Google Data Analytics Capstone: Complete A Case Study
Case Study 1: How does a bike-share navigate speedy success?
Author: Muhammad Shkir

A clear statement of business task

What is the problem you are trying to solve?

Understand how casual riders and annual riders use Cyclistic bikes differently, in order to design a new marketing strategy.

How can your insights drive business decisions?

My team's insights and recommendations will guide Cyclistics to transform casual riders into annual members.

A description of all data sources used

Where is your data located?

We're working with data provided by Cyclistic (first-party), which means it is located in their servers (Cyclistic uses AWS servers).

How is the data organised?

Cyclistic provided the analytics team with CSV (comma separated values) files, so we have imported this data to SQL tables.

Are there issues with bias or credibility in this data?

The Cyclistic data set isn't ROCCC (Reliable Original Comprehensive Current Cited). At Its raw form it isn't reliable since it has missing and biased data.

How did you verify the data's integrity?

1. Removing duplicate entries: validating that each record has repeated once only.
2. Perform regular back-ups: storing a copy of all the data at separate device and applying security controls.
3. Validating data: removing incorrect, biased, Null values. Checking data types. Fixing text problems(grammar, extra spaces..etc)

Are there any problems with the data?

Actually yes, we have found these issues:

- Missing values in 'gender' and 'birth_year' tables

- ‘Start_time’ less than ‘end_time’- logically incorrect to end the trip before starting
- Incorrect data types in ‘trip_duration’ table
- ‘birth_year’ values less than 1900

Documentation of any cleaning or manipulation of data

What tools are you choosing and why?

SQL- to import, manipulate, clean, and analyse data- it is capable to handle large data in no time!
DBeaver- to execute SQL queries.
VSCode- to document the cleaning process.
RStudio- to execute R queries in order to generate powerful visualisations.

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

Creating a new table containing ‘ready to analyse’ data- clean and ROCC data. Validating the cleaned table again.

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

Yes, I have a detailed log of my data cleaning process, including code snippets, brief explanation, date and time and the analyst who wrote the code. Plus how many overall records have been affected.

A summary of your analysis

```
Gender
total customers = 344,163
      female customers = 131,434 = 38%
      male customers = 212,729 = 62%
total subscribers = 2,914,215
      female customers = 726,527 = 25%%
      male customers = 2,187,688 = 75%
```

```
Trip duration
customer -->> 2,869 minute on average per ride
subscriber -->> 858 minutes on average per ride
```

```
Rent days
customers -->> Top three days: Saturday, Sunday
subscribers -->> Top three days: Tuesday, wednesday, Thursday
```

```
Birth year
Customer -->> 1988 on average
```

```
Subscriber -->> 1983 on average
```

1. Most Cyclistic non-subscriber users are males.
2. Regular customers rent bikes for a longer period of time.
3. Non-subscribers are renting more bikes on weekends while subscribers use them at the middle of the week.
4. Non-subscribers are older.

How will these insights help answer your business question?

These findings enable the advertising team to get to know the company's target audience, consequently, create a successful campaign to attract more customers to subscribe.

Supporting visualisations and key findings

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

Yes, I have concluded some interesting findings which can be employed to answer the business question.

How do your findings relate to your original question?

One of the facts that I have found is that about 62% of customers are males, which gives us a basic understanding of what advertisement idea we're going to choose. An additional key finding would be active week-days per member type; it turned out that casual riders tend to use Cyclistic on weekends more often, compared to annuals' who use it on Tuesdays and Wednesdays. Perhaps casual members are renting bikes for fun and annual members for work purposes, this hypothesis needs to be investigated! Another data-driven inference (which supports the previous one) is that subscribers use the service for less time- 858 compared to 2,869 for casuals (still not a hard proof and more digging needs to be done). Finally, annual members are younger, so youth advertisements would be more acceptable.

How Is Your Audience?

Lily Moreno- The director of marketing and your manager. Moreno is responsible for the development of campaigns and
initiatives to promote the bike-share program.
Cyclistic executive team- The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program

Can data visualisation help you share your findings?
Yes, all of these findings are built on data only, so visualisations will provide a graphical, easy-to read explanation.

Are your visualisations accessible to your audience?
I have taken into account the colour blind audience so they can easily read and identify trends from visuals. Besides, I have included Titles, Subtitles, Caption and legends.

Your top three recommendations based on your analysis?

What is your final conclusion based on your analysis?
By the end of this study we have got a better idea of our target customers- “non-subscribers / casual members
People who have used once or been using Cyclistic bikes frequently but didn’t pay for annual subscription (casual member). According to our analysis this group consists of 62% males and 38% females, They are in their mid thirties, and use Cyclistic mostly on weekends.”
Now we have the chance to affect and push them to subscribe. This could achieve two goals, one: Increase customer satisfaction and two: increase our profits.

How could your team and business apply your insight?
As a member of the marketing analyst team I would use this insight to launch an advertising campaign targeting customers who are not subscribed yet but used Cyclistic, offer discounts for long rides or on weekends.

What next steps would you or your stakeholders take based on your findings?
I would recommend stakeholders to collect more data about users in general and non-subscribers specifically. Although we have conducted a study aimed to get to know them better, much information is still missing!
To persuade more people we need to know why they use Cyclistic, How far they travel, and why they like/dislike it. ..etc

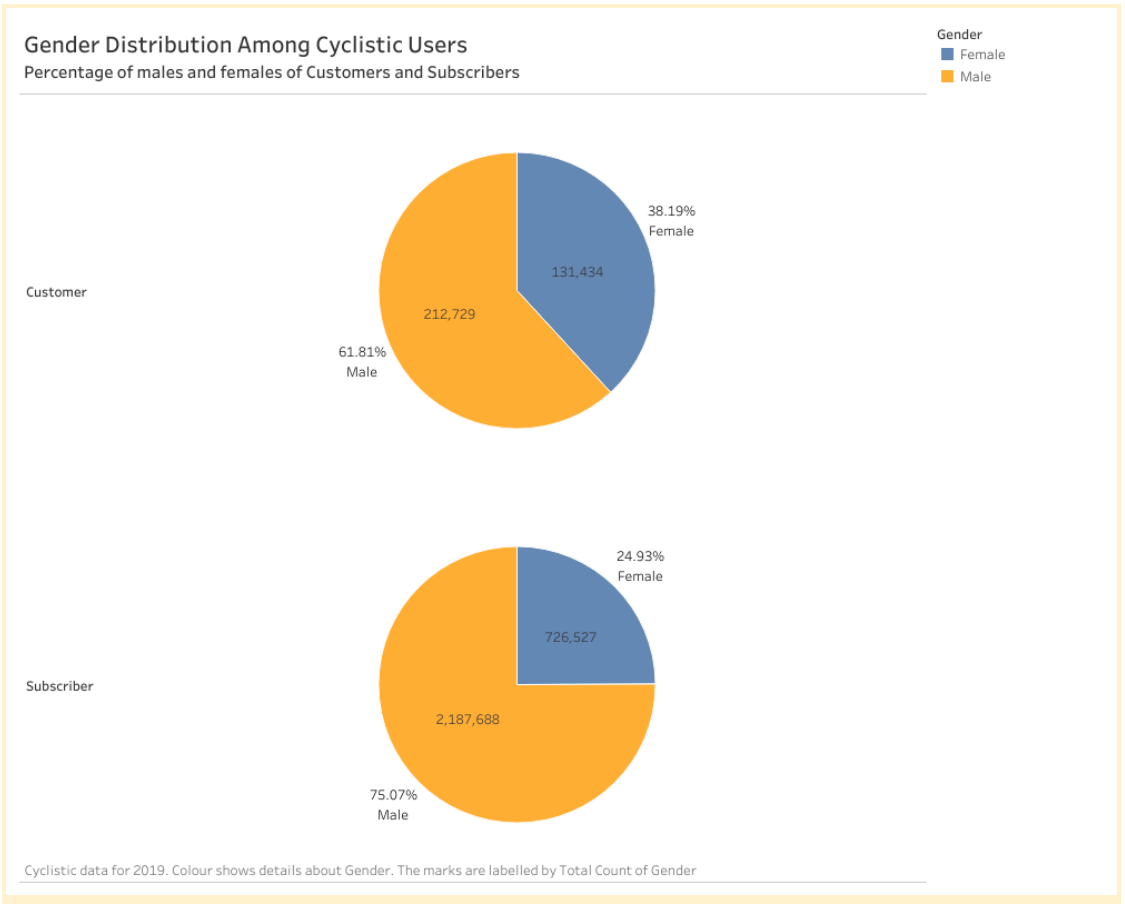
Is there additional data you could use to expand on your findings?

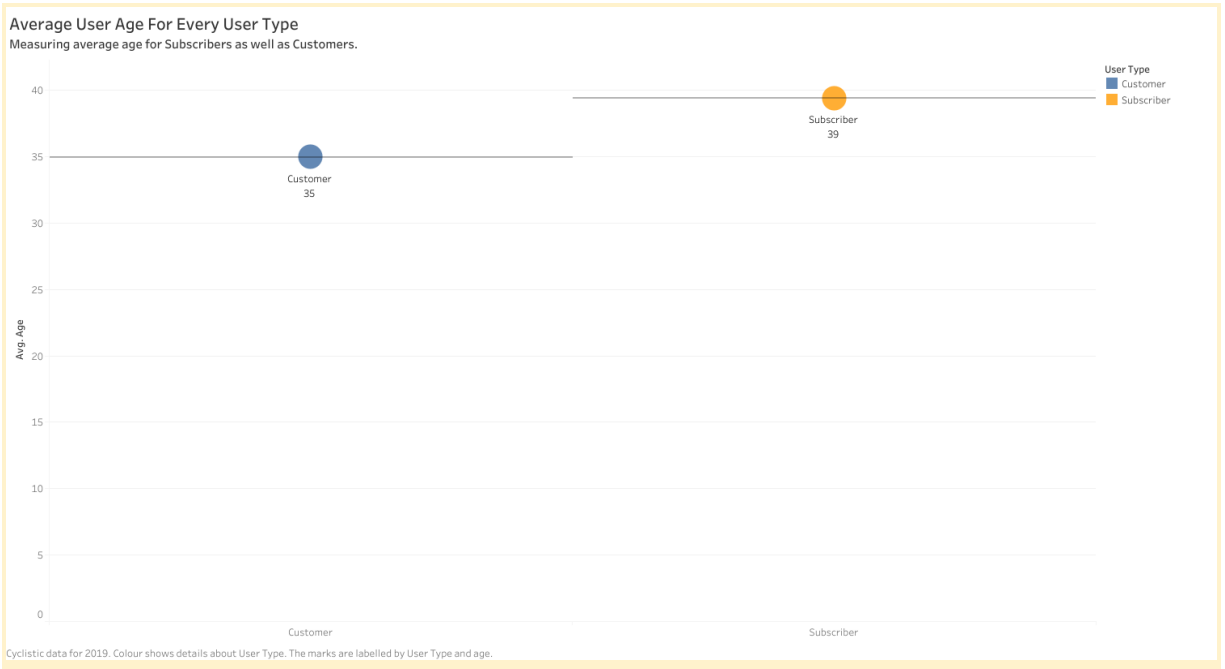
If there is publically available data for competitor companies in the same are, we could use it to deeply understand our customers.

Appendix

Customers / non-subscribers / casual members:
People who have used once or been using Cyclistic bikes frequently but didn’t pay for annual subscription (casual member). According to our analysis this group consists of 62% males and 38% females, They are in their mid thirties, and use Cyclistic mostly on weekends.

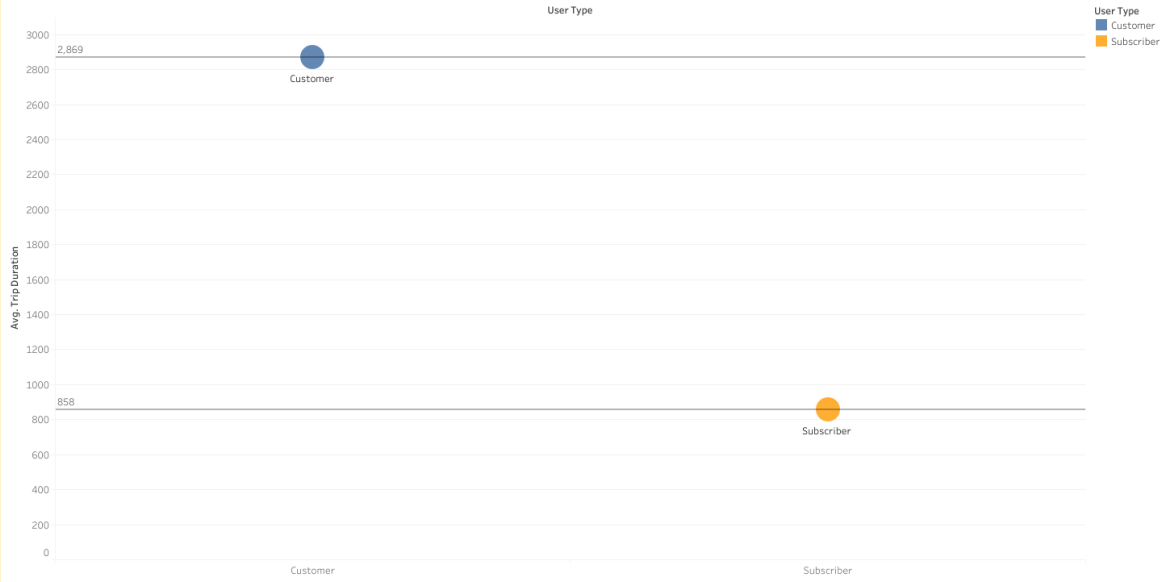
Visualisations:





Average Trip Duartion For Every User Type

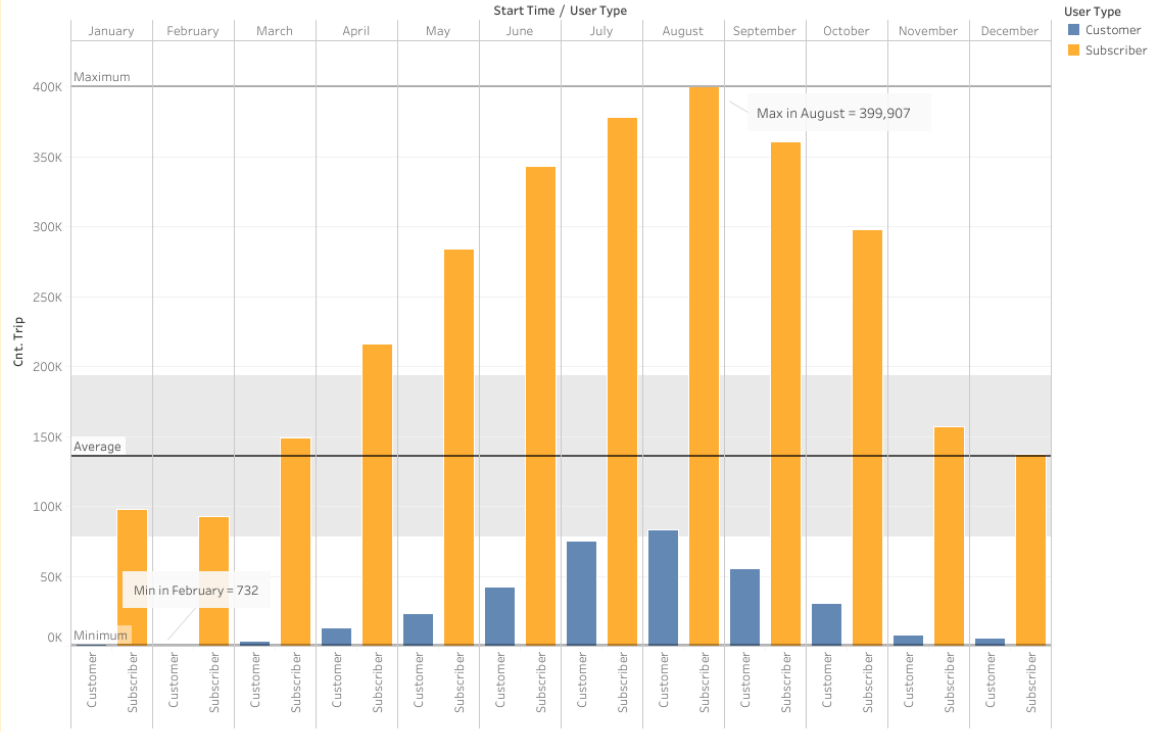
Measuring average rent duration for Subscribers as well as Customers.



Cyclistic data for 2019. Colour shows details about User Type. The marks are labelled by User Type.

Month And User Type vs. Number Of Trips

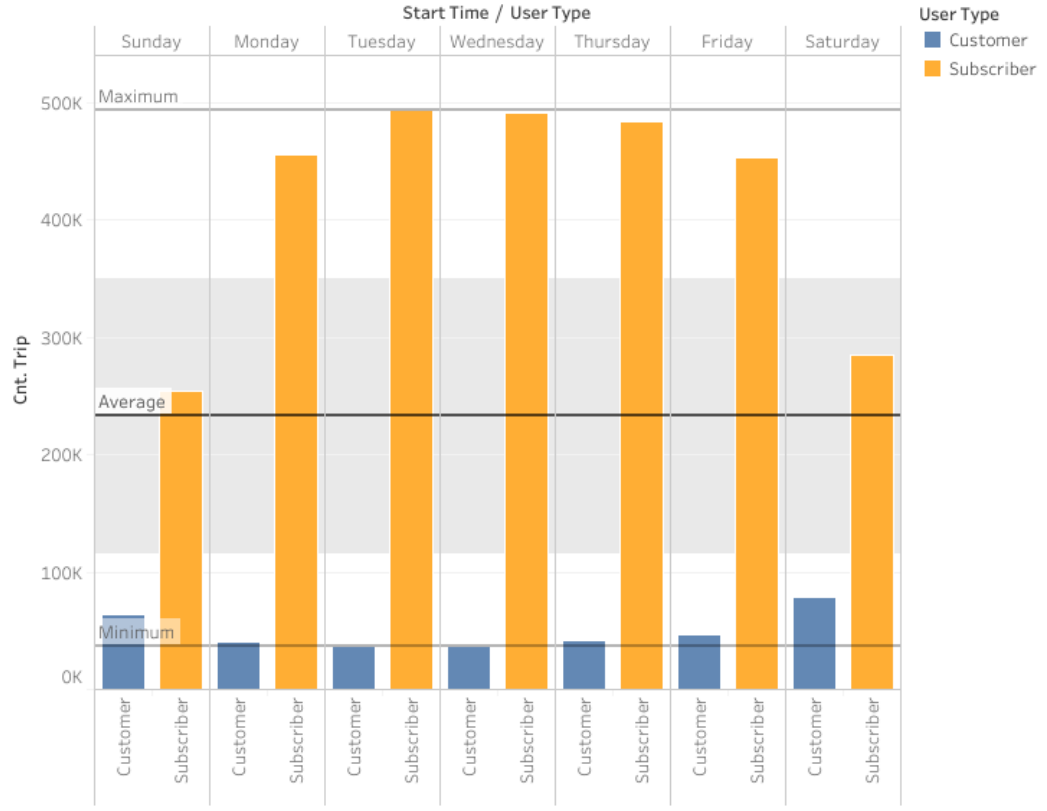
Compairing number of trips per month for Customers and Subscribers



Cyclistic data for 2019. Colour shows details about User Type.

Day Of Week And User Type vs. Number Of Trips

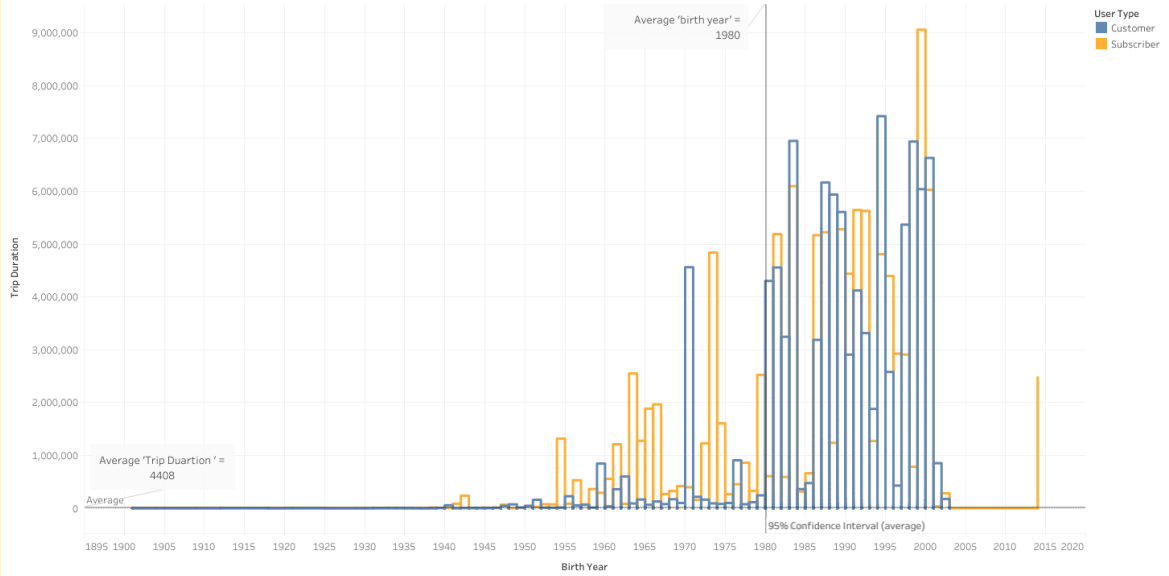
Compairing number of trips per Day Of Week for Customers and Subscribers



Cyclistic data for 2019. Colour shows details about User Type.

Birth Year vs. Trip Duartion

compairing users' age with rent duration



Cyclistic data for 2019. Colors shows user type. Vertical and horizontal lines shows averages.