Cyclist Case Study - Journal

Botond Koroknai

July 2025

1 Brief structure

Throughout my case study I will follow the recommended framework by the Google Data Analytics course, which consists of 6 steps:

1. Ask
2. Prepare
3. Process
4. Analyze
5. Share
6. Act

The data files I was using gave me a brief summarization about the trips that were taken in 2019/Q1, Q2, Q3 and Q4.

2 Ask

The guiding questions for the ask phase were:

- What is the problem you are trying to solve?
- How can your insights drive business decisions?

To briefly answer the questions:

- Our task is to come up with solutions on how to convert "casual" users into annual subscribers and to reinforce these insights using various data analytical techniques.
- If I can clearly identify underlying trends among annual subscribers and find the motive behind their decision, my "company" can launch a directed marketing campaign to motivate "casual" users to subscribe to the service.

The key tasks in this phase are to:

- Identify the business task
- Consider the stakeholders

Last but not least we need to deliver a clear statement of the business task.

3 Prepare

During this project I have used R, and the following chapter I will describe the steps that I took based on the guiding questions to prepare the data for analysis.

Where is your data located?
I have uploaded the two provided data sets to posit cloud, in order to use the cloud based R-studio.

How is the data organized?
To check the structure of the data I have used the *str()* R function which stands for structure. By executing this code I made sure that the data in each column are stored in a appropriate data type.

Are there issues with bias or credibility in this data?
As this was an internally collected data I would say that it is credible since it is original, cited, current and comprehensive. Also based on my findings I would say that this dataset is unbiased, because it is representative for the whole population which were observed (both "casual" users or subscribers)

How are you addressing licensing, privacy, security, and accessibility?
Privacy: The data set, that I am working with was manipulated previously, therefore it does not hold sensitive information like credit card numbers, or name of the user, which can be directly linked to the customers.
Licensing: Under the license of the company the data can be used as source material, as applicable, in analyses, reports, or studies published or distributed for non-commercial purposes. (*source: https://divvybikes.com/data-license-agreement*)
Security: As security measurements the cloud based RStudio notebook accessibility is set to private, therefore only I can view or edit the files which were used for the analysis.
Acessibility: If needed, thanks to the cloud based service the data can be shared to trusted users, which allows cooperation.

How did you verify the data's integrity?
It is the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle. Completeness: As first step I have checked for NA values using the *sum(is.na(data))* function. This step gave me a valuable insight on one of the data set which has 18023 NA values! After I inspected the table with the *View()* function it became clear that the "gender" and "birthyear" columns were the faulty ones, as there is a high chance that these fields were not mandatory to fill out when you register yourself in the system. I have taken note of this problem and will get back with a solution in the **Process** phase of the analysis.
Accuracy: As we are still working with first party data, I think if we clean the data properly, and we make sure that every column has the proper data types and values, we can call our data accurate.
Trustworthiness: The same is true in this case.

4 Process

In the cleaning process I have made the following changes:

- Corrected the columns with faulty data type using the *as.datatype()* function.
- created a trip duration column as it can be useful in the analyzation step.
- Uploaded the data into google sheet and also performed a trim to remove any unnecessary spaces.
- Removed duplications
- To get a clearer view I have organized the data into ascending order by date.
- Created a column under the name "Weekday" which tells us the name of the day based on the date.
- Removed the rows which had negative trip duration time, since it is clearly corrupted data.
- Removed "trips" which were longer than a week, since that is highly unlikely.

Data integrity: To ensure data integrity I have used this journal to document the changes so that it would be clear to anybody who is working with the cleaned data set. The second safety measurement that I took was that I never modified the raw data directly only the copy of it, so in case I have make a mistake in the cleaning process I can easily reload the raw and complete data.

To make sure that the data is ready to be analyzed I checked again if I still have unwanted duplications, or NA values.

5 Analyze

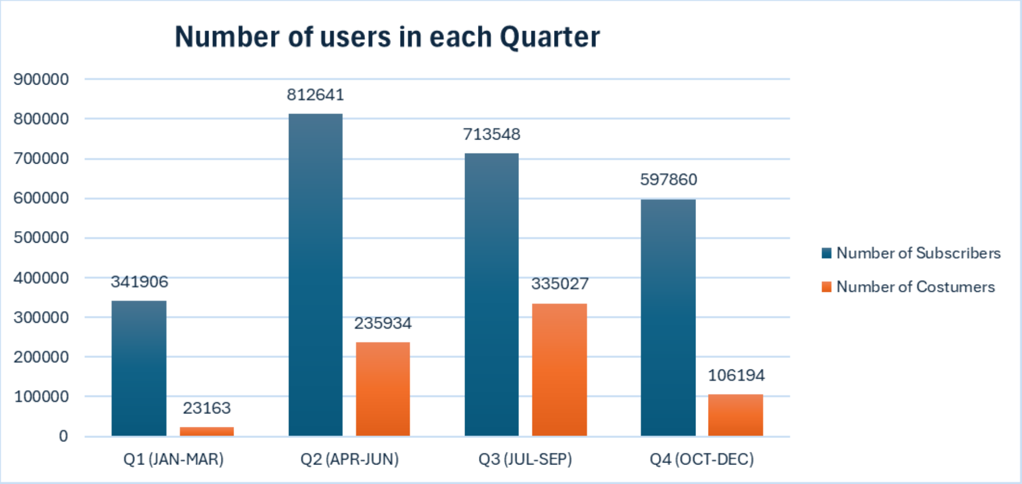
How should you organize the data to perform analysis on it?
My first step was to unify the structure for the 4 data files in order to perform the same analytical steps on them.

Has your data been properly formatted? I have made some minor formatting mistakes, like selecting faulty data types for some columns, but it was quickly corrected using Excel.

What surprises did you discover in the data?
When I first read that the business goal was to attract more people to be subscribers, my first thought was that there are way less subscribers than ”casual” users. In reality there are way more subscribers than users, which made me dig deeper into the details in order to find underlying trends to attract more people to be a regular user of the service.

What trends or relationships did you find in the data?

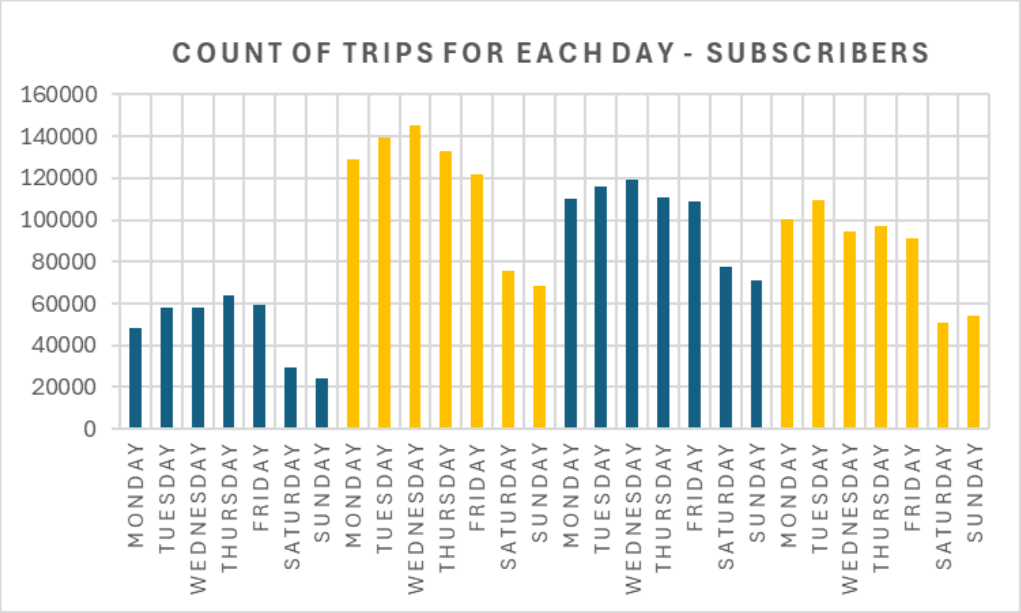
- Both the number of ”casual” users and subscribers are growing when the spring comes, however they peak at different times:



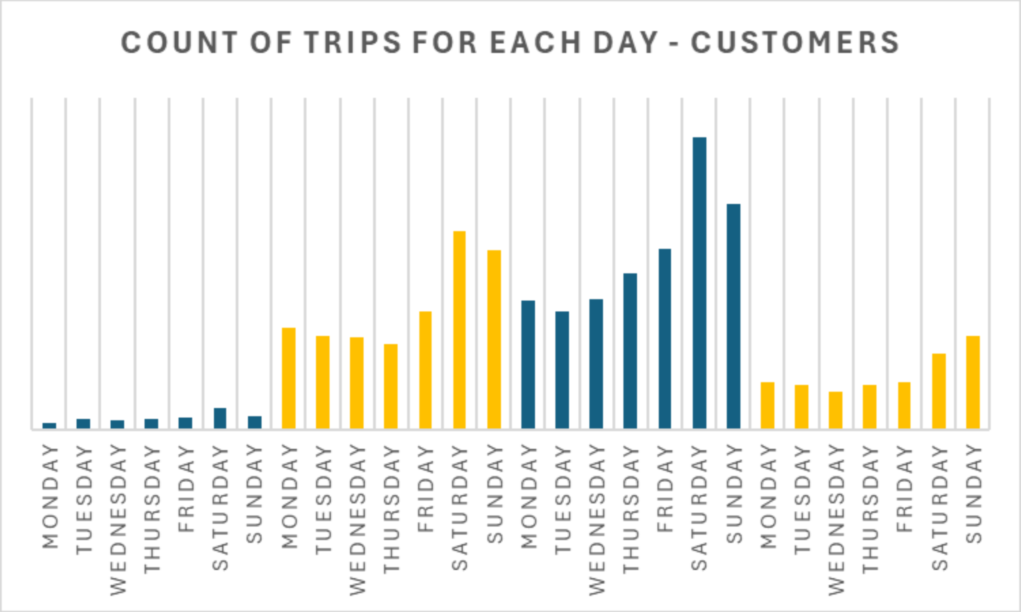
  - The peak of the subscribers come in Q2 (APR-JUN).
  - While the Costumers only reach their max in Q3 (JUL-SEP)

My hypothesis behind this trend is that once the weather is getting warmer both categories start to rise. Once it is summer the ”subscribers” mostly locals go away on holiday, while ”Costumers” mostly tourist come to the City, that is why it reaches its maximum later.

- By counting the number of trips for each day we can clearly see the differences between user types:
  - For the subscribers the majority of trips happens on weekdays, which indicates they are using the service for their everyday life like going to work/school.



  - On the other hand we can clearly see that most of the trips for the ”Costumers” happen in Q2 and Q3 with peaks on Saturday and Sunday which I think happens due to the extra ”residential” use who are not subscribers.



- What I consider also important is the geographical location of the trips: In case of costumers most trips was from / to the most popular attractions around Chicago

Start station	Count of trips
Streeter Dr & Grand Ave	22229
Lake Shore Dr & Monroe St	14445
Lake Shore Dr & North Blvd	9740
Michigan Ave & Oak St	9524
Millennium Park	8526
Shedd Aquarium	6949
Theater on the Lake	6908
Dusable Harbor	4746
Michigan Ave & Washington St	4548
Adler Planetarium	4278
Michigan Ave & 8th St	3847
Montrose Harbor	3703
Indiana Ave & Roosevelt Rd	3166

How will these insights help answer your business questions?
The insights reveal clear behavioral patterns that can inform strategies to convert casual users into subscribers. For example:

- **Seasonal trends** show that casual users peak in summer, suggesting that targeted marketing and subscription incentives during Q2–Q3 could be effective.
- **Day-of-week usage** indicates that subscribers mostly ride on weekdays, likely for commuting, while casual users prefer weekends. This implies that subscribers are typically locals, while casual users may include tourists or occasional riders.
- **Popular start stations** for casual users are located near major attractions and leisure areas. This highlights an opportunity to place targeted promotions or subscription ads at these high-traffic, tourist-friendly stations.

These insights help bridge the gap between casual and regular use by identifying when, where, and how casual users engage with the service and suggest touch points where we can introduce compelling reasons to subscribe.

6 Share

What story does your data tell?
The data tells a story of two distinct user groups, who interact with the bike-sharing service in very different ways.
Based on my findings: Subscribers are primarily local residents who rely on the service for routine, weekday travel, likely commuting to work or school. Their usage is consistent and peaks in the spring (Q2), aligning with better weather and return-to-work periods after winter.
In contrast, casual users show strong seasonal and weekend-based patterns. Their usage peaks during the summer (Q3), especially on weekends, and is concentrated around Chicago's most popular tourist destinations. This suggests that many of them are tourists or occasional recreational riders.
The difference in station usage, timing, and frequency paints a clear picture: the service successfully supports both daily urban mobility and leisure travel, but different engagement strategies are needed for each group. These patterns highlight where the business can focus efforts to convert occasional users into long-term subscribers—especially by targeting summer riders at popular landmarks with tailored promotions.

Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently? Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?
Yes, the analysis clearly shows that annual members (subscribers) and casual riders use the service in different ways.

- **Time of use:** Subscribers tend to use the service during weekdays, suggesting commuting or regular travel patterns. Casual riders, on the other hand, are most active on weekends, indicating recreational or occasional use.

- **Seasonal trends:** Subscribers peak in usage during Q2 (spring), whereas casual riders peak in Q3 (summer), likely due to tourism and outdoor leisure activities.
- **Trip locations:** Casual riders frequently start trips near popular tourist attractions such as Millennium Park, Shedd Aquarium.
- **Trip duration:** Casual riders generally take longer trips, which aligns with leisure or sightseeing behavior, whereas subscribers take shorter, more frequent trips aligned with commuting.

How do your findings relate to your original question?
The original question was: *How can we get more annual subscribers?* Based on my analysis, I can give a couple potential strategies to encourage more people to subscribe.

- **Trip locations:** Many casual trips start or end near popular attractions, which provides clear geographic targets for in-app promotions or station-based advertising encouraging riders to subscribe.
- **Seasonal opportunity:** Since casual ridership peaks in Q3, this is an ideal time to promote annual memberships with incentives like discounts, free trial periods, or ”upgrade now” campaigns.
- **Messaging insight:** Marketing efforts should focus on the benefits that matter to casual riders—unlimited rides, lower cost per trip, and year-round convenience.

Who is your audience? What is the best way to communicate with them?
Since my primary audience is the technical team, the best way to communicate findings is through:

- **Data visualizations and charts:** Clearly labeled graphs (e.g., usage by day, trip duration, and seasonal trends) help convey behavioral patterns quickly and effectively.
- **Concise reporting with reproducible methods:** Using tools like Python or R scripts, shared notebooks, and versioned datasets ensures transparency and reproducibility of analysis.
- **Structured summaries:** Presenting findings in structured formats—such as executive summaries, bullet-point insights, and pivot tables—makes it easier for the technical team to translate insights into features or experiments.
- **Recommendations with data backing:** Every recommendation should be supported by clear evidence from the data, along with visual or statistical justification where applicable.

7 Act

What is your final conclusion based on your analysis?
Deploy subscription ads at high-traffic tourist stations like Millennium Park and Shedd Aquarium during summer weekends, when casual ridership peaks.

What next steps would you or your stakeholders take based on your findings? Create promotions and offers targeting the costumers.