# Case Study



Cyclistic bike-share analysis.

## Introduction

Cyclistic is a fictional bike-share company based in Chicago, the director of marketing would love to maximize the number of annual memberships. The director of marketing, Lily Moreno is responsible for the development of campaigns to promote the bike-share program.

With this analysis, we are hoping to find out how casual users and members use the bike-sharing service differently and develop a business plan to maximize the number of annual memberships.

## Deliverable

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cylistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

We were assigned the first question: How do annual members and casual riders use Cyclistic bike differently?

**Deliverables:**

* A clear statement of the business task.
* A description of all data sources used.
* Documentation of any cleaning or manipulation of data.
* A summary of the analysis.
* Supporting visualizations and key findings.
* Top three recommendations based on the analysis.

## Phase 1: Ask

Questions:

* **What is the problem you are trying to solve?**

I am trying to find out how casual riders and annual members use the bike-sharing service differently; and what can be done to motivate the casual riders to become annual members.

* **How can your insights drive business decisions?**

With the data that was provided to us, we will be able to find patterns and make data-driven decisions.

Key tasks:

* **Identify the business task**

The primary goal for the analysis is to answer the question “How do annual members and casual riders use Cyclistic bike differently?”.

* **Who is or are the stakeholder(s)?**

The director of marking, Lily Moreno,

The marketing team of Cyclistic,

The executive team of Cyclistic.

Deliverable:

* A clear statement of the business task.

## Phase 2: Prepare

We will be using the previous 12 months (which is data from 2021.12 to 2022.11) of Cyclistic trip data to analyze and identify trends.

The data is available [here](https://divvy-tripdata.s3.amazonaws.com/index.html), and the data has been made available by Motivate International Inc. under this [license](https://ride.divvybikes.com/data-license-agreement).

**Question:**

* **Where is the data located?**

The data is stored on AWS(Amazon Web Services) S3, which I downloaded for the purpose of this analysis, therefore the data I used for this analysis also stored on my personal cloud space.

* **How is the data organized?**

The original data on the AWS, starting on April 2020, the data from each month in the format “yyyymm-divvy-tripdata.zip” where “yyyy” represents the year and “mm” represent the month.

The 12 months of raw data I downloaded from the source provided is stored in the subfolder “Data” which is in the folder for the project name “cyclistic\_analysis”.

* **Are there issues with bias or credibility in this data? Does your data ROCCC (Reliable, Original, Comprehensive, Current, Cited)?**

There are some missing values in the dataset, but overall meet the ROCCC rules.

* + Reliable: Yes, there are some missing data in the dataset, mostly station names and IDs, but with the geotracking feature of the bikes, most data have recorded the latitude and longitude.
  + Original: Yes, the data is first party data which will meet the original requirement.
  + Comprehensive: Yes, the data contains columns such as “ride\_id”, “rideable\_type”, “started\_at”, “end\_at”, “start\_lat”, “start\_lng”, “end\_lat”, “end\_lng”, etc. With more than 5 millions of data during the 12 months period for this analysis, there is sufficient amount of data for this analysis.
  + Current: Yes, the 12 months period we choose to use is from December 2021 to November 2022. But the source has data back to 2013.
  + Cited: Yes, we’re using the first-party data provided by Cyclistic.
* **How are you addressing licensing, privacy, security, and accessibility?**

As mentioned, the data is made available by Motivate International Inc. under this [license](https://ride.divvybikes.com/data-license-agreement).

The data used for this analysis will be store on my local drive as well as my personal cloud drive for back up and easy access for me.

* **How did you verify the data’s integrity?**

The data is first-party provided by Cyclistic, with more than 5 millions of data in the analysis period, it is sufficient amount of data, so when it comes to the null values of the data, I decided to drop the data with null values. On top of that, I also check if there are duplicated data (which it doesn’t) to ensure the data’s integrity.

* **How does it help you answer your question?**

The datasets contained data of how users used the bike-sharing service, and each data has the membership status under the “member\_casual” column, which will be helpful for use to gain insights of how annual members and casual users use the service differently.

* **Are there any problems with the data?**

For privacy and ethical reasons, the data doesn’t contain users’ personal information such as name, age, genders, payment info, etc. Therefore, if we try to gain insights such as if age affecting the users’ membership status? how different genders using the services? Such questions won’t be able to answer due to the lack of such information in the dataset.

The data is also lack of the duration of each trip, but since the data has started and ended time, we will be able to gain the data of trip duration with some data manipulation and transformation.

## Phase 3: Process

**Questions:**

* **What tools are you choosing and why?**

The data for this analysis has more than 5 millions rows of data, spreadsheet applications won’t be able to handle such massive dataset. So I choose to use python for data cleaning and processing since I have more experience with Python.

As for visualization, I choose to use Tableau. Tableau has many great features and easy to make effective visualizations for the analysis.

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

I combined the 12 months of data into one DataFrame. And change the data type for “started\_at” and “ended\_at” columns to datetime format. Removed rows that contains null values, checked for duplicated values. Created columns “trip\_duration” by using the values in the “ended\_at” column minus “started\_at” columns, but also noticed that some data has negative trip duration, which I decided to remove those data for more accurate analysis. One more step for the “trip\_duration” columns is that I convert the values from “x days hh:mm:ss” to seconds and then convert to datetime format in order to correctly use the values of this column in Tableau.

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

After data cleaning, I checked all the columns are in the correct data type; checked for null values which returned 0 null values; and check for duplicated values which also returned 0 duplicated values; and removed the rows that contained negative values in the “trip\_duration” column.

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

Yes, the document for data cleaning is named “cleaning\_log” and stored with the “data\_cleaning” folder.

## Phase 4: Analyze

**Question:**

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

During the cleaning, I combined the 12 months of data into one DataFrame, performed the necessary data cleaning, manipulating, and transforming. After I exported the processed data to a csv file under the folder “Data” with other csv files and named the file “all\_trips\_data.csv”.

* **Has your data been properly formatted?**

Yes, during the data cleaning phase I ensured all the columns are in the correct data types; removed null values; removed duplicated values; and the data type of “trip\_duration” column is in datetime format which is what I need in order to correctly use it in Tableau.

* **What trends or relationship did you find in the data?**

Annual members use the service more for everyday life activities such as commuting to work, home, etc. And casual members use the service more for leisure.

Both casual users and annual members use the service more during spring to fall. Annual members use the service more on the weekdays and casual users use the service mor on the weekends.

* **How will these insights help answer your business questions?**

With these insights we can try to come up with marketing strategies or promotions more tailored to specific user groups.

## Phase 5: Share

**Questions:**

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

Yes, Casual riders use the service more for leisure, they have longer average trip duration and use the service more on the weekends. Whereas annual members use the service more for commute, they have shorter average trip duration, and they tend to use the service during the week.

* **How does your findings relate to your original question?**

The question was “How do annual members and casual riders use Cyclistic bikes differently?”, and with the analysis we were able to find different pattern of how they use the service.

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

The director of marking, Lily Moreno, the marketing team of Cyclistic,, the executive team of Cyclistic are the audience and the stakeholders of this analysis. The best way to communicate with them would be to have a meeting either in person or virtual meeting and present all the finding from this analysis.

* **Can data visualization help you share your findings?**

Yes, people are better taking in information with the form of visuals, so with the visualization the stakeholders will be able to easily see how the casual riders and annual members use the service differently.

* **Is your presentation accessible to your audience?**

Yes, the code, report, and images as well as visualization are included on my GitHub repository for this analysis.

## Phase 6: Act

**Questions:**

* **What is your final conclusion based on your analysis?**

The question we were asked to answer was “How do casual riders and annual members use Cyclistic bikes differently?”. With the analysis we can see that casual riders use Cyclistic bikes more for leisure. They tend to have longer trip duration, time of the day for them to use the service falls in late morning/early evening. For annual members, they use the service more for daily commuting purpose, they have higher frequency but shorter average trip duration. Peak time for annual members are 7-8 am and 4-6 pm, they tend to use the service more during the week and use the service more during off seasons compare to casual riders.

* **How could your team and business apply your insights?**

With the insights, we can have different tiers of charging for the time the casual riders using the service. For example, casual riders have average of 24 mins of trip duration, we can have the first 10 minutes at higher rates and lower the rate on the second 10 mins and so on.

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

We will need to discuss the marketing strategies such as the pricing by the trip length. Social medias are powerful platforms, we can develop stories of how riders using the service and how it makes a different in their lives, encouraging riders to use hashtags to share moments when they use the service.

And since the riders are using the service more during Spring to Fall, we can have different promotions for peak seasons and low seasons.

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

If we could have data such as age or genders and other information about the users (but still able to protect the identity of the users), we will be able to have more personalized promotion for different demographic of users.