**Analysis Document – Motor Vehicle Collisions.**

***NYC***

Introduction:

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC.

Data profiling is required to understand the data quality, its structure, and characteristics. It involves examining characteristics such as data types, distributions, completeness, and uniqueness to identify patterns, anomalies, and potential issues, thereby laying the foundation for effective data analysis, cleansing, and preprocessing.

Dataset Details:

The dataset contains 2.08M rows and 29 columns where each row is a Motor Vehicle Collision.

The dataset comprises detailed information on each crash event, including crash date, time, location, borough, zip code, latitude, and longitude coordinates.

Source: <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data>

Column Analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column\_Name** | **Data\_Type** | **Distinct Values** | **Missing\_Count** | **Min/Max\_Values** |
| Crash Date | Object | 4283 | 0 | Min: 2012-07-01 00:00:00  Max: 2024-03-22 00:00:00 |
| Crash Time | Object | 1440 | 0 | Min: 2024-03-28 00:00:00  Max: 2024-03-28 23:59:00 |
| Borough | Object | 5 | 645746 | Brooklyn: 454727  Staten Island: 60012 |
| Zip Code | Object |  | 645996 |  |
| Latitude | Float | 126594 | 233626 | Min: 0  Max: 43.344444 |
| Longitude | Float | 98351 | 233626 | Min: -201.35999  Max: 0 |
| Location | Object | 283006 | 233626 |  |
| On Street Name | Object | 18410 | 440569 |  |
| Cross Street Name | Object | 20236 | 784436 |  |
| Off Street Name | Object | 225845 | 1727231 |  |
| Number of Persons Injured | Float | 32 | 18 | Min: 0  Max: 43 |
| Number of Persons Killed | Float | 7 | 31 | Min: 0  Max: 8 |
| Number of Pedestrians Injured | Integer | 14 | 0 | Min: 0  Max: 27 |
| Number of Pedestrians Killed | Integer | 4 | 0 |  |
| Number of Cyclist Injured | Integer | 5 | 0 |  |
| Number of Cyclist Killed | Integer | 3 | 0 |  |
| Number of Motorist Injured | Integer | 31 | 0 | Min: 0  Max: 43 |
| Number of Motorist Killed | Integer | 6 | 0 | Min: 0  Max: 5 |
| Contributing Factor Vehicle 1 | Object | 61 | 6802 |  |
| Contributing Factor Vehicle 2 | Object | 61 | 321736 |  |
| Contributing Factor Vehicle 3 | Object | 51 | 1927163 |  |
| Contributing Factor Vehicle 4 | Object | 41 | 2041953 |  |
| Contributing Factor Vehicle 5 | Object | 30 | 2066358 |  |
| Collision\_ID | Integer | 2075427 | 0 | Min: 22  Max: 4712252 |
| Vehicle Type Code 1 | Object | 1631 | 13691 |  |
| Vehicle Type Code 2 | Object | 1819 | 396691 |  |
| Vehicle Type Code 3 | Object | 260 | 1932530 |  |
| Vehicle Type Code 4 | Object | 101 | 2043115 |  |
| Vehicle Type Code 5 | Object | 70 | 2066635 |  |

A screenshot of a computer

Description automatically generated

Inference:

* 'Collision\_ID' is a unique identifier for each incident.
* Large portion of data is missing from the dataset(about 30%)
* ‘Borough', 'Zip Code', and 'Location' have a large number of missing values, this might give incorrect geographic analysis.
* No duplicate rows are present.
* The data related to injury and fatality counts are mostly numeric, hence it is easier for aggregation and statistic analysis.
* 'Latitude' and 'Longitude' have min values that do not make sense geographically (latitude 0 and longitude -201.35999), which could indicate data entry errors.

***Chicago***

Introduction:

Crash data shows information about each traffic crash on city streets within the City of Chicago limits and under the jurisdiction of the Chicago Police Department (CPD).

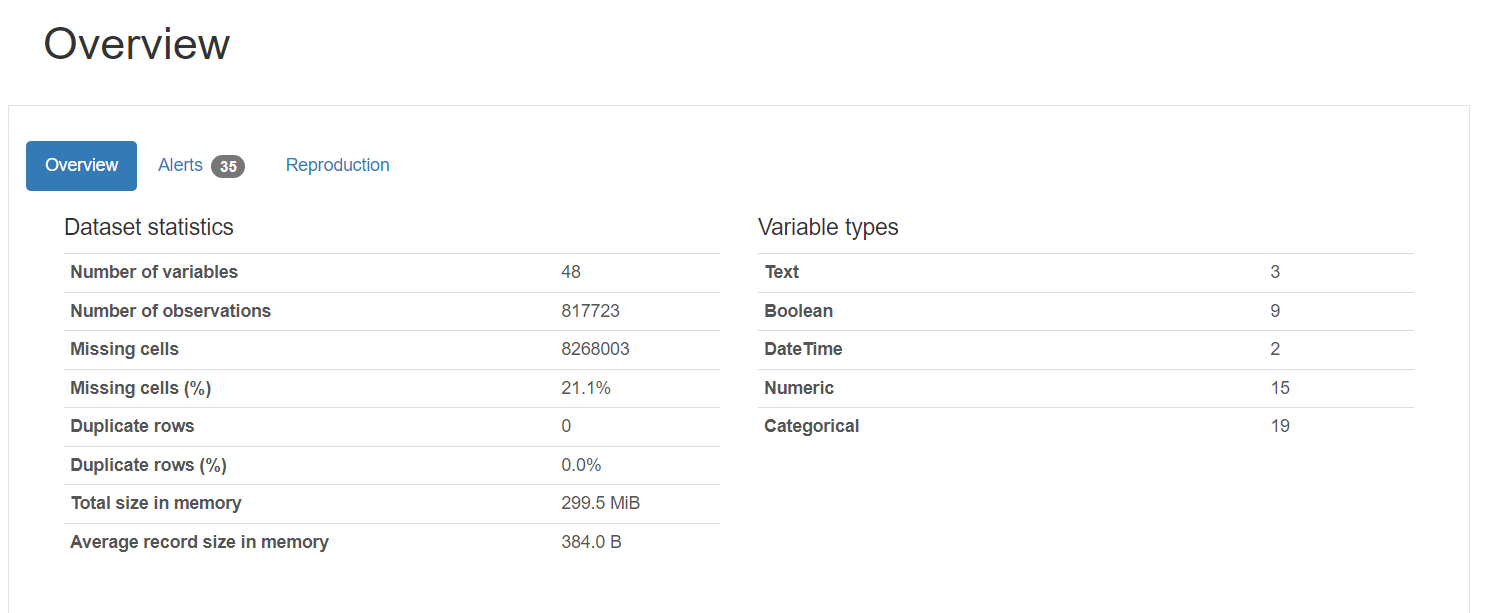
Dataset Details:

The dataset contains 818K rows and 48 columns where each row is a traffic crash.

Source: <https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data>

Column Analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column\_Name** | **Data\_Type** | **Distinct\_Count** | **Missing Count** | **Min/Max\_Values** |
| Crash\_Record\_ID | Object | 817723 | 0 |  |
| Crash\_Date\_Est\_I | Object | 2 | 756594 |  |
| Crash\_date | Object | 536888 | 0 | 2013-03-03 16:48:00/ 2024-03-23 01:40:00 |
| Posted\_Speed\_Limit | Integer | 46 | 0 | 0/99 |
| Traffic\_Control\_Device | Object | 19 | 0 |  |
| Device\_Condition | Object | 8 | 0 |  |
| Weather\_Condition | Object | 12 | 0 |  |
| Lighting\_Condition | Object | 6 | 0 |  |
| First\_Crash\_Type | Object | 18 | 0 |  |
| Trafficway\_Type | Object | 20 | 0 |  |
| Lane\_cnt | Float | 41 | 618714 | 0/1191625 |
| Alignment | Object | 6 | 0 |  |
| Roadway\_Surface\_cond | Object | 7 | 0 |  |
| Road\_Defect | Object | 7 | 0 |  |
| Report\_Type | Object | 3 | 24314 |  |
| Crash\_Type | Object | 2 | 0 |  |
| Intersection\_Related\_I | Object | 2 | 630174 |  |
| Not\_Right\_Of\_Way\_I | Object | 2 | 780015 |  |
| Hit\_And\_Run\_I | Object | 2 | 561774 | True-244955/False-10994 |
| Damage | Object | 3 | 0 |  |
| Date\_Police\_Notified | Object | 620545 | 0 | 2013-06-01 20:31:00/ 2024-03-26 01:42:00 |
| Prim\_Contributory\_Cause | Object | 40 | 0 |  |
| Sec\_Contributory\_Cause | Object | 40 | 0 |  |
| Street\_No | Integer | 11728 | 0 | 0/451100 |
| Street\_Direction | Object | 4 | 4 | E-55621/W-292260 |
| Street\_Name | Object | 1641 | 1 |  |
| Beat\_Of\_Occurrence | Float | 276 | 5 | 111/6100 |
| Photos\_Taken\_I | Object | 2 | 806948 | False-2665/True-8110 |
| Statements\_Taken\_I | Object | 2 | 799465 | False-3384/True-14874 |
| Dooring\_I | Object | 2 | 815211 | False-824/True-1688 |
| Work\_Zone\_I | Object | 2 | 813053 | False-1052/True-3618 |
| Work\_Zone\_Type | Object | 4 | 814105 |  |
| Workers\_Present\_I | Object | 2 | 816529 | False-138/True-1056 |
| Num\_units | Integer | 17 | 0 | 1/18 |
| Most\_Severe\_Injury | Object | 5 | 1792 |  |
| Injuries\_Total | Float | 20 | 1780 | 0/21 |
| Injuries\_Fatal | Float | 5 | 1780 |  |
| Injuries\_Incapacitating | Float | 10 | 1780 | 0/10 |
| Injuries\_Non\_Incapacitating | Float | 19 | 1780 | 0/21 |
| Injuries\_Reported\_Not\_Evident | Float | 13 | 1780 | 0/15 |
| Injuries\_No\_Indication | Float | 48 | 1780 | 0/61 |
| Injuries\_Unknown | Float | 1 | 1780 |  |
| Crash\_Hour | Integer | 24 | 0 | 0/23 |
| Crash\_Day\_Of\_Week | Integer | 7 | 0 | 1/7 |
| Crash\_Month | Integer | 12 | 0 | 1/12 |
| Latitude | Float | 300091 | 5615 | 0/42.02278 |
| Longitude | Float | 300054 | 5615 | -87.936193/0 |
| Location | Object | 300265 | 5615 |  |
|  |  |  |  |  |



Inference:

* There are 48 columns in total and about 818K rows.
* Considerable amount of data is missing from the dataset(21%)
* The columns ‘Lane\_cnt, Intersection\_Related\_I, Not\_Right\_Of\_Way\_I, and Hit\_And\_Run\_I’ have high number of missing values, whereas the column ‘Crash\_Date\_Est\_I’ column has highest number of missing values.
* The Lane\_cnt maximum value is extraordinarily high, which could be an outlier or an error.
* As the dataset covers dates over a decade(from 2013 to 2024) it can be used for trend analysis.
* The traffic and environment related columns have complete data, (Traffic\_Control\_Device, Device\_Condition, Weather\_Condition,). It means the details are well recorded and it can help in understanding the crash contexts.

***Austin***

Introduction:

The dataset contains crash-level records for crashes that have occurred in the last ten years for Austin.

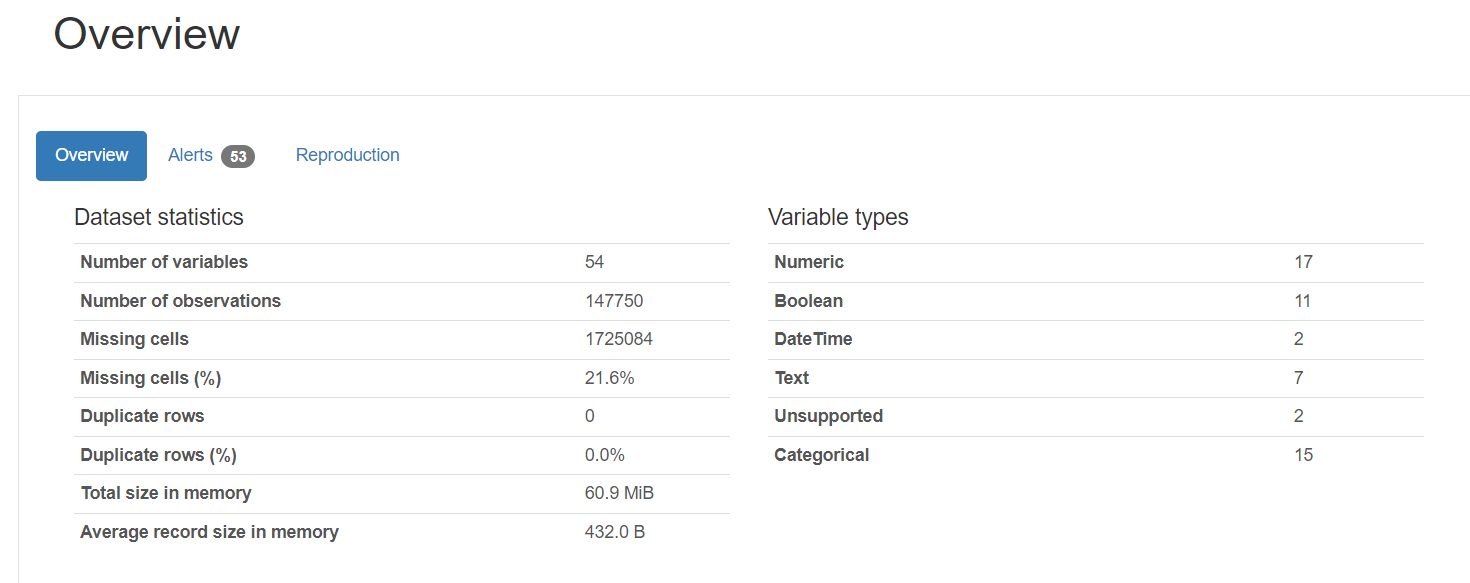
Dataset Details:

The dataset contains 148K rows and 54 columns where each row is a crash.

Source: https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5/about\_data

Column Analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Distinct Count** | **Missing count** | **Min/Max Values** |
| Crash\_id | Integer | 147750 | 0 | 1001/1.8029054\*10^8 |
| Crash\_fatal\_fl | Object | 2 | 0 |  |
| Crash\_date | Object | 144667 | 0 | 2014-03-26 06:41:00/ 2024-03-11 22:05:00 |
| Crash\_time | Object | 1440 | 0 | 2024-03-27 00:00:00/2024-03-27 23:59:00 |
| Case\_id | Object | 145678 | 1858 |  |
| Rpt\_latitude | Float | 7976 | 137456 | 25.83746/36.50048 |
| Rpt\_longitude | Float | 7264 | 137456 | -106.64592/-93.50795 |
| Rpt\_block\_num | Object |  | 19611 |  |
| Rpt\_street\_pfx | Object | 8 | 67805 |  |
| Rpt\_street\_name | Object | 9794 | 3 |  |
| Rpt\_street\_sfx | Object | 18 | 50340 |  |
| Crash\_speed\_limit | Float | 28 | 2 | -1/85 |
| Road\_constr\_zone\_fl | Object | 2 | 2 |  |
| Latitude | Float | 96355 | 2243 | 30.098737/30.511625 |
| Longitude | Float | 96230 | 2243 | -97.926789/-97.570148 |
| Street\_name | Object | 4630 | 2 |  |
| Street\_nbr | Float | 9826 | 87038 | 2/21146 |
| Street\_name\_2 | Object | 3396 | 81474 |  |
| Street\_nbr\_2 | Float |  | 147750 |  |
| Crash\_sev\_id | Integer | 8 | 0 | 0/99 |
| Sus\_serious\_injry\_cnt | Integer | 7 | 0 | 0/10 |
| Nonincap\_injry\_cnt | Float | 14 | 1 | 0/14 |
| poss\_injry\_cnt | Float | 16 | 1 | 0/20 |
| non\_injry\_cnt | Float | 46 | 1 | 0/56 |
| unkn\_injry\_cnt | Float | 16 | 2 | 0/41 |
| tot\_injry\_cnt | Float | 18 | 2 | 0/21 |
| death\_cnt | Integer | 5 | 0 |  |
| contrib\_factr\_p1\_id | Float | 70 | 119143 | 1/80 |
| contrib\_factr\_p2\_id | Float | 65 | 143235 | 1/79 |
| units\_involved | Object | 1112 | 7 |  |
| atd\_mode\_category\_metadata | Object | 147743 | 7 |  |
| pedestrian\_fl | Object | 1 | 144245 |  |
| motor\_vehicle\_fl | Object | 1 | 1116 |  |
| motorcycle\_fl | Object | 1 | 144148 |  |
| bicycle\_fl | Object | 1 | 145306 |  |
| other\_fl | Object | 1 | 142905 |  |
| point | Object | 97739 | 2243 |  |
| apd\_confirmed\_fatality | Object | 2 | 0 |  |
| apd\_confirmed\_death\_count | Integer | 5 | 0 |  |
| motor\_vehicle\_death\_count | Integer | 5 | 0 |  |
| motor\_vehicle\_serious\_injury\_count | Integer | 6 | 0 | 0/5 |
| bicycle\_death\_count | Integer | 2 | 0 |  |
| bicycle\_serious\_injury\_count | Integer | 4 | 0 |  |
| pedestrian\_death\_count | Integer | 3 | 0 |  |
| pedestrian\_serious\_injury\_count | Integer | 5 | 0 |  |
| motorcycle\_death\_count | Integer | 3 | 0 |  |
| motorcycle\_serious\_injury\_count | Integer | 3 | 0 |  |
| other\_death\_count | Integer | 1 | 0 |  |
| other\_serious\_injury\_count | Integer | 3 | 0 |  |
| onsys\_fl | Object | 2 | 0 |  |
| private\_dr\_fl | Integer | 1 | 0 |  |
| micromobility\_serious\_injury\_count | Integer | 3 | 0 |  |
| micromobility\_death\_count | Integer | 2 | 0 |  |
| micromobility\_fl | Object | 1 | 147439 |  |



Inference:

* There are 54 columns in total, allowing for comprehensive analysis of the city’s crash data.
* Substantial amount of missing data(about 21%)
* There are no duplicate rows.
* ‘Crash\_id’ has no missing values, and can be a unique identifier.
* ‘Street\_nbr\_2’ column is entirely missing.