PROFILING DOCUMENTATION FOR NEW YORK DATASET

The crash table from the Motor Vehicle Collisions dataset details each crash event, where every row is a distinct crash event. This data, sourced from all police-reported motor vehicle collisions in NYC, holds critical information for fulfilling business objectives. Key insights from the dataset include:

1. Total Number of Fields: The dataset is composed of 29 unique fields, including a special field 'COLLISION_ID' that serves as a unique identifier.
2. Number of Observations: There are total 2.08M(20,75,427) records, each record representing an individual Motor Vehicle Collision event.
3. Missing Cells: The dataset has a substantial number of missing entries, with 177,615,79 cells lacking data. This makes up about 29.5% of the data, a substantial proportion that may affect analyses.
4. Duplicate Rows: There is an absence of duplicate rows, with the count being zero, confirming the uniqueness of each data entry.
5. Total Size in Memory: It has a memory size of 459.2 MiB, indicating the storage space it occupies.
6. Average Record Size in Memory: On average, a single record takes up 232.0 bytes in memory, a consideration for evaluating data handling and processing performance.

COLUMN WISE ANALYSIS :

The dataset includes various variable types: 2 Date Time, 6 Categorical, 8 Numeric, and 13 Text variables.

1. CRASH DATE:

- The data field 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.
- No. of Distinct: 4283
- Min Value: 2012-07-01 00:00:00
- Max Value: 2024-03-22 00:00:00

2. CRASH TIME:

- It has 0.1% which is 1440 unique values for CRASH TIME(hh:mm).
- There are 0 missing values for the crash time variable.
- No. of Distinct: 1440
- Minimum: 00:00

- Maximum: 23:59

3. BOROUGH:

- It is a categorical field with 5 distinct values namely Brooklyn, Queens, Manhattan, Bronx, and Staten Island.
- It has 31.1 % (645746) missing values reported.
- No. of Distinct: 5
- Categories: 5
- a.Brooklyn b.Queens c.Manhattan d.Bronx e.Staten Island
- The missing values would be replaced by "NA"

4. LATITUDE:

- There are 126594 unique records present for the Latitude variable.
- It has 11.3% missing values which are 233626 missing values.
- No. of Distinct: 126594
- Min Value: 0
- Max Value: 43.34444

5. LONGITUDE:

- There are 98351 unique records present for the Latitude variable.
- It has 11.3% missing values which are 233626 missing values.
- No. of Distinct: 98351
- Min Value: -201.3599
- Max Value: 0
- Both LATITUDE and LONGITUDE would be converted to datatype float with precision 4 at the cleaning stage and the missing values would be replaced by -1.

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.
- No. of Distinct: 283006
- The missing values would be replaced by "NA"

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.
- No. of Distinct: 18410
- The missing values would be replaced by “NA”
- The special characters present in this column is . / & ‘ # , @

8. CROSS STREET NAME:

- Cross street name has 20236 unique values.
- It contains 784436 missing records which is 37.8% of missing values.
- No. of Distinct: 20236
- The missing values would be replaced by “NA”
- The special characters present in this column is . / & ‘ ? , @

9. OFF STREET NAME:

- The off-street name has 225845 unique values.
- It contains 1727231 missing records which is 83.2% of missing values.
- It contains a record for COLLISION ID: 3291249 where a tab space is contained between the text.
- No. of Distinct: 225845
- The missing values would be replaced by “NA”
- The special characters present in this column is . / & ‘ # , @

10. NUMBER OF PERSONS INJURED:

- It has 32 distinct values present.
- There are 18 missing values in the number of persons injured.
- The number of persons injured has a correlation with the number of motorists injured.
- No. of Distinct: 32
- Min Value: 0
- Max Value: 43
- The missing values in this column would be replaced by -1

11. NUMBER OF PERSONS KILLED:

- It has 7 distinct and 31 missing values present overall in the dataset.
- No. of Distinct: 7
- Min Value: 0
- Max Value: 8

- The missing values in this column would be replaced by -1
- It has a correlation with the number of motorists killed and the number of pedestrians killed.

12. NUMBER OF PEDESTRIANS INJURED:

- It has 14 distinct records present and has zero missing values.
- No. of Distinct: 14
- Min Value: 0
- Max Value: 27
- The missing values in this column would be replaced by -1

13. NUMBER OF PEDESTRIANS KILLED:

- It has 4 distinct values, and 0 missing values present.
- No. of Distinct: 4
- Min Value: 0
- Max Value: 6
- The missing values in this column would be replaced by -1
- This numeric variable shows a correlation with the number of cyclists killed and the number of persons killed variable.

14. NUMBER OF CYCLISTS INJURED:

- It is a numeric variable with 5 distinct values and 0 missing values present in the dataset.
- No. of Distinct: 5
- Min Value: 0
- Max Value: 4
- The missing values in this column would be replaced by -1

15. NUMBER OF CYCLISTS KILLED:

- This is a numeric variable with 3 distinct values and zero missing values
- No. of Distinct: 3
- Min Value: 0
- Max Value: 2
- The missing values in this column would be replaced by -1

16. NUMBER OF MOTORISTS INJURED:

- This is a numeric variable with 31 distinct values and zero missing values
- No. of Distinct: 31
- Min Value: 0
- Max Value: 43
- The missing values in this column would be replaced by -1

17. NO OF MOTORISTS KILLED:

- The no of motorists killed has 6 distinct values and zero missing values
- No. of Distinct: 6
- Min Value: 0
- Max Value: 5
- The missing values in this column would be replaced by -1

18. CONTRIBUTING FACTOR VEHICLE 1:

- This variable has 61 distinct records.
- The number of missing values for this variable is 6802.
- No. of Distinct: 61
- The missing values in this column would be replaced by "NA"
- The special characters present in this column is - / ( )
- It also contains a numeric value which will be mapped later using the contribution factor mapping document

19. CONTRIBUTING FACTOR VEHICLE 2:

- There are 61 distinct values.
- There are 321736 missing values reported which is 15.5%.
- No. of Distinct: 61
- The missing values in this column would be replaced by "NA"
- The special characters present in this column is - / ( )
- It also contains a numeric value which will be mapped later using the contribution factor mapping document

20. CONTRIBUTING FACTOR VEHICLE 3:

- There are 51 distinct values.
- There are 1927163 missing values reported which is 92.9%.
- No. of Distinct: 51
- The missing values in this column would be replaced by "NA"
- The special characters present in this column is - / ( )

- It also contains a numeric value which will be mapped later using the contribution factor mapping document

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%.
- No. of Distinct: 41
- The missing values in this column would be replaced by "NA"
- The special characters present in this column is - / ( )
- It also contains a numeric value which will be mapped later using the contribution factor mapping document

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%.
- No. of Distinct: 30
- The missing values in this column would be replaced by "NA"
- The special characters present in this column is - / ( )
- It also contains a numeric value which will be mapped later using the contribution factor mapping document

We have five columns related to the contributing factors of a vehicle, which we plan to normalize. For each collision ID, all contributing factors will be listed in new rows. This restructuring is carried out during the data cleansing stage. Additionally, we are verifying the contributing factor values against a mapping document. The corresponding codes and descriptions from this document will be utilized for the final staging process.

23. COLLISION ID:

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

- No. of Distinct: 2075427
- Min Value: 22
- Max Value: 4712252

24. VEHICLE TYPE CODE 1:

- This variable has 1631 distinct values present.
- The number of missing values is 13691 which is 0.7%.
- No. of Distinct: 1631
- The missing values in this column would be replaced by “NA”
- The special characters present in this column is . # & ? ‘ - / ( )

25. VEHICLE TYPE CODE 2:

- This variable has 1819 distinct values present.
- The number of missing values is 396691 which is 19.1%.
- No. of Distinct: 1819
- The missing values in this column would be replaced by “NA”
- The special characters present in this column is . # & ? ‘ - / ( )

26. VEHICLE TYPE CODE 3:

- This variable has 260 distinct values present.
- The number of missing values is 1932530 which is 93.1%.
- No. of Distinct: 260
- The missing values in this column would be replaced by “NA”
- The special characters present in this column is . # & ? ‘ - / ( )

27. VEHICLE TYPE CODE 4:

- This variable has 101 distinct values present.
- The number of missing values is 2043115 which is 98.4%.
- No. of Distinct: 101
- The missing values in this column would be replaced by “NA”
- The special characters present in this column is . # & ? ‘ - / ( )

28. VEHICLE TYPE CODE 5:

- This variable has 70 distinct values present.
- The number of missing values is 2066635 which is 99.6%.
- No. of Distinct: 70
- The missing values in this column would be replaced by “NA”
- The special characters present in this column is . # & ? ‘ - / ( )

We have five columns related to the type of a vehicle involved in the collision, which we plan to normalize. For each collision ID, all vehicle types will be listed in new rows. This restructuring is carried out during the data cleansing stage. Additionally, we are verifying the vehicle type values against a mapping document. We have created the mapping document for vehicle type. The corresponding codes and descriptions from this document will be utilized for the final staging process.

29. ZIPCODE:

- No. of Distinct: 235
- Missing Values: 645996
- Values with Blank spaces: 42

The dataset contains zip codes that are sometimes missing or represented by blank spaces. Consequently, we are treating the zip code as a string column during the staging phase. This strategy facilitates the initial data loading and will also allow for subsequent transformations to align with specific business needs.