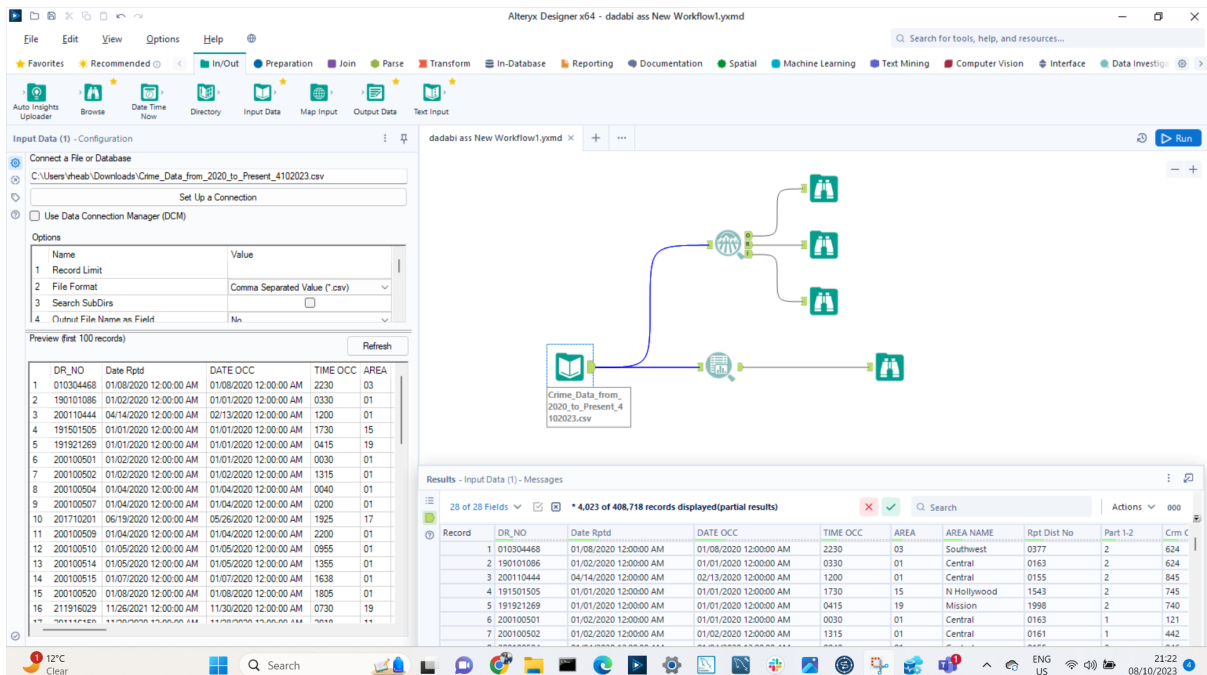
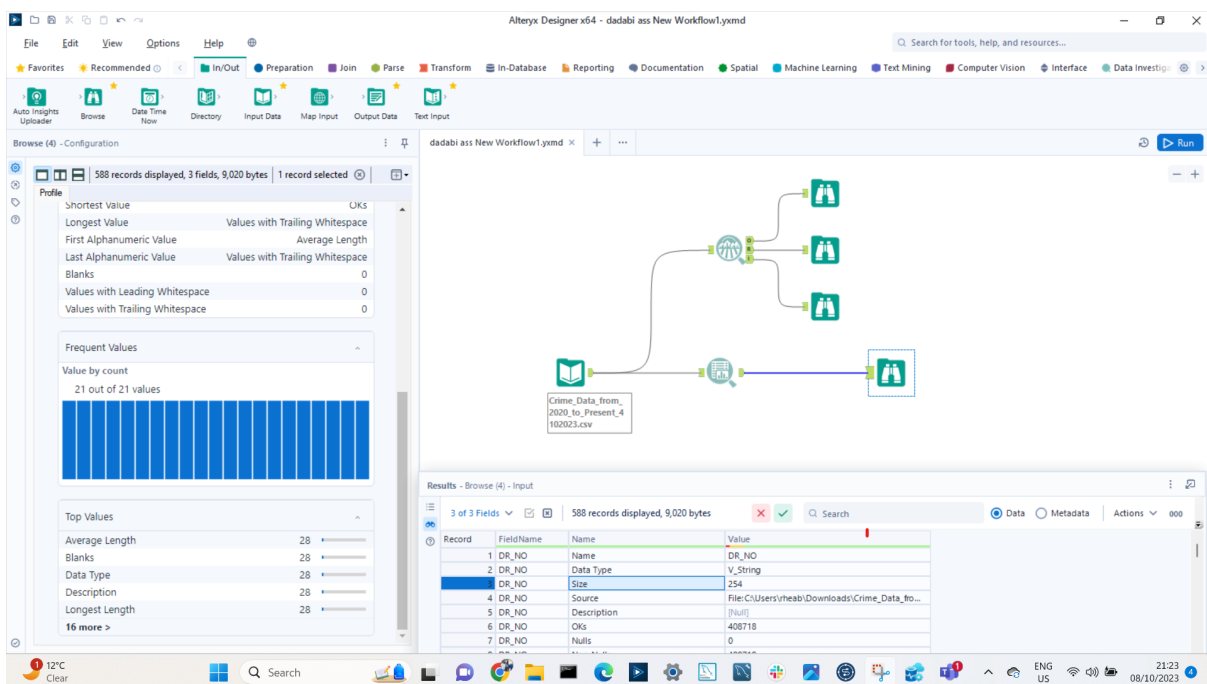


Alteryx Documentation



The screenshot shows the Alteryx Designer interface with a workflow named "dadabi ass New Workflow1.yxmd". The workflow consists of an "Input Data" tool connected to a "Join" tool, which is then connected to a "Parse" tool. The "Input Data" tool is configured to connect to a file at "C:\Users\vhieb\Downloads\Crime_Data_from_2020_to_Present_4102023.csv". The "Join" tool is configured to join on the "DR_NO" field. The "Parse" tool is configured to parse the "DATE OCC" field. The "Results - Input Data (1) - Messages" pane shows a table with 28 fields and 4,023 records displayed (partial results).

Record	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm C
1	010304468	01/08/2020 12:00:00 AM	01/08/2020 12:00:00 AM	2230	03	Southwest	0377	2	624
2	190101086	01/02/2020 12:00:00 AM	01/01/2020 12:00:00 AM	0330	01	Central	0163	2	624
3	200110444	04/14/2020 12:00:00 AM	02/13/2020 12:00:00 AM	1200	01	Central	0155	2	845
4	191501505	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	1730	15	N Hollywood	1543	2	745
5	191921269	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	0415	19	Mission	1998	2	740
6	200100501	01/02/2020 12:00:00 AM	01/01/2020 12:00:00 AM	0030	01	Central	0163	1	121
7	200100502	01/02/2020 12:00:00 AM	01/02/2020 12:00:00 AM	1315	01	Central	0161	1	442



The screenshot shows the Alteryx Designer interface with the same workflow. The "Browse (4) - Configuration" pane is open, showing the "Profile" tab. The "Profile" tab displays statistics for the "DR_NO" field, including "Shortest Value", "Longest Value", "First Alphanumeric Value", "Last Alphanumeric Value", "Blanks", "Values with Leading Whitespace", and "Values with Trailing Whitespace". The "Frequent Values" section shows a bar chart for "Value by count" with 21 out of 21 values. The "Top Values" section shows a list of values for "Average Length", "Blanks", "Data Type", "Description", and "Longest Length". The "Results - Browse (4) - Input" pane shows a table with 3 fields and 588 records displayed (9,020 bytes).

Record	FieldName	Name	Value
1	DR_NO	Name	DR_NO
2	DR_NO	Data Type	V_String
3	DR_NO	Size	254
4	DR_NO	Source	File:C:\Users\vhieb\Downloads\Crime_Data_fo...
5	DR_NO	Description	[Null]
6	DR_NO	OKs	408718
7	DR_NO	Nulls	0



File Edit View Options Help

★ Favorites Recommended < In/Out Preparation

> Browse > Input Data > Output Data > Text Input > Data Cleansing

Browse (2) - Configuration



408,718 records displayed, 28 fields, 37 MB

Profile

Type	Records	Data Type Size
V_String	408,718	254

● Ok	408,718	100.00%
Unique	21	0.01%
● Null	0	0.00%
Not		
● Ok	0	0.00%
● Empty	0	0.00%

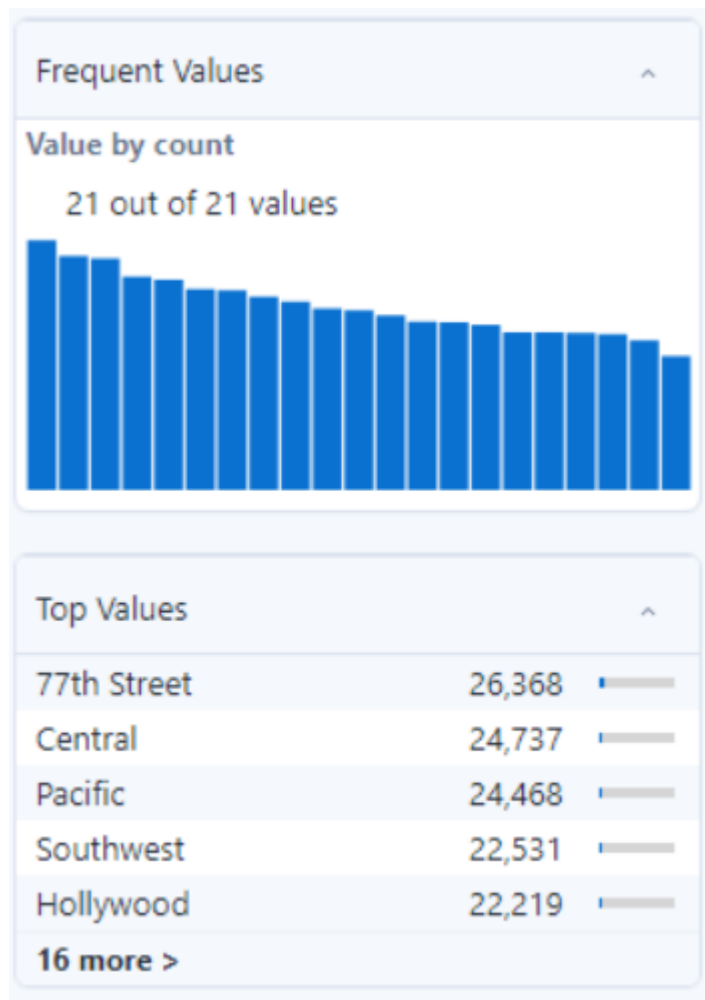
Length Statistics ^

Min	6
Max	11
Average	8.30
Shortest Value	Harbor
Longest Value	N Hollywo...
First Alphanumeric Value	77th Street
Last Alphanumeric Value	Wilshire
Blanks	0
Values with Leading Whitespace	0
Values with Trailing Whitespace	0

Frequent Values ^

Value by count

21 out of 21 values



Make observations on data, and write down the problems that you see with the data like missing values, and inconsistencies. Also, document how you plan to clean those using your staging data pipeline.

Based on the observations i see in the data i came to the following conclusion

- 1) **Missing Values:** Some columns have missing values indicated by the "Nulls" count. For example, "Crm Cd 2," "Crm Cd 3," "Crm Cd 4," "Crm Cd Desc," "Weapon Used Cd," "Weapon Desc," "Crm Cd 1," "Premis Desc," "Premis Cd," "Cross Street," and "LOCATION" columns have a significant number of missing values.
- 2) **Inconsistencies in Data Types:** Some columns have inconsistent data types. For instance, "Vict Age" is supposed to contain numerical values, but it is categorised as a V_String. Similarly, "TIME OCC" is categorised as V_String but should ideally be of a time-related data type.

- 3) **Potential Data Quality Issues:** The "DATE OCC" and "Date Rptd" columns have a mixture of date and time information. These columns should be split into separate date and time columns for clarity and consistency.
- 4) **Duplicate Column Names:** Some columns have duplicate names, such as "DR_NO" and "Date Rptd." It's important to clarify whether these columns hold different types of data or if this is a naming inconsistency.
- 5) **Data Length and Size Mismatch:** The "LOCATION" column has an average length of 35.6 characters, but the longest value is 40 characters. This indicates potential data truncation or inconsistencies in the data.
- 6) **Inconsistent Minimum and Maximum Values:** Some columns have minimum and maximum values that seem unusual or inconsistent with the data type. For example, "TIME OCC" has a minimum of 1 and a maximum of 2359, which may not represent valid time values.
- 7) **Uniqueness Issues:** Some columns have a high number of unique values, such as "LOCATION," "Crm Cd Desc," and "Mocodes," which may suggest potential data quality issues or the need for data cleaning.
- 8) **Whitespace Issues:** There are columns that report zero values for leading, trailing, or both leading and trailing whitespace. It's essential to verify if these values have been properly cleaned and formatted.
- 9) **Inconsistent Encoding:** The "AREA NAME" column has a mix of character encodings, which may lead to inconsistencies in data processing.
- 10) **Inconsistent Descriptions:** Some columns have descriptions labeled as "[Null]," which may indicate a lack of metadata or documentation for the data.

Data cleaning

To solve specific data quality problems, use Alteryx's data cleansing tools. Typical tasks include:

- deleting redundant records.
- Taking care of missing values (using methods like imputation, elimination, or replacement).
- standardising formats (such text, date, and time).
- data type corrections (such as converting text values to integers).
- deleting any white spaces or special letters.
- handling outliers (e.g., elimination or winsorization).

Data Transformation

Make the required changes to the data to prepare it for analysis.

- Aggregating data at several levels (such as summarizing sales data by month or region) is one example of this.
- creating derived columns or calculated fields.
- if necessary, combining and merging datasets.