**DATA ABSTRACT**

This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper; therefore, there may be some inaccuracies. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred blocks to maintain privacy. This data is as accurate as the data in the database. Please note questions or concerns in the comments.

**ABOUT THE DATA**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Type** |
| **DR\_NO** | Division of Records Number: Official file number made up of a 2-digit year, area ID, and 5-digit | Integer |
| **Date Rptd** | MM/DD/YYYY | DateTime |
| **DATE OCC** | MM/DD/YYYY | DateTime |
| **TIME OCC** | In 24-hour military time. | Integer |
| **AREA** | The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21. | String |
| **AREA NAME** | The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. | String |
| **Rpt Dist No** | A four-digit code that represents a sub-area within a Geographic Area. | Integer |
| **Part 1-2** |  | Integer |
| **Crm Cd** | Indicates the crime committed. (Same as Crime Code 1) | Integer |
| **Crm Cd Desc** | Defines the Crime Code provided. | Integer |
| **Mocodes** | Modus Operandi: Activities associated with the suspect in the commission of the crime. | Integer |
| **Vict Age** | Two characters numeric | Integer |
| **Vict Sex** | F - Female M - Male X - Unknown | String |
| **Vict Descent** | Race of the victim | String |
| **Premis Cd** | The type of structure, vehicle, or location where the crime took place. | Integer |
| **Premis Desc** | Defines the Premise Code provided. | String |
| **Weapon Used Cd** | The type of weapon used in the crime. | String |
| **Weapon Desc** | Defines the Weapon Used Code provided. | String |
| **Status** | Status of the case. (IC is the default) | String |
| **Status Desc** | Defines the Status Code provided. | String |
| **Crm Cd 1** | Indicates the crime committed. Crime Code 1 is the primary and most serious one. | Integer |
| **Crm Cd 2** | May contain a code for an additional crime, less serious than Crime Code 1. | Integer |
| **Crm Cd 3** | May contain a code for an additional crime, less serious than Crime Code 1. | Integer |
| **Crm Cd 4** | May contain a code for an additional crime, less serious than Crime Code 1. | Integer |
| **LOCATION** | Street address of crime incident rounded to the nearest hundred block to maintain anonymity. | String |
| **Cross Street** | Cross Street of rounded Address | String |
| **LAT** | Latitude | Integer |
| **LON** | Longitude | Integer |

**DATA PROFILING IN ALTRYX**

**A screenshot of a computer

Description automatically generated**

**TOOLS USED IN ALTRYX**

Using the below tools, I analyzed the Los Angeles Crime dataset in Alteryx, and gained valuable insights into the data's structure, quality, and content:

**Input Data Tool**: I use this to import the Los Angeles Crime dataset into my Alteryx workflow. Configure the tool to read the dataset from the desired source, such as a CSV file.

**Browse Tool:** Connect a Browse Tool to the output of the Input Data Tool which allows you to inspect the dataset interactively in the Alteryx Results window. You can review the data structure and quality, check for any anomalies, and get an initial understanding of its content.

**Unique Tool**: To identify unique values in specific columns, add Unique Tools to your workflow. For instance, you can use a Unique Tool to see all unique crime types in the "Crm Cd Desc" column.

**Basic Data Profile Tool**: Add a Basic Data Profile Tool connecting it to the output of the Input Data Tool. Configure the tool to profile the dataset and generate summary statistics, data type information, and visualizations.

**The Filter Tool**: This allows you to control which records pass through your pipeline based on specific criteria.

**ALTERYX DATA PROFILING AND EXPLORATION**

**MISSING VALUES:**

|  |  |
| --- | --- |
| Mocodes | 56391 |
| Vict Sex | 53797 |
| Vict Descent | 53802 |
| Premis Cd | 5 |
| Premis Desc | 160 |
| Weapon Desc | 262017 |
| Crm Cd 1 | 4 |
| Crm Cd 2 | 376241 |
| Crm Cd 3 | 407604 |
| Crm Cd 4 | 408683 |
| Cross Street | 337616 |

**DATA TYPES USED FOR STAGING PIPELINES IN TALEND**

Alteryx provides tools for data profiling like the Summarize Tool and Browse Tool to explore and analyze the data types present in your dataset. Using these datatypes, we can set the datatypes in Talend to avoid errors from source to target load.

|  |  |  |
| --- | --- | --- |
| Name | Type | Size |
| DR\_NO | String | 9 |
| Date Rptd | String | 22 |
| DATE OCC | String | 22 |
| TIME OCC | String | 4 |
| AREA | String | 2 |
| AREA NAME | String | 11 |
| Rpt Dist No | String | 4 |
| Part 1-2 | Byte | 1 |
| Crm Cd | Int16 | 2 |
| Crm Cd Desc | V\_String | 56 |
| Mocodes | V\_String | 49 |
| Vict Age | Int16 | 2 |
| Vict Sex | String | 1 |
| Vict Descent | String | 1 |
| Premis Cd | Int16 | 2 |
| Premis Desc | V\_String | 63 |
| Weapon Used Cd | Int16 | 2 |
| Weapon Desc | V\_String | 46 |
| Status | String | 2 |
| Status Desc | String | 12 |
| Crm Cd 1 | Int16 | 2 |
| Crm Cd 2 | Int16 | 2 |
| Crm Cd 3 | Int16 | 2 |
| Crm Cd 4 | Int16 | 2 |
| LOCATION | String | 40 |
| Cross Street | V\_String | 34 |
| LAT | Double | 8 |
| LON | Double | 8 |

**INSIGHTS ON THE LA CRIME DATA**

**1. Unique DR Numbers:** All DR numbers (Division of Records) within the dataset are unique, indicating that each crime incident is assigned a distinct identifier. This uniqueness ensures that no duplicate records exist for the same incident.

**2. Victim Age Analysis:** The "vict\_age" column contains some notable insights:

* **Outliers:** The dataset includes several outliers in the "vict\_age" column, indicating that there are extreme age values that may require further investigation or data validation.
* **Zero Values:** Approximately 98,709 rows have zero values in the "vict\_age" column, which could represent missing or unknown age information. Handling these zeros is essential for accurate analysis.
* **Negative Values:** The presence of negative values in the "vict\_age" column does not align with practical expectations, suggesting potential data entry errors.

**3. Victim Gender Distribution:** Analysis of the "vict\_sex" column reveals the following:

* **Male Dominance:** Approximately 42% of victims are recorded as male, making them the largest group among recorded victim genders.
* **Female Victims:** Around 36% of victims are identified as female.
* **Null Values:** There are 13% of null values in the "vict\_sex" column, which may require imputation or handling.
* **Unknown Gender:** The value "H" in the "vict\_sex" column, representing an unknown gender, could be imputed to "X" to signify an unknown gender category.

**4. Common Crime Descriptions:** Analysis of crime descriptions reveals that the most prevalent type of crime in the dataset is "vehicle being stolen." This insight provides an understanding of the predominant criminal activity in the area.

**5. Crime Resolution Status:** 77% of crimes have been marked as "solved" or "Investigation completed." This suggests that a significant portion of recorded crimes have reached a resolution or undergone investigation.

**CHALLENGES ENCOUNTERED WITH THE LA CRIME DATASET**

**1. Missing Values:** The dataset contains missing or incomplete data in some columns, which affects analysis or modeling efforts.

**2. Inconsistent Data:** Inconsistent data entry or formatting errors in date, time, or other fields can lead to discrepancies in the dataset. For example, the mocodes for a particular row have a list of codes mentioned in a single data entry just by giving space in between them. This indicates that the data is inconsistent because the value of the codes cannot be in a single data entry.

**3. Geographic Bias:** The dataset is biased towards specific geographic areas within Los Angeles, which impacts the generalizability of models.

**4. Outliers:** Outliers in the data, such as extreme values or anomalies need to be identified and addressed to avoid skewing analysis results. For example, the dataset includes several outliers in the "vict\_age" column, indicating that there are extreme age values that may require further investigation or data validation.

**DATA CLEANING USING STAGING PIPELINES**

**1. tMap Component:** Replace missing values in particular columns using the tMap component. The NVL() function can be used to substitute a default value for null or outliers.

A screenshot of a computer

Description automatically generated

**2. tUniqRow Component:** To eliminate duplicates from your collection, use the tUniqRow component. You can set it up to just keep the first or last instance of a duplicate and delete duplicates depending on certain columns.

**3. tReplace Component:** To find specific values or patterns in your data and replace them with desired values, use the tReplace Component. For erasing inconsistencies or typographical errors, this is helpful.