PROFILING DOCUMENTATION FOR CHICAGO DATASET

The Chicago dataset 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)
- This is the unique identifier for the crashes in the Chicago city. But since the data is of type alphanumeric, a surrogate key using Talend's Numeric.sequence would be created and indicated by the column **CRASH_ID** to accurately represent them and use them for further analysis.

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 column has Boolean data type, it accepts Y or N as input.

- The missing values would be replace by “NA” and hence the column would be changed to of datatype varchar

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
- The min and max values indicate that the crash dates ranges from the year 2013 to 2024

Crash_Time would be extracted from this field at the time of data cleaning into a separate column of type varchar

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
- The speed limit ranges from 0 to 99 for the reported crashes

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
- The column consists of categorical data and the length of the data in this column ranges from 5 to 24.This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘/’. ex:- POLICE/FLAGMAN,STOP SIGN/FLASHER.

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
- The column consists of categorical data and the length of the data in this column ranges from 5 to 24.This information will be helpful at the time of stagging and will help avoid value truncated errors.

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
- The column consists of categorical data and the length of the data in this column ranges from 4 to 24.This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column are ' , ' ' / '. ex:- BLOWING SAND,SOIL,DIRT
- CLOUDY/OVERCAST

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
- The column consists of categorical data and the length of the data in this column ranges from 4 to 22.This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ' , ' . ex:- DARKNESS, LIGHTED ROAD

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

- The column consists of categorical data and the length of the data in this column ranges from 5 to 28.This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The column IS_PEDESTRIAN, which will be of datatype varchar, is to be created to identify whether a pedestrian was involved in the crash. It will categorize entries into two values: 'Y' for yes, indicating pedestrian involvement, and 'N' for no, indicating no pedestrian involvement. This addition is necessary to meet specific business requirements.

trafficway_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.
- The column consists of categorical data and the length of the data in this column ranges from 4 to 31.This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column are ‘ , ’ ‘ / ’ ‘ – ’ ‘ (’ ‘) ’ . ex:- DIVIDED - W/MEDIAN (NOT RAISED)

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
- The missing values would be replaced by -1 since the column is of datatype int

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
- The column consists of categorical data and the length of the data in this column ranges from 12 to 21.This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ , ’ ex:- CURVE, LEVEL

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
- The column consists of categorical data and the length of the data in this column ranges from 3 to 15. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ , ’ ex:- SAND, MUD, DIRT

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
- The column consists of categorical data and the length of the data in this column ranges from 5 to 17. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ , ’ ex:- RUT,HOLES

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
- The column consists of categorical data and the length of the data in this column ranges from 7 to 26. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ (’ ‘)’ ex:- NOT ON SCENE (DESK REPORT)
- The missing values would be replace by “NA”

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
- The column consists of categorical data and the length of the data in this column ranges from 22 to 32. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ / ’ ex:- NO INJURY / DRIVE AWAY

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replace by “NA” and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replace by “NA” and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replace by “NA” and hence the column would be changed to of datatype varchar

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

- The column consists of categorical data and the length of the data in this column ranges from 11 to 13. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ $ ’ ‘ , ’ ‘ - ’ ex:- $501 - $1,500

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
- This also indicates that the police notified rates range from 2013-2024

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
- The column consists of categorical data and the length of the data in this column ranges from 6 to 80. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ / ’ ‘ , ’ ‘ . ’ ex:- IMPROPER TURNING/NO SIGNAL

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
- The column consists of categorical data and the length of the data in this column ranges from 6 to 80. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ / ’ ‘ , ’ ‘ . ’ ex:- DRIVING SKILLS/KNOWLEDGE/EXPERIENCE

PRIM_CONTRIBUTORY_CAUSE and SEC_CONTRIBUTORY_CAUSE would be split or normalized into 2 rows in the cleaning phase such that every crash has 2 associated contributory cause. Duplicate rows would be avoided. This will then be checked with the contributory factor list shared to get the associated contributory codes. 2 new columns, one of datatype int for contributory code and one of datatype varchar for the common contributory description would be created in the cleaning phase.

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
- The column consists of categorical data and the min and max length is 1. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The missing values would be replace by “NA”

STREET_NAME :

- The data field contains 1641 unique values for STREET_NAME.
- There is 1 missing values for the STREET_DIRECTION column.
- No. of Distinct: 1641
- The column consists of categorical data and the length of the data in this column ranges from 4 to 31. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ / ’ ‘ . ’ ‘ - ’
- The missing value would be replace by “NA”

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
- The missing values would be replaced by -1 in the cleaning phase as it is a numerical column

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replaced by "NA" and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replaced by "NA" and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replaced by "NA" and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replaced by "NA" and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replaced by "NA" and hence the column would be changed to of datatype varchar

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
- This column has Boolean data type, and hence takes Y and N as the input values.
- The missing values would be replaced by "NA" and hence the column would be changed to of datatype varchar

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
- This column is of datatype int

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
- The column consists of categorical data and the length of the data in this column ranges from 5 to 24. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ' , ' . ex: REPORTED,NOT EVIDENT
- The missing values would be replace by "NA"

INJURIES_TOTAL :

- The data field contains 20 unique values for INJURIES_TOTAL.
- There are 1780 missing values for the INJURIES_TOTAL.
- No. of Distinct: 20
- Minimum Value : 0
- Maximum Value : 21
- This column is of datatype int
- The missing values in this column would be replaced by -1

INJURIES_FATAL :

- The data field contains 5 unique values for INJURIES_TOTAL.
- There are 1780 missing values for the INJURIES_TOTAL.
- No. of Distinct: 20
- Minimum Value : 0
- Maximum Value : 4
- This column is of datatype int
- The missing values in this column would be replaced by -1

INJURIES_INCAPACITATING :

- The data field contains 10 unique values for INJURIES_TOTAL.
- There are 1780 missing values for the INJURIES_TOTAL.
- No. of Distinct: 20
- Minimum Value : 0
- Maximum Value : 10
- This column is of datatype int
- The missing values in this column would be replaced by -1

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
- This column is of datatype int
- The missing values in this column would be replaced by -1

INJURIES_REPORTED_NOT_EVIDENT :

- The data field contains 13 unique values for INJURIES_REPORTED_NOT_EVIDENT.
- There are 1780 missing values for the INJURIES_REPORTED_NOT_EVIDENT.
- No. of Distinct: 13
- Minimum Value : 0
- Maximum Value : 15
- This column is of datatype int
- The missing values in this column would be replaced by -1

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
- This column is of datatype int
- The missing values in this column would be replaced by -1

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
- This column is of datatype int
- The missing values in this column would be replaced by -1

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
- This column is of datatype int

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
- This column is of datatype int

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
- This column is of datatype int

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

Both LATITUDE and LONGITUDE would be converted to datatype float at the cleaning stage and the missing values would be replaced by -1.

LOCATION :

- The data field contains 300054 unique values for LOCATION.
- There are 5615 missing values for the LOCATION.
- No. of Distinct values : 300054
- The column consists of categorical data and the length of the data in this column ranges from 11 to 40. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ . ’ ‘ (’ ‘) ’ ‘ - ’. ex: POINT (-87.665902342962 41.854120262952)
- The missing values would be replace by “NA”