This is a list of steps taken for the Case\_Study\_1.

**# Download**

Downloaded 12 csv files from <https://divvy-tripdata.s3.amazonaws.com/index.html>

Saved all the .csv files to .xlsx files into their respective folders.

Saved the twelve .csv files to a single .db file.

Saved the .db file to a .csv file.

Saved the .csv file as another .csv file labeled “clean\_cyclistic\_data.csv”

**# Formatting**

Formatted all the files for easier viewing. (Excel)

Centered all text, bold/14 Font Size/angled/froze top text row for the attribute row, and autofit text width for all columns with small manual size adjustments.

**# Cleaning Documentation**

Excel – Looked for duplicate rows within the files. Check for any spelling errors within the files (some addresses would register as misspelt words, so I checked if that was a correct address then ignored all exact matches). Used filters to make sure there were only two or three choices in a couple columns. Find all empty/blank cells only turned up addresses that didn’t exist per the longitude and latitude that was imported for the specific ride.

SQLite – After combining all files into one database, turned the ride\_id column into a primary key, not null, and unique column to eliminate duplicates from the database (duplicates were the column attributes from the other 11 files). Ran queries to make sure the correct number of observations were present, i.e count(distinct ride\_id). Ran queries to check for nulls in the data. Ran queries to check and make sure the columns with 3 choices had only 3, columns with 2 choices only had 2, and the ride\_id column had the correct 16-character length in all observations.

RStudio – Used the head() function to make sure all the data was imported correctly and summary() function to verify the data. Created two new columns ride\_length, day\_of\_week, and distance\_miles. The ride\_length is in the format 15M 24S format and only displays minutes, hours, and days only if needed. The day\_of\_week column shows what day of the week the ride was from the point the ride was started. The distance\_miles column shows the distance in a straight line from lat and lng 1 to lat and lng 2 in miles. Saved the clean file as clean\_cycilistic\_data.csv. The month\_of\_year column shows the month of the ride from the start of the ride.

**# Combined all .csv files for easier use as a .db file. (SQLite)**

Could not combined all .csv files within Excel because Excel can only have 1,048,576 rows

Took all 12 .csv files and combined them into one table in the DB Browser for SQLite.

Made the ride\_id column a primary key, not null, and unique column.

The one table has all 13 columns and 5,779,444 rows.

**# Exported the now one table .db file into a .csv file**

**# Loaded the complete .csv file into RStudio for more cleaning and saved new .csv file call “clean\_cyclistic\_data”.**

**# Analysis performed on “clean\_cyclistic\_data”.**