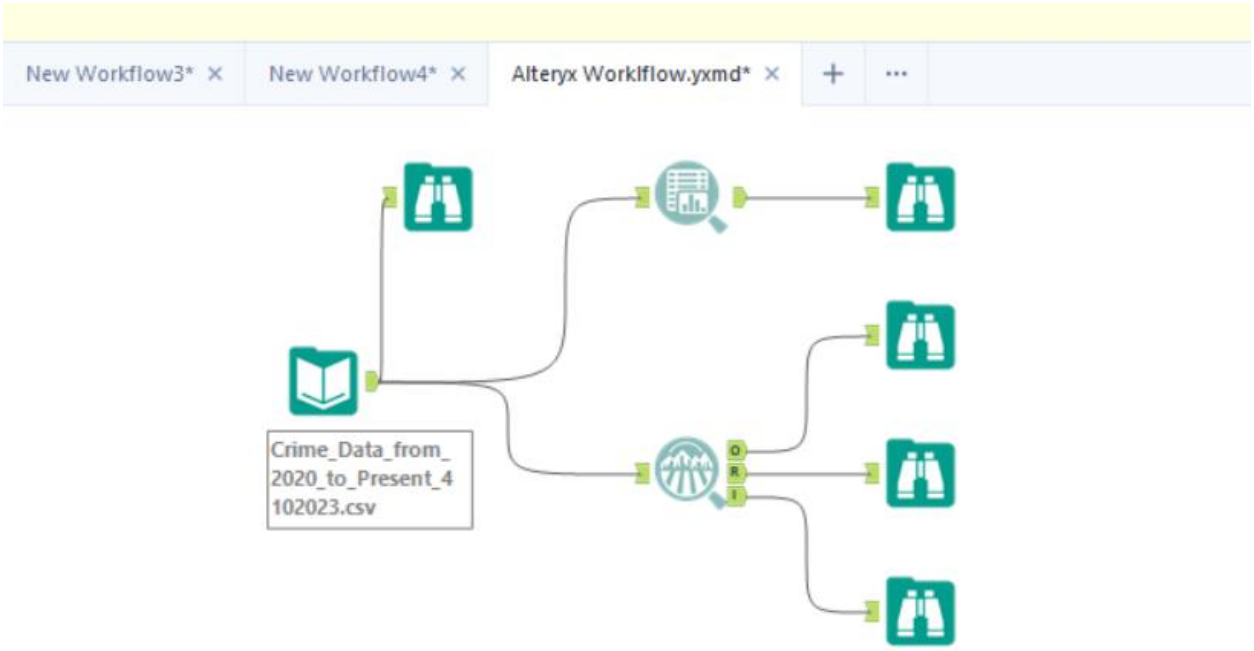
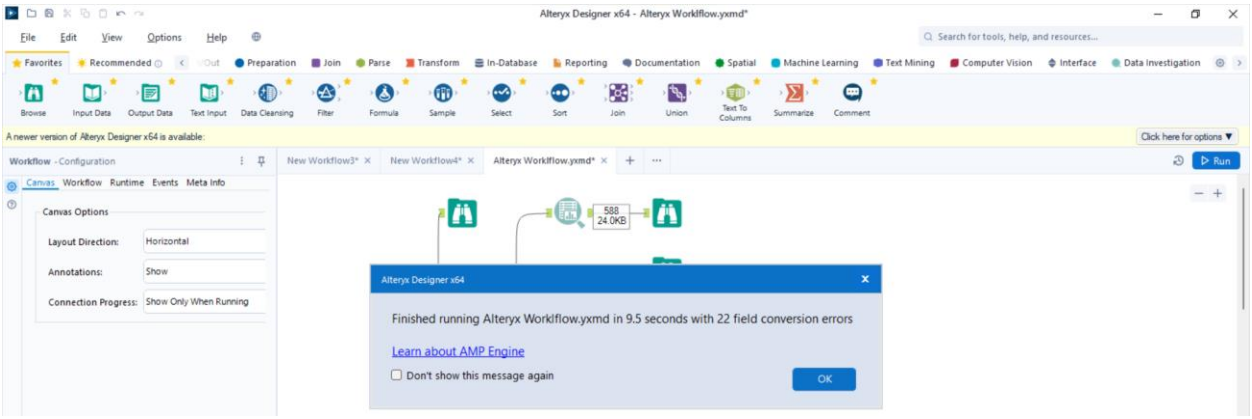
Individual Project 1 : LA Crime Data (Part 1) : Crime Data from 2020 to Present

1. Alteryx Data Profiling



The **Basic Data Profiling** tool in Alteryx offers a snapshot of your dataset's key attributes, **including data types, unique values, missing values, and basic statistics for each column**. It aids in understanding the data's structure and quality, serving as a valuable starting point for data cleaning and analysis tasks.

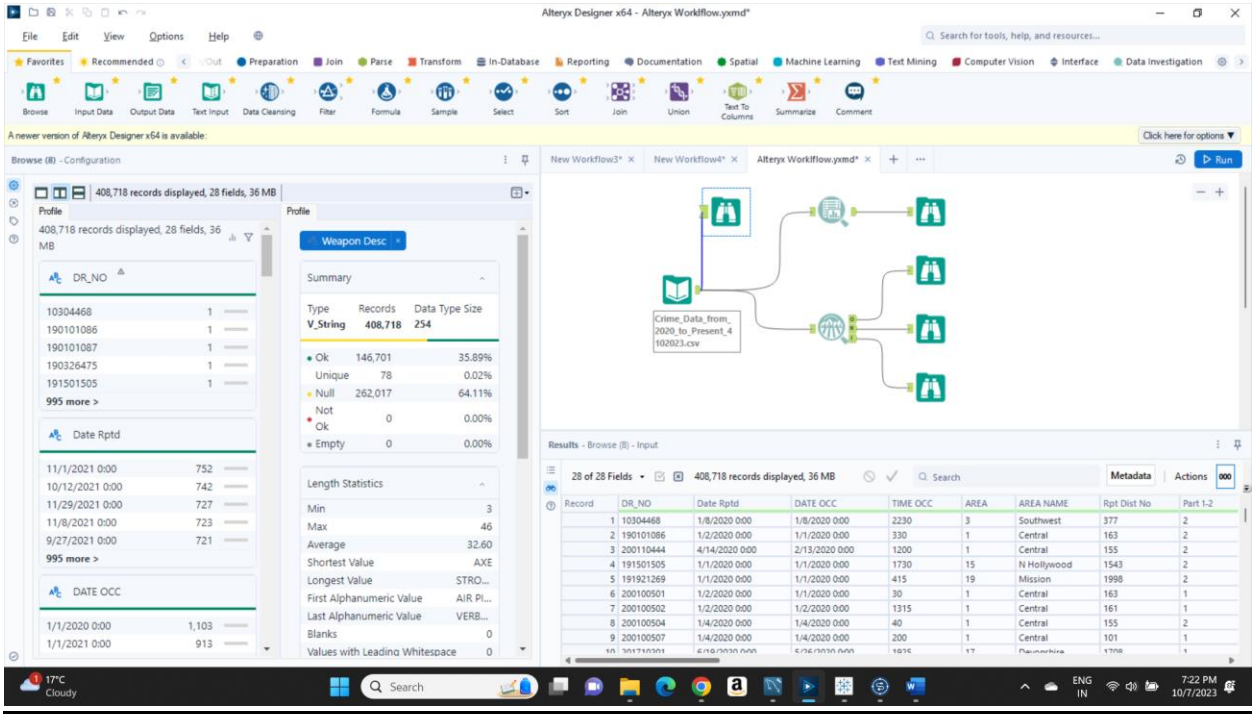
1. **Field Summary (O):**
   - Field Summary (O) stands for "Overall Field Summary."
   - It provides a comprehensive summary of a selected field within your dataset.
   - The output includes statistics, data distribution, and characteristics specific to that field.
2. **Field Summary (R):**
   - Field Summary (R) stands for "Record-By-Record Field Summary."
   - It generates a detailed summary report for each individual record or row in your dataset.
3. **Field Summary (I):**
   - Field Summary (I) stands for "Interactive Field Summary."
   - It allows interactive exploration and analysis of field-level statistics and characteristics.
   - When connected to a Browse tool, it provides an easy-to-use interface for exploring specific field summaries, unique values, and more.



Run Time for Alteryx Workflow Data Profiling: 9.5 seconds.

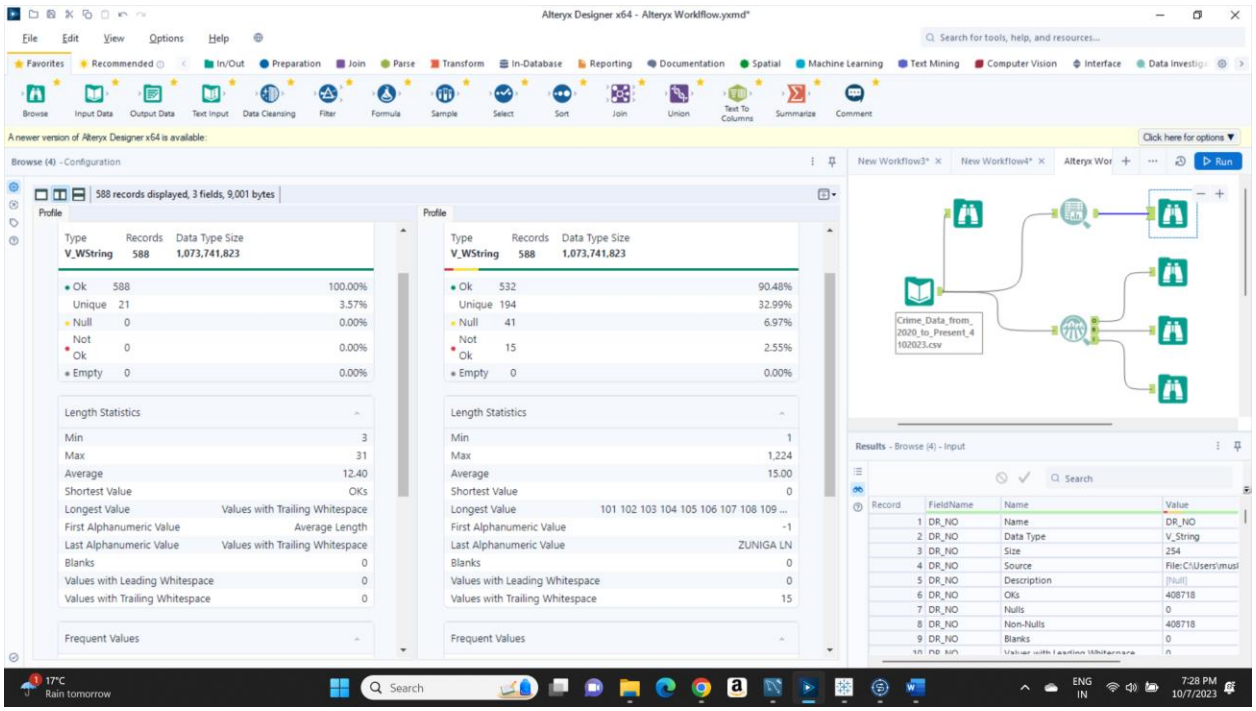
Requirement: Document indicating problems with the dataset, and how you plan to clean it before storing it in stage tables

A simple Browse of the Dataset



A simple browse through the Dataset gives us an interface that provides an interactive table view of the dataset. It allows us to visually inspect, sort, filter, and explore your data, making it easy to understand the dataset's content and structure before proceeding with data profiling, cleaning or analysis.

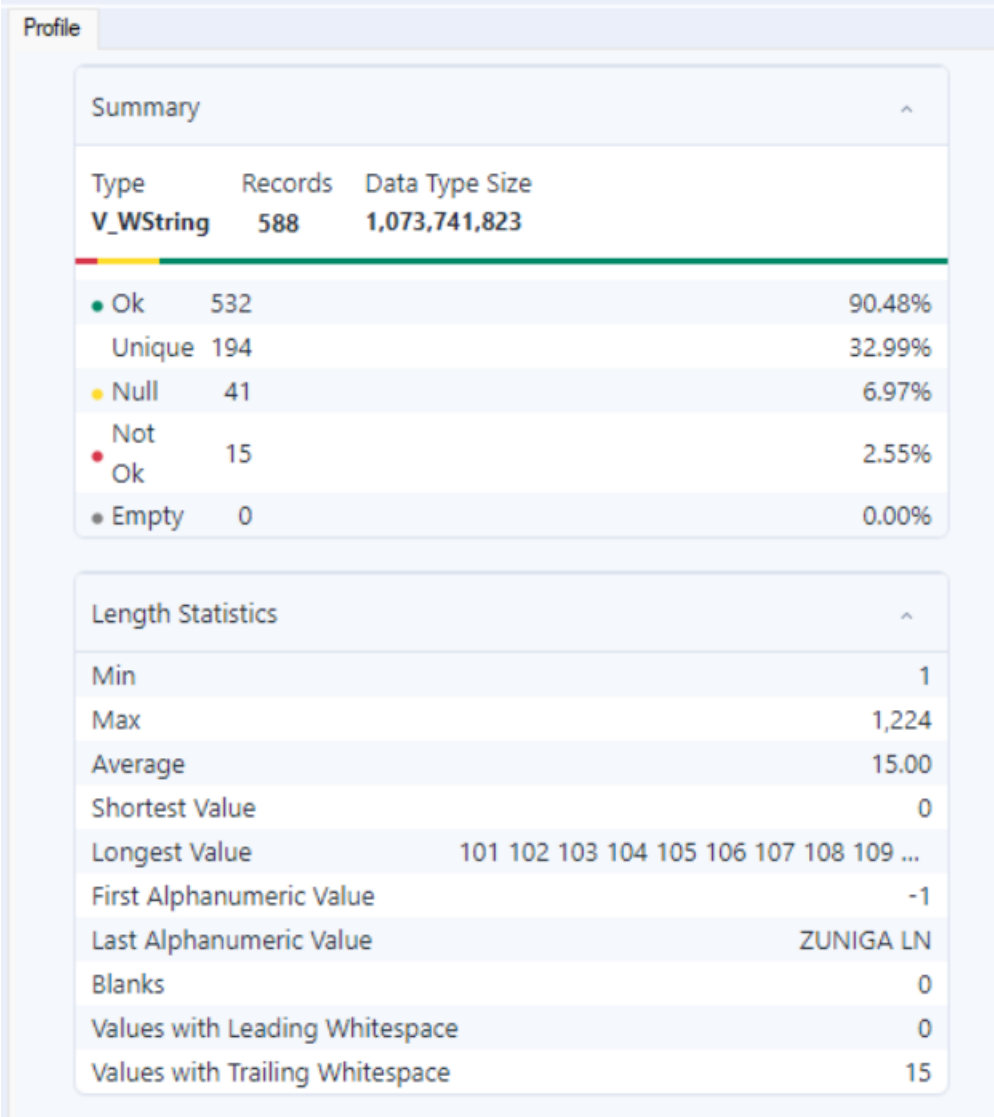
Basic Analysis: The data set contains a number of missing values and quite a number of outliers.



Basic Data Profile:

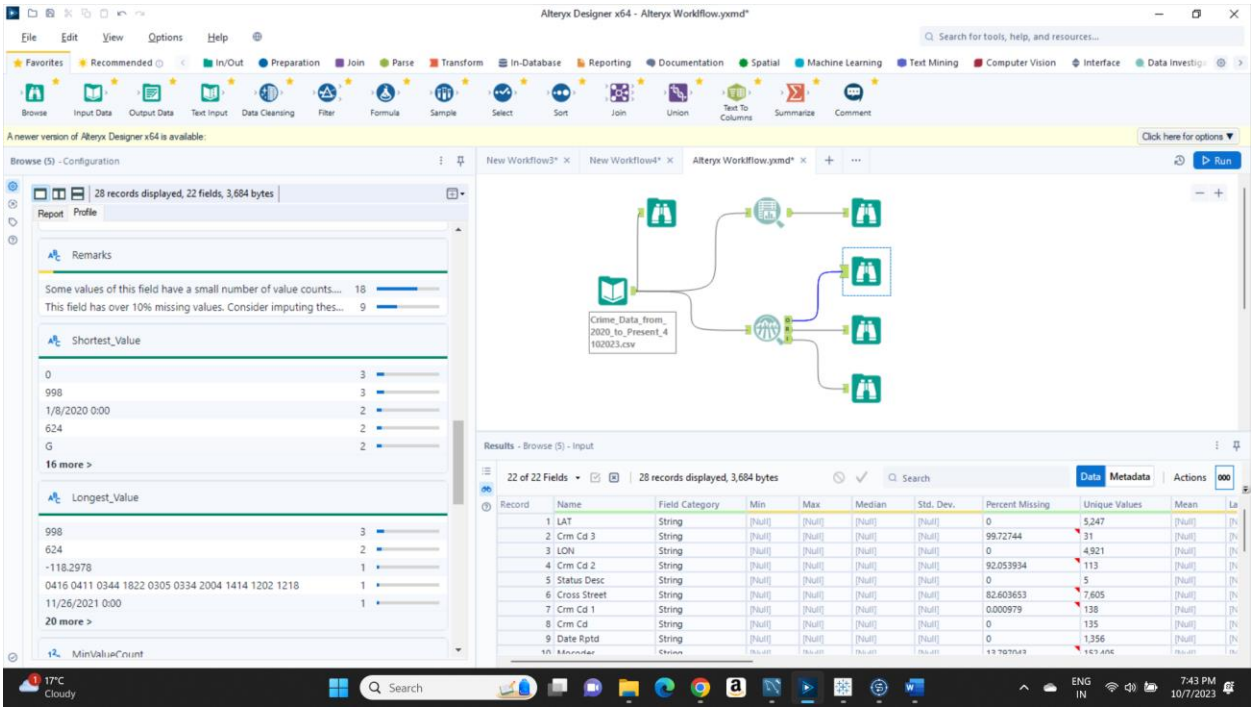
Length Statistics	
Min	3
Max	31
Average	12.40
Shortest Value	OKs
Longest Value	Values with Trailing Whitespace
First Alphanumeric Value	Average Length
Last Alphanumeric Value	Values with Trailing Whitespace
Blanks	0
Values with Leading Whitespace	0
Values with Trailing Whitespace	0

- The Name Profile for the dataset is as given above.

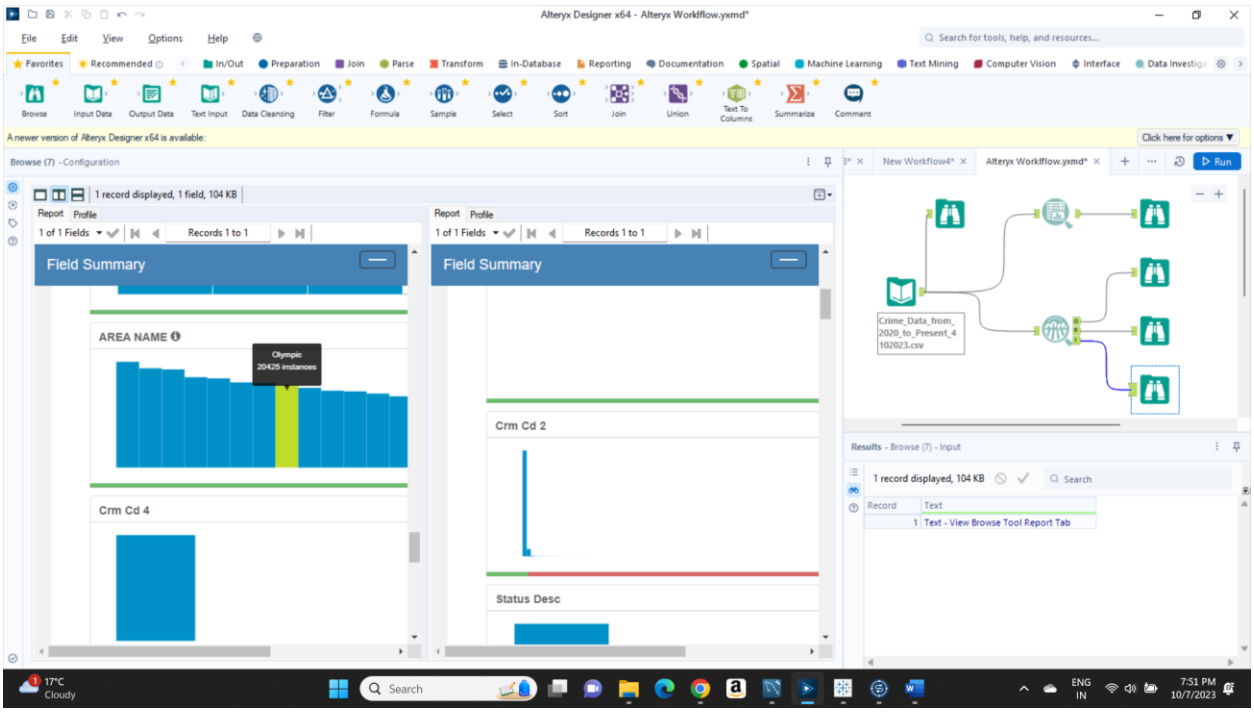


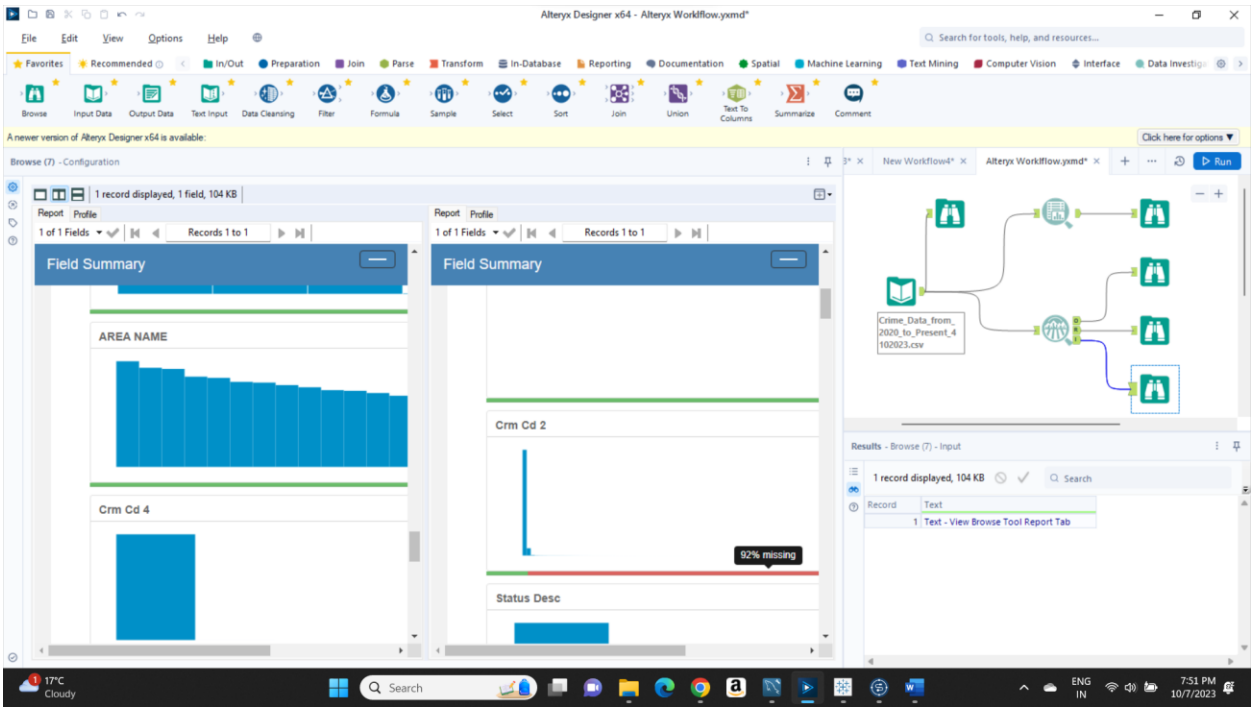
- The Basic Data Profile for the values is as shown above.
    - Few observations are as follows:
        - There is a total of 6.97% null values.
        - Overall data seems to be okay – which is around 90% of the data.
        - There are 32% unique values in the dataset.
        - Around 2.5% of the data record does not meet the required criteria or standards set for the dataset. It indicates that there might be an issue with that particular data point, such as missing or incomplete information, data format errors, outliers, or any other data quality problems.

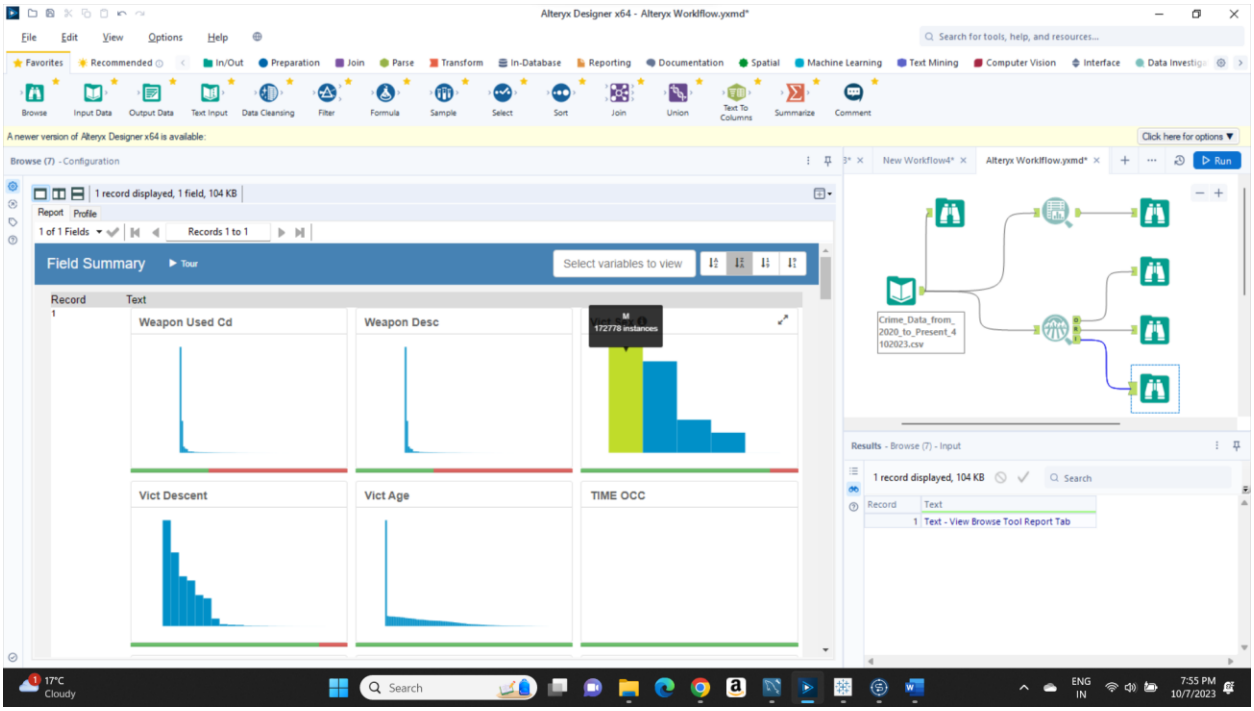
Overall Field Summary is shown below:



The Interactive Field Summary shown below displays the percentage of missing values, field-level statistics and characteristics etc.




Alteryx Designer x64 - Alteryx Workflow.yxmd*
Field Summary
AREA NAME
Crm Cd 4
Crm Cd 2
92% missing
Status Desc
Crime_Data_from_2020_to_Present_4102023.csv
1 record displayed, 104 KB
Text - View Browse Tool Report Tab


Alteryx Designer x64 - Alteryx Workflow.yxmd*
Field Summary
Select variables to view
Weapon Used Cd
Weapon Desc
172778 instances
Vict Descent
Vict Age
TIME OCC
Crime_Data_from_2020_to_Present_4102023.csv
1 record displayed, 104 KB
Text - View Browse Tool Report Tab

Record
1

Report

String/Character Fields

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
LAT	0.0%	5,247	0	34.0141	1	2,259	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Crm Cd 3	99.7%	31	998	998	1	407,604	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
LON	0.0%	4,921	0	-118.2978	1	2,998	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Crm Cd 2	92.1%	113	998	998	1	376,241	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Status Desc	0.0%	5	Juv Other	Adult Arrest	653	315,268	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Cross Street	82.6%	7,605	G	S SPRING ST	1	337,616	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

Record Report

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Crm Cd 1	0.0%	138	624	624	1	44,431	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Crm Cd	0.0%	135	624	624	1	44,438	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Date Rptd	0.0%	1,356	1/8/2020 0:00	11/26/2021 0:00	1	752	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Mocodes	13.8%	152,405	329	0416 0411 0344 1822 0305 0334 2004 1414 1202 1218	1	56,391	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
DR_NO	0.0%	408,718	817	190101086	1	1	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Premis Cd	0.0%	307	501	501	1	104,309	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
DATE OCC	0.0%	731	1/8/2020 0:00	11/30/2020 0:00	415	1,103	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

Record Report

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Rpt Dist No	0.0%	1,181	377	1543	1	2,122	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Vict Descent	13.2%	20	B	B	19	124,972	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
LOCATION	0.0%	53,880	G	14400 TITUS ST	1	760	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Premis Desc	0.0%	304	BANK	VEHICLE STORAGE LOT (CARS, TRUCKS, RV'S, BOATS, TRAILERS, ETC.)	1	104,309	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Vict Sex	13.2%	5	F	F	46	172,778	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Crm Cd Desc	0.0%	135	ARSON	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD	1	44,438	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Status	0.0%	5	AO	AO	653	315,268	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

Record Report

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
AREA NAME	0.0%	21	Harbor	N Hollywood	14,120	26,368	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Crm Cd 4	100.0%	6	998	998	1	408,683	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Weapon Desc	64.1%	79	AXE	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	1	262,017	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Vict Age	0.0%	103	0	120	1	98,709	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
TIME OCC	0.0%	1,439	1	2230	3	15,660	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Weapon Used Cd	64.1%	79	400	400	1	262,017	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

Record Report

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
AREA	0.0%	21	3	15	14,120	26,368	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Part 1-2	0.0%	2	2	2	169,498	239,220	

Overall Analysis:

1. There are a lot of Missing values and null values in the dataset which could affect the analysis of the data. We should either remove the missing data points with missing data or find a calculated mean/avg value to impute the missing data. We can also use various techniques like KNN imputation or MICE imputation method to get the missing data. Some of the data which contains missing values are:

   Morcodes, Vict_Age, Vict_Sex, Vict_Descent, Weapon_Used_Cd, Crm Cd 1/2/3/4, Cross_Street
2. Outliers: There are many outliers in the dataset as well. We saw the not okay values above. We should do various analysis techniques to identify these outliers and check how can we remove/include these outlier values so that it does not impact the data analysis.
3. Data Types: There are various variables that are in different formats like the Date value is represented in the String format etc. Therefore, when we are visualizing the dataset, we need to change the data types to get the correct visualization. Therefore we should change the schema of the dataset with correct data type and length.