# **Profiling Report- New York Motor Vehicles Collision Report**

This profiling report offers a comprehensive overview of a dataset concerning traffic collisions in New York, encapsulating a wide range of data points from crash dates and times to specific details about the involved vehicles. The detailed analysis is provided below

**Dataset Statistics:**

1. Number of Variables:

The dataset contains 2,075,427 observations across 29 variables, indicating a broad dataset covering numerous aspects of traffic collisions

2. Duplicate rows:

There are no duplicate rows which ensures the uniqueness of each record.

3. Memory allocation:

The dataset occupies 459.2 MiB in memory, with an average record size of 232 bytes.

4. Missing Data:

Approximately 29.5% of cells are missing, totaling 17,761,579 missing values, which is significant and could impact the analysis

5. Variable Types:

The dataset includes a variety of variable types: 2 Date Time, 6 Categorical, 1 Unsupported, 8 Numeric, and 12 Text variables.

**Unique Values:-**

The COLLISION\_ID variable has unique values, indicating each entry represents a distinct collision incident.

**Following are the variables in the New York dataset with their observations:**

1. Crash date:-  
The dataset contains 4283 unique values for the crash date.  
There are no missing values for the crash date variable as indicated by a count of 0.

2. Crash time:-  
It has 0.1% which is 1440 unique values for crash time.  
There are 0 missing values for the crash time variable.

3. Borough:-  
It is categorical with 5 distinct values namely Brooklyn, Queens, Manhattan, Bronx, and Staten Island.  
It has 31.1 % which is 645746 missing values reported.

4. Latitude:-  
There are 126594 unique records present for the Latitude variable.  
It has 11.35 missing values which are 233626 missing values.

5. Longitude:-  
There are 98351 unique records present for the Latitude variable.  
It has 11.3% missing values which are 233626 missing values.

6. Location:-  
There are 283006 unique records present for the Latitude variable.  
It has 11.35% of missing values which is 233626 of missing values.

7. On-street name:-  
This variable contains 1.1% of distinct values which is 18410.  
It contains 21.2% of missing data which is 440569 values.

8. Cross street name:-  
Cross street name has 20236 unique values.  
It contains 784436 missing records which is 37.8% of missing values.

9. Off street name:-  
The off-street name has 225845 unique values.  
It contains 1727231 missing records which is 83.21% of missing values.

10. Number of persons injured:-  
The number of persons injured has a high correlation with no of motorists injured.  
It has 32 distinct values present.  
There are 18 missing values in the number of persons injured.

11. Number of persons killed:-  
it has an overall high correlation with the number of motorists killed and the number of pedestrians killed.  
It has 7 distinct and 31 missing values present overall in the dataset.

12. Number of pedestrians injured:-  
It has 14 unique records present.  
No of pedestrians injured have zero missing values.

13. Number of pedestrians killed:-  
This variable shows a high correlation with the number of cyclists killed and the number of persons killed variable.  
It has 4 distinct values and 0 missing values present.

14. Number of cyclists injured:-  
It is a categorical variable with values (0, 1, 2, 3, and 4) with distinct and 0 missing values present in the dataset.

15. Number of cyclists killed:-  
This is a categorical variable with values (0, 1, and 2).  
A high correlation with the number of pedestrians killed is present.  
There are 3 distinct and 0 missing values present.

16. Number of motorists injured:-  
31 distinct records are present for the Number of motorists killed.  
There are 0 missing values detected for this particular column.

17. No of motorists killed:-  
The no of motorists killed has 6 distinct values.

There are 0 missing g values reported on the number of motorists killed in the dataset.

18. Contributing factor Vehicle 1:-  
This particular variable has 61 distinct records.  
The no of missing values for this variable is 6802.

19. Contributing factor Vehicle 2:-  
There are 61 distinct values.  
There are 321736 missing values reported which is 15.5%.

20. Contributing factor Vehicle 3:-  
There are 51 distinct values.  
There are 1927163 missing values reported which is 92.9%.

21. Contributing factor Vehicle 4:-  
It is a categorical variable that has a high correlation with contributing factor vehicle 5.  
There are 41 distinct values.  
There are 2041953 missing values reported which is 98.4%.

22. Contributing factor Vehicle 5:-  
It is a categorical variable that has a high correlation with contributing factor vehicle 4.  
There are 30 distinct values.  
There are 2066358 missing values reported which is 98.4%.

23. Collision Id:-  
Collision ID has 2075427 unique values present which is 100% distinct.  
It has 0 missing values reported.

24. Vehicle Type code 1:-  
This variable has 1631 distinct values present.  
The number of missing values is 13691 which is 0.7%.

25. Vehicle Type code 2:-  
This variable has 1819 distinct values present.  
The number of missing values is 396691 which is 19.1%.

26. Vehicle Type code 3:-  
This variable has 260 distinct values present.  
The number of missing values is 1932530 which is 93.1%.

27. Vehicle Type code 4:-  
This variable has 101 distinct values present.  
The number of missing values is 2043115 which is 98.4%.

28. Vehicle Type code 5:-  
This variable has 70 distinct values present.  
The number of missing values is 2066635 which is 99.6%.

**Observations:-**

1. **Observation1( Zip Code)**

The zip code has blank space entries for **42** numbers of rows.

**Suggestion**- Take Zip code as a string and convert it into int data type after replacing the blank spaces with **NA** (Cleaning).

1. **Observation2 (Address)**

The address has a Tab space in between the characters for 1 row.

**Suggestion**- Use the Escape character setting in the delimiter file. Use String. trim() to access blank spaces.

1. **Observation3 (Contributing Factor Vehicle)**

Contributing Factor vehicle has a number instead of a proper description (80).

**Suggestion-** Ithas been handled contribution factor mapping document provided by Tas.

1. **Observation4 (CFV -Vehicle Type Code)**

There is a relation between Contributing Factor Vehicle and Vehicle Type code. Which needs to be ignored.

**Overall Insights:-**

This data profiling document serves as a foundational analysis of the New York Traffic Accident dataset. It highlights the dataset's comprehensive coverage of traffic incidents across various dimensions while also pointing out areas requiring attention, such as missing data and variable imbalances, with significant correlations found between the number of individuals harmed in incidents and specific contributing factors.