**Chicago Profiling Document**

The Chicago dataset under consideration comprises of various datapoints that are critical to answering the business requirements. Following are the key observations about the dataset:

1. Number of fields: The dataset consists of 48 distinct fields, one of which is alpha-numeric field (crash) to keep it consistent with the other corresponding datasets we added an additional column “ crash\_id ” using a sequence generator.
2. Number of Observations: There are total 617.7k (817723) records, where each record corresponds to a Motor Vehicle Collision incident.
3. Missing Cells: A count of 8.2M (8268003) cells within the dataset have missing data, which constitutes approximately 21.1% of the total data. This is relatively high and can impact analysis.
4. Duplicate Rows: There are no duplicate rows, as indicated by a count of 0, which means that each observation is unique.
5. Total Size in Memory: The dataset occupies 299.5  MiB (Megabytes) in memory, providing an idea of the data's footprint in terms of storage.
6. Average Record Size in Memory: Each record, on average, occupies 384.0 bytes in memory, which can be useful for understanding the load and performance implications for data processing tasks.

**COLUMN WISE ANALYSIS :**

**crash\_record\_id :**

The data field contains 817723 unique values.

There are zero missing values for the crash\_record\_id as indicated by a count of 0.

No. of Distinct: 817723

Length : 128 characters (consistent)

**crash\_date\_estm :**

The data field contains 2 unique values.

There are 756594 missing values for the crash\_date\_est.

No. of Distinct values : 2

As the columns has Boolean data type, hence it accepts 1 and 0 as input.

**crash\_date (date and time) :**

The data field contains 53688 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: 817723

Min Value: 2013-03-03 16:48:00

Max Value: 2024-03-26 01:40:00

**posted\_speed\_limit :**

The data field contains 46 unique values for the posted\_speed\_limit.

There are no missing values for the posted\_speed\_limit variable as indicated by a count of 0.

No. of Distinct: 46

Min Value: 0

Max Value: 99

**traffic\_control\_device :**

The data field contains 19 unique values for the traffic\_control\_device

There are no missing values for the traffic\_control\_device variable as indicated by a count of 0.

No. of Distinct: 19

As this column consists of categorical data, it has no minimum and maximum values.

**device\_condition :**

The data field contains 8 unique values for the device\_condition.

There are no missing values for the device\_condition variable as indicated by count of 0.

No. of Distinct: 8

As this column consists of categorical data, it has no minimum and maximum values.

**weather\_condition :**

The data field contains 12 unique values for the weather\_condition.

There are no missing values for the weather\_condition variable as indicated by count of 0.

No. of Distinct: 12

As this column consists of categorical data, it has no minimum and maximum values.

**lighting\_condition** :

The data field contains 6 unique values for the lighting\_condition.

There are no missing values for the lighting\_condition variable

No. of Distinct: 6

As this column consists of categorical data, it has no minimum and maximum values.

**first\_crash\_type** :

The data field contains 18 unique values for the first\_crash\_type.

There are no missing values for the first\_crash\_type variable as indicated by count of 0.

No. of Distinct: 18

As this column consists of categorical data, it has no minimum and maximum values.

**traffic\_type** :

The data field contains 20 unique values for the traffic\_type.

There are no missing values for traffic\_type variable as indicated by count of 0.

No. of Distinct: 20

As this column consists of categorical data, it has no minimum and maximum values.

**lane\_count :**

The data field contains 41 unique values for the lane\_count.

There are 618714 missing values for the column lane\_count

No. of Distinct: 41

Minimum value : 0

Maximum value : 1191625

**alignment :**

The data field contains 6 unique values for the alignment.

There are no missing values for the alignment variable as indicated by count of 0.

No. of Distinct: 6

As this column consists of categorical data, it has no minimum and maximum values.

**Roadway\_surface\_condition** :

The data field contains 7 unique values for the Roadway\_surface\_condition.

There are no missing values for the Roadway\_surface\_condition variable as indicated by count of 0.

No. of Distinct: 7

As this column consists of categorical data, it has no minimum and maximum values.

**road\_defect:**

The data field contains 7 unique values for the road\_defect.

There are no missing values for the road\_defect variable as indicated by count of 0.

No. of Distinct: 7

As this column consists of categorical data, it has no minimum and maximum values.

**report\_type :**

The data field contains 3 unique values for the report\_type.

There are 24314 missing values for the report\_type variable.

No. of Distinct: 3

As this column consists of categorical data, it has no minimum and maximum values.

**crash\_type** :

The data field contains 2 unique values for the crash\_type.

There are no missing values for the road\_defect variable.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

**intersection\_related\_I :**

The data field contains 2 unique values for the crash\_type.

There are 630174 missing values for the crash\_type variable.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes 0 and 1 as the input values.

**NOT\_RIGHT\_OF\_WAY\_I :**

The data field contains 2 unique values for the NOT\_RIGHT\_OF\_WAY\_I.

There are 780015missing values for the NOT\_RIGHT\_OF\_WAY\_I variable .

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**HIT\_AND\_RUN\_I :**

The data field contains 2 unique values for the HIT\_AND\_RUN\_I.

There are 561774 missing values for the HIT\_AND\_RUN\_I variable .

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

**Damage :**

The data field contains 3 unique values for the Damage.

There are no missing values for the Damage variable as indicated by count of 0.

No. of Distinct: 3

As this column consists of categorical data, it has no minimum and maximum values.

**DATE\_POLICE\_NOTIFIED** :

The data field contains 620545 unique values for the DATE\_POLICE\_NOTIFIED.

There are no missing values for the DATE\_POLICE\_NOTIFIED variable as indicated by count of 0.

No. of Distinct: 620545

Minimum value: 2013-06-01 20:31:00

Maximum value : 2024-03-26 01:42:00

**PRIM\_CONTRIBUTORY\_CAUSE** :

The data field contains 40 unique values for the PRIM\_CONTRIBUTORY\_CAUSE.

There are no missing values for the PRIM\_CONTRIBUTORY\_CAUSE variable as indicated by count of 0.

No. of Distinct: 40

As this column consists of categorical data, it has no minimum and maximum values.

**SEC\_CONTRIBUTORY\_CAUSE :**

The data field contains 40 unique values for the SEC\_CONTRIBUTORY\_CAUSE.

There are no missing values for the SEC\_CONTRIBUTORY\_CAUSE variable as indicated by count of 0.

No. of Distinct: 40

As this column consists of categorical data, it has no minimum and maximum values.

**STREET\_NO :**

The data field contains 11728 unique values for the STREET\_NO.

There are no missing values for the STREET\_NO variable as indicated by count of 0.

No. of Distinct: 11728

Minimum Value : 0

Maximum Value : 451100

**STREET\_DIRECTION :**

The data field contains 4 unique values for STREET\_DIRECTION.

There are 4 missing values for the STREET\_DIRECTION column.

No. of Distinct: 4

As this column consists of categorical data, it accepts values : North, South, East, West

**STREET\_NAME :**

The data field contains 1641 unique values for STREET\_NAME.

There are 1 missing values for the STREET\_DIRECTION column.

No. of Distinct: 1641

Minimum Length : 4

Maximum Length : 31

**BEAT\_OF\_OCCURRENCE :**

The data field contains 276 unique values for BEAT\_OF\_OCCURRENCE.

There are 5 missing values for the BEAT\_OF\_OCCURRENCE column.

No. of Distinct: 276

Minimum Value : 111

Maximum Length : 6100

**PHOTOS\_TAKEN\_I :**

The data field contains 2 unique values for PHOTOS\_TAKEN\_I.

There are 806948 missing values for the PHOTOS\_TAKEN\_I column.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**STATEMENTS\_TAKEN\_I :**

The data field contains 2 unique values for STATEMENTS\_TAKEN\_I.

There are 799465 missing values for the STATEMENTS\_TAKEN\_I column.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**DOORING\_I :**

The data field contains 2 unique values for DOORING\_I.

There are 815211 missing values for the DOORING\_I column.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**WORK\_ZONE\_I :**

The data field contains 2 unique values for WORK\_ZONE\_I.

There are 813053 missing values for the WORK\_ZONE\_I column.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**WORK\_ZONE\_TYPE :**

The data field contains 4 unique values for WORK\_ZONE\_I.

There are 814105 missing values for the WORK\_ZONE\_I column.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**WORKERS\_PRESENT\_I :**

The data field contains 2 unique values for WORK\_ZONE\_I.

There are 816529 missing values for the WORK\_ZONE\_I column.

No. of Distinct: 2

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes True and False as the input values.

**NUM\_UNITS :**

The data field contains 17 unique values for NUM\_UNITS.

There are zero missing values for the NUM\_UNITS.

No. of Distinct: 17

Minimum Value : 1

Maximum Value : 18

**MOST\_SEVERE\_INJURY :**

The data field contains 5 unique values for MOST\_SEVERE\_INJURY.

There are 1792 missing values for the MOST\_SEVERE\_INJURY.

No. of Distinct: 5

As this column consists of categorical data, it has no minimum and maximum values.

**INJURIES\_TOTAL :**

The data field contains 20 unique values for MOST\_SEVERE\_INJURY.

There are 1780 missing values for the MOST\_SEVERE\_INJURY.

No. of Distinct: 20

Minimum Value : 0

Maximum Value : 21

**INJURIES\_FATAL :**

The data field contains 5 unique values for MOST\_SEVERE\_INJURY.

There are 1780 missing values for the MOST\_SEVERE\_INJURY.

No. of Distinct: 20

As this column consists of categorical data, it has no minimum and maximum values.

This column has Boolean data type, and hence takes 0,1,2,3,4 as the input values.

**INJURIES\_INCAPACITATING :**

The data field contains 10 unique values for INJURIES\_INCAPACITATING.

There are 1780 missing values for the INJURIES\_INCAPACITATING.

No. of Distinct: 10

Minimum Value : 1

Maximum Value : 10

**INJURIES\_NON\_INCAPACITATING :**

The data field contains 19 unique values for INJURIES\_NON\_INCAPACITATING.

There are 1780 missing values for the INJURIES\_NON\_INCAPACITATING.

No. of Distinct: 19

Minimum Value : 0

Maximum Value : 21

**INJURIES\_REPORTED\_NOT\_EVIDENT :**

The data field contains 13 unique values for INJURIES\_NON\_INCAPACITATING.

There are 1780 missing values for the INJURIES\_NON\_INCAPACITATING.

No. of Distinct: 13

Minimum Value : 0

Maximum Value : 15

**INJURIES\_NO\_INDICATION** :

The data field contains 48 unique values for INJURIES\_NO\_INDICATION.

There are 1780 missing values for the INJURIES\_NO\_INDICATION.

No. of Distinct: 48

Minimum Value : 0

Maximum Value : 61

**INJURIES\_UNKNOWN :**

The data field contains 1 unique value for INJURIES\_UNKNOWN.

There are 1780 missing values for the INJURIES\_UNKNOWN.

No. of Distinct: 1

As this column consists of categorical data, it has no minimum and maximum values.

**CRASH\_HOUR :**

The data field contains 24 unique values for INJURIES\_NO\_INDICATION.

There are zero missing values for the INJURIES\_NO\_INDICATION.

No. of Distinct: 24

Minimum Value : 0

Maximum Value : 23

**CRASH\_DAY\_OF\_WEEK** :

The data field contains 7 unique values for CRASH\_DAY\_OF\_WEEK.

There are zero missing values for the CRASH\_DAY\_OF\_WEEK.

No. of Distinct: 7

Minimum Value : 1

Maximum Value : 7

**CRASH\_MONTH :**

The data field contains 12 unique values for CRASH\_MONTH.

There are zero missing values for the CRASH\_MONTH.

No. of Distinct: 12

Minimum Value : 1

Maximum Value : 12

**LATITUDE :**

The data field contains 300091 unique values for LATITUDE.

There are 5615 missing values for the LATITUDE.

No. of Distinct: 300091

Minimum Value : 0

Maximum Value : 42.02278

**LONGITUDE :**

The data field contains 300054 unique values for LONGITUDE.

There are 5615 missing values for the LONGITUDE.

No. of Distinct: 300054

Minimum Value : -87.9361

Maximum Value : 0

**LOCATION :**

The data field contains 300054 unique values for LOCATION.

There are 5615 missing values for the LOCATION.

No. of Distinct values : 300054

This column has text data type .

**crash\_id :**

This column is generated using the sequence generator in the staging table with starting value of 100 and incremented by 1. It has integer data type.