ETL Project: Technical Write-Up
Team 20 - NY Registrations and NYC Collisions

OVERVIEW:
We were interested in uncovering insights on car accidents in New York City. To do this we sourced data from NY State DMV vehicle registrations and NYPD vehicle collision records. At the intersection of these datasets one could answer questions like: Are certain vehicle types disproportionately involved in collisions? Is there a relationship between number of registered vehicles and number of collisions in a neighborhood?

NY REGISTRATIONS:

EXTRACT:
Data Source:
NY State DMV
https://data.ny.gov/Transportation/Vehicle-Snowmobile-and-Boat-Registrations/w4pv-hbkt

CSV file with data: Vehicle__Snowmobile__and_Boat_Registrations.csv

PDF with body type definitions:
https://data.ny.gov/api/views/w4pv-hbkt/files/AUsdC2Y0iEymGyebFASIjDxZ7irrm1-_yS-o9qFzWTQ?download=true&filename=NYSDMV_VehicleSnowmobileAndBoat_Registration_Data%20Dictionary.pdf'

TRANSFORM:
- Imported registration CSV into pandas (12 million rows)
- Eliminated any rows other than those for vehicles (down to 10 million rows)
- Dropped columns not relevant to our analysis
- Aggregated and renamed the “body type” values
    - used DMV body type PDF to decode
    - In excel created a list of body types and identified those that should be combined and renamed
    - Final list of body type names and combinations were then used to rename the types in pandas.
    - The same types were also then used for collision dataframe transformation so we could use to join the tables.
- Eliminated any rows other than those for NYC counties (down to 2 million rows)
- Renamed columns in snake case
- Note: Throughout transformation, we had to troubleshoot trailing space errors in column values. Ultimately we used .str.strip() to eliminate this issue where necessary.

LOAD:
Created SQL database using Postgres and TablePlus
Created connection and loaded database from python using SQLalchemy

NYC COLLISIONS:

EXTRACT
Data Sources:
NYC Open Data
https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
CSV file with data:
NYPD_Motor_Vehicle_Collisions.csv

TRANSFORM:

- Imported registration CSV into pandas (1.5 million rows)
- Renamed the rows for easier indexing
- Renamed index to id
- Dropped unwanted columns
- Removed row with blanks using .dropna()
- Filtered out the data only keeping vehicle types based on value counts
- Aggregated and renamed the “vehicle type” values
    - Matched the vehicle types with the body types from NY Registrations data to be able to join tables.

LOAD:
Created SQL database using Postgres and TablePlus
Troubleshoot errors importing to TablePlus

FURTHER ANALYSIS:
With these tables now in our database, one could perform a number of operations.

Crucially, now that we have transformed the data in both tables, we can now join the data geographically (on zip code) or by vehicle type.