Objective & Dataset Introduction

This analysis aims to predict whether a traffic crash results in injury using a historical dataset from Chicago. The dataset includes various environmental, vehicular, and situational attributes (e.g., weather, lighting, control device) that may influence injury outcomes. The goal is to:

- Preprocess and clean the data
- Explore and visualize key relationships
- Engineer features and prepare them for modeling
- Train and evaluate classification and regression models

Step 1: Load and Inspect the Dataset

We begin by loading the dataset, displaying the shape, the first few records, and a summary of data types and statistical information to understand the structure and content.

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

# Load dataset
file_path = 'Dataset for Data Mining.csv'  # replace with your actual path
df = pd.read_csv(file_path)

# Display basic info
print("💾 Shape:", df.shape)
print("\n📌 First 5 rows:")
print(df.head())

print("\n📄 Dataset Info:")
print(df.info())

print("\n📊 Summary Statistics:")
print(df.describe(include='all').T)
```

```
💾 Shape: (235978, 48)

📌 First 5 rows:
                                   CRASH_RECORD_ID CRASH_DATE_EST_I  \
0  ec987eca9d84e964c10e750c95b8dca890036480463020...              NaN
1  deafa5f8bdd844d7154eae177cbd86bf5754991a0c7e77...              NaN
2  96577eb19233057b1bf0b40befc1acbcc26b21fe7abb93...              NaN
3  31e890ddbd9c76afdcd32bcda5f8e885311cf486b8afae...              NaN
4  49dcf06cc6debf56fea45cbdd5c07156bd0bc6a475657c...              NaN

               CRASH_DATE  POSTED_SPEED_LIMIT TRAFFIC_CONTROL_DEVICE  \
0  11/19/2019 09:20:00 AM                30.0            NO CONTROLS
1  02/12/2025 01:00:00 PM                35.0         TRAFFIC SIGNAL
2  12/30/2022 09:50:00 PM                30.0            NO CONTROLS
3  07/15/2018 10:00:00 PM                25.0            NO CONTROLS
4  10/27/2024 10:30:00 PM                30.0            NO CONTROLS

       DEVICE_CONDITION WEATHER_CONDITION      LIGHTING_CONDITION  \
0           NO CONTROLS   CLOUDY/OVERCAST                DAYLIGHT
1  FUNCTIONING PROPERLY             CLEAR                DAYLIGHT
2           NO CONTROLS             CLEAR  DARKNESS, LIGHTED ROAD
3           NO CONTROLS             CLEAR  DARKNESS, LIGHTED ROAD
4           NO CONTROLS             CLEAR  DARKNESS, LIGHTED ROAD

           FIRST_CRASH_TYPE                  TRAFFICWAY_TYPE  ...  \
0                  TURNING                      NOT DIVIDED  ...
1                 REAR END                          ONE-WAY  ...
2      PARKED MOTOR VEHICLE                      NOT DIVIDED  ...
3  SIDESWIPE SAME DIRECTION  DIVIDED - W/MEDIAN (NOT RAISED)  ...
4               PEDESTRIAN                      NOT DIVIDED  ...

   INJURIES_NON_INCAPACITATING INJURIES_REPORTED_NOT_EVIDENT  \
0                          0.0                          0.0
1                          0.0                          0.0
2                          0.0                          0.0
3                          0.0                          0.0
4                          0.0                          0.0

  INJURIES_NO_INDICATION INJURIES_UNKNOWN CRASH_HOUR CRASH_DAY_OF_WEEK  \
0                    2.0              0.0        9.0               3.0
1                    2.0              0.0       13.0               4.0
2                    1.0              0.0       21.0               6.0
3                    3.0              0.0       22.0               1.0
4                    3.0              0.0       22.0               1.0

  CRASH_MONTH   LATITUDE  LONGITUDE                                 LOCATION
0        11.0  41.997278 -87.709407  POINT (-87.709407357178 41.997278474417)
1         2.0  41.736209 -87.624306  POINT (-87.624305519124 41.736209469449)
2        12.0  41.762562 -87.683039  POINT (-87.683038501875 41.762561538156)
3         7.0  41.960890 -87.742650  POINT (-87.742649505184 41.960889822368)
4        10.0  41.865306 -87.659382   POINT (-87.65938195693 41.865306036308)

[5 rows x 48 columns]

📋 Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235978 entries, 0 to 235977
Data columns (total 48 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
```

```
 0   CRASH_RECORD_ID                235945 non-null  object
 1   CRASH_DATE_EST_I               17364 non-null   object
 2   CRASH_DATE                     235948 non-null  object
 3   POSTED_SPEED_LIMIT             235941 non-null  float64
 4   TRAFFIC_CONTROL_DEVICE         235940 non-null  object
 5   DEVICE_CONDITION               235942 non-null  object
 6   WEATHER_CONDITION              235933 non-null  object
 7   LIGHTING_CONDITION             235944 non-null  object
 8   FIRST_CRASH_TYPE               235939 non-null  object
 9   TRAFFICWAY_TYPE                235942 non-null  object
 10  LANE_CNT                       49710 non-null   float64
 11  ALIGNMENT                      235942 non-null  object
 12  ROADWAY_SURFACE_COND           235941 non-null  object
 13  ROAD_DEFECT                    235940 non-null  object
 14  REPORT_TYPE                    228323 non-null  object
 15  CRASH_TYPE                     235942 non-null  object
 16  INTERSECTION_RELATED_I         54033 non-null   object
 17  NOT_RIGHT_OF_WAY_I             10733 non-null   object
 18  HIT_AND_RUN_I                  73949 non-null   object
 19  DAMAGE                         235943 non-null  object
 20  DATE_POLICE_NOTIFIED           235938 non-null  object
 21  PRIM_CONTRIBUTORY_CAUSE        235935 non-null  object
 22  SEC_CONTRIBUTORY_CAUSE         235938 non-null  object
 23  STREET_NO                      235940 non-null  float64
 24  STREET_DIRECTION               235933 non-null  object
 25  STREET_NAME                    235944 non-null  object
 26  BEAT_OF_OCCURRENCE             235942 non-null  float64
 27  PHOTOS_TAKEN_I                 3288 non-null    object
 28  STATEMENTS_TAKEN_I             5473 non-null    object
 29  DOORING_I                      730 non-null     object
 30  WORK_ZONE_I                    1296 non-null    object
 31  WORK_ZONE_TYPE                 1004 non-null    object
 32  WORKERS_PRESENT_I              315 non-null     object
 33  NUM_UNITS                      235934 non-null  float64
 34  MOST_SEVERE_INJURY             235435 non-null  object
 35  INJURIES_TOTAL                 235442 non-null  float64
 36  INJURIES_FATAL                 235446 non-null  float64
 37  INJURIES_INCAPACITATING        235447 non-null  float64
 38  INJURIES_NON_INCAPACITATING    235439 non-null  float64
 39  INJURIES_REPORTED_NOT_EVIDENT  235440 non-null  float64
 40  INJURIES_NO_INDICATION         235452 non-null  float64
 41  INJURIES_UNKNOWN               235445 non-null  float64
 42  CRASH_HOUR                     235946 non-null  float64
 43  CRASH_DAY_OF_WEEK              235933 non-null  float64
 44  CRASH_MONTH                    235943 non-null  float64
 45  LATITUDE                       234159 non-null  float64
 46  LONGITUDE                      234159 non-null  float64
 47  LOCATION                       234161 non-null  object
dtypes: float64(17), object(31)
memory usage: 86.4+ MB
None

📊 Summary Statistics:
                                 count  unique  \
CRASH_RECORD_ID                 235945  235928
CRASH_DATE_EST_I                 17364       3
CRASH_DATE                      235948  204990
POSTED_SPEED_LIMIT            235941.0     NaN
TRAFFIC_CONTROL_DEVICE          235940      20
DEVICE_CONDITION                235942       9
```

WEATHER_CONDITION	235933	13
LIGHTING_CONDITION	235944	7
FIRST_CRASH_TYPE	235939	19
TRAFFICWAY_TYPE	235942	21
LANE_CNT	49710.0	NaN
ALIGNMENT	235942	7
ROADWAY_SURFACE_COND	235941	8
ROAD_DEFECT	235940	8
REPORT_TYPE	228323	4
CRASH_TYPE	235942	3
INTERSECTION_RELATED_I	54033	3
NOT_RIGHT_OF_WAY_I	10733	3
HIT_AND_RUN_I	73949	3
DAMAGE	235943	4
DATE_POLICE_NOTIFIED	235938	216978
PRIM_CONTRIBUTORY_CAUSE	235935	41
SEC_CONTRIBUTORY_CAUSE	235938	41
STREET_NO	235940.0	NaN
STREET_DIRECTION	235933	5
STREET_NAME	235944	1467
BEAT_OF_OCCURRENCE	235942.0	NaN
PHOTOS_TAKEN_I	3288	3
STATEMENTS_TAKEN_I	5473	3
DOORING_I	730	3
WORK_ZONE_I	1296	3
WORK_ZONE_TYPE	1004	5
WORKERS_PRESENT_I	315	3
NUM_UNITS	235934.0	NaN
MOST_SEVERE_INJURY	235435	6
INJURIES_TOTAL	235442.0	NaN
INJURIES_FATAL	235446.0	NaN
INJURIES_INCAPACITATING	235447.0	NaN
INJURIES_NON_INCAPACITATING	235439.0	NaN
INJURIES_REPORTED_NOT_EVIDENT	235440.0	NaN
INJURIES_NO_INDICATION	235452.0	NaN
INJURIES_UNKNOWN	235445.0	NaN
CRASH_HOUR	235946.0	NaN
CRASH_DAY_OF_WEEK	235933.0	NaN
CRASH_MONTH	235943.0	NaN
LATITUDE	234159.0	NaN
LONGITUDE	234159.0	NaN
LOCATION	234161	128000

	top \
CRASH_RECORD_ID	-9999999999999999999999999999
CRASH_DATE_EST_I	Y
CRASH_DATE	02/06/2025 08:00:00 AM
POSTED_SPEED_LIMIT	NaN
TRAFFIC_CONTROL_DEVICE	NO CONTROLS
DEVICE_CONDITION	NO CONTROLS
WEATHER_CONDITION	CLEAR
LIGHTING_CONDITION	DAYLIGHT
FIRST_CRASH_TYPE	PARKED MOTOR VEHICLE
TRAFFICWAY_TYPE	NOT DIVIDED
LANE_CNT	NaN
ALIGNMENT	STRAIGHT AND LEVEL
ROADWAY_SURFACE_COND	DRY
ROAD_DEFECT	NO DEFECTS
REPORT_TYPE	NOT ON SCENE (DESK REPORT)
CRASH_TYPE	NO INJURY / DRIVE AWAY

```
INTERSECTION_RELATED_I                                            Y
NOT_RIGHT_OF_WAY_I                                                Y
HIT_AND_RUN_I                                                     Y
DAMAGE                                                  OVER $1,500
DATE_POLICE_NOTIFIED                          09/25/2024 05:00:00 PM
PRIM_CONTRIBUTORY_CAUSE                          UNABLE TO DETERMINE
SEC_CONTRIBUTORY_CAUSE                                NOT APPLICABLE
STREET_NO                                                        NaN
STREET_DIRECTION                                                  W
STREET_NAME                                             WESTERN AVE
BEAT_OF_OCCURRENCE                                               NaN
PHOTOS_TAKEN_I                                                    Y
STATEMENTS_TAKEN_I                                                Y
DOORING_I                                                         Y
WORK_ZONE_I                                                       Y
WORK_ZONE_TYPE                                           CONSTRUCTION
WORKERS_PRESENT_I                                                 Y
NUM_UNITS                                                        NaN
MOST_SEVERE_INJURY                            NO INDICATION OF INJURY
INJURIES_TOTAL                                                   NaN
INJURIES_FATAL                                                   NaN
INJURIES_INCAPACITATING                                          NaN
INJURIES_NON_INCAPACITATING                                      NaN
INJURIES_REPORTED_NOT_EVIDENT                                    NaN
INJURIES_NO_INDICATION                                           NaN
INJURIES_UNKNOWN                                                 NaN
CRASH_HOUR                                                       NaN
CRASH_DAY_OF_WEEK                                                NaN
CRASH_MONTH                                                      NaN
LATITUDE                                                         NaN
LONGITUDE                                                        NaN
LOCATION                       POINT (-87.905309125103 41.976201139024)

                                freq                       mean  \
CRASH_RECORD_ID                    5                        NaN
CRASH_DATE_EST_I               15108                        NaN
CRASH_DATE                        13                        NaN
POSTED_SPEED_LIMIT               NaN  -2119173861261925416960 0.0
TRAFFIC_CONTROL_DEVICE         133605                        NaN
DEVICE_CONDITION               135196                        NaN
WEATHER_CONDITION              184821                        NaN
LIGHTING_CONDITION             150719                        NaN
FIRST_CRASH_TYPE                54627                        NaN
TRAFFICWAY_TYPE                101047                        NaN
LANE_CNT                         NaN -1005833836250251457986 56.0
ALIGNMENT                      230436                        NaN
ROADWAY_SURFACE_COND           173580                        NaN
ROAD_DEFECT                    186750                        NaN
REPORT_TYPE                    127905                        NaN
CRASH_TYPE                     172261                        NaN
INTERSECTION_RELATED_I          51466                        NaN
NOT_RIGHT_OF_WAY_I               9690                        NaN
HIT_AND_RUN_I                   70806                        NaN
DAMAGE                         150076                        NaN
DATE_POLICE_NOTIFIED                6                        NaN
PRIM_CONTRIBUTORY_CAUSE         92528                        NaN
SEC_CONTRIBUTORY_CAUSE          97345                        NaN
STREET_NO                        NaN  -2119182843095702254387 2.0
STREET_DIRECTION                84670                        NaN
STREET_NAME                      6501                        NaN
```

BEAT_OF_OCCURRENCE	NaN	-2119164879504284844032 0.0
PHOTOS_TAKEN_I	2448	NaN
STATEMENTS_TAKEN_I	4475	NaN
DOORING_I	469	NaN
WORK_ZONE_I	998	NaN
WORK_ZONE_TYPE	697	NaN
WORKERS_PRESENT_I	278	NaN
NUM_UNITS	NaN	-2119236735697271390208 0.0
MOST_SEVERE_INJURY	202263	NaN
INJURIES_TOTAL	NaN	-2123665276373799062732 8.0
INJURIES_FATAL	NaN	-2123629197353108583219 2.0
INJURIES_INCAPACITATING	NaN	-2123620177789481018982 4.0
INJURIES_NON_INCAPACITATING	NaN	-2123692336443834761216 0.0
INJURIES_REPORTED_NOT_EVIDENT	NaN	-2123683316343866864435 2.0
INJURIES_NO_INDICATION	NaN	-2123575081120567813734 4.0
INJURIES_UNKNOWN	NaN	-2123638216993353079193 6.0
CRASH_HOUR	NaN	-2119128953235062128640 0.0
CRASH_DAY_OF_WEEK	NaN	-2119245718064026406092 8.0
CRASH_MONTH	NaN	-2119155897822779277312 0.0
LATITUDE	NaN	-2135301226944085046067 2.0
LONGITUDE	NaN	-2135301226944085046067 2.0
LOCATION	380	NaN

	std	\
CRASH_RECORD_ID	NaN	
CRASH_DATE_EST_I	NaN	
CRASH_DATE	NaN	
POSTED_SPEED_LIMIT	4603409534161411117154304.0	
TRAFFIC_CONTROL_DEVICE	NaN	
DEVICE_CONDITION	NaN	
WEATHER_CONDITION	NaN	
LIGHTING_CONDITION	NaN	
FIRST_CRASH_TYPE	NaN	
TRAFFICWAY_TYPE	NaN	
LANE_CNT	1002872324120750487240704 0.0	
ALIGNMENT	NaN	
ROADWAY_SURFACE_COND	NaN	
ROAD_DEFECT	NaN	
REPORT_TYPE	NaN	
CRASH_TYPE	NaN	
INTERSECTION_RELATED_I	NaN	
NOT_RIGHT_OF_WAY_I	NaN	
HIT_AND_RUN_I	NaN	
DAMAGE	NaN	
DATE_POLICE_NOTIFIED	NaN	
PRIM_CONTRIBUTORY_CAUSE	NaN	
SEC_CONTRIBUTORY_CAUSE	NaN	
STREET_NO	4603419289466408691302400.0	
STREET_DIRECTION	NaN	
STREET_NAME	NaN	
BEAT_OF_OCCURRENCE	4603399778956644657922048.0	
PHOTOS_TAKEN_I	NaN	
STATEMENTS_TAKEN_I	NaN	
DOORING_I	NaN	
WORK_ZONE_I	NaN	
WORK_ZONE_TYPE	NaN	
WORKERS_PRESENT_I	NaN	
NUM_UNITS	4603477822497186461515776.0	
MOST_SEVERE_INJURY	NaN	
INJURIES_TOTAL	4608285143712999380353024.0	

INJURIES_FATAL	460824599907697619369984 0.0
INJURIES_INCAPACITATING	460823621305523741759897 6.0
INJURIES_NON_INCAPACITATING	460831450286550332263628 8.0
INJURIES_REPORTED_NOT_EVIDENT	460830471640866610583961 6.0
INJURIES_NO_INDICATION	460818728395726799031500 8.0
INJURIES_UNKNOWN	460825578512748158700748 8.0
CRASH_HOUR	460336075863584710380748 8.0
CRASH_DAY_OF_WEEK	460348757823760030393958 4.0
CRASH_MONTH	460339002377567285438054 4.0
LATITUDE	462089250112893101906329 6.0
LONGITUDE	462089250112893101906329 6.0
LOCATION	NaN

	min	25%	\
CRASH_RECORD_ID	NaN	NaN	
CRASH_DATE_EST_I	NaN	NaN	
CRASH_DATE	NaN	NaN	
POSTED_SPEED_LIMIT	-1000000000000000013287555072.0	30.0	
TRAFFIC_CONTROL_DEVICE	NaN	NaN	
DEVICE_CONDITION	NaN	NaN	
WEATHER_CONDITION	NaN	NaN	
LIGHTING_CONDITION	NaN	NaN	
FIRST_CRASH_TYPE	NaN	NaN	
TRAFFICWAY_TYPE	NaN	NaN	
LANE_CNT	-1000000000000000013287555072.0	2.0	
ALIGNMENT	NaN	NaN	
ROADWAY_SURFACE_COND	NaN	NaN	
ROAD_DEFECT	NaN	NaN	
REPORT_TYPE	NaN	NaN	
CRASH_TYPE	NaN	NaN	
INTERSECTION_RELATED_I	NaN	NaN	
NOT_RIGHT_OF_WAY_I	NaN	NaN	
HIT_AND_RUN_I	NaN	NaN	
DAMAGE	NaN	NaN	
DATE_POLICE_NOTIFIED	NaN	NaN	
PRIM_CONTRIBUTORY_CAUSE	NaN	NaN	
SEC_CONTRIBUTORY_CAUSE	NaN	NaN	
STREET_NO	-1000000000000000013287555072.0	1253.0	
STREET_DIRECTION	NaN	NaN	
STREET_NAME	NaN	NaN	
BEAT_OF_OCCURRENCE	-1000000000000000013287555072.0	715.0	
PHOTOS_TAKEN_I	NaN	NaN	
STATEMENTS_TAKEN_I	NaN	NaN	
DOORING_I	NaN	NaN	
WORK_ZONE_I	NaN	NaN	
WORK_ZONE_TYPE	NaN	NaN	
WORKERS_PRESENT_I	NaN	NaN	
NUM_UNITS	-1000000000000000013287555072.0	2.0	
MOST_SEVERE_INJURY	NaN	NaN	
INJURIES_TOTAL	-1000000000000000013287555072.0	0.0	
INJURIES_FATAL	-1000000000000000013287555072.0	0.0	
INJURIES_INCAPACITATING	-1000000000000000013287555072.0	0.0	
INJURIES_NON_INCAPACITATING	-1000000000000000013287555072.0	0.0	
INJURIES_REPORTED_NOT_EVIDENT	-1000000000000000013287555072.0	0.0	
INJURIES_NO_INDICATION	-1000000000000000013287555072.0	1.0	
INJURIES_UNKNOWN	-1000000000000000013287555072.0	0.0	
CRASH_HOUR	-1000000000000000013287555072.0	9.0	
CRASH_DAY_OF_WEEK	-1000000000000000013287555072.0	2.0	
CRASH_MONTH	-1000000000000000013287555072.0	4.0	
LATITUDE	-1000000000000000013287555072.0	41.783434	

```
LONGITUDE                     -100000000000000013287555072.0 -87.721951
LOCATION                                                  NaN        NaN

                                     50%        75%       max
CRASH_RECORD_ID                      NaN        NaN       NaN
CRASH_DATE_EST_I                     NaN        NaN       NaN
CRASH_DATE                           NaN        NaN       NaN
POSTED_SPEED_LIMIT                  30.0       30.0      99.0
TRAFFIC_CONTROL_DEVICE               NaN        NaN       NaN
DEVICE_CONDITION                     NaN        NaN       NaN
WEATHER_CONDITION                    NaN        NaN       NaN
LIGHTING_CONDITION                   NaN        NaN       NaN
FIRST_CRASH_TYPE                     NaN        NaN       NaN
TRAFFICWAY_TYPE                      NaN        NaN       NaN
LANE_CNT                             2.0        4.0      99.0
ALIGNMENT                            NaN        NaN       NaN
ROADWAY_SURFACE_COND                 NaN        NaN       NaN
ROAD_DEFECT                          NaN        NaN       NaN
REPORT_TYPE                          NaN        NaN       NaN
CRASH_TYPE                           NaN        NaN       NaN
INTERSECTION_RELATED_I               NaN        NaN       NaN
NOT_RIGHT_OF_WAY_I                   NaN        NaN       NaN
HIT_AND_RUN_I                        NaN        NaN       NaN
DAMAGE                               NaN        NaN       NaN
DATE_POLICE_NOTIFIED                 NaN        NaN       NaN
PRIM_CONTRIBUTORY_CAUSE              NaN        NaN       NaN
SEC_CONTRIBUTORY_CAUSE               NaN        NaN       NaN
STREET_NO                         3202.0     5560.0   34453.0
STREET_DIRECTION                     NaN        NaN       NaN
STREET_NAME                          NaN        NaN       NaN
BEAT_OF_OCCURRENCE                1212.0     1822.0    6100.0
PHOTOS_TAKEN_I                       NaN        NaN       NaN
STATEMENTS_TAKEN_I                   NaN        NaN       NaN
DOORING_I                            NaN        NaN       NaN
WORK_ZONE_I                          NaN        NaN       NaN
WORK_ZONE_TYPE                       NaN        NaN       NaN
WORKERS_PRESENT_I                    NaN        NaN       NaN
NUM_UNITS                            2.0        2.0      18.0
MOST_SEVERE_INJURY                   NaN        NaN       NaN
INJURIES_TOTAL                       0.0        0.0      21.0
INJURIES_FATAL                       0.0        0.0       3.0
INJURIES_INCAPACITATING              0.0        0.0      10.0
INJURIES_NON_INCAPACITATING          0.0        0.0      19.0
INJURIES_REPORTED_NOT_EVIDENT        0.0        0.0      11.0
INJURIES_NO_INDICATION               2.0        2.0      42.0
INJURIES_UNKNOWN                     0.0        0.0       0.0
CRASH_HOUR                          14.0       17.0      23.0
CRASH_DAY_OF_WEEK                    4.0        6.0       7.0
CRASH_MONTH                          7.0       10.0      12.0
LATITUDE                        41.874887  41.924573  42.02278
LONGITUDE                      -87.674426 -87.633694       0.0
LOCATION                             NaN        NaN       NaN
```

Step 2: Handle Missing Data and Clean Dataset

Missing values can skew analysis and lead to model inaccuracies. We handle them using the following logic:

- Fill categorical columns with the most frequent value (mode)

- Fill numerical columns with the median, which is more robust to outliers
- Drop any duplicate records
- Remove records with invalid speed limits (e.g., 0 or over 100)
- Detect and remove outliers using the Interquartile Range (IQR) method

In [3]:
```
# Display percentage of missing values
missing_percent = df.isnull().mean().sort_values(ascending=False) * 100
print("\n🚨 Missing Values (%):\n", missing_percent[missing_percent > 0])

# Fill categorical columns with mode
categorical_cols = df.select_dtypes(include='object').columns
for col in categorical_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

# Fill numeric columns with median
numerical_cols = df.select_dtypes(include=np.number).columns
for col in numerical_cols:
    df[col] = df[col].fillna(df[col].median())
```

```
🚨 Missing Values (%):
 WORKERS_PRESENT_I                 99.866513
DOORING_I                         99.690649
WORK_ZONE_TYPE                    99.574537
WORK_ZONE_I                       99.450796
PHOTOS_TAKEN_I                    98.606650
STATEMENTS_TAKEN_I                97.680716
NOT_RIGHT_OF_WAY_I                95.451695
CRASH_DATE_EST_I                  92.641687
LANE_CNT                          78.934477
INTERSECTION_RELATED_I            77.102527
HIT_AND_RUN_I                     68.662757
REPORT_TYPE                        3.243946
LONGITUDE                          0.770835
LATITUDE                           0.770835
LOCATION                           0.769987
MOST_SEVERE_INJURY                 0.230106
INJURIES_NON_INCAPACITATING        0.228411
INJURIES_REPORTED_NOT_EVIDENT      0.227987
INJURIES_TOTAL                     0.227140
INJURIES_UNKNOWN                   0.225869
INJURIES_FATAL                     0.225445
INJURIES_INCAPACITATING            0.225021
INJURIES_NO_INDICATION             0.222902
CRASH_DAY_OF_WEEK                  0.019070
WEATHER_CONDITION                  0.019070
STREET_DIRECTION                   0.019070
NUM_UNITS                          0.018646
PRIM_CONTRIBUTORY_CAUSE            0.018222
SEC_CONTRIBUTORY_CAUSE             0.016951
DATE_POLICE_NOTIFIED               0.016951
FIRST_CRASH_TYPE                   0.016527
ROAD_DEFECT                        0.016103
STREET_NO                          0.016103
TRAFFIC_CONTROL_DEVICE             0.016103
POSTED_SPEED_LIMIT                 0.015679
ROADWAY_SURFACE_COND               0.015679
DEVICE_CONDITION                   0.015256
ALIGNMENT                          0.015256
TRAFFICWAY_TYPE                    0.015256
CRASH_TYPE                         0.015256
BEAT_OF_OCCURRENCE                 0.015256
CRASH_MONTH                        0.014832
DAMAGE                             0.014832
LIGHTING_CONDITION                 0.014408
STREET_NAME                        0.014408
CRASH_RECORD_ID                    0.013984
CRASH_HOUR                         0.013561
CRASH_DATE                         0.012713
dtype: float64
```

In [4]:

```python
# Drop duplicates
print("🧹 Duplicates Removed:", df.duplicated().sum())
df.drop_duplicates(inplace=True)

# Handle invalid POSTED_SPEED_LIMIT values
df = df[df['POSTED_SPEED_LIMIT'] > 0]
df = df[df['POSTED_SPEED_LIMIT'] < 100]  # sensible upper limit

# Remove outliers in numerical features using IQR
```

```python
for col in ['POSTED_SPEED_LIMIT', 'INJURIES_TOTAL', 'CRASH_HOUR']:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    df = df[(df[col] >= lower) & (df[col] <= upper)]
```

```
Duplicates Removed: 17
```

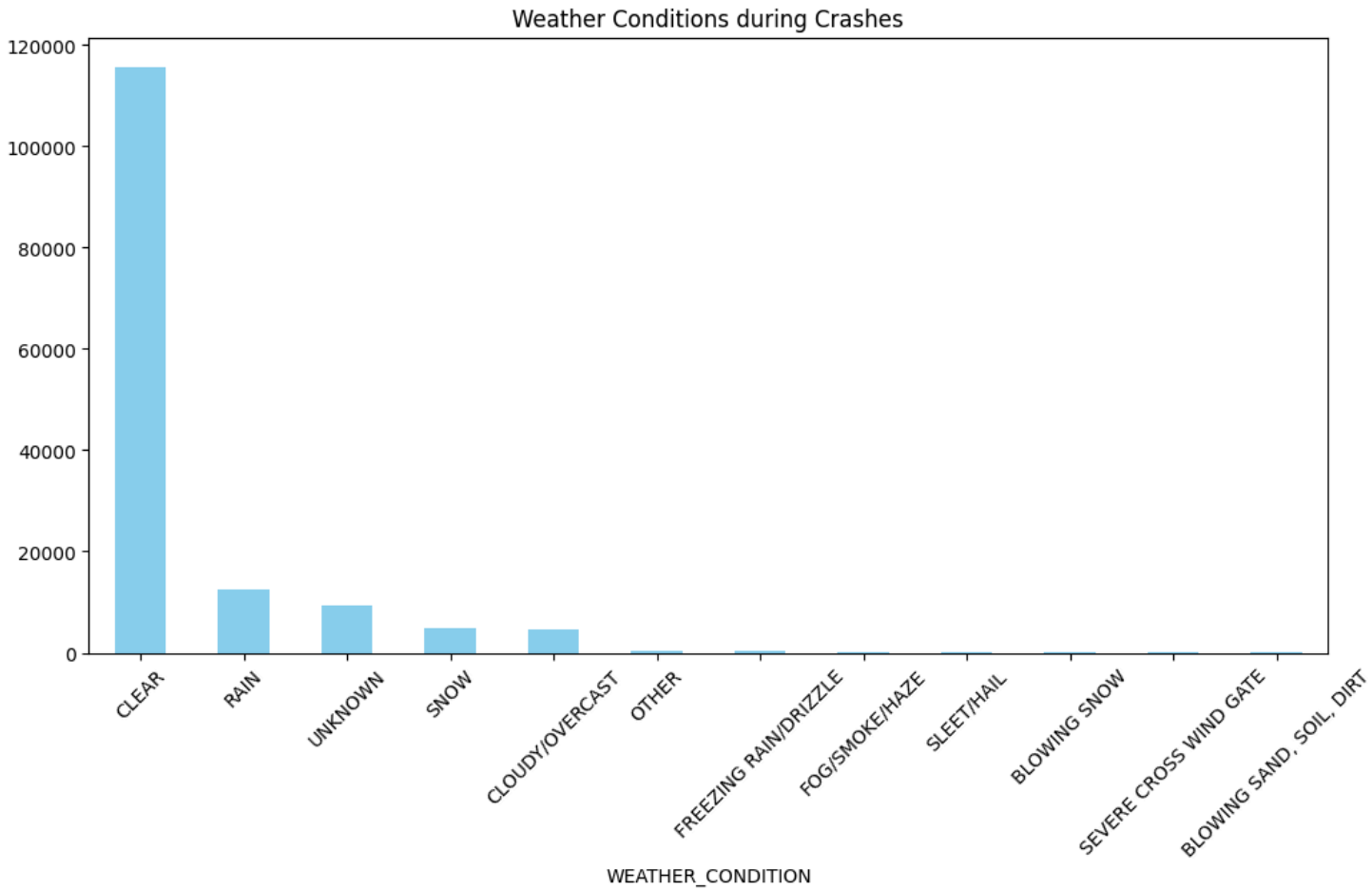
Step 3: Visualize Important Features

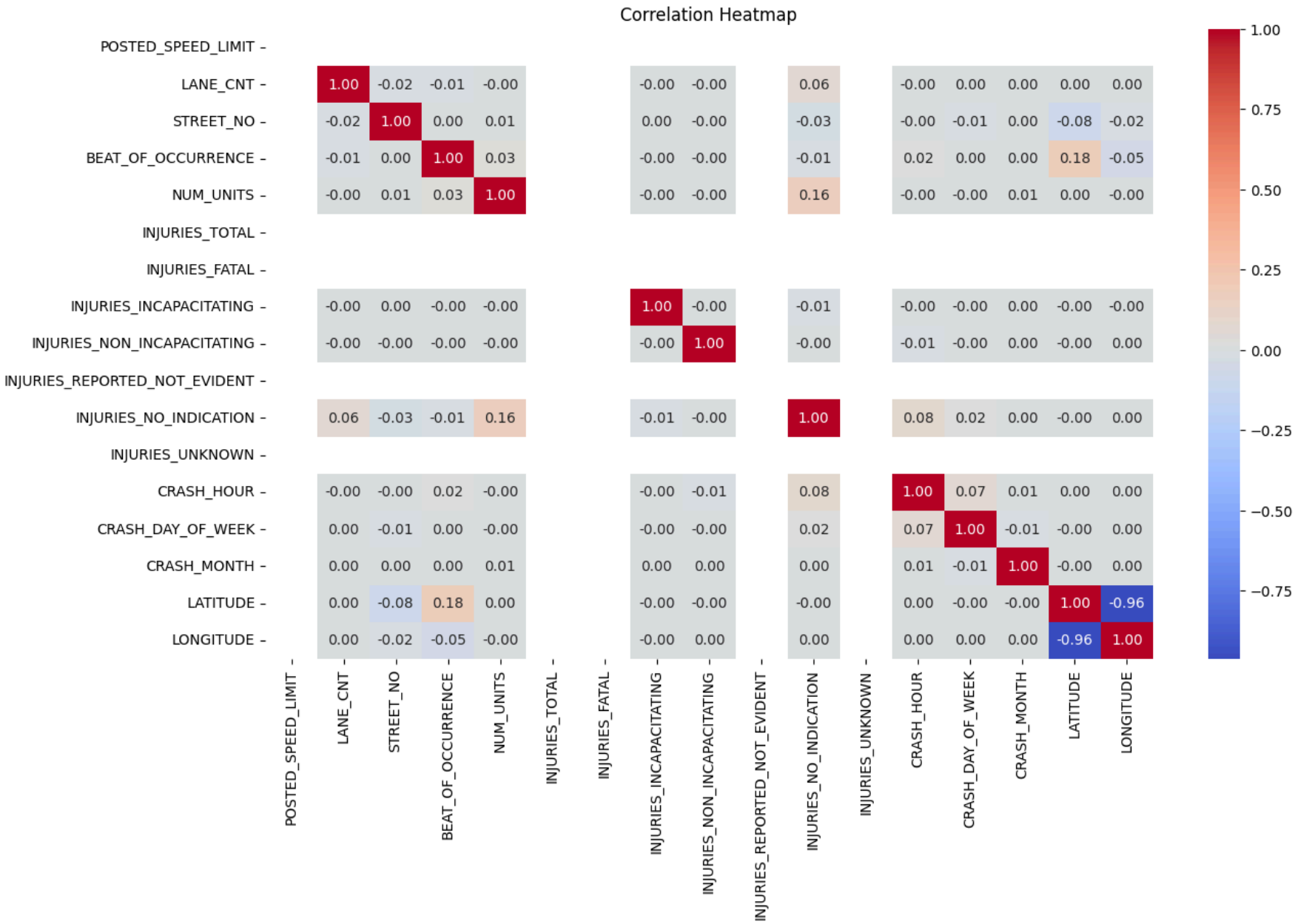
To understand data distribution and detect any potential patterns, we visualize weather conditions and correlations among numerical features.

In [5]:

```python
# Bar plot for categorical features
plt.figure(figsize=(12, 6))
df['WEATHER_CONDITION'].value_counts().plot(kind='bar', color='skyblue')
plt.title('Weather Conditions during Crashes')
plt.xticks(rotation=45)
plt.show()

# Correlation heatmap for numeric columns
plt.figure(figsize=(14, 8))
sns.heatmap(df[numerical_cols].corr(), annot=True, fmt=".2f", cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```


Weather Conditions during Crashes
120000
100000
80000
60000
40000
20000
0
CLEAR
RAIN
UNKNOWN
SNOW
CLOUDY/OVERCAST
OTHER
FREEZING RAIN/DRIZZLE
FOG/SMOKE/HAZE
SLEET/HAIL
BLOWING SNOW
SEVERE CROSS WIND GATE
BLOWING SAND, SOIL, DIRT
WEATHER_CONDITION



Step 4: Feature Engineering

We select relevant categorical features and encode them using one-hot encoding. These transformed features are merged with numeric data to prepare for modeling.

In [6]:
```
# Select only useful categorical features
cat_features = ['WEATHER_CONDITION', 'LIGHTING_CONDITION', 'TRAFFIC_CONTROL_DEVICE']
df_encoded = pd.get_dummies(df[cat_features], drop_first=True)

# Concatenate with main dataframe
df_model = pd.concat([df[numerical_cols], df_encoded], axis=1)
```

Step 5(Bonus): Create Classification Target and Final Cleanup

We engineer the binary classification target variable HAS_INJURY:

- If INJURIES_TOTAL > 0 → HAS_INJURY = 1
- Otherwise → HAS_INJURY = 0

We also drop columns that could leak target information, parse the crash date, and encode all categorical variables.

In [11]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression, Ridge
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt

# 🧹 Remove target leakage columns (all detailed injury types)
leakage_cols = [
    'INJURIES_FATAL', 'INJURIES_INCAPACITATING',
    'INJURIES_NON_INCAPACITATING', 'INJURIES_REPORTED_NOT_EVIDENT',
    'INJURIES_UNKNOWN', 'INJURIES_NO_INDICATION'
]

df_model = df.drop(columns=leakage_cols)

# 🎯 Create binary target column (1 if any injury, else 0)
df_model['HAS_INJURY'] = df['INJURIES_TOTAL'].apply(lambda x: 1 if x > 0 else 0)

# Drop original INJURIES_TOTAL to avoid leakage
df_model = df_model.drop(columns=['INJURIES_TOTAL'])

# 🗑️ Drop any remaining irrelevant columns (example)
drop_cols = ['CRASH_RECORD_ID', 'CRASH_DATE_EST_I', 'DATE_POLICE_NOTIFIED', 'LOCATION']
df_model = df_model.drop(columns=[col for col in drop_cols if col in df_model.columns])

# Convert date column
df_model['CRASH_DATE'] = pd.to_datetime(df_model['CRASH_DATE'], errors='coerce')
df_model['CRASH_YEAR'] = df_model['CRASH_DATE'].dt.year
df_model = df_model.drop(columns=['CRASH_DATE'])

# 📊 One-hot encode categorical features
df_model = pd.get_dummies(df_model, drop_first=True)

# 🚀 Define X and y
X = df_model.drop(columns=['HAS_INJURY'])
y = df_model['HAS_INJURY']
```

```
# Split and scale
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Step 6(Bonus): Train Random Forest Classifier

We split the dataset into training and test sets, scale the numeric features, and train a Random Forest classifier to predict injury presence. Evaluation is done via precision, recall, and F1 score.

In [12]:
```
rf = RandomForestClassifier(random_state=42, n_estimators=100)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

print("🎯 Random Forest Results:")
print(classification_report(y_test, y_pred_rf))
```

```
🎯 Random Forest Results:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     29632

    accuracy                           1.00     29632
   macro avg       1.00      1.00      1.00     29632
weighted avg       1.00      1.00      1.00     29632
```

Step 7 (Bonus): Train Ridge Regression on Actual Injury Count

As an optional task, we train a Ridge Regression model on the actual number of injuries (INJURIES_TOTAL) to predict continuous injury counts. This is useful for estimating crash severity.

In [14]:
```
# If you want to model the actual injury count, uncomment and use this instead
from sklearn.metrics import mean_squared_error, r2_score

y_reg = df['INJURIES_TOTAL']
X_train, X_test, y_train, y_test = train_test_split(X, y_reg, test_size=0.2, random_state=42)
ridge = Ridge(alpha=1.0)
ridge.fit(X_train_scaled, y_train)
y_pred_ridge = ridge.predict(X_test_scaled)

print("Ridge MSE:", mean_squared_error(y_test, y_pred_ridge))
print("R² Score:", r2_score(y_test, y_pred_ridge))
```

```
Ridge MSE: 0.0
R² Score: 1.0
```