# **Chicago Crash Data Profiling Report**

***DAMG7370 Fall 2024 – Final Project – Motor Vehicle Collision Analysis***

***Description of the dataset***:

Chicago's traffic crash data, available through the city's data portal, provides detailed information on incidents under the jurisdiction of the Chicago Police Department (CPD). The data, sourced from CPD's electronic crash reporting system (E-Crash), omits personal identifiers and follows Illinois Department of Transportation’s SR1050 format.

**Key Details:**

* **Coverage**: Partial data is available from 2015, with comprehensive citywide data starting in September 2017.
* **Reporting**: Crashes are documented either on-scene by officers or self-reported at police districts.
* **Accuracy**: Certain details, such as weather or road conditions, are officer assessments and may not align with official records.
* **Exclusions**: Incidents handled by other agencies, such as crashes on freeways, are not included.
* **Criteria**: Includes crashes with property damage ≥$1,500 or injuries, involving at least one moving vehicle on public roads.
* **Comprehensiveness**: CPD records all reported crashes, sometimes exceeding state reporting requirements.

The profiling report on Chicago's traffic data offers an in-depth analysis of various facets of traffic incidents, such as crash severity, locations, and timings. This dataset includes 48 variables and 896,756 records, providing valuable insights into the patterns and dynamics of traffic incidents in the city. Here's a detailed summary based on the dataset's features.

**Description of columns (As observed from the dataset, Metadata):**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| CRASH\_RECORD\_ID | This number can be used to link to the same crash in the Vehicles and People datasets. This number also serves as a unique ID in this dataset. |
| CRASH\_DATE\_EST\_I | Crash date estimated by desk officer or reporting party (only used in cases where crash is reported at police station days after the crash) |
| CRASH\_DATE | Date and time of crash as entered by the reporting officer |
| POSTED\_SPEED\_LIMIT | Posted speed limit, as determined by reporting officer |
| TRAFFIC\_CONTROL\_DEVICE | Traffic control device present at crash location, as determined by reporting officer |
| DEVICE\_CONDITION | Condition of traffic control device, as determined by reporting officer |
| WEATHER\_CONDITION | Weather condition at time of crash, as determined by reporting officer |
| LIGHTING\_CONDITION | Light condition at time of crash, as determined by reporting officer |
| FIRST\_CRASH\_TYPE | Type of first collision in crash |
| TRAFFICWAY\_TYPE | Trafficway type, as determined by reporting officer |
| LANE\_CNT | Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection) |
| ALIGNMENT | Street alignment at crash location, as determined by reporting officer |
| ROADWAY\_SURFACE\_COND | Road surface condition, as determined by reporting officer |
| ROAD\_DEFECT | Road defects, as determined by reporting officer |
| REPORT\_TYPE | Administrative report type (at scene, at desk, amended) |
| CRASH\_TYPE | A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away |
| INTERSECTION\_RELATED\_I | A field observation by the police officer whether an intersection played a role in the crash. Does not represent whether or not the crash occurred within the intersection. |
| NOT\_RIGHT\_OF\_WAY\_I | Whether the crash begun or first contact was made outside of the public right-of-way. |
| HIT\_AND\_RUN\_I | Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid |
| DAMAGE | A field observation of estimated damage. |
| DATE\_POLICE\_NOTIFIED | Calendar date on which police were notified of the crash |
| PRIM\_CONTRIBUTORY\_CAUSE | The factor which was most significant in causing the crash, as determined by officer judgment |
| SEC\_CONTRIBUTORY\_CAUSE | The factor which was second most significant in causing the crash, as determined by officer judgment |
| STREET\_NO | Street address number of crash location, as determined by reporting officer |
| STREET\_DIRECTION | Street address direction (N,E,S,W) of crash location, as determined by reporting officer |
| STREET\_NAME | Street address name of crash location, as determined by reporting officer |
| BEAT\_OF\_OCCURRENCE | Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74 |
| PHOTOS\_TAKEN\_I | Whether the Chicago Police Department took photos at the location of the crash |
| STATEMENTS\_TAKEN\_I | Whether statements were taken from unit(s) involved in crash |
| DOORING\_I | Whether crash involved a motor vehicle occupant opening a door into the travel path of a bicyclist, causing a crash |
| WORK\_ZONE\_I | Whether the crash occurred in an active work zone |
| WORK\_ZONE\_TYPE | The type of work zone, if any |
| WORKERS\_PRESENT\_I | Whether construction workers were present in an active work zone at crash location |
| NUM\_UNITS | Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory. |
| MOST\_SEVERE\_INJURY | Most severe injury sustained by any person involved in the crash |
| INJURIES\_TOTAL | Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer |
| INJURIES\_FATAL | Total persons sustaining fatal injuries in the crash |
| INJURIES\_INCAPACITATING | Total persons sustaining incapacitating/serious injuries in the crash as determined by the reporting officer. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they were capable of performing before the injury occurred. Includes severe lacerations, broken limbs, skull or chest injuries, and abdominal injuries. |
| INJURIES\_NON\_INCAPACITATING | Total persons sustaining non-incapacitating injuries in the crash as determined by the reporting officer. Any injury, other than fatal or incapacitating injury, which is evident to observers at the scene of the crash. Includes lump on head, abrasions, bruises, and minor lacerations. |
| INJURIES\_REPORTED\_NOT\_EVIDENT | Total persons sustaining possible injuries in the crash as determined by the reporting officer. Includes momentary unconsciousness, claims of injuries not evident, limping, complaint of pain, nausea, and hysteria. |
| INJURIES\_NO\_INDICATION | Total persons sustaining no injuries in the crash as determined by the reporting officer |
| INJURIES\_UNKNOWN | Total persons for whom injuries sustained, if any, are unknown |
| CRASH\_HOUR | The hour of the day component of CRASH\_DATE. |
| CRASH\_DAY\_OF\_WEEK | The day of the week component of CRASH\_DATE. Sunday=1 |
| CRASH\_MONTH | The month component of CRASH\_DATE. |
| LATITUDE | The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash |
| LONGITUDE | The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash |
| LOCATION | The crash location, as determined by reporting officer, as derived from the reported address of crash, in a column type that allows for mapping and other geographic analysis in the data portal software |

**Dataset Profiling**:

NUMERIC COLUMNS:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Data Type** | **% Missing** | **Unique Values** | **Min** | **Max** | **Nulls** |
| LANE\_CNT | INT 64 | 77.8% | 42 | 0.000 | 1,191,625.000 | 697735 |
| INJURIES\_FATAL | INT 64 | 0.2% | 6 | 0.000 | 4.000 | 1972 |
| INJURIES\_UNKNOWN | INT 64 | 0.2% | 2 | 0.000 | 0.000 | 1972 |
| CRASH\_DAY\_OF\_WEEK | INT 64 | 0.0% | 7 | 1.000 | 7.000 | 0 |
| INJURIES\_NO\_INDICATION | INT 64 | 0.2% | 51 | 0.000 | 61.000 | 1972 |
| LONGITUDE | INT 64 | 0.7% | 318,271 | -87.936 | 0.000 | 6467 |
| NUM\_UNITS | INT 64 | 0.0% | 17 | 1.000 | 18.000 | 0 |
| STREET\_NO | INT 64 | 0.0% | 11,827 | 0.000 | 451,100.000 | 0 |
| LATITUDE | INT 64 | 0.7% | 318,306 | 0.000 | 42.023 | 6467 |
| BEAT\_OF\_OCCURRENCE | INT 64 | 0.0% | 277 | 111.000 | 6,100.000 | 0 |
| INJURIES\_REPORTED\_NOT\_EVIDENT | INT 64 | 0.2% | 14 | 0.000 | 15.000 | 1972 |
| CRASH\_MONTH | INT 64 | 0.0% | 12 | 1.000 | 12.000 | 0 |
| INJURIES\_NON\_INCAPACITATING | INT 64 | 0.2% | 21 | 0.000 | 21.000 | 1972 |
| CRASH\_HOUR | INT 64 | 0.0% | 24 | 0.000 | 23.000 | 0 |
| INJURIES\_INCAPACITATING | INT 64 | 0.2% | 11 | 0.000 | 10.000 | 1972 |
| INJURIES\_TOTAL | INT 64 | 0.2% | 21 | 0.000 | 21.000 | 1972 |
| POSTED\_SPEED\_LIMIT | INT 64 | 0.0% | 46 | 0.000 | 99.000 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Data Type** | **% Missing** | **Unique Values** | **Latest Date** | **Earliest Date** | **Nulls** |
| Police\_Notified\_Date | DATETIME | 0.0% | 614,454 | 11/27/2024 12:00 | 06/01/2013 08:00 | 0 |
| Crash\_Date\_Date | DATETIME | 0.0% | 505,602 | 11/27/2024 04:00 | 03/03/2013 04:00 | 0 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Data Type** | **% Missing** | **Unique Values** | **Shortest Value** | **Longest Value** | **Min Value Count** | **Max Value Count** | **Nulls** |
| WEATHER\_CONDITION | V\_W String | 0.0% | 12 | SNOW | BLOWING SAND, SOIL, DIRT | 7 | 705,199 | 0 |
| CRASH\_DATE\_EST\_I | V\_W String | 92.6% | 3 | Y | Y | 8,505 | 830,549 | 830549 |
| LIGHTING\_CONDITION | V\_W String | 0.0% | 6 | DUSK | DARKNESS, LIGHTED ROAD | 14,953 | 576,096 | 0 |
| DOORING\_I | V\_W String | 99.7% | 3 | Y | Y | 931 | 893,914 | 893914 |
| ROAD\_DEFECT | V\_W String | 0.0% | 7 | OTHER | DEBRIS ON ROADWAY | 659 | 714,691 | 0 |
| STREET\_NAME | V\_W String | 0.0% | 1,649 | 82ND | MICHIGAN AVE 175 E CHESTNUT AVE | 1 | 24,473 | 0 |
| CRASH\_RECORD\_ID | V\_W String | 0.0% | 896,756 | 23a79931ef555d54118f64dc9be2cf2dbf59636ce253f7a1179c4a1c091442a6eeab8352220c7c56ca1ff7c4b4b0fc345c74e3e85ecb9d43deeb66b5f803d4a0 | 23a79931ef555d54118f64dc9be2cf2dbf59636ce253f7a1179c4a1c091442a6eeab8352220c7c56ca1ff7c4b4b0fc345c74e3e85ecb9d43deeb66b5f803d4a0 | 1 | 1 | 0 |
| FIRST\_CRASH\_TYPE | V\_W String | 0.0% | 18 | ANGLE | SIDESWIPE OPPOSITE DIRECTION | 46 | 207,635 | 0 |
| REPORT\_TYPE | V\_W String | 3.1% | 4 | AMENDED | NOT ON SCENE (DESK REPORT) | 240 | 488,342 | 27829 |
| TRAFFICWAY\_TYPE | V\_W String | 0.0% | 20 | RAMP | DIVIDED - W/MEDIAN (NOT RAISED) | 185 | 386,429 | 0 |
| STATEMENTS\_TAKEN\_I | V\_W String | 97.7% | 3 | Y | Y | 3,738 | 876,130 | 876130 |
| STREET\_DIRECTION | V\_W String | 0.0% | 5 | S | S | 4 | 321,103 | 0 |
| ALIGNMENT | V\_W String | 0.0% | 6 | CURVE, LEVEL | STRAIGHT ON HILLCREST | 388 | 875,515 | 0 |
| TRAFFIC\_CONTROL\_DEVICE | V\_W String | 0.0% | 19 | OTHER | PEDESTRIAN CROSSING SIGN | 34 | 507,733 | 0 |
| LOCATION | V\_W String | 0.7% | 318,508 | POINT (0 0) | POINT (-87.665902342962 41.854120262952) | 1 | 6,467 | 6467 |
| HIT\_AND\_RUN\_I | V\_W String | 68.6% | 3 | Y | Y | 12,052 | 615,492 | 615492 |
| WORK\_ZONE\_I | V\_W String | 99.4% | 3 | Y | Y | 1,139 | 891,740 | 891740 |
| PHOTOS\_TAKEN\_I | V\_W String | 98.6% | 3 | Y | Y | 3,023 | 884,518 | 884518 |
| DEVICE\_CONDITION | V\_W String | 0.0% | 8 | OTHER | WORN REFLECTIVE MATERIAL | 99 | 513,745 | 0 |
| MOST\_SEVERE\_INJURY | V\_W String | 0.2% | 6 | FATAL | NONINCAPACITATING INJURY | 981 | 768,812 | 1986 |
| NOT\_RIGHT\_OF\_WAY\_I | V\_W String | 95.4% | 3 | Y | Y | 3,752 | 855,911 | 855911 |
| SEC\_CONTRIBUTORY\_CAUSE | V\_W String | 0.0% | 40 | ANIMAL | OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER | 58 | 369,647 | 0 |
| PRIM\_CONTRIBUTORY\_CAUSE | V\_W String | 0.0% | 40 | ANIMAL | OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER | 23 | 350,757 | 0 |
| INTERSECTION\_RELATED\_I | V\_W String | 77.0% | 3 | Y | Y | 9,818 | 690,843 | 690843 |
| WORKERS\_PRESENT\_I | V\_W String | 99.9% | 3 | Y | Y | 147 | 895,461 | 895461 |
| WORK\_ZONE\_TYPE | V\_W String | 99.6% | 5 | UTILITY | CONSTRUCTION | 242 | 892,879 | 892879 |
| ROADWAY\_SURFACE\_COND | V\_W String | 0.0% | 7 | DRY | SAND, MUD, DIRT | 321 | 663,398 | 0 |
| DAMAGE | V\_W String | 0.0% | 3 | OVER $1,500 | $501 - $1,500 | 100,735 | 564,207 | 0 |
| CRASH\_TYPE | V\_W String | 0.0% | 2 | NO INJURY / DRIVE AWAY | INJURY AND / OR TOW DUE TO CRASH | 241,294 | 655,462 | 0 |

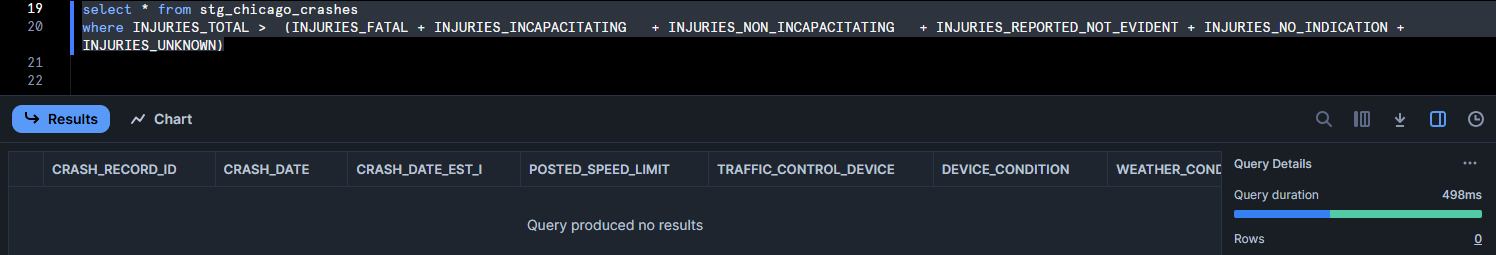
**Dataset Overview:**

1. **Number of fields and observations:** There are in total 896,756 observations and 48 variables or fields. This scope indicates a thorough data collection process, encompassing a diverse array of details about traffic incidents.
2. **Datatypes of Fields:** There are 2 Date Time columns, 9 Boolean columns, 20 String columns and 17 Numeric columns.

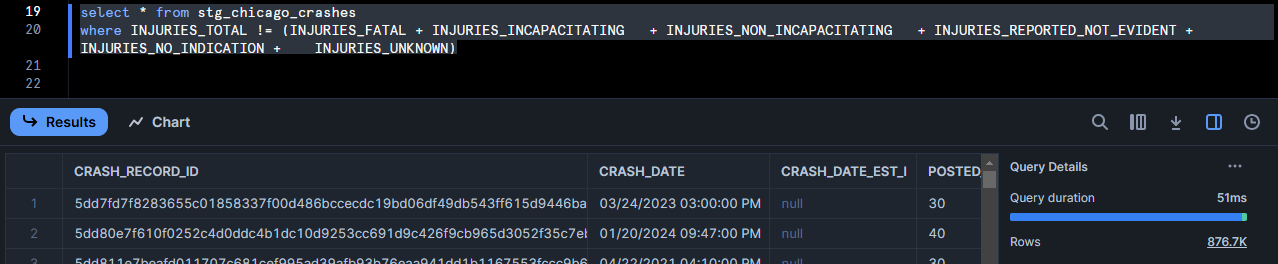
**Inferences from the dataset:**

1. Columns like DOORING\_I (99.7% missing), WORKERS\_PRESENT\_I (99.9%), WORK\_ZONE\_I (99.4%), and WORK\_ZONE\_TYPE (99.6%) have an overwhelmingly high percentage of missing values. These could potentially be dropped.
2. Although columns like LANE\_CNT and INTERSECTION\_RELATED\_I have 77.8 % and 77.0% nulls, it might be relevant for analysis, so these columns as well can be dropped if proven not worthy of analysis.
3. Columns such as STATEMENTS\_TAKEN\_I, PHOTOS\_TAKEN\_I, and HIT\_AND\_RUN\_I have Boolean-like values (Y/N) but include missing entries. These should be converted into binary format (1/0) for easier processing, and missing values should be handled explicitly like 0 in case of missing values.
4. For geographical analysis, apart from latitude and longitude, street no and street name, other columns such as street\_direction and location can be dropped as these columns doesn’t add much value.
5. CRASH\_DATE and DATE\_POLICE\_NOTIFIED can be split into separate DATE and TIME components for temporal analysis.
6. There are no records where the latitude, longitude is null and the location is not.
7. Injuries total column could be re- calculated by performing a sum operation on the columns that denotes injuries.

It was also observed that there are no records where the injuries total is greater than the aggregated value, denoting that there is no category that is left out.



Also, we find around 876k records where the injuries total does not match the sum of other injury columns.



So, the injuries total could be recalculated.