Data Profiling NYC MV COLLISIONS

1. **Motor\_Vehicle\_Collisions - Person:**

Graphical user interface, application

Description automatically generated

The data contains information about various fields in a data set related to vehicle crashes. Some of the fields have a small number of value counts, and it is suggested that combining some value levels together may be appropriate. Many of the fields have missing values, and it is recommended to consider imputing these values.

The fields with missing values above 10% include CONTRIBUTING\_FACTOR\_2, SAFETY\_EQUIPMENT, POSITION\_IN\_VEHICLE, PED\_LOCATION, PED\_ACTION, PERSON\_SEX, EMOTIONAL\_STATUS, BODILY\_INJURY, and COMPLAINT. The fields with missing values below 10% include VEHICLE\_ID and CRASH\_TIME.

Some fields have a very high percentage of missing values, such as CONTRIBUTING\_FACTOR\_2, PED\_LOCATION, and PED\_ACTION, which have over 98% of values missing.

The data set contains various fields related to crashes, including crash date, crash time, collision ID, vehicle ID, contributing factors, safety equipment, personal information, injury details, and emotional status.

1. **Motor\_Vehicle\_Collisions -Vehicles**

Graphical user interface, text, application

Description automatically generated

The data provides information about various fields in a dataset, likely related to vehicle crashes.

The fields include:

VEHICLE\_ID: This field has 2,175,578 values, all with a value count of 1. The data suggests that some value levels may be combined.

CRASH\_TIME: This field has 1,440 values, with some value levels that may be combined.

CONTRIBUTING\_FACTOR\_2: This field has over 10% missing values and some value levels that may be combined.

PUBLIC\_PROPERTY\_DAMAGE: This field has over 10% missing values and some value levels that may be combined.

VEHICLE\_MODEL: This field has over 10% missing values and some value levels that may be combined.

CRASH\_DATE: This field has 3,444 values, with some value levels that may be combined.

VEHICLE\_TYPE: This field has 4.8% missing values and some value levels that may be combined.

VEHICLE\_DAMAGE\_3: This field has over 10% missing values and some value levels that may be combined.

PRE\_CRASH: This field has over 10% missing values and some value levels that may be combined.

VEHICLE\_DAMAGE\_1: This field has over 10% missing values and some value levels that may be combined.

VEHICLE\_OCCUPANTS: This field has over 10% missing values and some value levels that may be combined.

PUBLIC\_PROPERTY\_DAMAGE\_TYPE: This field has over 10% missing values and some value levels that may be combined.

POINT\_OF\_IMPACT: This field has over 10% missing values and some value levels that may be combined.

TRAVEL\_DIRECTION: This field has over 10% missing values and some value levels that may be combined.

VEHICLE\_YEAR: This field has over 10% missing values and some value levels that may be combined.

UNIQUE\_ID: This field has 3,704,406 values, all with a value count of 1. The data suggests that some value levels may be combined.

VEHICLE\_DAMAGE: This field has over 10% missing values and some value levels that may be combined.

Overall, the data suggests that the dataset may need some data cleaning and preprocessing before it can be used for analysis. Specifically, missing values may need to be imputed and some value levels may need to be combined.

**3.MotorVehicleCollision\_NYC**

**Graphical user interface, application

Description automatically generated**

The data above provides information about several fields in a data set, including the percentage of missing values, the number of unique values, the shortest and longest values, and the minimum and maximum count of values for each field. Additionally, the data provides some remarks or suggestions regarding these fields.

Fields, including "contributing\_factor\_vehicle\_4," "vehicle\_type\_code\_5," "borough," "contributing\_factor\_vehicle\_3," "vehicle\_type\_code1," "contributing\_factor\_vehicle\_1," "vehicle\_type\_code\_3," "contributing\_factor\_vehicle\_5," "vehicle\_type\_code2," "contributing\_factor\_vehicle\_2," "cross\_street\_name," "location," "on\_street\_name," "off\_street\_name," and "vehicle\_type\_code\_4."

The second column displays the percentage of missing values for each field, ranging from 0.3% to 99.6%. Some fields, such as "vehicle\_type\_code1" and "contributing\_factor\_vehicle\_1," have only a small percentage of missing values, while others, such as "vehicle\_type\_code\_5" and "contributing\_factor\_vehicle\_5," have a high percentage of missing values.

The third column, Unique Values, indicates the number of unique values for each field. Some fields have a small number of unique values, such as "contributing\_factor\_vehicle\_4" with only 41 unique values, while others have a larger number of unique values, such as "vehicle\_type\_code2" with 1,669 unique values.

The fourth column, Shortest Value, displays the shortest value for each field, while the fifth column shows the longest value. For example, the shortest value for "vehicle\_type\_code1" is ".", while the longest value is "Enclosed Body - Nonremovable Enclosure."

The sixth and seventh columns Min Value Count and Max Value Count provide the minimum and maximum count of values for each field. For example, "contributing\_factor\_vehicle\_1" has a minimum count of 10 and a maximum count of 684,568.

Finally, the remarks section provides some suggestions for the fields. For instance, the data recommends inputting missing values for all fields with over 10% missing values. Additionally, for some fields with a small number of value counts, such as "vehicle\_type\_code1" and "contributing\_factor\_vehicle\_1," the data suggests combining some value levels together if appropriate. Overall, the data provides insights and recommendations for how to handle missing values and values with low counts in the listed fields.