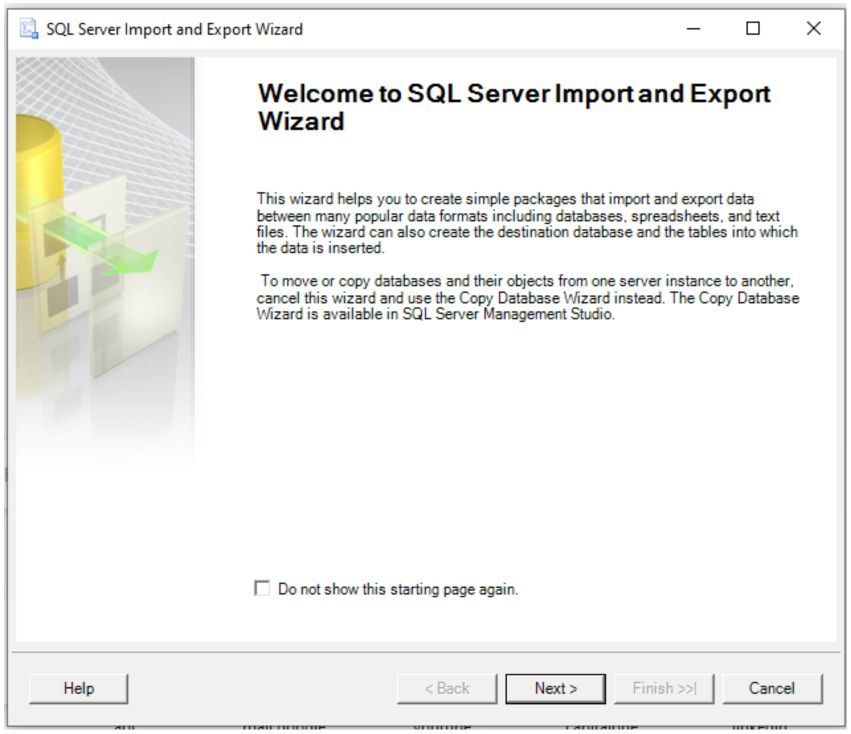
Google Data Analysis Certificate Capstone Project

Matija Mosunic

January 2022

R Studio Cloud timed out uploading Cyclistic files after 100 mb, approximately five months of ridership data. I uploaded the 12 months of data into MS SQL server, performing data cleaning and analysis here. The original data can be found here: Kaggle.



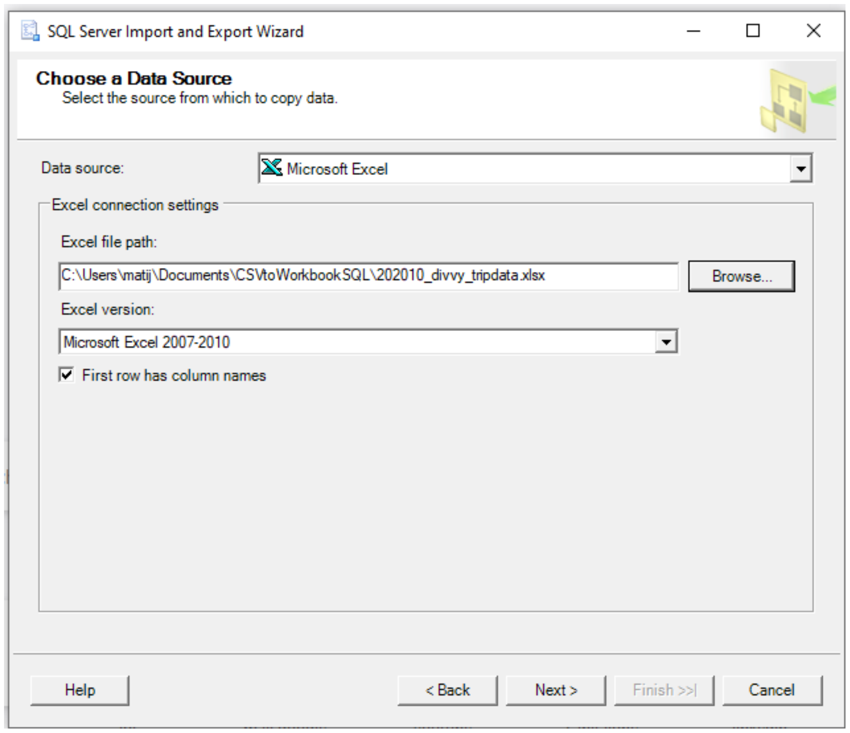
SQL Server Import and Export Wizard

Welcome to SQL Server Import and Export Wizard

This wizard helps you to create simple packages that import and export data between many popular data formats including databases, spreadsheets, and text files. The wizard can also create the destination database and the tables into which the data is inserted.

To move or copy databases and their objects from one server instance to another, cancel this wizard and use the Copy Database Wizard instead. The Copy Database Wizard is available in SQL Server Management Studio.

☐ Do not show this starting page again.


SQL Server Import and Export Wizard

Choose a Data Source
Select the source from which to copy data.

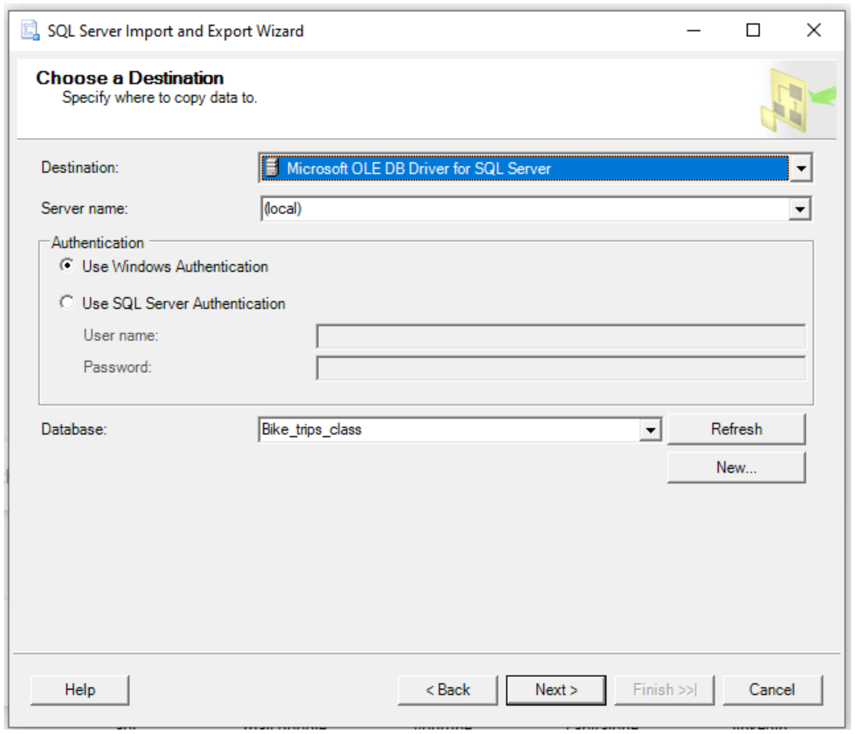
Data source: Microsoft Excel

Excel connection settings

Excel file path: C:\Users\matij\Documents\CS\toWorkbookSQL\202010_divvy_tripdata.xlsx

Excel version: Microsoft Excel 2007-2010

☑ First row has column names


SQL Server Import and Export Wizard

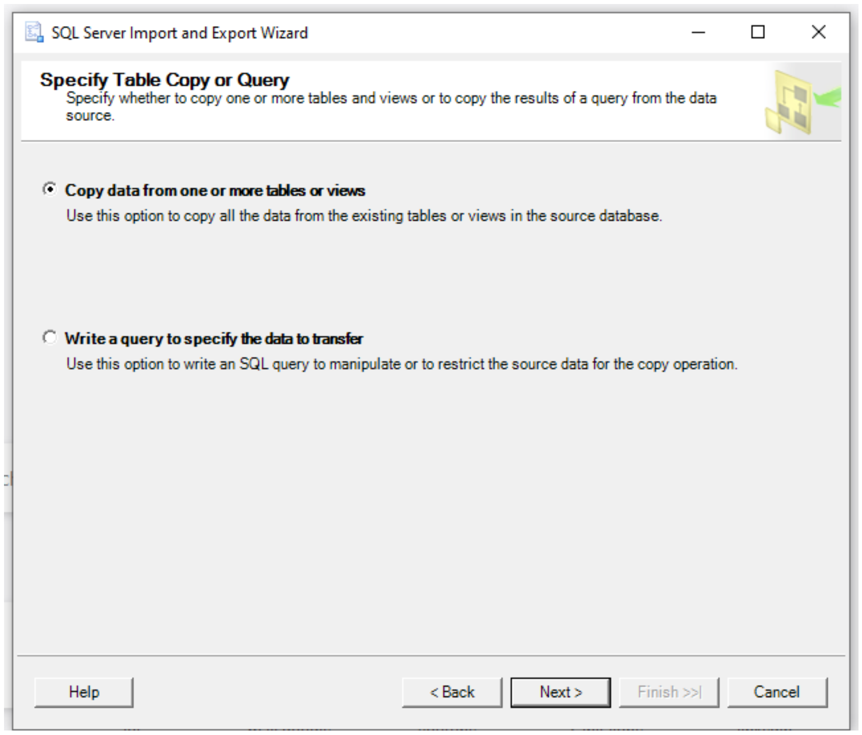
Choose a Destination
Specify where to copy data to.

Destination: Microsoft OLE DB Driver for SQL Server

Server name: (local)

Authentication
[✓ Use Windows Authentication]
Use SQL Server Authentication
User name:
Password:

Database: Bike_trips_class


SQL Server Import and Export Wizard

Specify Table Copy or Query

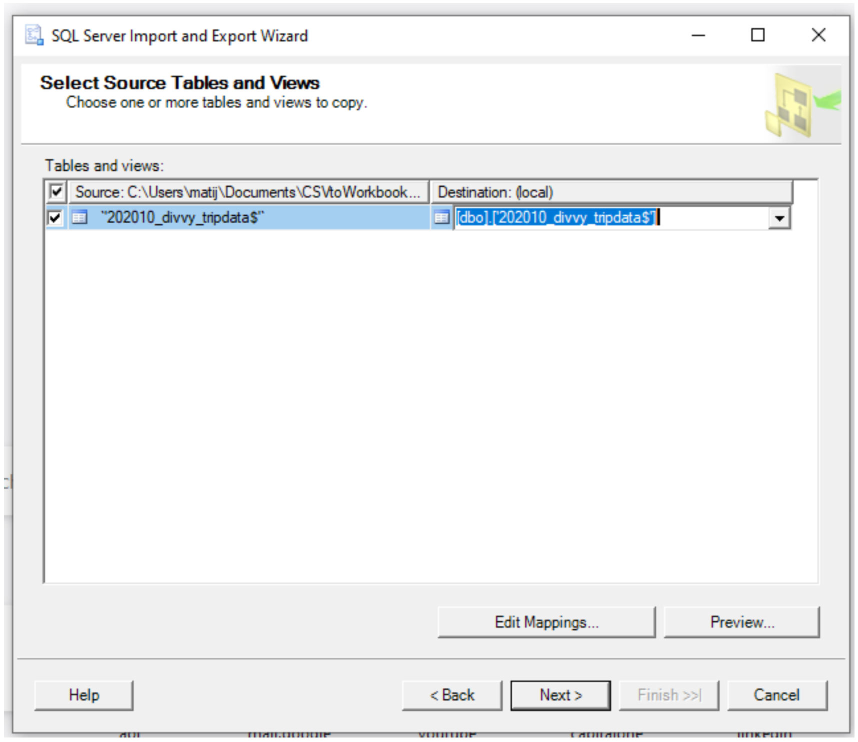
Specify whether to copy one or more tables and views or to copy the results of a query from the data source.

(●) **Copy data from one or more tables or views**

Use this option to copy all the data from the existing tables or views in the source database.

( ) **Write a query to specify the data to transfer**

Use this option to write an SQL query to manipulate or to restrict the source data for the copy operation.

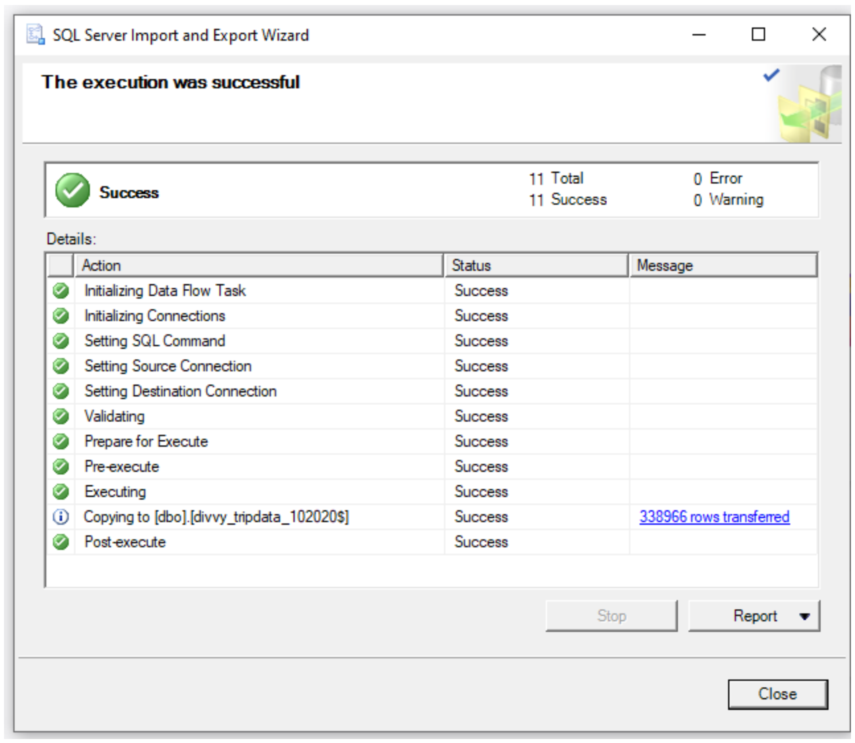

SQL Server Import and Export Wizard

Select Source Tables and Views

Choose one or more tables and views to copy.

Tables and views:

✓ Source: C:\Users\matij\Documents\CSVtoWorkbook...	Destination: (local)
✓ `202010_divvy_tripdata$`	[dbo].['202010_divvy_tripdata$']



SQL Server Import and Export Wizard

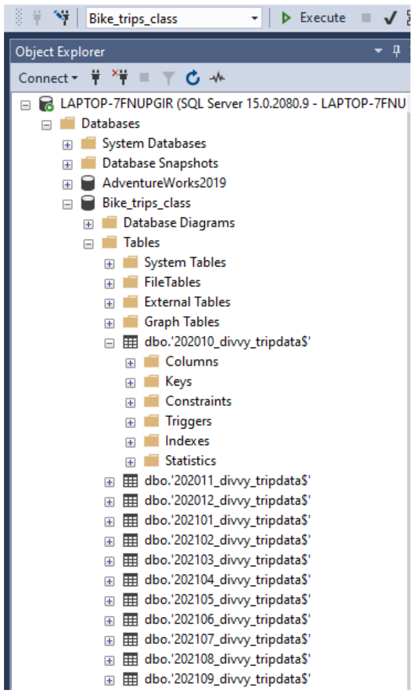
The execution was successful

Success — 11 Total, 11 Success, 0 Error, 0 Warning

Details:

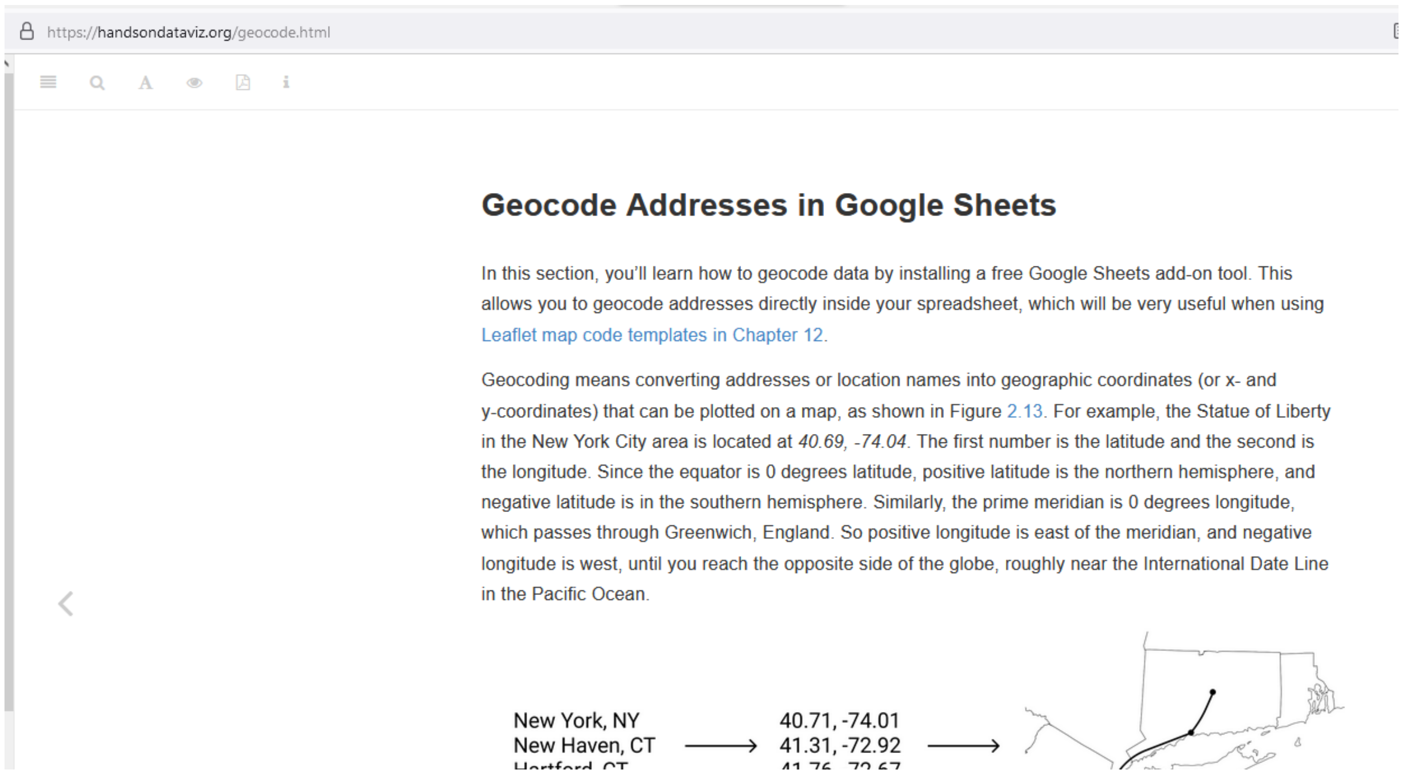
	Action	Status	Message
✓	Initializing Data Flow Task	Success	
✓	Initializing Connections	Success	
✓	Setting SQL Command	Success	
✓	Setting Source Connection	Success	
✓	Setting Destination Connection	Success	
✓	Validating	Success	
✓	Prepare for Execute	Success	
✓	Pre-execute	Success	
✓	Executing	Success	
ⓘ	Copying to [dbo].[divvy_tripdata_102020$]	Success	338966 rows transferred
✓	Post-execute	Success	

I uploaded all to Bike_trips_class, a database I created.



```
LAPTOP-7FNUPGIR (SQL Server 15.0.2080.9 - LAPTOP-7FNU
  Databases
    System Databases
    Database Snapshots
    AdventureWorks2019
    Bike_trips_class
      Database Diagrams
      Tables
        System Tables
        FileTables
        External Tables
        Graph Tables
        dbo.'202010_divvy_tripdata$'
          Columns
          Keys
          Constraints
          Triggers
          Indexes
          Statistics
        dbo.'202011_divvy_tripdata$'
        dbo.'202012_divvy_tripdata$'
        dbo.'202101_divvy_tripdata$'
        dbo.'202102_divvy_tripdata$'
        dbo.'202103_divvy_tripdata$'
        dbo.'202104_divvy_tripdata$'
        dbo.'202105_divvy_tripdata$'
        dbo.'202106_divvy_tripdata$'
        dbo.'202107_divvy_tripdata$'
        dbo.'202108_divvy_tripdata$'
        dbo.'202109_divvy_tripdata$'
```

I noted inconsistencies with latitudinal and longitudinal axes, and downloaded correct data via the Geocoding Add-in (Smart Monkey) for Google Sheets.

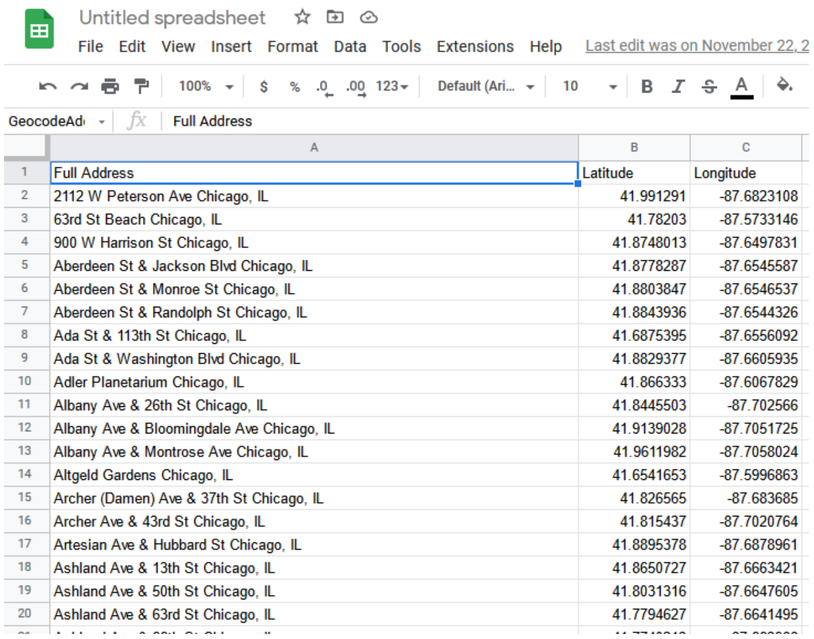


Geocode Addresses in Google Sheets

In this section, you'll learn how to geocode data by installing a free Google Sheets add-on tool. This allows you to geocode addresses directly inside your spreadsheet, which will be very useful when using Leaflet map code templates in Chapter 12.

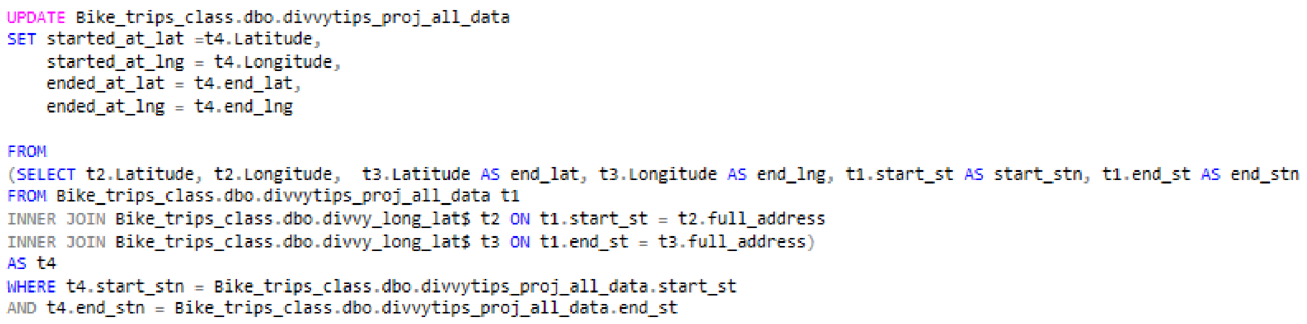
Geocoding means converting addresses or location names into geographic coordinates (or x- and y-coordinates) that can be plotted on a map, as shown in Figure 2.13. For example, the Statue of Liberty in the New York City area is located at *40.69, -74.04*. The first number is the latitude and the second is the longitude. Since the equator is 0 degrees latitude, positive latitude is the northern hemisphere, and negative latitude is in the southern hemisphere. Similarly, the prime meridian is 0 degrees longitude, which passes through Greenwich, England. So positive longitude is east of the meridian, and negative longitude is west, until you reach the opposite side of the globe, roughly near the International Date Line in the Pacific Ocean.

New York, NY
New Haven, CT
40.71, -74.01
41.31, -72.92


Untitled spreadsheet
File Edit View Insert Format Data Tools Extensions Help
Last edit was on November 22, 2
100% $ % .0 .00 123 Default (Ari... 10 B I S A
GeocodeAd fx Full Address

	A	B	C
1	Full Address	Latitude	Longitude
2	2112 W Peterson Ave Chicago, IL	41.991291	-87.6823108
3	63rd St Beach Chicago, IL	41.78203	-87.5733146
4	900 W Harrison St Chicago, IL	41.8748013	-87.6497831
5	Aberdeen St & Jackson Blvd Chicago, IL	41.8778287	-87.6545587
6	Aberdeen St & Monroe St Chicago, IL	41.8803847	-87.6546537
7	Aberdeen St & Randolph St Chicago, IL	41.8843936	-87.6544326
8	Ada St & 113th St Chicago, IL	41.6875395	-87.6556092
9	Ada St & Washington Blvd Chicago, IL	41.8829377	-87.6605935
10	Adler Planetarium Chicago, IL	41.866333	-87.6067829
11	Albany Ave & 26th St Chicago, IL	41.8445503	-87.702566
12	Albany Ave & Bloomingdale Ave Chicago, IL	41.9139028	-87.7051725
13	Albany Ave & Montrose Ave Chicago, IL	41.9611982	-87.7058024
14	Altgeld Gardens Chicago, IL	41.6541653	-87.5996863
15	Archer (Damen) Ave & 37th St Chicago, IL	41.826565	-87.683685
16	Archer Ave & 43rd St Chicago, IL	41.815437	-87.7020764
17	Artesian Ave & Hubbard St Chicago, IL	41.8895378	-87.6878961
18	Ashland Ave & 13th St Chicago, IL	41.8650727	-87.6663421
19	Ashland Ave & 50th St Chicago, IL	41.8031316	-87.6647605
20	Ashland Ave & 63rd St Chicago, IL	41.7794627	-87.6641495

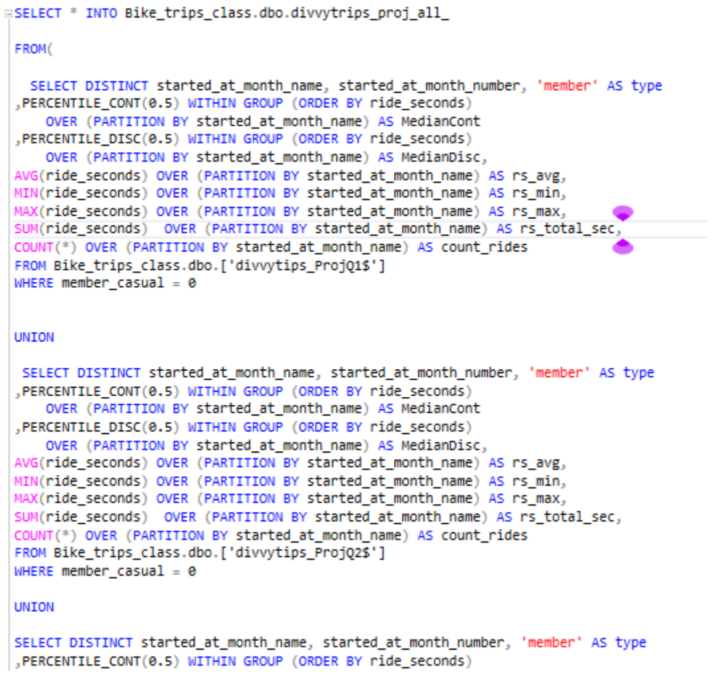
I transferred the resulting data into a table I created in my SQL database and uploaded it to my data tables.



```sql
UPDATE Bike_trips_class.dbo.divvytips_proj_all_data
SET started_at_lat =t4.Latitude,
    started_at_lng = t4.Longitude,
    ended_at_lat = t4.end_lat,
    ended_at_lng = t4.end_lng

FROM
(SELECT t2.Latitude, t2.Longitude,  t3.Latitude AS end_lat, t3.Longitude AS end_lng, t1.start_st AS start_stn, t1.end_st AS end_stn
FROM Bike_trips_class.dbo.divvytips_proj_all_data t1
INNER JOIN Bike_trips_class.dbo.divvy_long_lat$ t2 ON t1.start_st = t2.full_address
INNER JOIN Bike_trips_class.dbo.divvy_long_lat$ t3 ON t1.end_st = t3.full_address)
AS t4
WHERE t4.start_stn = Bike_trips_class.dbo.divvytips_proj_all_data.start_st
AND t4.end_stn = Bike_trips_class.dbo.divvytips_proj_all_data.end_st
```

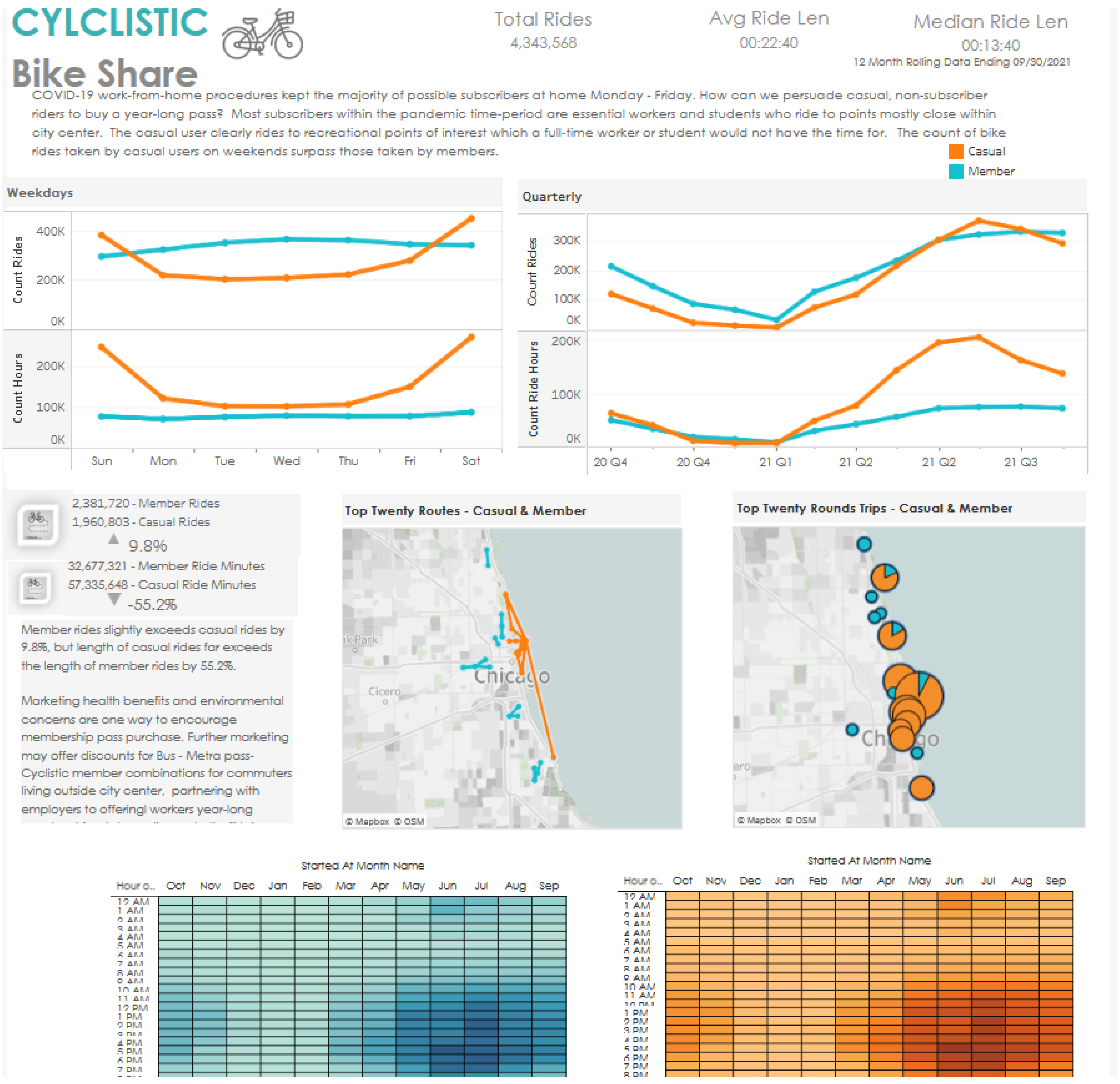
Further cleaning of data deleted rows without a start or end station, rows without a start or ending time, rows with ride times equal to or less than one minute, rows with ride times of equal to or greater than 24 hours, and rows with a start or end station of Test. I tested for repeats of ride ids and found approximately ten.

In my quest to find patterns within the data, I calculated the average, the minimum, the maximum, and the total ride seconds and the number of rides per month, for members and for casual users.



```sql
SELECT * INTO Bike_trips_class.dbo.divvytrips_proj_all_

FROM(

  SELECT DISTINCT started_at_month_name, started_at_month_number, 'member' AS type
,PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY ride_seconds)
    OVER (PARTITION BY started_at_month_name) AS MedianCont
,PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY ride_seconds)
    OVER (PARTITION BY started_at_month_name) AS MedianDisc,
AVG(ride_seconds) OVER (PARTITION BY started_at_month_name) AS rs_avg,
MIN(ride_seconds) OVER (PARTITION BY started_at_month_name) AS rs_min,
MAX(ride_seconds) OVER (PARTITION BY started_at_month_name) AS rs_max,
SUM(ride_seconds)  OVER (PARTITION BY started_at_month_name) AS rs_total_sec,
COUNT(*) OVER (PARTITION BY started_at_month_name) AS count_rides
FROM Bike_trips_class.dbo.['divvytips_ProjQ1$']
WHERE member_casual = 0


UNION

 SELECT DISTINCT started_at_month_name, started_at_month_number, 'member' AS type
,PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY ride_seconds)
    OVER (PARTITION BY started_at_month_name) AS MedianCont
,PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY ride_seconds)
    OVER (PARTITION BY started_at_month_name) AS MedianDisc,
AVG(ride_seconds) OVER (PARTITION BY started_at_month_name) AS rs_avg,
MIN(ride_seconds) OVER (PARTITION BY started_at_month_name) AS rs_min,
MAX(ride_seconds) OVER (PARTITION BY started_at_month_name) AS rs_max,
SUM(ride_seconds)  OVER (PARTITION BY started_at_month_name) AS rs_total_sec,
COUNT(*) OVER (PARTITION BY started_at_month_name) AS count_rides
FROM Bike_trips_class.dbo.['divvytips_ProjQ2$']
WHERE member_casual = 0

UNION

SELECT DISTINCT started_at_month_name, started_at_month_number, 'member' AS type
,PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY ride_seconds)
```

I entered data within Tableau and created a dashboard to answer the questions posed by the Marketing team. My suggestions for increasing membership purchases are included in the dashboard.



CYLCLISTIC Bike Share

Total Rides 4,343,568 | Avg Ride Len 00:22:40 | Median Ride Len 00:13:40

12 Month Rolling Data Ending 09/30/2021

COVID-19 work-from-home procedures kept the majority of possible subscribers at home Monday - Friday. How can we persuade casual, non-subscriber riders to buy a year-long pass? Most subscribers within the pandemic time-period are essential workers and students who ride to points mostly close within city center. The casual user clearly rides to recreational points of interest which a full-time worker or student would not have the time for. The count of bike rides taken by casual users on weekends surpass those taken by members.

2,381,720 - Member Rides
1,960,803 - Casual Rides
▲ 9.8%

32,677,321 - Member Ride Minutes
57,335,648 - Casual Ride Minutes
▼ -55.2%

Member rides slightly exceeds casual rides by 9.8%, but length of casual rides far exceeds the length of member rides by 55.2%.

Marketing health benefits and environmental concerns are one way to encourage membership pass purchase. Further marketing may offer discounts for Bus - Metra pass-Cyclistic member combinations for commuters living outside city center, partnering with employers to offeringl workers year-long