**Group 14 - Data Analysis Report**

Prasad Gavas

Shantanu Mahakal

Prarthana Shetty  
Soham Shah

**Project Overview:**

# **Austin Dataset:**

Crash data is obtained from the Texas Department of Transportation (TXDOT) Crash Record Information System (CRIS) database, which is populated by reports submitted by Texas Peace Officers throughout the state, including Austin Police Department (APD), and maintained by TXDOT.

This dataset contains crash-level records for crashes which have occurred in the last ten years. Crash data takes several days or weeks to be initially provided and finalized as it is furnished to the Austin Transportation & Public Works Department, therefore a two-week delay is observed that ensures more accurate and complete results.

# **Chicago Dataset:**

Crash data is obtained from the Texas Department of Transportation (TXDOT) Crash Record Information System (CRIS) database, which is populated by reports submitted by Texas Peace Officers throughout the state, including Austin Police Department (APD), and maintained by TXDOT.

This dataset contains crash-level records for crashes which have occurred in the last ten years. Crash data may take several days or weeks to be initially provided and finalized as it is furnished to the Austin Transportation & Public Works Department, therefore a two-week delay is observed that ensures more accurate and complete results.

**NYC Dataset:**

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. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage. It should be noted that the data is preliminary and subject to change when the MV-104AN forms are amended based on revised crash details.

Police officers complete form MV-104AN for all vehicle collisions. The MV-104AN is a New York State form that has all of the details of a traffic collision.

### 

### **Step 1: Understanding the Dataset**

● Analyzed and discussed the Schema and Content of the provided dataset

● Identified key tables, fields, and relationships to establish a foundational understanding

● Noted and handled all unique characteristics or challenges within the dataset's organization

● Identified potential data quality issues, outliers, or anomalies

● Discussed a schema structure for both datasets

### **Step 2: Data Profiling using ydata\_profiling**

### 

● Imported the datasets

● Data Import and Initial Assessment:

● Data Quality Review

● Column Profiling

● Correlation Analysis

● Data Transformation Insight

**Aim of the Assignment:**

The aim of this project is to construct a comprehensive and advanced data architecture that will serve as the backbone for business intelligence and analytical solutions in the domain of traffic safety and vehicular incidents. Leveraging detailed crash data from three major cities—New York, Chicago, and Austin—this initiative strives to harmonize disparate datasets into a singular, unified schema that will facilitate in-depth analysis and reporting on various aspects of motor vehicle collisions.

The goal is to create a reliable and scalable data warehouse that will enable stakeholders to uncover critical insights such as accident hotspots, temporal trends in accident occurrences, the incidence of injuries and fatalities, and contributory factors to accidents.

**Data Source Observations:**

**NYC Dataset**

| **Name** | **Datatype** | **Unique values** | **Null values** | **Metadata** | **Min Value** | **Max Value** |
| --- | --- | --- | --- | --- | --- | --- |
| CRASH DATE | Date | 0.2% | 0.00% | Occurrence date of collision | 20212-07-01 | 2024-03-22 |
| CRASH TIME | Date | 0.1% | 0.00% | Occurrence time of collision | 2024-03-29 00:00:00 | 2024-03-29 23:59:00 |
| BOROUGH | Categorical | <0.1% | 31.1% | Borough where collision occurred | 5 | 13 |
| ZIP CODE | Text | <0.1% | 31.1% | Postal code of incident occurrence | 5 | 5 |
| LATITUDE | Real Number | 6.9% | 11.3% | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | 0 | 43.344 |
| LONGITUDE | Real Number | 5.3% | 11.3% | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326). | -201.359 | 0 |
| LOCATION | Text | 15.4% | 11.3% | Latitude , Longitude pair | 25 | 10 |
| ON STREET NAME | Text | 1.1% | 21.2% | Street on which the collision occurred | 2 | 32 |
| CROSS STREET NAME | Text | 1.6% | 37.8% | Nearest cross street to the collision | 1 | 32 |
| OFF STREET NAME | Text | 64.9% | 83.2% | Street address if known | 8 | 40 |
| NUMBER OF PERSONS INJURED | Real Number | <0.1% | 18 | Number of persons injured | 0 | 43 |
| NUMBER OF PERSONS KILLED | Real Number | <0.1% | <0.1% | Number of persons killed | 0 | 8 |
| NUMBER OF PEDESTRIANS INJURED | Real Number | <0.1% | 0.0% | Number of pedestrians injured | 0 | 27 |
| NUMBER OF PEDESTRIANS KILLED | Categorical- Wrong | <0.1% | 0% | Number of pedestrians killed | 0 | 6 |
| NUMBER OF CYCLIST INJURED | Categorical- Wrong | <0.1% | 0% | Number of cyclists injured | 0 | 4 |
| NUMBER OF CYCLIST KILLED | Categorical | <0.1% | 0% | Number of cyclists killed | 0 | 2 |
| NUMBER OF MOTORIST INJURED | Real Number | <0.1% | 0% | Number of vehicle occupants injured | 0 | 43 |
| NUMBER OF MOTORIST KILLED | Real Number | <0.1% | 0% | Number of vehicle occupants killed | 0 | 5 |
| CONTRIBUTING FACTOR VEHICLE 1 | Text | <0.1% | 0.3% | Factors contributing to the collision for designated vehicle | 1 | 53 |
| CONTRIBUTING FACTOR VEHICLE 2 | Text | <0.1% | 15.5% | Factors contributing to the collision for designated vehicle | 1 | 53 |
| CONTRIBUTING FACTOR VEHICLE 3 | Categorical | <0.1% | 92.9% | Factors contributing to the collision for designated vehicle | 1 | 53 |
| CONTRIBUTING FACTOR VEHICLE 4 | Categorical | <0.1% | 98.4% | Factors contributing to the collision for designated vehicle | 5 | 43 |
| CONTRIBUTING FACTOR VEHICLE 5 | Categorical | 0.3% | 99.6% | Factors contributing to the collision for designated vehicle | 5 | 43 |
| COLLISION\_ID | Real Number | 100% | 0% | Unique record code generated by system. Primary Key for Crash table. | 22 | 4712252 |
| VEHICLE TYPE CODE 1 | Text | 0.1% | 0.7% | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | 1 | 38 |
| VEHICLE TYPE CODE 2 | Text | 0.1% | 19.1% | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | 1 | 38 |
| VEHICLE TYPE CODE 3 | Text | 0.2% | 93.1% | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | 2 | 35 |
| VEHICLE TYPE CODE 4 | Text | 0.3% | 98.4% | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | 2 | 35 |
| VEHICLE TYPE CODE 5 | Text | 0.08% | 99.6% | Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other) | 2 | 35 |

**Austin Dataset**

| **Name** | **Datatype** | **Unique values** | **Null values** | **Metadata** | **Min Value** | **Max Value** |
| --- | --- | --- | --- | --- | --- | --- |
| crash\_id | Real\_number | 100% | 0.00% | TxDOT C.R.I.S. system-generated unique identifying number for a crash | 1001 | 1.802 \*10^8 |
| crash\_fatal\_fl | Boolean | <0.1% | 0% | Fatal Crash Identifier - Indicates that the crash involved one or more fatalities | False | True |
| crash\_date | Date | 97.9% | 0.0% | Crash Date | 2014-03-26 06:41:00 | 2024-03-11 22:05:00 |
| crash\_time | Date | 1% | 0.0% | Crash Time - Time crash occurred | 2024-03-29 00:00:00 | 2024-03-29 23:59:00 |
| case\_id | Text | 99.9% | 1.3% | Case ID | 1 | 20 |
| rpt\_latitude | Real Number | 77.5% | 93% | Reported Latitude | 25.83746 | 36.50048 |
| rpt\_longitude | Real Number | 70.6% | 93% | Reported Longitude | -106.645 | -93.50795 |
| rpt\_block\_num | Text | 3.7% | 13.3% | Reported Block Number (road on which crash occurred) | 1 | 9 |
| rpt\_street\_pfx | Categorical | <0.1% | 45.9% | Reported Street Prefix (road on which crash occurred: N, S, E, W, SW..) | 1 | 2 |
| rpt\_street\_name | Text | 6.6% | <0.1% | Reported Street Name (road on which crash occurred) | 1 | 49 |
| rpt\_street\_sfx | Categorical | <0.1% | 34.1% | Reported Street Suffix (road on which crash occurred) | 2 | 4 |
| crash\_speed\_limit | Real number | <0.1% | <0.1% | Speed Limit | -1 | 85 |
| road\_constr\_zone\_fl | Boolean | <0.1% | <0.1% | Construction Zone - Indicates whether the crash occurred in or was related to a construction, maintenance, or utility work zone, regardless of whether or not workers were actually present at the time of the crash | False | True |
| Latitude | Real Number | 66.2% | 1.5% | Derived Latitude map coordinate of the crash | 30.0987 | 30.5116 |
| Longitude | Real Number | 66.1% | 1.5% | Derived Longitude map coordinate of the crash | -97.926 | -97.570 |
| street\_name | Text | 3.1% | <0.1% | Derived Street Name - Name of the road crash occurred on, as determined by the Locator application. | 3 | 41 |
| street\_nbr | Real Number | 16.2% | 58.9% | Derived Street Number - Block number of primary street where crash occurred as determined by the Locator application | 0 | 21146 |
| street\_name\_2 | Text | 5.1% | 55.1% | Derived Street Name 2 - The road name for the secondary road related to the crash location (If applicable) | 3 | 38 |
| street\_nbr\_2 |  | <0.1% | 0.3% | Derived Street Number 2 - Block number of secondary street related to the crash location as determined by the Locator application (If applicable) | 1 | 53 |
| crash\_sev\_id | Real Number | <0.1% | 0% | Crash Severity - Most severe injury suffered by any one person involved in the crash ( 0=UNKNOWN, 1=INCAPACITATING INJURY, 2=NON-INCAPACITATING INJURY, 3=POSSIBLE INJURY, 4=KILLED, 5=NOT INJURED) | 0 | 99 |
| sus\_serious\_injry\_cnt | Real Number | <0.1% | 0% | Total Suspected Serious Injury Count | 0 | 10 |
| nonincap\_injry\_cnt | Real Number | <0.1% | <0.1% | Total Non-incapacitating Injury Count | 0 | 14 |
| poss\_injry\_cnt | Real Number | <0.1% | <0.1% | Total Possible Injury Count | 0 | 20 |
| non\_injry\_cnt | Real Number | <0.1% | <0.1% | Total Not Injured Count. | 0 | 56 |
| unkn\_injry\_cnt | Real Number | <0.1% | <0.1% | Total Unknown Injury Count | 0 | 41 |
| tot\_injry\_cnt | Real Number | <0.1% | <0.1% | Total Injury Count | 0 | 21 |
| death\_cnt | Categorical- Wrong | <0.1% | 0% | Total Death Count | 0 | 4 |
| contrib\_factr\_p1\_id | Real Number | 0.2% | 80.6% | The first factor for a given vehicle which the officer felt possibly contributed to the crash | 1 | 80 |
| contrib\_factr\_p2\_id | Real Number | 1.4% | 96.9% | The second factor for a given vehicle which the officer felt possibly contributed to the crash | 1 | 79 |
| units\_involved | Text | 0.8% | <0.1% | Mode of units involved in crash | 10 | 321 |
| atd\_mode\_category\_metadata | Text | 100% | <0.1% | Description of units involved in crash | 214 | 3936 |
| pedestrian\_fl | Boolean | <0.1% | 97.6% | Pedestrian involved crash flag | Missing | True |
| motor\_vehicle\_fl | Boolean | <0.1% | 0.8% | Motor vehicle involved crash flag | Missing | True |
| motorcycle\_fl | Boolean | <0.1% | 97.6% | Motorcycle involved crash flag | Missing | True |
| bicycle\_fl | Boolean | <0.1% | 98.3% | Bicyclist involved crash flag | Missing | True |
| other\_fl | Boolean | <0.1% | 96.7% | Other involved crash flag | Missing | True |
| point | Text | 67.2% | 1.5% | Point data type created with crash latitude and longitude to enable request of GeoJSON. | 22 | 45 |
| apd\_confirmed\_fatality | Boolean | <0.1% | 0% | APD Fatality flag | False | True |
| apd\_confirmed\_death\_count | Categorical-Wrong | <0.1% | 0% | APD Fatality Count | 0 | 4 |
| motor\_vehicle\_death\_count | Categorical-Wrong | <0.1% | 0% |  | 0 | 4 |
| motor\_vehicle\_serious\_injury\_count | Real Number | <0.1% | 0% |  | 0 | 5 |
| bicycle\_death\_count | Categorical-Wrong | <0.1% | 0% |  | 0 | 1 |
| bicycle\_serious\_injury\_count | Categorical-Wrong | <0.1% | 0% |  | 0 | 3 |
| pedestrian\_death\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 2 |
| pedestrian\_serious\_injury\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 9 |
| motorcycle\_death\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 2 |
| motorcycle\_serious\_injury\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 2 |
| other\_death\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 0 |
| other\_serious\_injury\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 3 |
| onsys\_fl | Boolean | <0.1% | 0% | Flag indicates whether primary road of crash was on the TxDOT highway system. | False | True |
| private\_dr\_fl | Boolean | <0.1% | 0% | Flag indicating whether crash occurred on a private drive or road/private property/parking lot. | False | False |
| micromobility\_serious\_injury\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 2 |
| micromobility\_death\_count | Categorical- Wrong | <0.1% | 0% |  | 0 | 1 |
| micromobility\_fl | Boolean | 0.3% | 99.8% |  | Missing | True |

**Chicago Dataset:**

| **Name** | **Datatype** | **Unique values** | **Null values** | **Metadata** | **Min Value** | **Max Value** |
| --- | --- | --- | --- | --- | --- | --- |
| CRASH\_RECORD\_ID | Text | 100% | 0.00% | 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 | 128 | 128 |
| CRASH\_DATE\_EST\_I | Boolean | <0.1% | 92.5% | Crash date estimated by desk officer or reporting party (only used in cases where crash is reported at police station days after the crash) | False | True |
| CRASH\_DATE | Date | 65.7% | 0.0% | Date and time of crash as entered by the reporting officer | 2013-03-03 16:48:00 | 2024-03-26 01:40:00 |
| POSTED\_SPEED\_LIMIT | Real Number | <0.1% | 0.0% | Posted speed limit, as determined by reporting officer | 0 | 99 |
| TRAFFIC\_CONTROL\_DEVICE | Categorical | <0.1% | 0% | Traffic control device present at crash location, as determined by reporting officer | 5 | 24 |
| DEVICE\_CONDITION | Categorical | <0.1% | 0% | Condition of traffic control device, as determined by reporting officer | 5 | 24 |
| WEATHER\_CONDITION | Categorical | <0.1% | 0% | Weather condition at time of crash, as determined by reporting officer | 4 | 24 |
| LIGHTING\_CONDITION | Categorical | <0.1% | 0% | Light condition at time of crash, as determined by reporting officer | 4 | 22 |
| FIRST\_CRASH\_TYPE | Categorical | <0.1% | 0% | Type of first collision in crash | 5 | 28 |
| TRAFFICWAY\_TYPE | Categorical | <0.1% | 0% | Trafficway type, as determined by reporting officer | 4 | 31 |
| LANE\_CNT | Real Number | <0.1% | 75% | Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection) | 0 | 1191625 |
| ALIGNMENT | Categorical | <0.1% | 0% | Street alignment at crash location, as determined by reporting officer | 12 | 21 |
| ROADWAY\_SURFACE\_COND | Categorical | <0.1% | 0% | Road surface condition, as determined by reporting officer | 3 | 15 |
| ROAD\_DEFECT | Categorical | <0.1% | 0% | Road defects, as determined by reporting officer | 5 | 17 |
| REPORT\_TYPE | Categorical | <0.1% | 3.0% | Administrative report type (at scene, at desk, amended) | 7 | 26 |
| CRASH\_TYPE | Categorical | <0.1% | 0% | A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away | 22 | 32 |
| INTERSECTION\_RELATED\_I | Boolean | <0.1% | 77.1% | 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. | FALSE | TRUE |
| NOT\_RIGHT\_OF\_WAY\_I | Boolean | <0.1% | 95.4% | Whether the crash begun or first contact was made outside of the public right-of-way. | FALSE | TRUE |
| HIT\_AND\_RUN\_I | Boolean | <0.1% | 68.7% | Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid | FALSE | TRUE |
| DAMAGE | Categorical | <0.1% | 0% | A field observation of estimated damage. | 11 | 13 |
| DATE\_POLICE\_NOTIFIED | Date | 75.9% | 0% | Calendar date on which police were notified of the crash | 2013-06-01 20:31:00 | 2024-03-26 01:42:00 |
| PRIM\_CONTRIBUTORY\_CAUSE | Categorical | <0.1% | 0% | The factor which was most significant in causing the crash, as determined by officer judgment | 6 | 80 |
| SEC\_CONTRIBUTORY\_CAUSE | Categorical | <0.1% | 0% | The factor which was second most significant in causing the crash, as determined by officer judgment | 6 | 80 |
| STREET\_NO | Real Number | 1.4% | 0% | Street address number of crash location, as determined by reporting officer | 0 | 451100 |
| STREET\_DIRECTION | Categorical | <0.1% | <0.1% | Street address direction (N,E,S,W) of crash location, as determined by reporting officer | 1 | 1 |
| STREET\_NAME | Text | <0.2% | <0.1% | Street address name of crash location, as determined by reporting officer | 4 | 31 |
| BEAT\_OF\_OCCURRENCE | Real Number | <0.1% | <0.1% | Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74 | 111 | 6100 |
| PHOTOS\_TAKEN\_I | Boolean | 2% | 98.7% | Whether the Chicago Police Department took photos at the location of the crash | FALSE | TRUE |
| STATEMENTS\_TAKEN\_I | Boolean | <0.1% | 97.8% | Whether statements were taken from unit(s) involved in crash | FALSE | TRUE |
| DOORING\_I | Boolean | 0.1% | 99.7% | Whether crash involved a motor vehicle occupant opening a door into the travel path of a bicyclist, causing a crash | FALSE | TRUE |
| WORK\_ZONE\_I | Boolean | <0.1% | 99.4% | Whether the crash occurred in an active work zone | FALSE | TRUE |
| WORK\_ZONE\_TYPE | Categorical | 0.1% | 99.6% | The type of work zone, if any | 7 | 12 |
| WORKERS\_PRESENT\_I | Boolean | 0.2% | 99.9% | Whether construction workers were present in an active work zone at crash location | FALSE | TRUE |
| NUM\_UNITS | Real Number | <0.1% | 0% | 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. | 1 | 18 |
| MOST\_SEVERE\_INJURY | Categorical | <0.1% | 0.2% | Most severe injury sustained by any person involved in the crash | 5 | 24 |
| INJURIES\_TOTAL | Real Number | <0.1% | 0.2% | Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer | 0 | 21 |
| INJURIES\_FATAL | Categorical- Wrong | <0.1% | 0.2% | Total persons sustaining fatal injuries in the crash. | 0 | 4 |
| INJURIES\_INCAPACITATING | Real Number | <0.1% | 0.2% | 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. | 0 | 10 |
| INJURIES\_NON\_INCAPACITATING | Real Number | <0.1% | 0.2% | 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. | 0 | 21 |
| INJURIES\_REPORTED\_NOT\_EVIDENT | Real Number | <0.1% | 0.2% | 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. | 0 | 15 |
| INJURIES\_NO\_INDICATION | Real Number | <0.1% | 0.2% | Total persons sustaining no injuries in the crash as determined by the reporting officer | 0 | 61 |
| INJURIES\_UNKNOWN | Categorical | <0.1% | 0.2% | Total persons for whom injuries sustained, if any, are unknown | 3 | 3 |
| CRASH\_HOUR | Real Number | <0.1% | 0% | The hour of the day component of CRASH\_DATE. | 0 | 23 |
| CRASH\_DAY\_OF\_WEEK | Real Number | <0.1% | 0% | The day of the week component of CRASH\_DATE. Sunday=1 | 1 | 7 |
| CRASH\_MONTH | Real Number | <0.1% | 0% | The month component of CRASH\_DATE. | 1 | 12 |
| LATITUDE | Real Number | 37% | 0.7% | The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash | 0 | 42.02278 |
| LONGITUDE | Real Number | 36.9% | 0.7% | The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash | -87.936193 | 0 |
| LOCATION | Text | 37% | 0.7% | 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 | 11 | 40 |

**NYC Dataset:**

**CRASH DATE & CRASH TIME:** These columns contain the date and time of the collision, respectively. They have been appropriately formatted and are crucial for analyzing trends over time.

**BOROUGH & ZIP CODE**: These categorical variables identify the location of the collision, with a significant percentage of null values. Boroughs where collisions occur are crucial for understanding geographic patterns.

**LATITUDE & LONGITUDE:** Real number coordinates of the collision location. They have a considerable percentage of null values and provide precise spatial information.

**LOCATION:** This column contains latitude and longitude pairs, which can be utilized for mapping and spatial analysis. However, it has a high percentage of null values.

**NUMBER OF PERSONS INJURED & NUMBER OF PERSONS KILLED:** These numerical variables provide information on the severity of the collisions, with a low percentage of null values.

**CONTRIBUTING FACTOR VEHICLE 1-5:** These categorical variables identify factors contributing to collisions, with varying percentages of null values.

**COLLISION\_ID:** This serves as a unique identifier for each collision record and is crucial for data management.

**VEHICLE TYPE CODE 1-5:** These categorical variables describe the types of vehicles involved in the collisions, with varying percentages of null values.

**Key Observations**  
Spatial Distribution:Spatial distribution analysis using latitude and longitude coordinates enables us to map and visualize the geographical spread of collisions across New York City. By plotting these coordinates on a map, we can identify collision hotspots, concentration areas, and spatial patterns of collision occurrence.   
Temporal Trends: I observed interesting temporal patterns in the crash date and time columns, indicating potential daily, weekly, or seasonal variations in collision occurrence. Understanding these trends is crucial for identifying high-risk periods and implementing targeted interventions to enhance road safety.  
Severity Analysis: By examining the number of persons injured and killed, I gained insight into the severity of collisions. Understanding the factors contributing to severe collisions is essential for developing effective strategies to reduce injuries and fatalities on NYC roads and to answer the business requirements.  
Contributing Factors: Analysis of contributing factor variables revealed common causes of collisions, such as distracted driving, speeding, or poor road conditions. Addressing these factors through targeted interventions and public awareness campaigns is crucial for improving overall road safety.  
Data Quality Concerns: The presence of null values and inconsistencies in data completeness across columns raises concerns about data quality. It's imperative to conduct thorough data cleansing and validation to ensure the reliability of analytical results and insights derived from the dataset.

**Austin Dataset:**

**crash\_id**: This column serves as a unique identifier.

**crash\_date & crash\_time**: These columns represent the date and time of the collision occurrence, crucial for temporal analysis.

**Latitude & Longitude**: Derived coordinates of the collision location, providing spatial information for mapping and analysis.

**crash\_sev\_id**: This numerical variable indicates the severity of the crash, ranging from unknown to fatal injuries.

**contrib\_factr\_p1\_id & contrib\_factr\_p2\_id**: These numerical variables identify primary and secondary contributing factors to the collision, respectively.

**Key Observations:**

Temporal Patterns: Analysis of crash date and time revealed distinct temporal trends in collision occurrence, including daily, weekly, or hourly patterns. This information is essential for implementing timely interventions and resource allocation during high-risk periods.

Spatial Distribution: With latitude and longitude coordinates available, I conducted spatial analysis to identify collision hotspots and distribution patterns across Austin. Understanding these patterns can inform urban planning and infrastructure improvements to enhance road safety.

Severity Assessment: By examining the crash\_sev\_id column, I gained insight into the severity of collisions. Understanding factors contributing to severe collisions is crucial for developing strategies to reduce injury rates and improve road safety.

Contributing Factors: Analysis of contributing factor variables highlighted common causes of collisions in Austin, such as impaired driving or adverse weather conditions. Addressing these factors through targeted interventions can mitigate collision risks and enhance road safety.

Data Quality Considerations: Similar to the NYC dataset, concerns regarding null values and data completeness necessitate thorough data cleansing and validation processes. Ensuring data accuracy and reliability is essential for deriving actionable insights from the dataset.

**Chicago Dataset:**

**CRASH\_RECORD\_ID**: These columns serve as identifiers

**CRASH\_DATE & CRASH\_TIME**: These columns represent the date and time of the collision occurrence, essential for temporal analysis.

**POSTED\_SPEED\_LIMIT & TRAFFIC\_CONTROL\_DEVICE:** These categorical variables provide information on speed limits and traffic control devices present at the crash location.

**WEATHER\_CONDITION & LIGHTING\_CONDITION:** These categorical variables describe weather and lighting conditions at the time of the collision, impacting visibility and road conditions.

**LANE\_CNT & ALIGNMENT**: These numerical and categorical variables offer insights into the number of lanes and street alignment at the crash location.

**PRIM\_CONTRIBUTORY\_CAUSE & SEC\_CONTRIBUTORY\_CAUSE**: These categorical variables identify primary and secondary contributing factors to the collision, respectively.

**NUM\_UNITS & MOST\_SEVERE\_INJURY**: These numerical and categorical variables provide information on the number of units involved and the most severe injury sustained in the collision.

**Key Observations:**

Temporal Analysis: Analysis of crash date and time revealed distinct temporal patterns in collision occurrence, including daily, weekly, or monthly trends. Understanding these patterns is essential for implementing timely interventions and allocating resources effectively.

Environmental Conditions: Variables such as weather\_condition and lighting\_condition provided insights into environmental factors influencing collision occurrence. Analyzing the impact of these factors on collision rates can inform strategies to improve road safety under adverse conditions.

Roadway Characteristics: By examining lane\_cnt and alignment variables, I gained insights into roadway features at crash locations. Assessing the relationship between these features and collision occurrence is crucial for guiding infrastructure improvements and traffic management strategies.

Contributing Factors: Analysis of contributing cause variables identified primary and secondary factors contributing to collisions in Chicago. Understanding these factors is essential for developing targeted interventions to address root causes of collisions.

Data Completeness and Accuracy: Concerns regarding null values and data completeness across columns underscore the importance of thorough data validation and cleansing. Addressing these issues is essential to ensure the reliability and validity of analytical insights derived from the dataset.

Note: In the process of profiling all 3 accident dataset, a critical observation has been made regarding the data types assigned to certain variables. A few variables have been inaccurately classified as categorical due to the presence of a limited number of distinct values within the dataset. However, these variables are inherently continuous and should not be confined to observed categories. These variables have been duly highlighted in yellow to signify their misclassification.  
For instance, variables such as NUMBER OF PERSONS INJURED and NUMBER OF PERSONS KILLED might exhibit a small range of distinct numbers due to the dataset's constraints (e.g., most accidents involve 0-5 injuries). However, their theoretical range is not bounded by these observed values, and they can take on any non-negative integer value.