DATA DOCUMENT

## Data Analyst: Arez Shahid

## Client/Sponsor:

Cyclistic

## Dataset:

Cyclistic Trip data of 12 months of the year 2019, distributed in 4 quarters, was acquired from the Cyclistic’s own Amazon S3 bucket and was in the .csv format.

* Divvy\_Trips\_2019\_Q1.csv
* Divvy\_Trips\_2019\_Q2.csv
* Divvy\_Trips\_2019\_Q3.csv
* Divvy\_Trips\_2019\_Q4.csv

## Divvy\_Trips\_2019\_Q1:

This dataset contains monthly trip data ranging from January to March. Following were the fields in this dataset:

| **Column Name** | **Data Type** |
| --- | --- |
| trip\_id | INT |
| start\_time | DATE-TIME |
| end\_time | DATE-TIME |
| bikeid | INT |
| tripduration | INT |
| from\_station\_id | INT |
| from\_station\_name | STRING |
| to\_station\_id | INT |
| to\_station\_name | STRING |
| usertype | STRING |
| gender | STRING |
| birthyear | INT |

## Divvy\_Trips\_2019\_Q2:

This dataset contains monthly trip data ranging from April to June. Following were the fields in this dataset:

| **Column Name** | **Data Type** |
| --- | --- |
| 01 - Rental Details Rental ID | INT |
| 01 - Rental Details Local Start Time | DATE-TIME |
| 01 - Rental Details Local End Time | DATE-TIME |
| 01 - Rental Details Bike ID | INT |
| 01 - Rental Details Duration In Seconds Uncapped | INT |
| 03 - Rental Start Station ID | INT |
| 03 - Rental Start Station Name | STRING |
| 02 - Rental End Station ID | INT |
| 02 - Rental End Station Name | STRING |
| User Type | STRING |
| Member Gender | STRING |
| 05 - Member Details Member Birthday Year | INT |

## Divvy\_Trips\_2019\_Q3:

This dataset contains monthly trip data ranging from July to September. Following were the fields in this dataset:

| **Column Name** | **Data Type** |
| --- | --- |
| trip\_id | INT |
| start\_time | DATE-TIME |
| end\_time | DATE-TIME |
| bikeid | INT |
| tripduration | INT |
| from\_station\_id | INT |
| from\_station\_name | STRING |
| to\_station\_id | INT |
| to\_station\_name | STRING |
| usertype | STRING |
| gender | STRING |
| birthyear | INT |

## Divvy\_Trips\_2019\_Q4:

This dataset contains monthly trip data ranging from October to December. Following were the fields in this dataset:

| **Column Name** | **Data Type** |
| --- | --- |
| trip\_id | INT |
| start\_time | DATE-TIME |
| end\_time | DATE-TIME |
| bikeid | INT |
| tripduration | INT |
| from\_station\_id | INT |
| from\_station\_name | STRING |
| to\_station\_id | INT |
| to\_station\_name | STRING |
| usertype | STRING |
| gender | STRING |
| birthyear | INT |

## cleaned\_cyclistic\_dataset\_2019:

This is the aggregated dataset that contains 12 months trip data from January to December having a total of approximate 37 million rows. Following were the fields in this dataset:

| **Column Name** | **Data Type** |
| --- | --- |
| trip\_id | INT |
| starting\_time | TIME |
| starting\_date | DATE |
| start\_month | STRING |
| ending\_time | TIME |
| ending\_date | DATE |
| end\_month | STRING |
| bikeid | INT |
| tripduration | INT |
| from\_station\_id | INT |
| from\_station\_name | STRING |
| to\_station\_id | INT |
| to\_station\_name | STRING |
| usertype | STRING |
| gender | STRING |
| birthyear | INT |

## Process:

1. Merge the quarterly Datasets into one complete yearly Dataset by matching the column field names of all 4 datasets.
2. Check null entries in all columns.
3. Check format of trip\_id (8 digit integer).
4. Check if end\_time is earlier than start\_time (If so, swap them).
5. Split DATE-TIME object to separate DATE and TIME object. And also create a separate MONTH field according to it. Example, start\_time (DATE-TIME) to starting\_time (TIME), starting\_date (DATE) and start\_month (STRING) and vice versa.
6. Remove ‘,’ from tripduration field and convert it from STRING to FLOAT. Check if there are any tripduration value less than 0.
7. Drop rows that have a tripduration less than 60 seconds.
8. Replace missing values in the gender column with N/A.
9. Check for outliers in the birthyear column (birthyear < 1949 and birthyear >=2005) and drop those rows. Also, replace missing values in the birthyear column with N/A.
10. Check the mapping of station id to station names to see if they are consistent and have no duplicates. If there are duplicates, fix them by mapping it to its correct id and name.
11. Exported the cleaned dataset as ‘cleaned\_cyclistic\_dataset\_2019.csv’.