ETL Project

By Dany Koudede

1. **Extraction**

I extracted 2 datasets (CSV files) from Data.Gov that contain details on the City of New-York vehicle crashes from April 2016 to July 2020. One of the datasets contains details of people involved in the crash, each row represents a person (driver, occupant, pedestrian, bicyclist) and the second dataset contains details on each vehicle involved in the crash. Each row represents a motor vehicle involved in the crash.

References:

* Motor Vehicle Collision-Vehicle from Data.Gov

<https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles>

* Motor Vehicle Collision-Person from Data.Gov

<https://catalog.data.gov/dataset/motor-vehicle-collisions-person>

1. **Transformation**

I had first started by extracting the desired data from the original datasets by dropping unwanted columns, and created three tables in the process:

* **Vehicle table**

The table is centered on the car 's information. This table took birth from the Motor Vehicle Collision-Vehicle ‘s dataset.

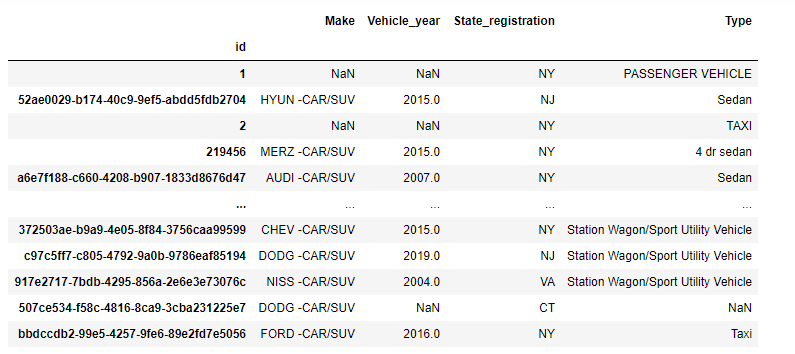
* **Person table**

The table is centered on the person's information. The Person table took birth from the Motor Vehicle Collision-Person’s dataset.

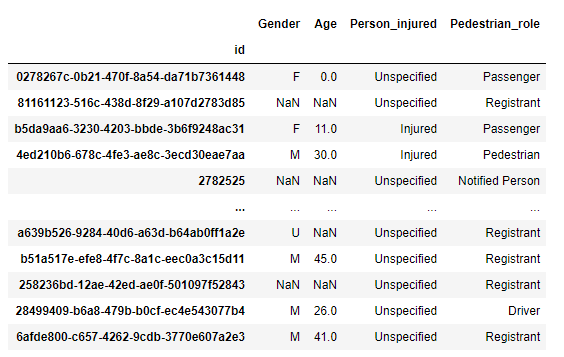
* **Collision table**

The table focuses on the data related to the event. The collision table from both the Motor Vehicle Collision-Vehicle and from the Motor Vehicle Collision-Person’s dataset.

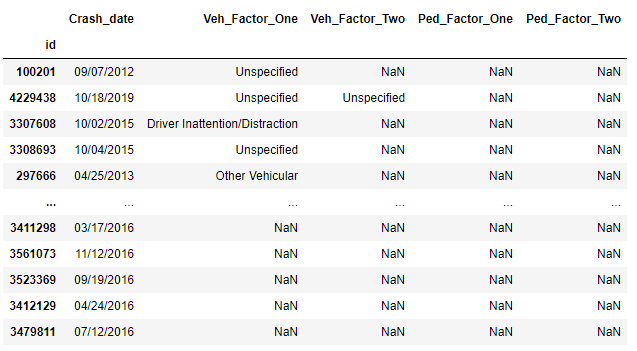
**Vehicle table**



* **Person table**

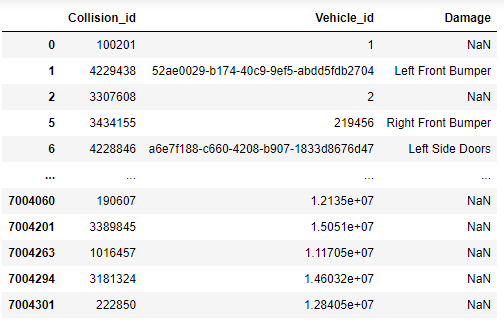


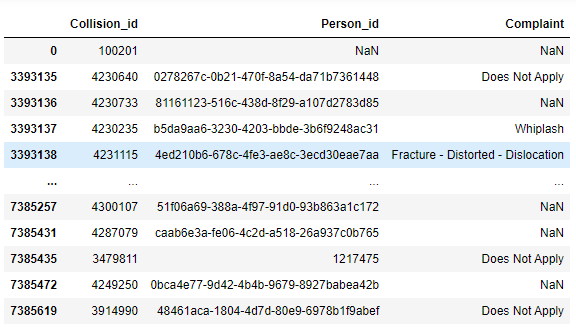
* **Collision table**

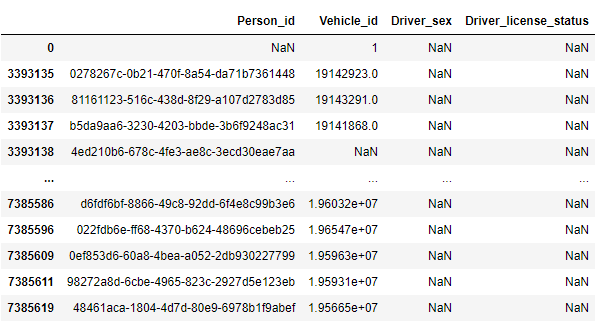


**Junction tables**

After getting new datasets from the principal ones I then created junction tables where I also stored attribute(s) of the relationship between tables before cleaning.







**Cleaning**

For the cleaning I followed the same process for every table:

-Rename the columns in a more suitable name and font

- remove all empty rows in non-unique ids (person\_id and vehicle\_id)

- remove any duplicates

- set indexes

1. **Load**

The last part consisted of transferring the final results to a database. For this process I used Postgres, and here are the steps followed:

I created a new database (NYC\_colllisin.db)

Created a schema based on the related tables with the proper data format and constraints

Connected to the local database with pandas

Checking the existence of the tables

Used pandas to load CSV converted DataFrame into the database

Confirm that data has been added by querying all tables in Jupiter notebook

* **Summary**

To sum up, the end goal of the work process was to create a consistent relationship between the two datasets in an initiative scheme that took place to see how the person table can relate to the vehicle table.