Omar Trejo

For my ETL Project, I was curious to see if the increasing number of bike sharing systems are having an effect on public transportation usage. I was unable to find data for the Bcycle bike system in Houston, but I found both bike sharing and public transportation data for Chicago. I figured that would work better also since Chicago has much higher public transportation usage. My two data sources were csv files from the Divvy Bike Sharing system website in Chicago and City of Chicago Portal.

The Divvy bike csv files were very large. Their data was split by quarters in the year and the csv files were almost 2 GBs for data from 2013-2018. Since we aren’t actually conducting an analysis for this project, I decided to only use two csv files for the first quarters in 2017 and 2018. I cleansed the data by removing columns that were not relevant. Every bike rental trip had its own trip ID record broken out by day, so I took the count of the trips and grouped them by day to get the total number of trips by day. The public transportation data was already pretty clean. It included data on total rides from 2001-2018, so I had to filter the pandas dataframe to only include the data from 2017-2018. I saved both dataframes as csv files and loaded them in MySql to join both tables and create the final database.

If I was conducting an analysis on this data, I would have included more years to get a much better comparison, but hypothetically this data can be used to perform an analysis that checks to see if bike mobility is becoming a viable transportation option in cities where public transportation has been dominant.