Data Profiling and Cleaning Documentation:

Dataset Overview: Crime Data in Los Angeles

Step 1: Initial Data Assessment

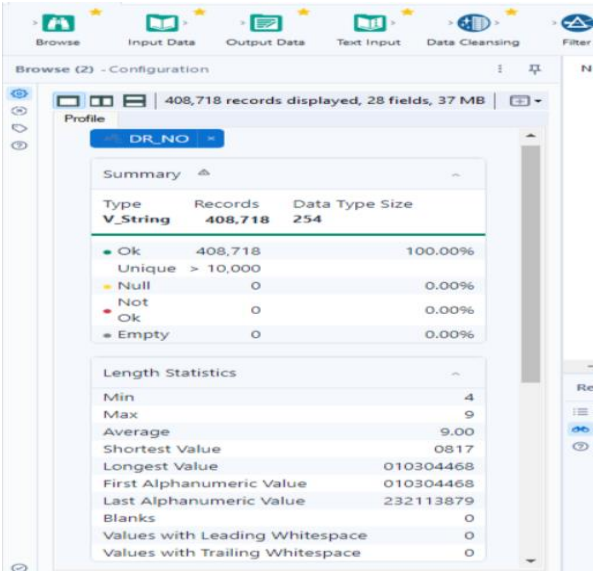
- Loaded the dataset comprising approximately 400,000 rows using the Input Data Tool in Alteryx
- Utilized the Browse Tool to visually inspect the data for further analysis

Step 2: Initial Data Evaluation

- Assessed metadata displayed in the left panel, encompassing data types, count of unique values, presence of null values, and length statistics for all 28 columns.

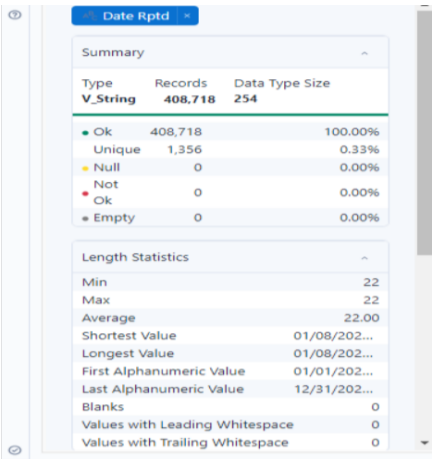
Step 3: Data Refinement Process

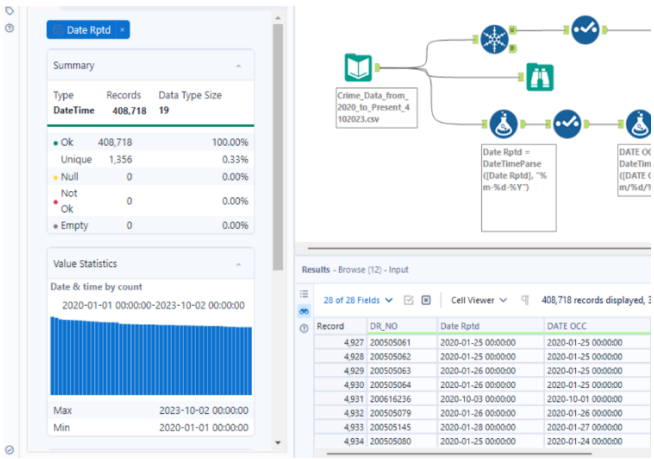
1) DR_NO: There are no null values or empty values in this column.
   - Confirmed the uniqueness of each row using the Unique Tool to guarantee distinct report numbers as this row can be considered as unique column.

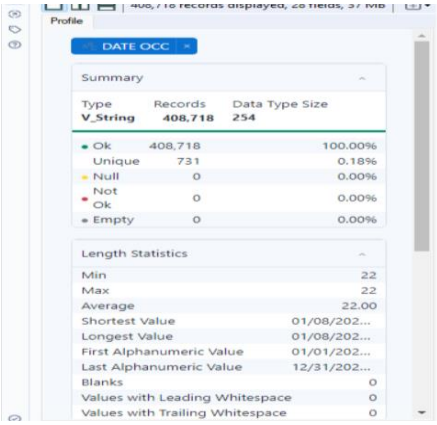
- 

2) Date_Rptd, DATE OCC: Here too, the data is clean, but the main issue is the data type. It is optimum to convert string to datetime type as the columns has date and time.
   - Transformed data type from v_string to datetime utilizing a Formula Tool with the expression **DateTimeParse([Column_name], "%m/%d/%Y")**.
   - Implemented a Select Tool to enforce the datetime data type consistently.

Output:

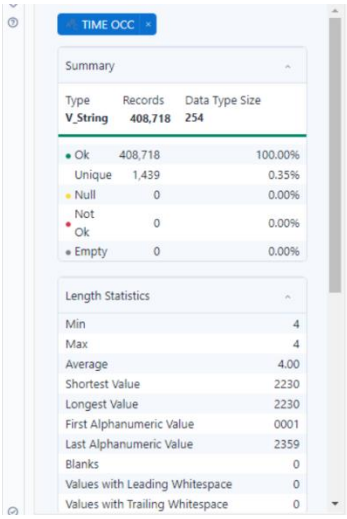






3) TIME OCC: In this column, the time in all the rows were not in the correct format such as hh:mm, by changing to the correct format will be useful during analysis and visualization.

Output:

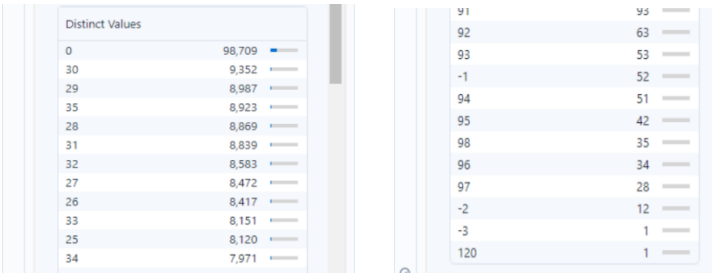


TIME OCC
2146
2125
2130
1800
0610
0531
1730
1900

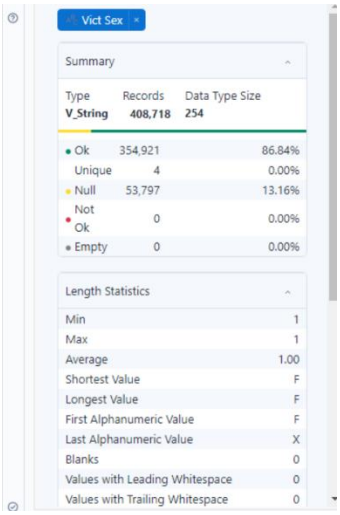
TIME OCC
21:46:00
21:25:00
21:30:00
18:00:00
06:10:00
05:31:00
17:30:00
19:00:00

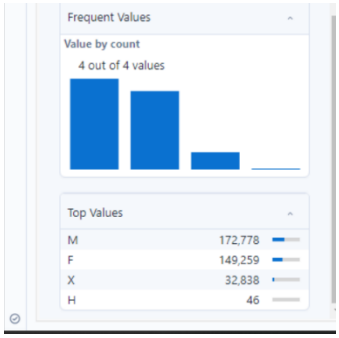
- Converted data type from v_string to time through a Formula Tool using the expression **ToDateTime(ToNumber(Left([TIME OCC], 2)) / 24 + ToNumber(Right([TIME OCC], 2)) / 1440)**.
- Employed a Select Tool to apply the time data type uniformly.

4) Vict Age: Observed that there a few negative and zero values which is incorrect w.r.t. age. It would be the best to confirm if the values are entered correct or get rid of those values because according to me imputation does not seem right as the data would be incorrect and fake.



5) Vict Sex: Here, along with multiple null values, there are unspecified values such as X with does not fall under any criteria. Like age, we could confirm if the values are entered correct or get rid of those values because according to me imputation does not seem right as the data would be incorrect and fake, moreover, it is not easy to impute string type data.





6) Null Value Considerations:
Managed columns with notable counts of NULL values (e.g., crm cd 2, crm cd 3, srm cd 4, weapons cd, weapons desc etc) by retaining their original state. Imputation was omitted to prevent potential data distortion. Further, Null could be replaced with 0 or NA for numbers or string type, respectively.

7) LOCATION, Cross Street:
Eliminated extraneous whitespace using the Data Cleansing Tool, enhancing data legibility.
Output:

LOCATION	Cross Street
1700 WESTMONT DR	[Null]
1500 N GAFFEY ST	[Null]
1700 ISLAND AV	[Null]
700 S PACIFIC AV	[Null]
1800 N LA BREA AV	[Null]
700 S PACIFIC AV	[Null]
WESTERN AV	228TH ST
1600 PACIFIC COAST HY	[Null]

LOCATION	Cross Street
1700 WESTMONT DR	[Null]
1500 N GAFFEY ST	[Null]
1700 ISLAND AV	[Null]
700 S PACIFIC AV	[Null]
1800 N LA BREA AV	[Null]
700 S PACIFIC AV	[Null]
WESTERN AV	228TH ST
1600 PACIFIC COAST HY	[Null]

8) Standardizing Data Types:
Updated data types of various columns using the Select Tool to ensure uniformity.