NY Motor Vehicle Collision Model Discrepancies/Observations and Cleaning

* We replaced null values with “No Value Provided” in all the Dimensions keeping -99 as SK.
* We observed discrepancies in vehicle occupants, for example 2000 occupants. Since most of the data was wrong, we did not proceed with correcting/replacing.
* Redundant directions in Travel\_Direction were changed to a homogeneous format.
* Vehicle\_Year with nulls and values less than 1900 or greater than 2023 with 9999.
* Respective excel sheets are maintained for vehicle\_type\_code, vehicle\_make, vehicle\_model, vehicle\_year with corrected data since these have large number of discrepancies. We implemented a lookup with these excel files instead of tReplace because it took a lot of memory on our systems.
* Contributing factors which had numbers and spelling errors were replaced with “Unknown” or the correct spelling.
* Driver\_License\_Jurisdiction format was maintained with a two-character format and invalid inputs such as special characters for example “-“, “PA’” etc were corrected or replaced with “Invalid Input”.
* Person\_Age with negative/large values have been replaced with 0.
* It was observed that certain Collision\_ID(s) in stg\_NYC\_MV\_Collision\_Person was not present in stg\_NYC\_MV\_Collision\_Vehicle but present in stg\_NYC\_MV\_Collision\_Crashes. So we have assumed that there is Vehicle data missing for these Person\_IDs

Assumptions

* Contributing Factor in Vehicle is assumed to be the correct version since we observed most of these values matching in both person and vehicle table, but for some collision IDs there were empty values in only person table but not in vehicle table.