Workflow Documentation – Blumen Data Engineer – NYC Taxi Trip Data Pipeline

1. Downloaded the data (Yellow and Green Taxi Trip records, Taxi Zone shapefile, and NTAs shapefile) and read the Trip Record User Guide and parquet file field guides.

2. Installed a parquet reader for pandas to read and manipulate the files.

3. Began by merging the parquet data, viewing to see what needed to be cleaned.

4. Merged the pickup and drop off fields from the yellow and green taxis.

5. Replaced all the NaN values with 0 for the calculations later on.

6. Looked at the taxi zones and NTAs in QGIS and realized that there were not going to be clean joins so when I did that join, I just chose whichever the join found first and removed the duplicates. (Only one neighborhood per taxi zone.)

7. I removed duplicate taxi record -- rides that were disputed or no charge had negative charges - so I removed those for the distance and trip counts.

8. I calculated the fares with the disputes/no charges since I figured they should count toward the overall cost.

9. Added a function to calculate the trip counts, average distance and average fares per NTA and added each of those to the corresponding NTA record in the shapefile so that it can be viewed in QGIS.

10. For peak times, I isolated the hour from the datetime pickup field, calculated the frequency of those hours, then whichever was the highest would be the peak times.

**The attached Summary Report CSV contains all the trip counts, average distances and fares for each of the NTAs.**

**The peak hours I found with the two taxi records** (green\_tripdata\_2024-01.parquet , yellow\_tripdata\_2024-01.parquet ) **were between 18-19:00.**