**Data Preparation strategies**

## Data source

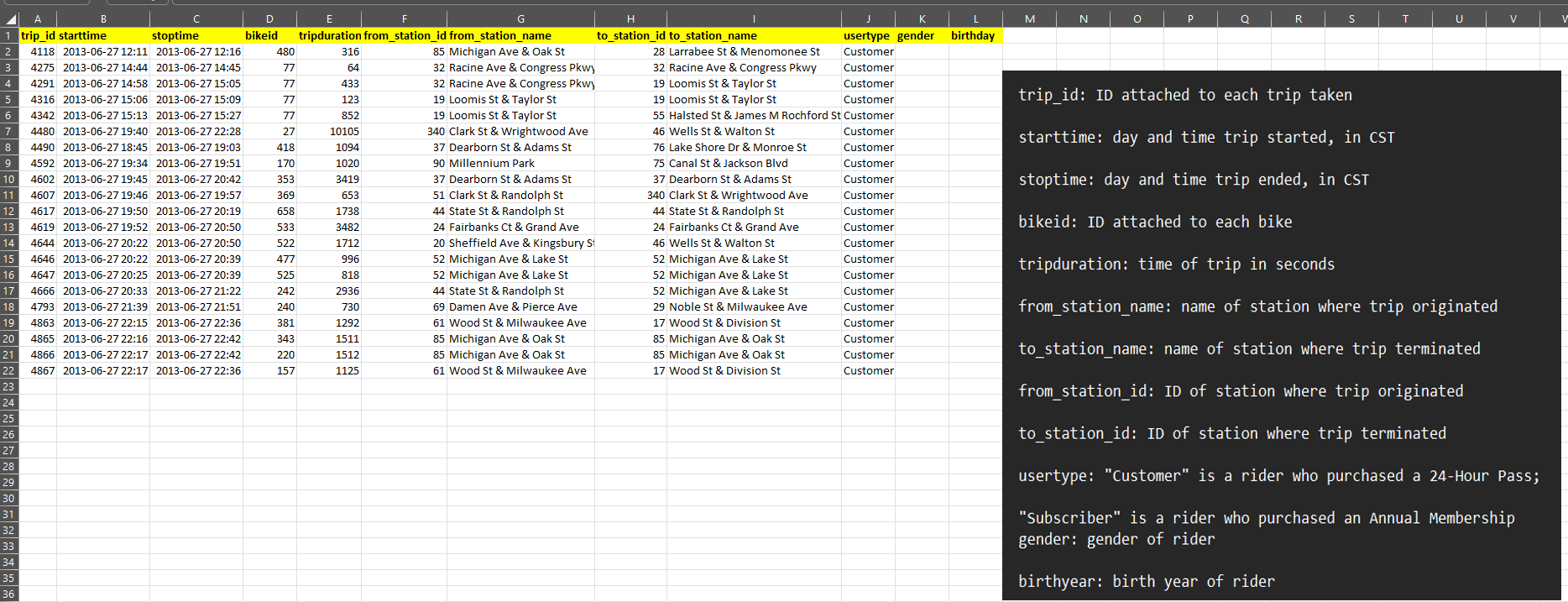
I will use Cyclistic’s historical trip data to analyze and identify trends. Click [here](https://divvy-tripdata.s3.amazonaws.com/index.html). This is a real public dataset suitable for exploring how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit you from using riders’ personally identifiable information. The data has been made available by Motivate International Inc. under this [license](https://divvybikes.com/data-license-agreement).)

## How data are organized

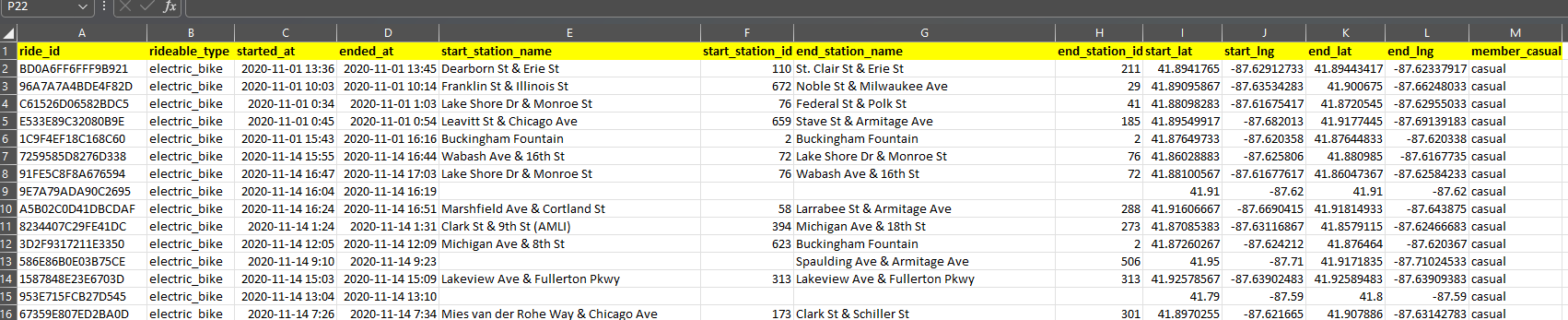
* There are various csv files that are in Zip folders.
* Data size ranges from 3 MB up to 57 MB
* Data are arranged in time periods (in years and quarters)
* Data are stored in long format.

To quickly capture the columns and shape of data, I manually snapshot some top rows for each table in Excel and compare similarities.

**The first form of data table Is from 2013 – 2019. These data tables include key features such as Trip duration and gender and birthday. These tables can go up to 1+ millions of rows.**

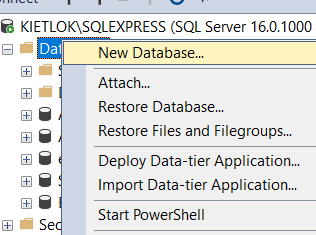


**The second form of data is from 2020 – August 2024. These tables include key features such as start and end time, member type but with no Gender and Birthday. These tables can go up to 1+ millions of rows.**



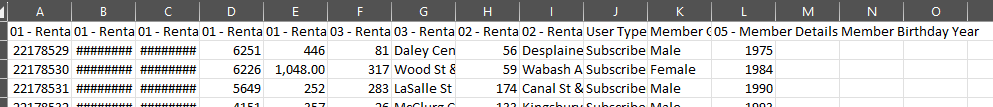
## Load Data

In this project, I will use SQL Server Studio to manage and query data. First, I will create a new database named Bike\_Share.



Then I will Load data csv from 2013 to 2023

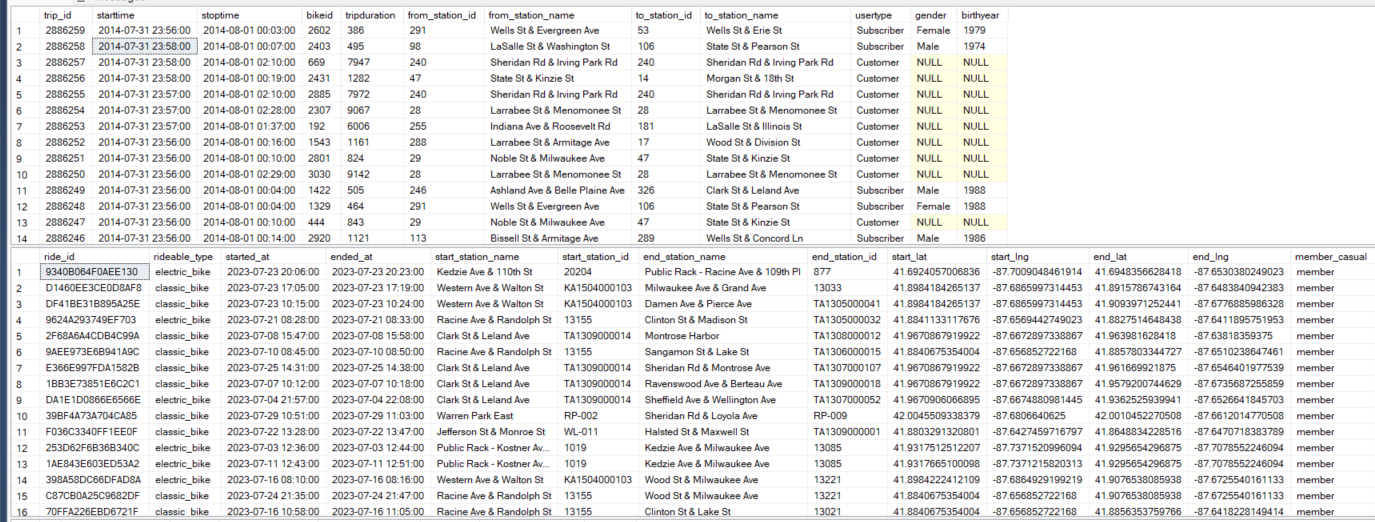
While loading the data to SQL Server Database, I encountered some tables have wrong columns name



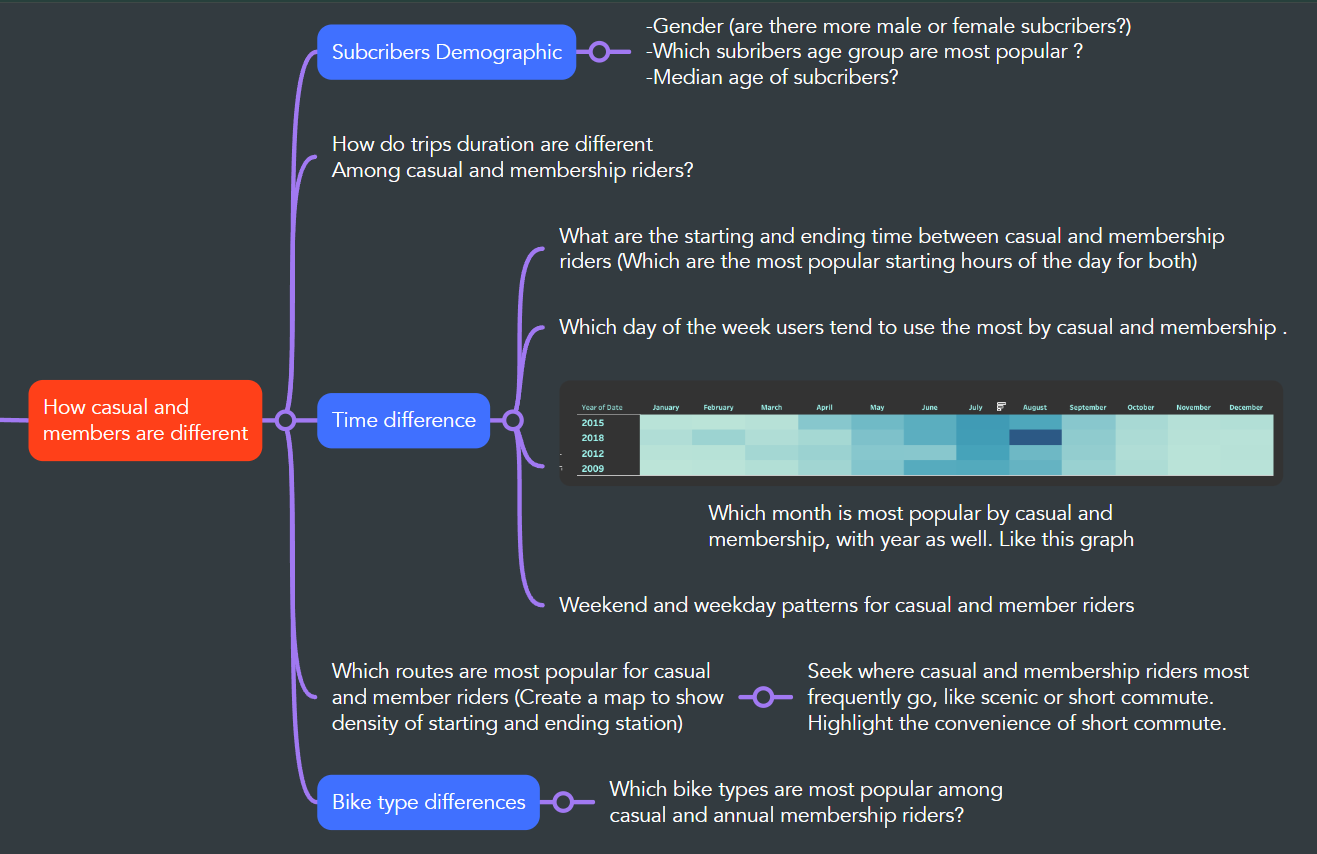
To resolve this, I simply change the columns name to align with the other tables.

**After that, I backup the database (Bike\_Share.bak)**

Before we proceed to any data modeling phase, I selected top 100 rows of all the form of data and visualize any potential valuable insights that I can extract from this data and then start the data modeling process.



Here are my initial insights. We might discover more in the EDA phase.



## Data Integration

Because of the uniqueness of this data, tables from 2013 to 2019 contain Gender and Birthday columns, but starting from 2020, these two columns are dropped and a new Bike Type column as well as the longitude and latitude are introduced. So, I will merge 2013-2019 and 2020-2023.

Note that when populating the temp table, I won’t need some Id columns to use less computer resources.

USE Bike\_Share;

/\* DATA INTEGRATION: MERGING TABLES \*/

-- Create a table called Trips\_2013\_2019

CREATE TABLE Trips\_2013\_2019

(

Start\_Time smalldatetime,

Stop\_Time smalldatetime,

Trip\_Duration int,

From\_Station\_Name nvarchar(70),

To\_Station\_Name nvarchar(70),

User\_Type nvarchar(50),

Gender char(10),

Birth\_Year smallint,

From\_Station\_Id smallint,

To\_Station\_Id smallint

);

-- Insert data from Trips\_2013

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthday, from\_station\_id, to\_station\_id

FROM [Trips\_2013];

--Insert data from Trips\_2014\_Q1\_Q2

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [Trips\_2014\_Q1\_Q2];

--Repeat the insertions for the remaining tables

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [Trips\_2014\_Q3\_07];

-- Insert [dbo].[Trips\_2014\_Q3\_0809]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [Trips\_2014\_Q3\_0809];

-- Insert [dbo].[Trips\_2014\_Q4]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2014\_Q4]

-- Insert [dbo].[Trips\_2015\_Q1]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2015\_Q1]

-- Insert [dbo].[Trips\_2015\_Q2]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2015\_Q2]

-- Insert [dbo].[Trips\_2015\_Q3\_07]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2015\_Q3\_07]

-- Insert [dbo].[Trips\_2015\_Q3\_08]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2015\_Q3\_08]

-- Insert [dbo].[Trips\_2015\_Q3\_09]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2015\_Q3\_09]

-- Insert [dbo].[Trips\_2015\_Q4]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2015\_Q4]

-- Insert [dbo].[Trips\_2016\_Q1]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2016\_Q1]

-- Insert [dbo].[Trips\_2016\_Q2\_04]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2016\_Q2\_04]

-- Insert [dbo].[Trips\_2016\_Q2\_05]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2016\_Q2\_05]

-- Insert [dbo].[Trips\_2016\_Q2\_06]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2016\_Q2\_06]

-- Insert [dbo].[Trips\_2016\_Q3]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2016\_Q3]

-- Insert [dbo].[Trips\_2016\_Q4]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2016\_Q4]

-- Insert [dbo].[Trips\_2017\_Q1]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2017\_Q1]

-- Insert [dbo].[Trips\_2017\_Q2]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2017\_Q2]

-- Insert [dbo].[Trips\_2017\_Q3]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2017\_Q3]

-- Insert [dbo].[Trips\_2017\_Q4]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2017\_Q4]

-- Insert [dbo].[Trips\_2018\_Q1]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthday, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2018\_Q1]

-- Insert [dbo].[Trips\_2018\_Q2]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2018\_Q2]

-- Insert [dbo].[Trips\_2018\_Q3]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2018\_Q3]

-- Insert [dbo].[Trips\_2018\_Q4]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2018\_Q4]

-- Insert [dbo].[Trips\_2019\_Q1]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2019\_Q1]

-- Insert [dbo].[Trips\_2019\_Q2]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

starttime, stoptime, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthday, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2019\_Q2]

-- Insert [dbo].[Trips\_2019\_Q3]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2019\_Q3]

-- Insert [dbo].[Trips\_2019\_Q4]

INSERT INTO Trips\_2013\_2019

(Start\_Time, Stop\_Time, Trip\_Duration, From\_Station\_Name, To\_Station\_Name, User\_Type, Gender, Birth\_Year, From\_Station\_Id, To\_Station\_Id)

SELECT

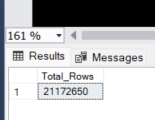
start\_time, end\_time, tripduration, from\_station\_name, to\_station\_name, usertype, gender, birthyear, from\_station\_id, to\_station\_id

FROM [dbo].[Trips\_2019\_Q4]

-- Finally, check the total rows for the completed table

SELECT COUNT(\*) AS Total\_Rows FROM Trips\_2013\_2019;

Our Temp table Trips\_2013\_2019 will have 21,172,650 rows. That’s huge!!



Next, we will repeat the merging process for tables from 2020 to 2023.

----Create a temp table merging data from 2020 to 2023

CREATE TABLE Trips\_2020\_2023 (

Ride\_Type nvarchar(50),

Start\_Time smalldatetime,

Stop\_Time smalldatetime,

From\_Station\_Name nvarchar(70),

To\_Station\_Name nvarchar(70),

Start\_Lat float,

Start\_lng float,

End\_Lat float,

End\_Lng float,

User\_Type nvarchar(50)

)

--Insert [dbo].[Trips\_2020\_Q1]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q1]

--Insert [dbo].[Trips\_2020\_Q2\_04]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q2\_04]

--Insert [dbo].[Trips\_2020\_Q2\_05]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q2\_05]

--Insert [dbo].[Trips\_2020\_Q2\_06]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q2\_06]

--Insert [dbo].[Trips\_2020\_Q3\_07]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q3\_07]

--Insert [dbo].[Trips\_2020\_Q3\_08]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q3\_08]

--Insert [dbo].[Trips\_2020\_Q3\_09]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q3\_09]

--Insert [dbo].[Trips\_2020\_Q4\_10]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q4\_10]

--Insert [dbo].[Trips\_2020\_Q4\_11]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q4\_11]

--Insert [dbo].[Trips\_2020\_Q4\_12]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2020\_Q4\_12]

--Insert [dbo].[Trips\_2021\_Q1\_01]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q1\_01]

--Insert [dbo].[Trips\_2021\_Q1\_02]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q1\_02]

--Insert [dbo].[Trips\_2021\_Q1\_03]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q1\_03]

--Insert [dbo].[Trips\_2021\_Q2\_04]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q2\_04]

--Insert [dbo].[Trips\_2021\_Q2\_05]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q2\_05]

--Insert [dbo].[Trips\_2021\_Q2\_06]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q2\_06]

--Insert [dbo].[Trips\_2021\_Q3\_07]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q3\_07]

--Insert [dbo].[Trips\_2021\_Q3\_08]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q3\_08]

--Insert [dbo].[Trips\_2021\_Q3\_09]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q3\_09]

--Insert [dbo].[Trips\_2021\_Q4\_10]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q4\_10]

--Insert [dbo].[Trips\_2021\_Q4\_11]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q4\_11]

--Insert [dbo].[Trips\_2021\_Q4\_12]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2021\_Q4\_12]

--Insert [dbo].[Trips\_2022\_Q1\_01]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q1\_01]

--Insert [dbo].[Trips\_2022\_Q1\_02]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q1\_02]

--Insert [dbo].[Trips\_2022\_Q1\_03]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q1\_03]

--Insert [dbo].[Trips\_2022\_Q2\_04]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q2\_04]

--Insert [dbo].[Trips\_2022\_Q2\_05]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q2\_05]

--Insert [dbo].[Trips\_2022\_Q2\_06]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q2\_06]

--Insert [dbo].[Trips\_2022\_Q3\_07]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q3\_07]

--Insert [dbo].[Trips\_2022\_Q3\_08]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q3\_08]

--Insert [dbo].[Trips\_2022\_Q3\_09]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q3\_09]

--Insert [dbo].[Trips\_2022\_Q4\_10]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q4\_10]

--Insert [dbo].[Trips\_2022\_Q4\_11]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q4\_11]

--Insert [dbo].[Trips\_2022\_Q4\_12]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2022\_Q4\_12]

--Insert [dbo].[Trips\_2023\_Q1\_01]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q1\_01]

--Insert [dbo].[Trips\_2023\_Q1\_02]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q1\_02]

--Insert [dbo].[Trips\_2023\_Q1\_03]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q1\_03]

--Insert [dbo].[Trips\_2023\_Q2\_04]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q2\_04]

--Insert [dbo].[Trips\_2023\_Q2\_05]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q2\_05]

--Insert [dbo].[Trips\_2023\_Q2\_06]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q2\_06]

--Insert [dbo].[Trips\_2023\_Q3\_07]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q3\_07]

--Insert [dbo].[Trips\_2023\_Q3\_08]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q3\_08]

--Insert [dbo].[Trips\_2023\_Q3\_9]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q3\_9]

--Insert [dbo].[Trips\_2023\_Q4\_10]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q4\_10]

--Insert [dbo].[Trips\_2023\_Q4\_11]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q4\_11]

--Insert [dbo].[Trips\_2023\_Q4\_12]

INSERT INTO Trips\_2020\_2023

(Ride\_Type, Start\_Time, Stop\_Time, From\_Station\_Name, To\_Station\_Name, Start\_Lat, Start\_lng, End\_Lat, End\_Lng, User\_Type)

SELECT

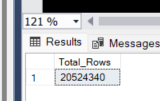
rideable\_type, started\_at, ended\_at, start\_station\_name, end\_station\_name, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual

FROM [dbo].[Trips\_2023\_Q4\_12]

-- Finally, check the total rows for the completed table

SELECT COUNT(\*) AS Total\_Rows FROM Trips\_2020\_2023;

Our Trips\_2020\_2023 table contain 20,524,340 rows.



After merging to these 2 new tables, I consider drop the individual tables since we have them all in the backup file.

## Data cleaning and validation

### For table Trips\_2013\_2019

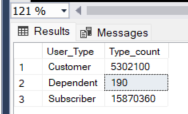
--See Total user type

SELECT User\_Type,

COUNT (User\_Type) AS Type\_count

FROM Trips\_2013\_2019

GROUP BY User\_Type



We have

* 5302100 customers (Day-pass, short rides)
* 15870360 Subscribers (Annual membership).
* 190 Dependent

Since in this scope of project we only focus on customer and membership riders, I will drop the dependent values.

--Drop Dependent Users

DELETE FROM Trips\_2013\_2019

WHERE User\_Type = 'Dependent'

Next, I will check for from and to station id nulls for Trips\_2013\_2019 table. Note that for Customer types, there will be no data for Gender and birth year.

--Check nulls for From and To station Id

SELECT \*

FROM Trips\_2013\_2019

WHERE From\_Station\_Id IS NULL OR To\_Station\_Id IS NULL;

We don’t have any nulls in these 2 columns.

With regard to the Birth\_Year columns, I consider removing any birthyear that less than 1920 as 100+ years old membership would be likely abnormal entries. However, some other data values are still valuable so I will just let it be for now.

--Identify any Birth year that < 1920

SELECT\*

FROM Trips\_2013\_2019

WHERE Birth\_Year < 1920

We have 5611 records.

### For table Trips\_2020\_2023

Firstly, I ensure we only have 2 types of customers, member and casual

--See Total user type for 2020\_2023

SELECT User\_Type,

COUNT (User\_Type) AS Type\_count

FROM Trips\_2020\_2023

GROUP BY User\_Type

We have 12247549 members and 8276791 casuals

Next, I will identify bicycle types

--identify Total bicycle types for Trips\_2020\_2023

SELECT Ride\_Type,

COUNT (Ride\_Type) AS \_count

FROM Trips\_2020\_2023

GROUP BY Ride\_Type

We have 3 types of bicycles: electric, classic, docked

Next, I will check null values for from station and to station name columns

--Check nulls for From\_Station\_Name to To\_Station\_Name for Trips\_2020\_2023

SELECT\*

FROM Trips\_2020\_2023

WHERE From\_Station\_Name is null or To\_Station\_Name is null

We have 3,844, 356 rows that contain nulls either in From\_Station\_Name or To\_Station\_Name.

--Check nulls for End\_Lng for Trips\_2020\_2023

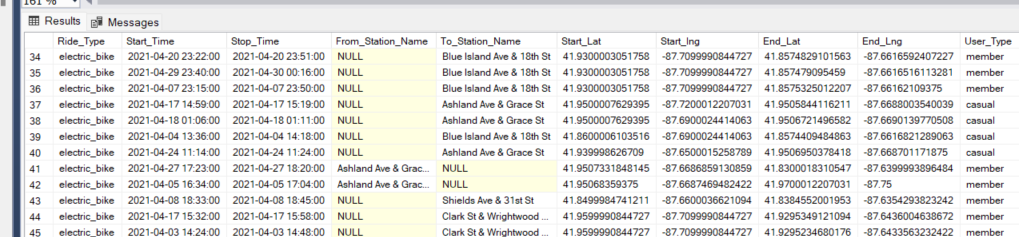
SELECT\*

FROM Trips\_2020\_2023

WHERE End\_Lng is null

We have 21885 for Ending Latitude

For our data, given we still have many other valuable data in other columns and the longitude and latitude is adequate, so I decided to leave the nulls value untouched.

****

## Data normalization

I noticed for Trips\_2013\_2019 the customer types are ‘Customer’ and ‘Subscriber” while for Trips\_2020\_2023, they are ‘member’ and ‘casual’. Therefore, to keep it consistent, I decided to keep **‘Customer’ and ‘Subscriber’.**

--Change user types to 'Casual' and 'Subscriber' for Trips\_2020\_2023

AlTER TABLE Trips\_2020\_2023

ADD User\_Type2 nvarchar(50);

UPDATE Trips\_2020\_2023

SET User\_Type2 = CASE

WHEN User\_Type = 'casual' THEN 'Customer'

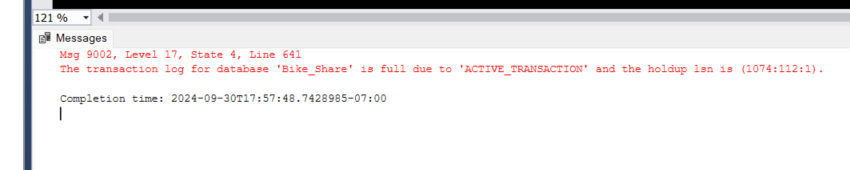
ELSE 'Subscriber'

END;

ALTER TABLE Trips\_2020\_2023

DROP COLUMN User\_Type;

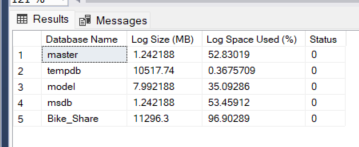
However, when I try to perform the above queries, I encountered the transaction log error.



So, to deal with this, I check the disk log transaction space

DBCC SQLPERF(LOGSPACE);

I can see that the Bike\_Share database takes nearly 97% of log space. So instead of directly shrink the log space which could slow the queries over time, I decided to use temp table to achieve this query.



--Try temp table

SELECT \*,

CASE

WHEN User\_Type = 'casual' THEN 'Customer'

ELSE 'Subscriber'

END AS User\_Type2

INTO #TempTrips\_2020\_2023

FROM Trips\_2020\_2023;

--View top 100 of the #tempTrips

SELECT top 100\* from #TempTrips

It works!!



And this is the final step for our data models. In summary, we would have the following tables:

* Station\_name (585 rows)
* Trips\_2013\_2019 (21,172,460 rows)
* Trips\_2020\_2023 (20,524,340 rows)
* A temp table #TempTrips (a copy of Trips\_2020\_2023 with normalized User\_Type).

**We will continue the Analysis in the Analysis section**