Name: Monalika Pradhan

NUID: 002768020

DAMG 7370 Designing Advanced Data Architectures for Business Intelligence

Individual Assignment:Report for LA Crime Data

About LA Crime Data:

This dataset provides information on crime incidents in the City of Los Angeles dating back to 2020. The data is sourced from original crime reports that were transcribed from paper records, which may contain some inaccuracies.

Description of Data:

No	Column Name	Description	DataType
1.	DR_NO	Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits	Long/BigInt
2.	Date Rptd	MM/DD/YYYY HH:mm:ss	Date & Time
3.	Date Occ	MM/DD/YYYY HH:mm:ss	Date & Time
4.	Time Occ	In 24 hour military time.	Long/BigInt
5.	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.	Integer
6.	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. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.	String
7.	Rpt Dist No	A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons. Find LAPD Reporting Districts on the LA City GeoHub at http://geohub.lacity.org/datasets/c4f83909b81d4786aa8ba8a74a4b4db1_4	Integer
8.	Part 1-2		Integer
9.	Crm Cd	Indicates the crime committed. (Same as Crime Code 1)	Integer

10.	Crm Cd Desc	Defines the Crime Code provided.	String
11.	Mocodes	Modus Operandi: Activities associated with the suspect in commission of the crime.See attached PDF for list of MO Codes in numerical order. https://data.lacity.org/api/views/y8tr-7khq/files/3a967fbd-f210-4857-bc52-60230efe256c?download=true&filename=MO%20CODES%20(numerical%20order).pdf	String
12.	Vict Age	Two character numeric	Integer
13.	Vict Sex	F - Female M - Male X - Unknown	String
14.	Vict Descent	Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian	String
15.	Premis Cd	The type of structure, vehicle, or location where the crime took place.	Integer
16.	Premis Desc	Defines the Premise Code provided.	String
17.	Weapon Used Cd	The type of weapon used in the crime.	BigInt/Long
18.	Weapon Desc	Defines the Weapon Used Code provided.	String
19.	Status	Status of the case. (IC is the default)	String
20	Status Desc	Defines the Status Code provided.	String
21.	Crm Cd 1	Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.	Integer
22.	Crm Cd 2	May contain a code for an additional crime, less serious than Crime Code 1.	Integer
23.	Crm Cd 3	May contain a code for an additional crime, less serious than Crime Code 1.	Integer
24.	Crm Cd 4	May contain a code for an additional crime, less serious than Crime Code 1.	Integer
25.	Location	Street address of crime incident rounded to the nearest hundred block to maintain anonymity.	String

26.	Cross Street	Cross Street of rounded Address	String
27.	LAT	Latitude	Float
28.	LON	Longtitude	Double

Altreyx Tools Used:

Data Investigation Tool: Basic Profile| Browse
Preparation Tool: Unique| Select
Following are my views on LA Crime Data:

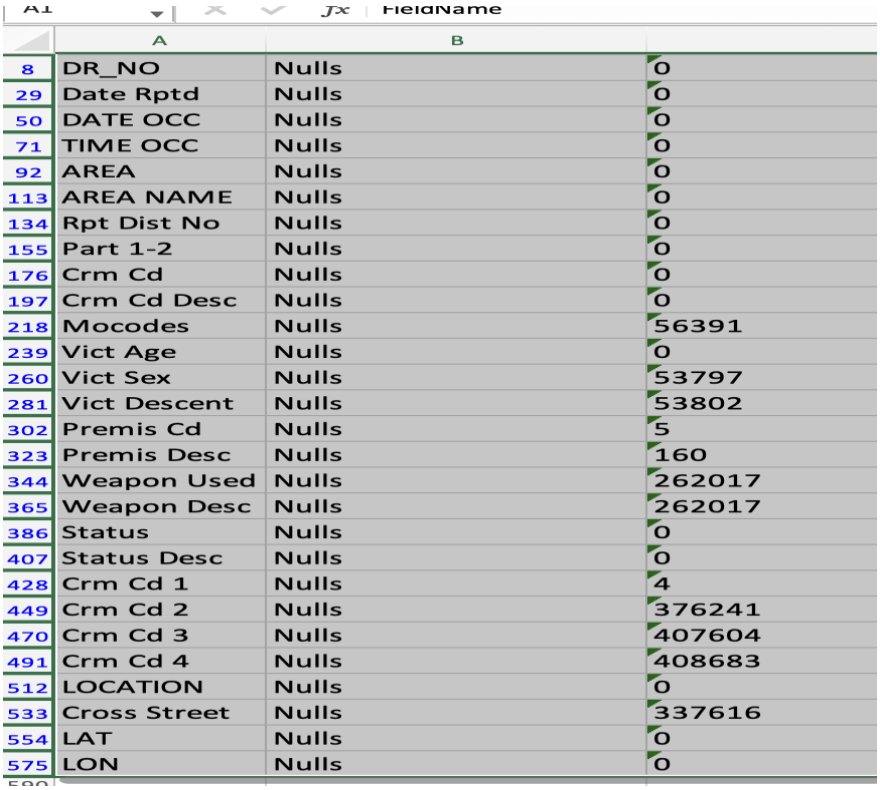
1. All the records in DR_No column are unique which can be set to primary key with data type as Long/Bigint
2. The Date_Rptd and Date_Occ has following date-time data type and the format of this columns are MM-DD-YYYY HH:MM:SS.
3. The Area Code with 12 and Area Name with 77th Street has the highest crime rate with 26368 crime happened that is around 6.45% of total crime
4. The column Part1-2 has 239220 (58.52%) records under part 1 and rest under part 2
5. From Crm Cd Desc we see that Vehicle stolen has the highest crime rate with 10.87% of total crime reported
6. The Vict Age column in the dataset displays some noticeable irregularities. Firstly, it contains numerous outliers, which are data points that significantly differ from the usual age range. Additionally, a sizable portion of the dataset, comprising approximately 98,709 entries, contains the value zero, indicating the absence of age information for these cases. Furthermore, it's important to highlight that the column includes negative values such as -1, -2, -3 which lack practical relevance when interpreting age-related data. This emphasizes the importance of assessing data quality and considering potential data cleaning measures to ensure the accuracy and meaningfulness of age-related information within the
7. Analyzing the Victim Sex column reveals that 42.27% of the cases are male, followed by 36.51% that are female, 13% containing null values and 8.03% unknown values. There are some values represented by H which is not clear to us so that column can be imputed to X.
8. The Weapon with Code 400 and Description Strong Arm (Hands, Fist, Feet or Body) is the highest used one.
9. Around 3,15,268 cases out of 408718 cases have been resolved, or the investigations have been concluded which is around 77% of total crime reported.

Basic Data Profiling using Altreyx:

1. Using the Basic Profile, I decided the data types of each column
2. Further Basic Profile was used to calculate the length of each column depending on the longest length of the column
3. Identify Unique columns/ duplicates

Missing Values & Inconsistencies:

- Used Basic Profile and exported to excel and added filter of null values



	A	B	
8	DR_NO	Nulls	0
29	Date Rptd	Nulls	0
50	DATE OCC	Nulls	0
71	TIME OCC	Nulls	0
92	AREA	Nulls	0
113	AREA NAME	Nulls	0
134	Rpt Dist No	Nulls	0
155	Part 1-2	Nulls	0
176	Crm Cd	Nulls	0
197	Crm Cd Desc	Nulls	0
218	Mocodes	Nulls	56391
239	Vict Age	Nulls	0
260	Vict Sex	Nulls	53797
281	Vict Descent	Nulls	53802
302	Premis Cd	Nulls	5
323	Premis Desc	Nulls	160
344	Weapon Used	Nulls	262017
365	Weapon Desc	Nulls	262017
386	Status	Nulls	0
407	Status Desc	Nulls	0
428	Crm Cd 1	Nulls	4
449	Crm Cd 2	Nulls	376241
470	Crm Cd 3	Nulls	407604
491	Crm Cd 4	Nulls	408683
512	LOCATION	Nulls	0
533	Cross Street	Nulls	337616
554	LAT	Nulls	0
575	LON	Nulls	0

- Victim age has 0 and negative values which are practically not possible which leads to inconsistency.

Data Cleaning in Stage Pipelines:

- From the Basic Profile, We can see that the age has some values like 0,-1,-2,-3 that are not practically possible. So we can use Mean/Median/Mode Imputation: Calculate the mean, median, or mode of a column and replace missing values with these statistics. You can use Talend's aggregation functions in combination with the tAggregateRow component.
- For Missing values the data cleaning can be done by Constant value imputation: The tMap component to replace missing values with a constant value.
- Use Formula in Altreyx to transform the inconsistent date to a standard value

Screenshot of Altreyx Workflow: