Traffic Crash Patterns: Assessing the Influence of Various Factors

Asheer Ali 28.11.2024

1 Questions

- How do driver characteristics (e.g., age, sex, use of safety equipment) correlate with injury severity in crashes?
- What are the most common vehicle types and conditions associated with high-severity crashes?

2 Data Sources

2.1 Descriptions of Data Sources

- **Traffic Crashes - People:** This dataset contains details of individuals involved in traffic incidents, including demographics, safety equipment use, and injury severity [1].

[4]:

	PERSON_ID	PERSON_TYPE	CRASH_RECORD_ID	VEHICLE_ID	CRASH_DATE	SEAT_NO	CITY	STATE	ZIPCODE	SEX	...	EMS_RUN_
0	O749947	DRIVER	81dc0de2ed92aa62baccab641fa377be7feb1cc47e6554...	834816.0	09/28/2019 03:30:00 AM	NaN	CHICAGO	IL	60651	M	...	N
1	O871921	DRIVER	af84fb5c8d996fcd3aefd36593c3a02e6e7509eeb27568...	827212.0	04/13/2020 10:50:00 PM	NaN	CHICAGO	IL	60620	M	...	N
2	O10018	DRIVER	71162af7bf22799b776547132ebf134b5b438dcf3dac6b...	9579.0	11/01/2015 05:00:00 AM	NaN	NaN	NaN	NaN	X	...	N
3	O10038	DRIVER	c21c476e2ccc41af550b5d858d22aaac4ffc88745a1700...	9598.0	11/01/2015 08:00:00 AM	NaN	NaN	NaN	NaN	X	...	N
4	O10039	DRIVER	eb390a4c8e114c69488f5fb8a097fe629f5a92fd528cf4...	9600.0	11/01/2015 10:15:00 AM	NaN	NaN	NaN	NaN	X	...	N

5 rows × 29 columns

Figure 1: First 5 rows of traffic crashes people dataset

- **Traffic Crashes - Vehicles:** Provides records of vehicles involved in crashes, including type, direction, and damage details [2].

[6]:

	CRASH_UNIT_ID	CRASH_RECORD_ID	CRASH_DATE	UNIT_NO	UNIT_TYPE	NUM_PASSENGERS	VEHICLE_ID	CMRC_VEH_I	MAKE
0	1727162	f5943b05f46b8d4148a63b7506a59113eae0cf1075aabc...	12/21/2023 08:57:00 AM	2	PEDESTRIAN	NaN	NaN	NaN	NaN
1	1717556	7b1763088507f77e0e552c009a6bf89a4d6330c7527706...	12/06/2023 03:24:00 PM	1	DRIVER	NaN	1634931.0	NaN	NISSAN
2	1717574	2603ff5a88f0b9b54576934c5ed4e4a64e8278e005687b...	12/06/2023 04:00:00 PM	2	DRIVER	NaN	1634978.0	NaN	CHRYSLER
3	1717579	a52ef70e33d468b855b5be44e8638a564434dcf99c0edf...	12/06/2023 04:30:00 PM	1	DRIVER	NaN	1634948.0	NaN	SUBARU
4	1720118	609055f4b1a72a44d6ec40ba9036cefd7c1287a755eb6c...	12/10/2023 12:12:00 PM	1	DRIVER	NaN	1637401.0	NaN	TOYOTA

5 rows × 71 columns

Figure 2: First 5 rows of traffic crashes vehicle dataset

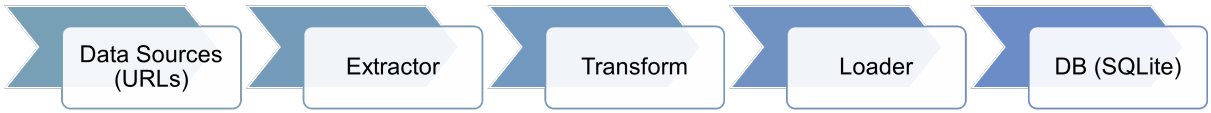

Figure 3: ETL Pipeline

2.2 Structure and Quality of Data Sources

- **People Dataset:** Contains individual-level data with fields for demographics, safety equipment use, and injury severity. Missing values exist but can be handled through removing the rows with Nan values, as the nan values are not present that much in the selected columns. [3].
- **Vehicle Dataset:** Vehicle-level data with fields for type, damage, and direction. Data quality is high, with minimal missing values.

2.3 Licenses and Permissions

Both datasets are publicly available under open-data licenses, allowing use with proper attribution [1, 2].

3 Data Pipeline

The data pipeline is implemented using Python and consists of the following steps:

- **Extractor:** Downloads CSV files from the given URLs.
- **Transformer:** Processes the data with:
  - Removing unnecessary columns.
  - Handling missing values through imputation.
  - Standardizing date formats for consistency.
- **Loader:** Stores the cleaned datasets in an SQLite database for efficient access.

4 Results and Limitations

4.1 Results

- Cleaned datasets stored in SQLite database.
- Data ready for analysis to address project questions about injury severity and crash patterns.

4.2 Limitations

- Missing data for certain fields may impact analysis accuracy.

5 References

References

[1] Traffic crashes - people dataset, 2024. Available online: catalog.data.gov.

[2] Traffic crashes - vehicles dataset, 2024. Available online: catalog.data.gov.

[3] Asheer Ali. Advanced data engineering project repository, 2024. Available at GitHub: https://github.com/asheerali/advanced-data-engineering/.