Google Data Analytics Certificate via **Coursera**

CAPSTONE PROJECT

TOPIC: Identify how casual riders and annual members use Cyclistic’s bikes differently.

Wilfried TCHOUMBOU MBOUWOU
24/01/2024

TABLE OF CONTENTS

Table des matières

1. Context

Cyclistic is a bike-share company with a fleet of 5824 bikes geotracked and locked into a network of 692 stations across Chicago in USA. It has two groups of customers: Casual riders and annual members.

Cyclistic's financial analysts established that annual members are much more profitable than casual riders so that the marketing manager wants to focus the new marketing strategies on maximizing the number of annual memberships. Moreno (the manager) believes that will be the success' key of the company's growth.

As junior data analyst, I am in charge of coming up with insights about how those 2 groups use Cyclistic's bikes differently.

2. Business task

Identify the reasons why some people subscribe as annual members while others prefer to stay as casual riders. Then, identify when becoming an annual member would be more profitable for a casual rider.

3. Description of all data sources used.

This dataset provides a detailed compilation of variables related to trips done during the last 12 months of the previous year in Chicago. These attributesinclude the ID of each ride (*ride_id*), the type of ride (*rideable_type*), the started and ended datetime, the ID of the starting station and the ID of the ending station (*start_station_id & end_station_id*), the geographic position of the start (*start_lat & end_lat; start_lng &end_lng)* and the type of user (*member_casual*). It's owned by the city of Chicago and collected by divvy bikeshare. The last update was completed on January 4th, 2024. It is currently available to the public under license. It's a sample of a large collection of 10 years of data which contains 41767,684 rides. Using the sample size calculator, for a 99%'s confidence level and 1%'s margin of error, we found out that the right sample size would be 16,635 rides and this dataset got 5,719,877 rides which is far upper than what is required. In addition, it has the same distribution (64% of members against 36% of casual riders) compared with the whole (67% of members against 33% of casual riders). Regarding all the information listed above, I can assume that this dataset is unbiased, reliable, original, current and comprehensive.

4. Cleaning and transforming data.

4.1) Cleaning data

Ensuring data's integrity requires cleaning the dataset. That means checking if that dataset has:

- Inconsistent labels
- Inconsistent formatting
- Duplicates values
- Misspelling values
- Empty spaces
- Null values.

I chose to work in SQL because this dataset is large so it’s difficult to manipulate it in spreadsheet.

4.1.1) Troubleshooting inconsistent labels.

I checked each table to ensure that labels are consistent across all the datasets. I found out all are correct.

4.1.2) Troubleshooting inconsistent formatting.

I checked if each variable is correctly formatted. According to their definition, *ride_id* column should be in string, *started_at* & *ended_at* columns should be in date format, *member_casual* column should be in string. I checked the 12 datasets and found out that each variable was correctly formatted.

4.1.3) Troubleshooting duplicates values

To do that, first only the variables needed to solve the problem should be selected. Such varables are *ride_id*, *started_at*, *ended_at* and *member_casual columns*. Then, because ride_id identifies each ride taken, if there is duplicate, it would show up in the *ride_id* column.~~.~~ I thus decided to check duplicates in the *ride_id* column.

To find it out, I found the initial number of rows including eventual duplicates, then used the DISTINCT function to display only distinct values and see if the number of rows would be different. Each table had the same number of rows for both cases. So, I concluded that there were no duplicates values.

4.1.4) Troubleshooting misspelling values and extra spaces

Here, for each table, Firstly, I checked distinct values of each variable. Then I looked for misspelled values. It was quite difficult to get all the rows due to the high number of observations. Secondly, I calculated the length of each variable and tried to see if one among them had different values. It was still difficult to catch all the values due to the high number of observations. Thirdly, I calculated the max and the min of the length of each variable to see if they are different. Fortunately, those values were the same. So, I concluded that, for each variable there was neither misspelled value nor extra space.

4.1.5) Null values

Here, I updated all those twelve tables with only the 4 variables we need to solve the problem to make it easier to manipulate. Then, for each table, I ran the SQL query to filter the table with at least one null value. I found out nothing for everyone. I concluded therefore that there was not null value.

To wrap up, I concluded that these datasets are accurate, complete, consistent, and ready to analyze.

4.2) Transforming data

- I split one column into 4 columns (ride_id, started_at, enden_at, member_casual)

- I created a conditional column to calculate the duration of each trip. Solving the problem requires knowing the duration made by the two distinct types of users.

- I created a conditional column to show up the day of the week for each trip's started date (*started_at*).

5. Analyzing phase.

5.1) Initial hypothesis

Regarding the details of each offer, we can make some initial hypothesis:

- (h1) Annual members do not care about the limited time set on the offers.
- (h2) The number of rides increases during the summer and drops during the winter.

5.2) The analysis' summary

In this part, we will assess the accuracy of each hypothesis based on data-driven decisions.

5.2.1) The average of ride's length calculation

To assess h1, we calculated the average of *ride_length* column for each month per type of user (casual or member) and compared.

To make it possible and display it into one table, we merged the 12 recent divvy trips 2023 months tables. Here are following steps:

1. We tried that regard in spreadsheet, and it was not possible to import it as the number of rows of each table was greater than the acceptable limit.

2. Then, we loaded those datasets in Google drive and imported them into SQL (Big query).

3. We used the UNION DISTINCT command to merge all those 12 tables into one called full_year_divvy_trips_2023.

4. We used the TIMESTAMP_DIFF function to calculate the difference between the end date of ride and the start date of ride to get the length of ride (only where the ended date was greater than the started date. Anything else would be an error using the FILTER command) We recorded that as *ride_length* column.

5. We extracted the month from the start date through the EXTRACT function and got a new column called month.

6. Then, we aggregated the full_year_divvy_trips_2023 dataset to calculate the monthly average of ride's length using the AVG function grouping by month.

7. At last, we updated the full_year_divvy_trips_2023 including the average of ride's length by modifying the request as a permanent ending table.

5.2.2) Number of rides and gap's calculation

To assess h2, we calculated the number of rides taken by both members and casual riders during the hot season and compared the gap. We did the same thing during the cold season to show how members lose their money while casual riders save theirs. Here are following steps:

1. On the full_year_divvy_trips_2023 table, we extracted the month from the started date through the EXTRACT function and got a new column called month.

2. We aggregated the full_year_divvy_trips_2023 dataset to count the number of rides using the COUNT function grouping by the type of user. We did it both for months of cold season in one hand, and for months of hot season in the other hand.

3. To get the number of rides' gap between members and casual riders, we aggregated the full_year_divvy_trips_2023 dataset to calculate the difference between their numbers grouping them by month. We also added the gap's percentage of each of them. We did it for each season (cold and hot) using the FILTER command and logical operators. We created new tables named *gap_cold_season_2023* and *gap_hot_season_2023* to store them. We rounded the value of the gap's percentage in 2-decimal value using the ROUND function.

5.2.3) Total duration calculation

I also calculated the total duration of members and casual riders to know which types of riders are the most likely user of the divvy bikes. We used the following steps:

We aggregated the full_year_divvy_trips_2023 dataset to count the ride's length of each ride over a year, grouping by type of user (*member_casual*) using Tableau.

5.2.4) Number of rides each day of a week.

To assess h3, we calculated the number of trips taken each day of a week. We followed these steps:

- We extracted the day of the week from the *started_at* column. Then we counted the total number of trips taken each day of the week over a year.

6. Share phase

6.1) business objective:

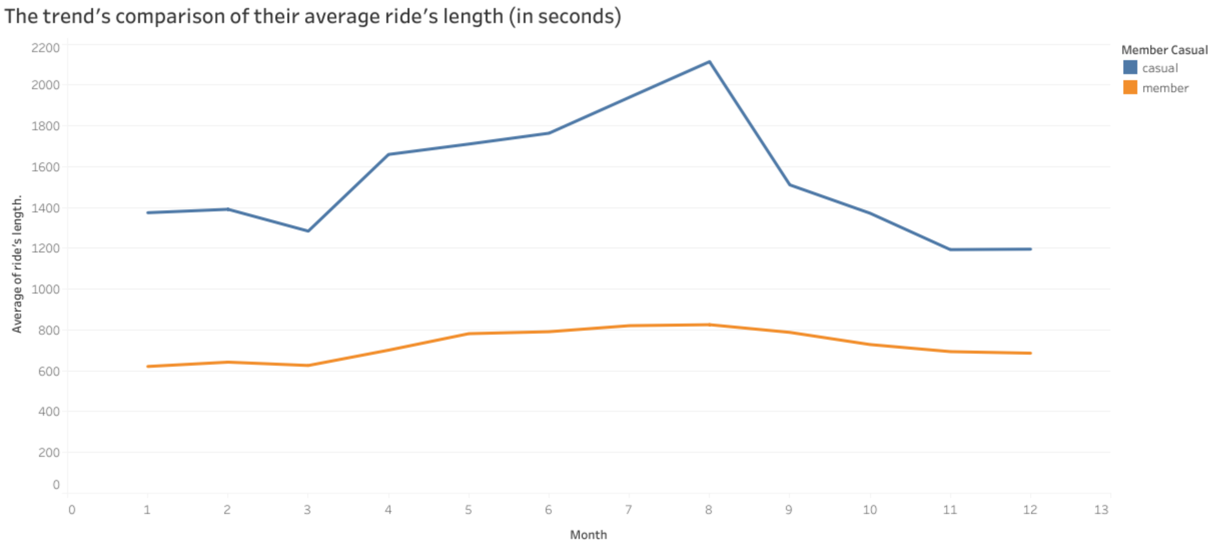
Identify how members and casual riders use Cyclistic bikes differently in order to help build an innovative marketing strategy aimed at converting casual riders into members.

6.2) Storytelling

6.2.1) About the hypothesis h1.

Let's see if annual members do not care about the limited time set on the offers. To figure it out, I built a line chart to compare the trends of each user type about the average ride length over a year. Here are the results.

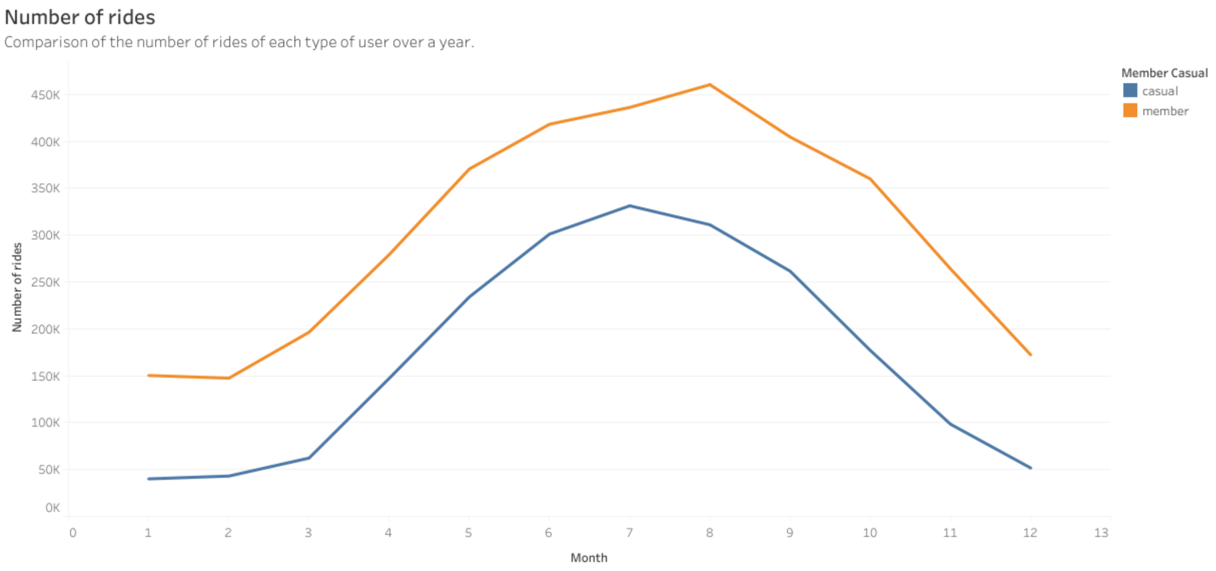
Chart 1: The trend's comparison of their average ride's length.



As you can see, casual riders took longer trips than members. The gap is significant. That means they are more likely to cause higher maintenance charges.

We can also think that the impact would be minimized if they took a smaller number of rides than members. Let's compare the number of rides of both and see what it says.

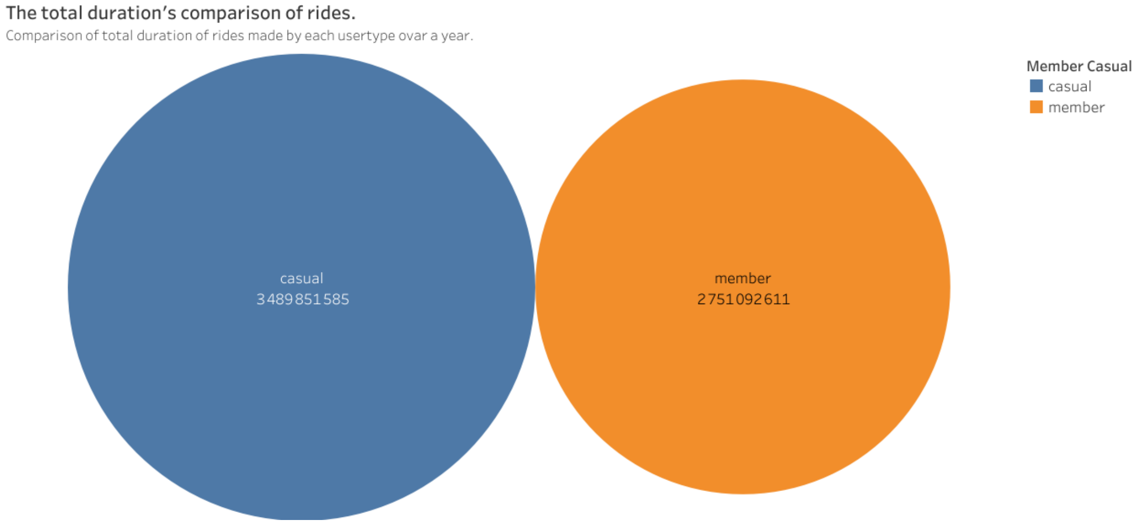
Chart 2: The trend’s comparison of their number of rides.



As you can see, members have the number of rides greater than casual riders’ one. It seems comprehensible as members can take rides limitless. We thus have a case where members take a high number of rides with a lesser trip time. That proves sufficiently that members do not care about the trip’s time. Someone registered as a member may take at least two rides to make a distance where a casual rider takes one.

In addition, to clearly draw on what user type would cause higher maintenance charges, let’s see, through the total duration comparison of trips taken by both user types in a year, what data tell.

Chart 3: The total duration’s comparison of rides.

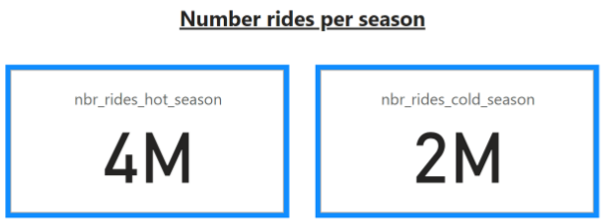


Even though members have the highest number of trips, we can see that casual riders have the highest duration of rides. They thus cause higher maintenance costs. The more you have a longer ride, the more you are likely to cause a maintenance's issue.

6.2.2) About the hypothesis h2.

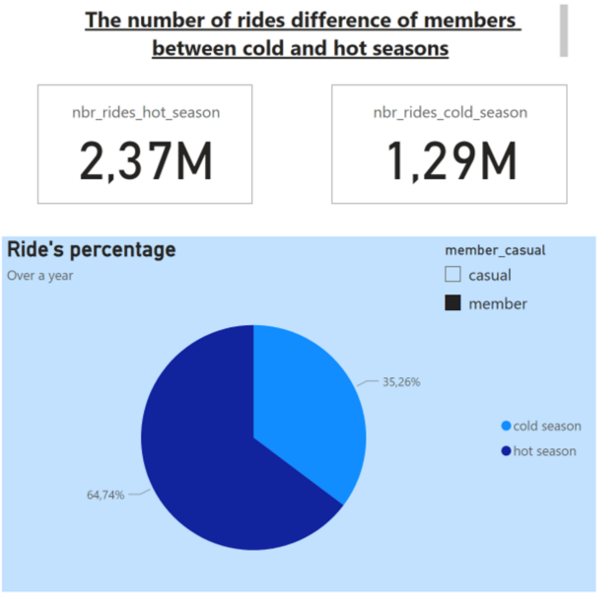
Let's see through data if the number of rides increases during the summer and drops during the winter.

Here is the total number of rides during the hot and the cold season.

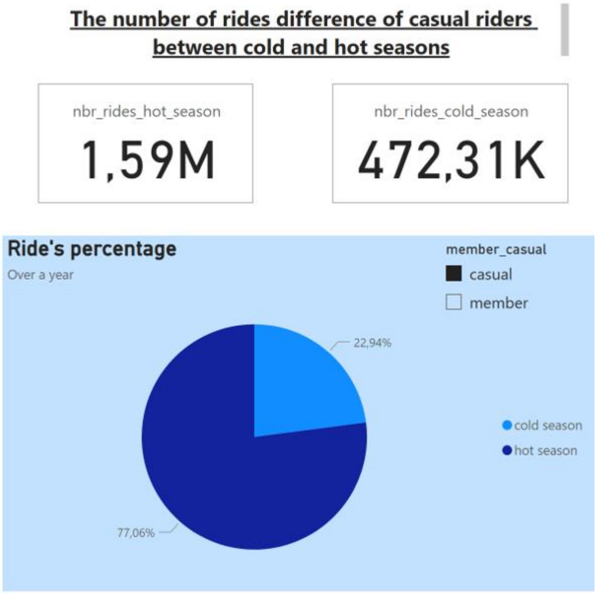


hot season: April, May, June, July, August, September
Cold season: October, November, December, January, February, March

As you can see, the number of rides in the hot season is twice as large as that of the cold season. Plus, when you see these others charts below:



hot season: April, May, June, July, August, September
Cold season: October, November, December, January, February, March



hot season: April, May, June, July, August, September
Cold season: October, November, December, January, February, March

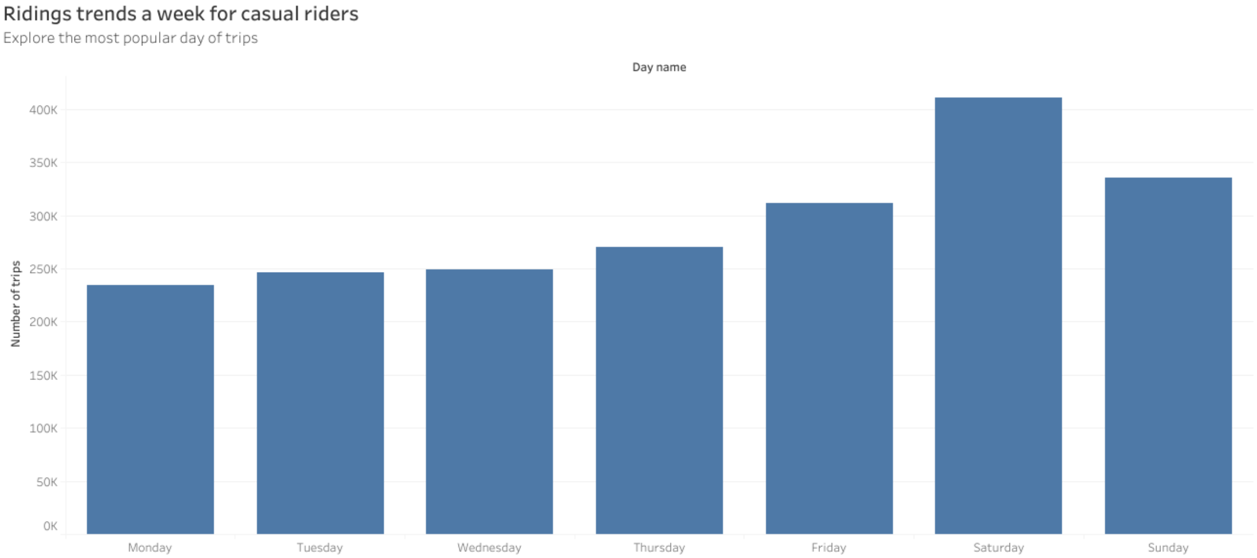
You can easily guess that casual riders think of losing money in the cold season. That could be the reason leading them to not subscribe to an annual membership.

6.2.3) About the hypothesis h3

Let’s see, for each type of user, through data, if the most popular day is on the weekend

- For casual riders

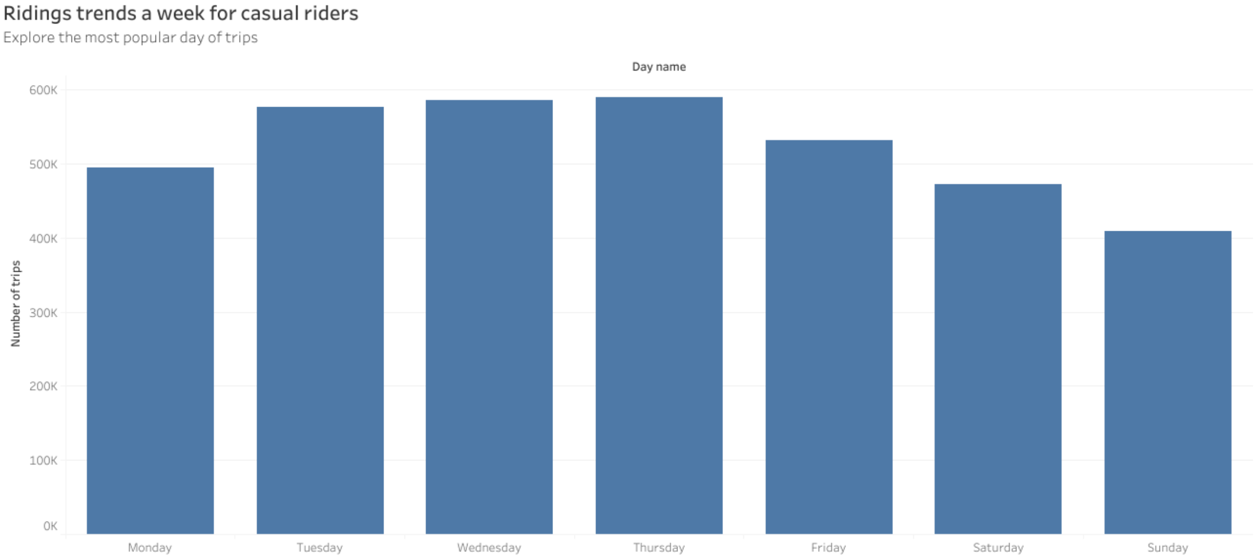
  Here are our findings



As you can see, the most popular days are Friday, Saturday and Sunday.

- For members

  Here are findings



The most popular days for trips taken by members are Tuesdays, Wednesdays and Thursdays.

6.3) Key findings

- ➢ The total trips' duration of casual riders exceeds that of members by about 10%. That means that casual riders are more likely to cause higher maintenance costs.
- ➢ The number of rides in the hot season is twice as high as in the cold season. Plus, the Cyclistic bikes' use of casual riders is almost 12% less than that of members. That clearly shows up why people hesitate to subscribe as annual membership.
- ➢ Casual riders and members take rides most of theirs rides different days.

7. Recommendations

In this case study, the business problem was to identify how members and casual riders use Cyclistic bikes differently to help draw up an innovative marketing strategy aimed at converting casual riders. Based on findings from these datasets, I clearly figured out that people hesitate to subscribe to an annual membership because they want to use their money efficiently. They do not want to feel like they are wasting their money. So, I recommend:

- To build a strategy which turns casual riders into members.
- To offer a discount on the annual membership offer to casual riders during the cold season.
- To optimize resources, the company should focus advertising campaigns on casual riders Thursdays, Fridays, Saturdays and Sundays instead of every day.