**Assignment 2: Data Preparation**

**Safe LA: Enhancing Traffic Safety Through Data-Driven Insights**

**Team 8**

Alexandra Gladkova

Christina Saju

Hao Lun Rong

Kushwanth Sai Kolli

Ladan Asempour

**Capstone Project**

**BIA-5450-0LA**

**Samer Al-Obaidi**

**Group Member Contribution**

|  |  |
| --- | --- |
| **Topic** | **Contributor** |
| Data sources | Alexandra Gladkova |
| Data dictionary | Christina Saju - Hao Lun Rong |
| Data cleaning | Christina Saju - Hao Lun Rong |
| Data storage | Ladan Asempour |
| Data output | Kushwanth Sai Kolli |
| Challenges | Kushwanth Sai Kolli - Hao Lun Rong |

**Data sources**

The primary data source for this project is the **Traffic Collision Data from 2010 to Present**, available on Data.gov *(*[*https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present*](https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present) *).* This dataset contains detailed records of traffic collisions in Los Angeles, including location, time, involved parties, and contributing factors.

To understand the structure and meaning of the dataset’s fields, we referred to the official data dictionary and column descriptions provided by the City of Los Angeles *(*[*https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w/about\_data*](https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w/about_data)*)*.

Additionally, for incident classification and interpretation of MO (Modus Operandi) codes, we used the official MO Codes document from the city’s open data portal, accessible in MO\_CODES\_Numerical\_20180627.pdf on the <https://data.lacity.org/> website.

These sources together provide a comprehensive foundation for analyzing traffic collisions, identifying trends, and deriving insights to enhance road safety in Los Angeles.

**Data dictionary – a glossary of terms to define the data elements used for reference**

Name of the final dataset – collision\_data

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Date\_Occured | Date Time | Year, Month, and Day of collision occurrence. |
| Time\_Occured | Object | Time the collision occurred in military format (24 hours). |
| Area\_Name | Object | The district or neighbourhood the collision occurred. |
| MO\_Codes | Object | Modus Operandi Code – Numerical codes representing the details of the incident. |
| MO\_Code01 | Object | Description for each individual MO Code |
| MO\_Code02 | Object | Description for each individual MO Code |
| MO\_Code03 | Object | Description for each individual MO Code |
| MO\_Code04 | Object | Description for each individual MO Code |
| MO\_Code05 | Object | Description for each individual MO Code |
| MO\_Code06 | Object | Description for each individual MO Code |
| MO\_Code07 | Object | Description for each individual MO Code |
| MO\_Code08 | Object | Description for each individual MO Code |
| MO\_Code09 | Object | Description for each individual MO Code |
| MO\_Code10 | Object | Description for each individual MO Code |
| Victim\_Age | Integer | Age of the victim involved in the collision. |
| Victim\_Sex | Object | Gender of the victim involved in the collision. |
| Victim\_Descent | Object | The ethnicity of the victim involved in the collision. |
| Premise\_Description | Object | Type of location where the collision took place. |
| Address | Object | The primary street where the collision took place. |
| Cross\_Street | Object | Nearest intersection to the location of the collision. |
| Intersection | Object | Combination of the given address and cross street, in a consistent format. |
| Age\_Group | Category | Victim Age grouped into categories including missing values. |

**Data cleaning**

Data cleaning was performed on the dataset to ensure accuracy and relevance in analysis. First, we filtered the dataset to only include accidents from the years 2023 and 2024.

Several columns were dropped, including DR Number, Date Reported, Area ID, Reporting District, Crime Code, and Premise Code, as they did not provide direct value for the analysis and hence were deemed unnecessary or redundant.

The MO\_Codes column, originally stored as a single string with multiple values seperated by space, was split into separate columns (MO\_Code01 to MO\_Code10) to allow for better individual analysis of contributing factors.

The MO Codes were then cross referenced with another CSV file (MO\_Code\_Ref), containing the description of all the MO Codes used by LAPD. Using SQL (using a specialized package to run SQL queries in Python), we matched the MO Codes from the collision dataset to its corresponding definitions in the MO\_Code\_Ref CSV.

Once the definitions of the MO Codes were cross referenced, we remapped all the codes into the respective descriptions. Doing so provided us with more meaning to the common MO codes present in traffic collision reports.

For the Location Data, an intersection column was created by combining the main address of the collision with the nearest intersection. We joined both columns, separated by a comma to create the intersection column with entries containing the first street and second street names.

To ensure consistency of data and not confuse entries due to the order of the streets (example: 1st and 3rd street is the same as 3rd and 1st street), the street combinations were sorted in alphanumeric order before joining for each entry in the dataset.

The Date\_Occurred column was converted into a standard DateTime format (%m/%d/%Y) to maintain consistency and facilitate easy time-based queries.

Similarly, the Time\_Occurred column was reformatted into the HH: MM 24-hour format to ensure consistency in time-related analysis.

The demographic fields required careful handling of missing values. The Victim\_Sex and Victim\_Descent columns were cleaned, and missing values were filled with X (unknown) to ensure uniformity.

The Victim\_Age column was binned into predefined age categories to standardize age-related analysis and improve interpretability.

**Data storage**

Once we retrieved the raw data from the data source mentioned above, it was stored on our local machine to be accessed when needed.

The semi-cleaned dataset was stored in a SQLite database named temp\_database.db for structured data management and query execution when cleaning the data. Two main tables were created: UMO, which holds unique MO codes, and RMO, which contains reference MO codes along with descriptions. SQL joins were performed to merge the UMO table with RMO, allowing retrieval of MO descriptions to provide better interpretability of the dataset.

After processing, the cleaned dataset was converted to a pandas dataset for further cleaning. This structured approach ensures that the data remains accessible and well-organized for both database querying and external analysis tools.

**Data output**

The final dataset is now structured and cleaned, making it ready for in-depth analysis. The MO\_Codes have been mapped to their descriptions, allowing for better understanding of contributing factors in each accident. The data is formatted correctly, ensuring smooth integration with visualization tools for reporting and analysis.

Key insights can be drawn from this structured data, including trends related to victim demographics, high-risk geographical zones, and prevalent causes of collisions. The dataset is optimized for exploratory data analysis (EDA), predictive modeling, and decision-making processes that aim to enhance traffic safety.

**Challenges**

During the data preparation processes we encountered several challenges:

* **Handling NULL Values in Pandas vs. SQL**

One of the major challenges in data cleaning was the handling of missing values. Pandas treats missing values as NaN, whereas SQL uses NULL, leading to inconsistencies when performing SQL joins. To address this, missing values were explicitly handled in Pandas before exporting the data to SQL, ensuring uniformity across both environments.

* **Complexity of MO Codes**

Due to the LAPD having over 2000 MO Codes referenced, learning how to extract this info was a difficult task. While our dataset only contains roughly 150 of the MO codes possibly listed, the LAPD did not provide the MO Code descriptions in an accessible database format. We had to convert this into a csv file on our own to then have the ability to cross reference and sort our codes out using SQL.

* **Intersection Data Accuracy**

Some records lacked a Cross\_Street value, making it difficult to form complete intersections. In such cases, the dataset defaulted to using the primary Address column to maintain data integrity and avoid loss of information. However, this limitation means that not all intersection data may be fully accurate, requiring additional location-based verification.