Cleaning Data Log

A. Data Analyst: Stefanus Yudi Irwan

B. Cleaning Date: September 15th 2022

C. Data Set

No	Raw Data	Clean Data
1	202108-divvy-tripdata.csv	202108_bike-data.xlsx
2	202109-divvy-tripdata.csv	202109_bike-data.xlsx
3	202110-divvy-tripdata.csv	202110_bike-data.xlsx
4	202111-divvy-tripdata.csv	202111_bike-data.xlsx
5	202112-divvy-tripdata.csv	202112_bike-data.xlsx
6	202201-divvy-tripdata.csv	202201_bike-data.xlsx
7	202202-divvy-tripdata.csv	202202_bike-data.xlsx
8	202203-divvy-tripdata.csv	202203_bike-data.xlsx
9	202204-divvy-tripdata.csv	202204_bike-data.xlsx
10	202205-divvy-tripdata.csv	202205_bike-data.xlsx
11	202206-divvy-tripdata.csv	202206_bike-data.xlsx
12	202207-divvy-tripdata.csv	202207_bike-data.xlsx

D. Cleaning Steps

1. Import csv data to excel
2. format “started_at” and “ended_at” column into date, in the format : yyyy-mm-dd hh:mm:ss
3. convert "start_lat", "start_lng", "end_lat", "end_lng" from text data to numeric data
4. round "start_lat", "start_lng", "end_lat", "end_lng" to 5 digit behind comma for spatial accuracy +- 1m on the ground
5. check blank value for every column. Found blank at "start_station_name", "start_station_id", "end_station_name", "end_station_id", "end_lat", and "end_lng". Do not drop the data because time related data is still complete
6. check duplication for every column, found no duplicate value
7. check spelling error for category "rideable_type" and "member_casual" column, found all spelling for category are all correct
8. create the column "started_days" which consist of day name from Sunday to Saturday for every rows
9. create the column "duration" by calculating the delta time in minute between “ended_at” and “started_at”. Make the duration data into interger