Katheryn Raatz, Jen Barny, Rajitha Baddam

ETL Project Summary

For our ETL project we chose to focus on the subject of NYC taxis in relation to the weather going on at that time. The data was extracted from the following sources:

1. NYC Yellow and Green Taxi <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
2. Taxi zones <https://chih-ling-hsu.github.io/2018/05/14/NYC>
3. Weather Underground (<https://Wunderground.com>)

Four different files were used: NYC Yellow Taxi data, NYC Green Taxi data, Taxi Zones data, and Weather Underground data. The first three datasets are all csv, while the last dataset from Wunderground was acquired using web-scraping. All of the datasets were acquired specifically for the month of June 2019 (6/1/19-6/30/19). Given that the datasets for NYC Yellow Taxis and NYC Green Taxis was very large it made the most sense to focus on only one month.

The Yellow Taxi trip data and the Green Taxi trip data include the fields pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, location ID, itemized fares, rate types, payment types, driver-reported passenger counts, etc. The data was collected by the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The date and time are included together as one column up to the second of pick-up and drop-off.

The Taxi Zone data includes the location ID, the borough, the Zone, and service zone. This data set was utilized in order to add meaning to the location ID values in the Yellow Taxi and Green Taxi trip data by connecting it to a physical location.

The Wunderground data was scraped off the website wunderground.com in order to collect the historical weather conditions that occur during the month of June 2019. The data set includes the temperature, dew point, humidity, wind, wind speed, wind gust, pressure, precipitation, and weather condition of each hour and day in the month of June.

The csv files utilized for this project are all located within the Resources folder. We created two jupyter files, in wunder\_scraper.ipynb we scraped the historical weather data from Weather Underground (wunderground.com) and in extract\_transform\_load.ipynb we used the datasets from all sources and formed them into one dataframe.

To begin, the datasets for the NYC Yellow Taxi data and NYC Green Taxi data were each uploaded onto pandas. They were both converted into datetime format and cleaned to only included data in the date range of June 2019. A new column was also created called Taxi Type to distinguish the yellow (Y) and green (G) taxis. Columns that were seen as not useful were dropped on both datasets and the columns datetime, LocationID, passenger\_count, trip\_distance, fare\_amount, total\_amount, congestion\_surcharge, payment\_type, and Taxi Type were kept. Then after the datasets were both individually formatted, they were then added together, appending the Green Taxi data to the Yellow Taxi data. The datetime column is the pickup time and date. We chose to focus on the pickup datetime in our analysis and decided to drop the drop-off datetime column because we believed looking at both pickup and drop-off would make our final dataframe too hectic and more difficult to draw conclusions from.

Then following that the Taxi Zone csv file was uploaded as a dataframe, then merged together with the previous dataframe that includes the yellow and green taxi data all together. An inner merge was performed, only keeping the common values for both dataframes based on the column LocationID in order to match the borough with the given LocationID from the NYC taxi data. The Zone and service zone variables were dropped from the dataset because they were not seen to add any additional value to our research that was not already provided by the borough column.

Finally, the weather dataset was uploaded and merged with the last dataframe to create a single dataframe with all our information. However, before doing that it was necessary to match the format of datetime as was done for the taxi data by concatenating the two columns Date and Time. By combining date and time into one variable it allowed us to merge the two dataframes using that column. Differing from our first merge we had to utilize and asof merge so that the datetimes would match on the nearest key rather than on equal keys. This made it so that even though the pickup time and datetime for the weather didn’t match exactly it would match them to the nearest value allowing a merge to occur between the two dataframes.

The final dataframe includes the information from all our sources, with information on NYC taxis and the weather that occurred during those trips. Possible analyses that may be conducted could be comparing columns between the green and yellow taxis, making conclusions based on weather and determining if weather affects the amount of taxi trips given, etc. These are just a couple of the possible angles that can be explored using the dataframe we created.