# Data source

Cyclistic trip data (<https://divvy-tripdata.s3.amazonaws.com/index.html>)

* This dataset describes data collected by Divvy Bikes, a bike-sharing company that operates in Chicago.
* The dataset contains data about individual trips by Divvy Bikes users and also data about Divvy Bikes’s bike sharing stations that allow users to collect and release their bike.
* There are data from 2013 to 2023 in the dataset. In this case study, I will take the most recent 5 years of data to do my analysis. This selection of range is because data from too long ago has lost its credibility to deduce insight/conclusion for current time. Also, the COVID-19 pandemic outrage clearly has a huge impact on people’s mobilization and commute. As the pandemic has now become endemic, the society is experiencing a huge change. The data from 2020-2021 may not be that reliable to deduct conclusion about current and the future. Including data from 2019 also allows me to compare it with data from 2022 to 2023.

# Data organization

* The dataset contains multiple CSV files, each CSV file describes the data of trips/stations for a certain period of time (a season or a year).
* There are 3 types of organizations of columns in the CSVs data that I include in my analysis, for the readability of the documentation, I will leave the information about columns organization in the next section.
* The data is huge, and it contains too many rows to manipulate with spreadsheets, thus, a database will be used.

# Column organization

Data of 2019 : Divvy\_Trips\_2019\_QX.csv , X = {1,2,3,4}

|  |  |  |
| --- | --- | --- |
| Column name | Data Type | Description |
| trip\_id | qualitative | Unique ID for an individual trip |
| start\_time | quantitative | Starting time of the trip |
| end\_time | quantitative | Ending time of the trip |
| bikeid | qualitative | Unique ID for each bike |
| tripduration | quantitative | Duration of the trip |
| from\_station\_id | qualitative | The ID of starting station |
| from\_station\_name | qualitative | The name of starting station |
| to\_station\_id | qualitative | The ID of ending station |
| to\_station\_name | qualitative | The name of ending station |
| usertype | qualitative | Type of the user who underwent this trip |
| gender | Boolean | Gender of the user who underwent this trip |
| birthyear | quantitative | Birthyear of the user who underwent this trip |

Data of 2020 to 2023 : YYYYMM-divvy-tripdata.csv, YYYYMM=the year and specific month of which the data is about

|  |  |  |
| --- | --- | --- |
| Column name | Data Type | Description |
| ride\_id | qualitative | Unique ID for an individual trip |
| rideable\_type | qualitative | The type of the bike used in this trip |
| started\_at | quantitative | Starting time of the trip |
| ended\_at | quantitative | Ending time of the trip |
| start\_station\_name | qualitative | The ID of starting station |
| start\_station\_id | qualitative | The name of starting station |
| end\_station\_id | qualitative | The ID of ending station |
| end\_station\_name | qualitative | The name of ending station |
| start\_lat | quantitative | The latitude of the starting position |
| start\_lng | quantitative | The longitude of the ending position |
| end\_lat | quantitative | The latitude of the ending position |
| end\_lng | quantitative | The longitude of the ending position |
| member\_casual | qualitative | Whether or not the user is classified as member or as a casual user |

# Data Integrity

* Accuracy : this dataset was made available by Divvy Bikes, which is real-life data of their users historical trips, the data was anonymized so that it does not contain any personal information.
* Completeness : This data is fairly complete, given that it gives a well summarized information about individual trips. Still, there are some data that I would like to have to do a more informed analysis, such as the purchase data of users, the usual route of which the users take. However, there are some data privacy concerns that might be preventing the company from releasing such data. So, in the scope of publicly available data, it is fair to say that this data is complete.
* Consistency : The historical data has not experienced much update after the first publication(mostly published in 2020), so it should be consistent for anyone with access to this public data.
* Trustworthiness : The trustworthiness is evaluated in the next section.

# Dataset credibility assessment (ROCCC)

|  |  |  |
| --- | --- | --- |
| **Feature** | **Assessment** | **Score** |
| Reliable | This data is made public by Divvy Bikes, as the data is huge and thus the sample size, this data is fairly reliable to do inferential analysis. However, Divvy Bikes is a company that operates mainly in Chicago, so we should assume that the data is not reliable enough to make conclusions on broader population group. | 4 |
| Original | This is a 3-rd party data that is neither from the fictional company Cyclistic nor from our own data collections, so it has low originality. | 2 |
| Comprehensive | The data comprises of numerous data of trips, which is enough to do analysis on trip patterns. However, there can be more data to do relevant analysis, such as data about users subscription and payments. | 4.0 |
| Current | The latest time period of the data is July of 2023, which is recent enough for the insights derived to be applied to current situations. | 4.5 |
| Cited | Yes | 4.5 |
| **Overall score of credibility** | | 3.8 |

# Problems with the data

* The data is stored separately in several CSV files, in which some of them have different organizations from others, so a data merging process needs to be taken out.
* Missing values presented, some inconsistency in station name observed(some data adds ‘\*’ after station name),needs data cleaning.
* Inconsistency between user type across year, also inconsistency between matching of station id to station name, all this needs to be addressed when doing data merging.
* Also some data was not collected in years before 2020(gender, birthyear), and some not collected after it(longitude and latitude).
* The dataset is lack of demographic data about the users, which may be due to a privacy concern.

# About licensing, privacy, security, and accessibility

* Divvy Bikes grants to the public a non-exclusive, royalty-free, limited, perpetual license to access, reproduce, analyse, copy, modify, distribute in one’s product or service and use the Data for any lawful purpose (“License”).
* As for the privacy, the dataset does not contain any personal information that would lead to the identification of sample individuals. However, a full attention about data privacy is to be paid if there will be other dataset(s) included into the project.
* As for the security and accessibility, regarding data ethics, also to show case my work in my portfolio, I will make all related materials public, including documentations, code, cleaned data, data visualizations, report.