DAMG7370 Fall 2024 – Final Project – Motor Vehicle Collision Analysis

New York City

Description about the data:

The Motor Vehicle Collisions Crash Dataset contains details of police-reported motor vehicle collisions in NYC. Each row represents a crash event documented using the MV-104AN form, required for incidents involving injuries, fatalities, or damages over $1,000.

The data is sourced from the Finest Online Records Management System (FORMS), implemented in 2016 to electronically record and store all crash details, enabling detailed traffic safety analyses. This dataset supports initiatives like Vision Zero, which aims to eliminate traffic fatalities.

Source: <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data>

Last Updated: November 28, 2024

No. of Rows: 2139048 Records

No. of Columns: 29

Description of columns (as observed in Metadata):

|  |  |
| --- | --- |
| Name | Description |
| BOROUGH | Borough where collision occurred |
| COLLISION\_ID | Unique record code generated by system. Primary Key for Crash table. |
| CONTRIBUTING FACTOR VEHICLE 1 | Factors contributing to the collision for designated vehicle |
| CONTRIBUTING FACTOR VEHICLE 2 | Factors contributing to the collision for designated vehicle |
| CONTRIBUTING FACTOR VEHICLE 3 | Factors contributing to the collision for designated vehicle |
| CONTRIBUTING FACTOR VEHICLE 4 | Factors contributing to the collision for designated vehicle |
| CONTRIBUTING FACTOR VEHICLE 5 | Factors contributing to the collision for designated vehicle |
| CRASH DATE | Occurrence date of collision |
| CRASH TIME | Occurrence time of collision |
| CROSS STREET NAME | Nearest cross street to the collision |
| LATITUDE | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |
| LOCATION | Latitude , Longitude pair |
| LONGITUDE | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |
| NUMBER OF CYCLIST INJURED | Number of cyclists injured |
| NUMBER OF CYCLIST KILLED | Number of cyclists killed |
| NUMBER OF MOTORIST INJURED | Number of vehicle occupants injured |
| NUMBER OF MOTORIST KILLED | Number of vehicle occupants killed |
| NUMBER OF PEDESTRIANS INJURED | Number of pedestrians injured |
| NUMBER OF PEDESTRIANS KILLED | Number of pedestrians killed |
| NUMBER OF PERSONS INJURED | Number of persons injured |
| NUMBER OF PERSONS KILLED | Number of persons killed |
| OFF STREET NAME | Street address if known |
| ON STREET NAME | Street on which the collision occurred |
| VEHICLE TYPE CODE 1 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) |
| VEHICLE TYPE CODE 2 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) |
| VEHICLE TYPE CODE 3 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) |
| VEHICLE TYPE CODE 4 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) |
| VEHICLE TYPE CODE 5 | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) |
| ZIP CODE | Postal code of incident occurrence |

Data Profiling Details:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Data Type | % Missing | Nulls | Unique Values | Shortest / Min Value |
| BOROUGH | V\_WString | 31.10% | 663,744 | 7 | 0 |
| COLLISION\_ID | Int64 | 0.00% | 2 | 2,137,457 | 22 |
| CONTRIBUTING FACTOR VEHICLE 1 | V\_WString | 0.30% | 7,234 | 62 | 1 |
| CONTRIBUTING FACTOR VEHICLE 2 | V\_WString | 15.70% | 336,056 | 62 | 1 |
| CONTRIBUTING FACTOR VEHICLE 3 | V\_WString | 92.80% | 1,983,694 | 53 | 1 |
| CONTRIBUTING FACTOR VEHICLE 4 | V\_WString | 98.40% | 2,102,495 | 43 | Glare |
| CONTRIBUTING FACTOR VEHICLE 5 | V\_WString | 99.60% | 2,127,923 | 32 | Glare |
| CRASH DATE | Date | 0.00% | 2,137,458 | 4,531 | 7/27/2012 |
| CRASH TIME | Time | 0.00% | 2,137,458 | 1,442 | 0 |
| CROSS STREET NAME | V\_WString | 38.10% | 814,863 | 15,291 | 1 |
| LATITUDE | Float | 11.20% | 239,426 | 127,595 | 0 |
| LOCATION | V\_WString | 11.20% | 239,426 | 296,850 | 0 |
| LONGITUDE | Float | 11.20% | 239,426 | 99,039 | 0 |
| NUMBER OF CYCLIST INJURED | Int64 | 0.00% | 1 | 7 | 0 |
| NUMBER OF CYCLIST KILLED | Int64 | 0.00% | 2 | 5 | 0 |
| NUMBER OF MOTORIST INJURED | Int64 | 0.00% | 2 | 33 | 2 |
| NUMBER OF MOTORIST KILLED | Int64 | 0.00% | 2 | 7 | 0 |
| NUMBER OF PEDESTRIANS INJURED | Int64 | 0.00% | 2 | 15 | 0 |
| NUMBER OF PEDESTRIANS KILLED | Int64 | 0.00% | 2 | 6 | 0 |
| NUMBER OF PERSONS INJURED | Int64 | 0.00% | 20 | 34 | 2 |
| NUMBER OF PERSONS KILLED | Int64 | 0.00% | 33 | 8 | 0 |
| OFF STREET NAME | V\_WString | 82.90% | 1,771,529 | 208,857 | LIE |
| ON STREET NAME | V\_WString | 21.40% | 458,302 | 14,444 | 0 |
| VEHICLE TYPE CODE 1 | V\_WString | 0.70% | 14,689 | 1,741 | . |
| VEHICLE TYPE CODE 2 | V\_WString | 19.50% | 417,109 | 1,932 | 0 |
| VEHICLE TYPE CODE 3 | V\_WString | 93.10% | 1,989,458 | 278 | PK |
| VEHICLE TYPE CODE 4 | V\_WString | 98.40% | 2,103,743 | 109 | PK |
| VEHICLE TYPE CODE 5 | V\_WString | 99.60% | 2,128,218 | 74 | C3 |
| ZIP CODE | Int64 | 31.10% | 664,048 | 237 | 0 |

Additional Notes / Inferences:

1. The Record at row number 1269803 has a missing close quote and this has been fixed in the dataset.
2. For columns which might help in finding the address or location such as latitude, longitude, zipcode, on\_street\_name, cross\_street\_name, off\_street\_name all have over 10 % null values on average. These columns need to tended, we can derive the accident\_reported\_address with a combination of these columns. Even then, around 49k records are present where all the said columns have null values.
3. Crash\_Time column follows 24 hr format, we can include AM/PM based on the time so that this column can be standardized and uniform with other dataset’s data.
4. Collision\_ID is a unique identifier column in this dataset. However, upon careful observation it is found that few records are having nulls, few records doesn’t have the uniform length ID. This can be addressed by coming up with our own PK standard and format so that each record is uniquely identified.
5. We can derive the zip code values from Borough column.
6. Vehicle type codes column needs to be categorized based on broad category.
7. There is no record where the location, longitude is null but the location is not null.
8. If on street name is null, then the off street name could be considered. If the off street name is also null, then we can consider the cross street name.
9. For around 8.9k records, the number of injured person does not match the sum of injured motorists / pedestrians / cyclists. This could be because there could be someone who got injured – but did not fit into the mentioned three categories.