duration.

- null values in gender/birth years, occurring in about 8% of records: since this information doesn't affect trip duration or station destination, the record will be kept, but this might affect the precision of a demographic analysis.

- duplicate records: since trip_id was defined as primary key, there are no duplicate transactions in the database.

● What steps have you taken to ensure that your data is clean?

There were errors when trying to import the first CSV file into Postgres. After some investigation, it was concluded that the trip duration information was causing this error, because of the presence of a comma as a thousand separator. I had to change the formatting of the column to hide it. After that, importing proceed without any issues.

● How can you verify that your data is clean and ready to analyze?

After all the above steps were performed, data is clean and ready.

● Have you documented your cleaning process so you can review and share those results?

This document and the history of SQL queries provide a thorough documentation.

Deliverable

● Documentation of any cleaning or manipulation of data

The procedure will be detailed in a following document.

Case Study Roadmap - Analyze

Guiding questions

● How should you organize your data to perform analysis on it?

The data has been clean, formatted and stored in Postgres DB under table name Trips_Information.

● Has your data been properly formatted?

Yes, all fields are correctly formatted and ready to be analyzed.

● What surprises did you discover in the data?

The trip_duration field caused problems when importing to Postgres because of the thousand separator. This was unexpected, since a trial run in BigQuery did not result in any errors. The end date being inconsistent with the start date in some records was also a surprise.

● What trends or relationships did you find in the data?

By using SQL queries, we find out that:

- Casual members use bikes, on average, for longer periods of time (37 minutes) and most commonly on Sundays. Annual members use bikes for shorter runs (11.5 minutes) and most commonly on Tuesdays.

- Performing more queries, we find out that the number of casual members on weekends is almost twice more than on weekdays. However, the average ride length stays almost constant - weekend rides are less than 10% longer.

- The trend is reversed with annual members: there are almost twice as more users on weekdays than on weekends. But the ride length also stays constant, only about one minute longer on weekends.

● How will these insights help answer your business questions?

The above insights are enough to help answer our business question, as they show a distinct difference in how casual and annual members use Cyclistic.

Deliverable

● A summary of your analysis

The queries' results will be pasted, formatted, and further explained in a following document.

Case Study Roadmap - Share

Guiding questions

● Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?

Yes, the data show clear patterns that will help answer the business question.

● What story does your data tell?

The data shows that casual members use Cyclistic services much more frequently on weekends (Saturday and Sunday) and ride for more than 3 times longer than subscribers (36 min vs 11.5 min). Subscribers tend to favor shorter rides and weekdays use, although they are still significant on weekends. Casual users represent around 5% of the total user base. Male users represent 78% and female 22%. Subscribers tend to be a little bit older than casual users.

● How do your findings relate to your original question?

The findings show a clear picture on how casual and annual members use Cyclistic differently.

● Who is your audience? What is the best way to communicate with them?

The director of marketing and the executive team are my audience. The best way to communicate with them is using high-level material, like clear visualizations and tables.

● Can data visualization help you share your findings?

Yes, charts are a useful tool to present the findings.

● Is your presentation accessible to your audience?

Yes, I used high-contrast colors on the visualizations.

Deliverable

● **Supporting visualizations and key findings**

They are contained in a spreadsheet.

Guiding questions

● What is your final conclusion based on your analysis?

The conclusion is that casual users are much more active in weekends than weekdays (89% variation) and they also ride for long periods of time (average of 36 minutes). Annual users are most of the user base (95%), ride more on weekdays when compared to weekends (49% variation) and prefer shorter trips (average of 11.5 minutes). Male users are the majority regardless of the user type, consisting of 78% of the userbase. Subscriber are a little bit older (DOB 1982-1984) when compared to casual users (DOB 1988-1989).

● How could your team and business apply your insights?

● What next steps would you or your stakeholders take based on your findings?

Described in the deliverable.

● Is there additional data you could use to expand on your findings?

More information on the profile of users: reason to use the service (go to work, leisure, exercise, etc.), frequency of use, etc.

Deliverable

● Your top three recommendations based on your analysis

Based on my analysis, the top three recommendations are:

- *Creation of plans that benefit weekend users and users that ride for longer. This covers most casual users.*
- *Marketing campaign to attract more female users. Both casual and subscribers are composed majority by men; there is a huge market opportunity here and can greatly expand Cyclistic userbase*
- *Customer retention. Although attracting casual members is an important part of business, annual members still are the majority of userbase. Keeping these customers satisfied is key for a successful business.*