End-to-End LA Crime Analytics (Medallion Architecture)

Alteryx Data Profiling

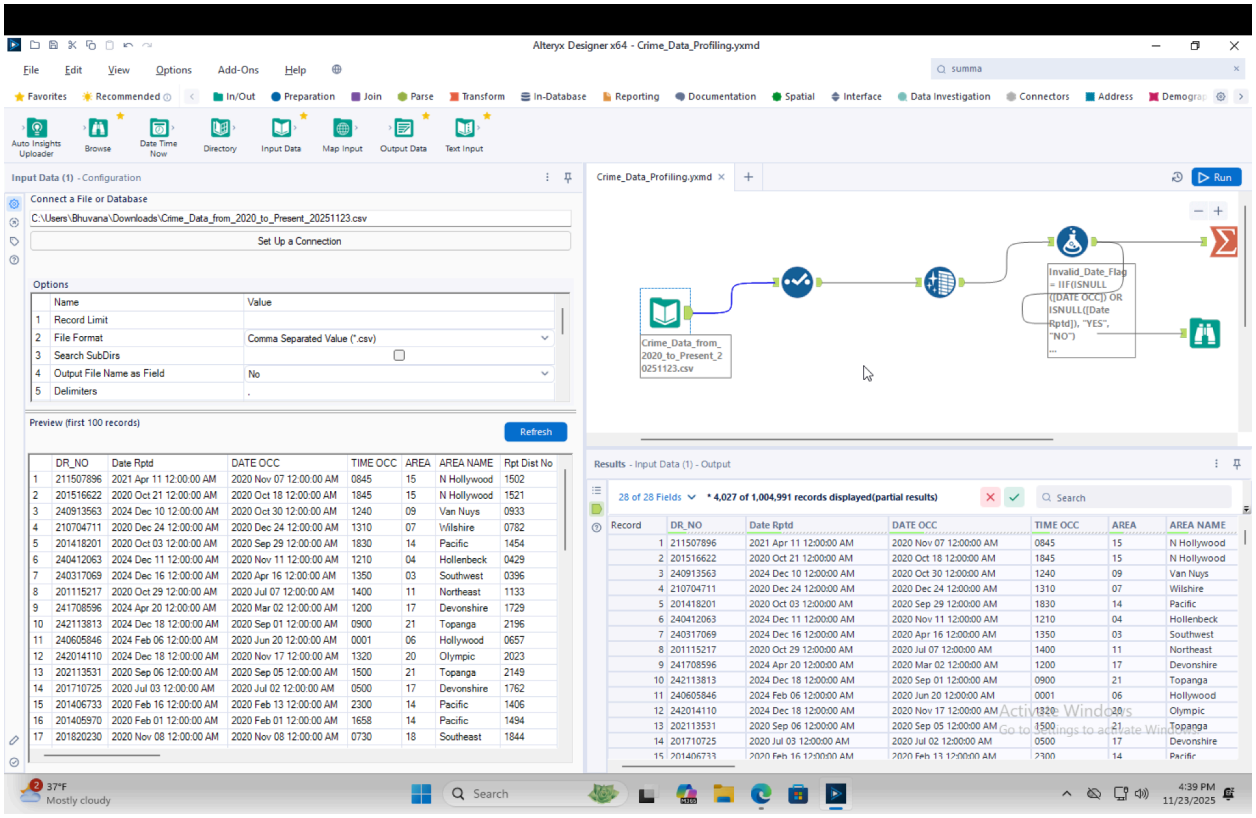
1. Input & Select Tool
• Loaded the LA Crime dataset (2020–Present).
• Standardized data types: Date fields → Date, numeric fields → Int/Double, descriptive fields → String.

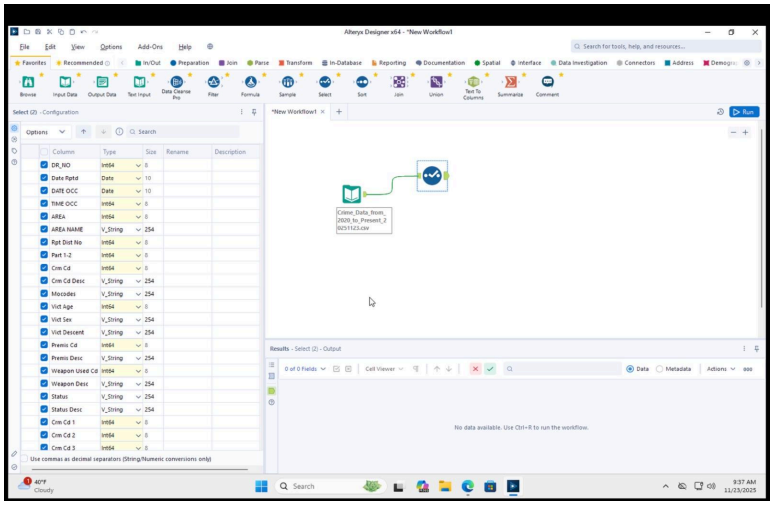
2. Data Cleansing
• Replaced nulls in numeric fields with 0 and in string fields with blanks.
• Removed leading/trailing whitespace.

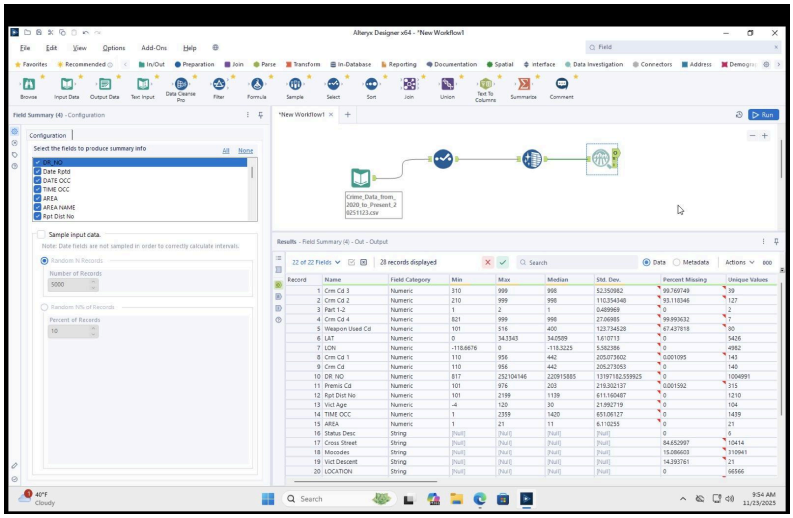
3. Field Summary & Browse Tools
• Generated automated profiling for:
- Numeric fields (min, max, unique counts)
- Date fields (missingness, earliest/latest)
- String fields (unique values, length)

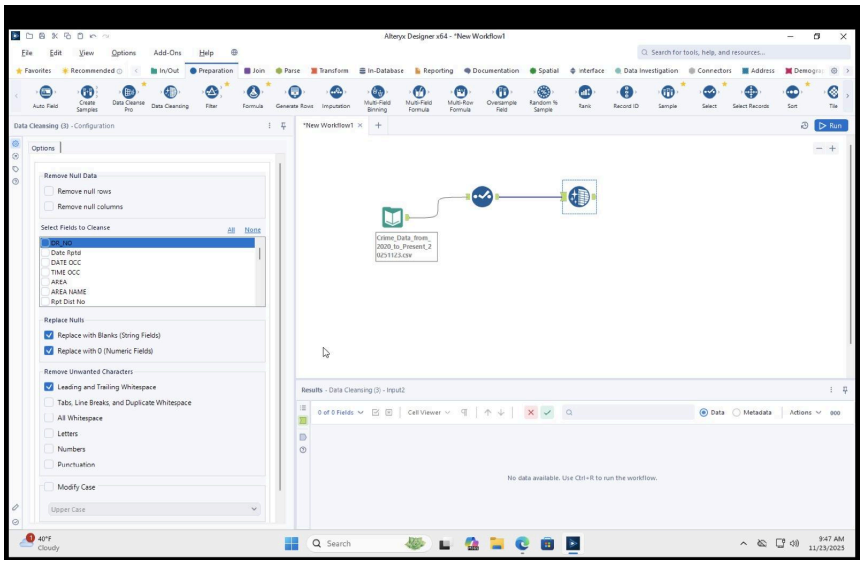
4. Formula Tool – Data Quality Flags Added
Created four validation columns:
• Invalid_Date_Flag
• Invalid_Time_Flag
• Invalid_Geo_Flag
• Invalid_Age_Flag
These flags help identify missing or out of range values.

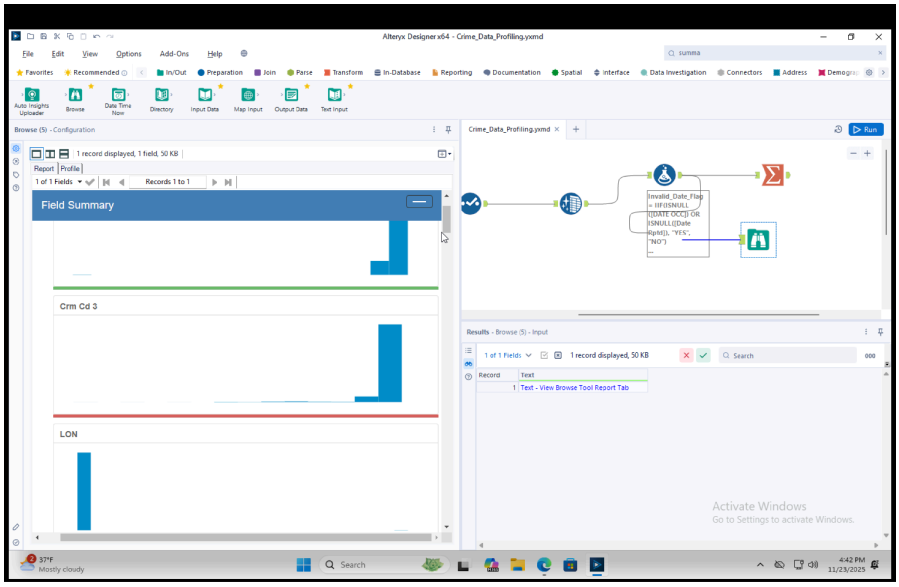
5. Summarize Tool
• Counted invalid occurrences for each flag
.• All four flags returned counts equal to total records, indicating significant data quality issues.

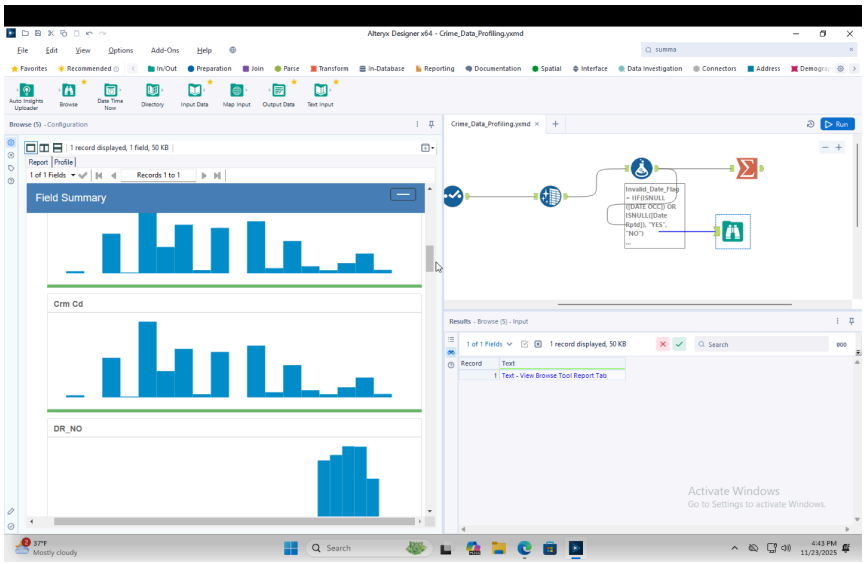

Alteryx Designer x64 - Crime_Data_Profiling.yxmd
Input Data (1) - Configuration
Connect a File or Database
C:\Users\Bhuvana\Downloads\Crime_Data_from_2020_to_Present_20251123.csv
Set Up a Connection
Options
Name
Value
Record Limit
File Format
Comma Separated Value (*.csv)
Search SubDirs
Output File Name as Field
No
Delimiters
Preview (first 100 records)
Refresh
DR_NO
Date Rptd
DATE OCC
TIME OCC
AREA
AREA NAME
Rpt Dist No
Crime_Data_Profiling.yxmd
Run
Crime_Data_from_2020_to_Present_20251123.csv
Invalid_Date_Flag = IIF(ISNULL([DATE OCC]) OR ISNULL([Date Rptd]), "YES", "NO")
Results - Input Data (1) - Output
28 of 28 Fields
* 4,027 of 1,004,991 records displayed(partial results)
Search
Record
DR_NO
Date Rptd
DATE OCC
TIME OCC
AREA
AREA NAME
37°F
Mostly cloudy
Search
4:39 PM
11/23/2025

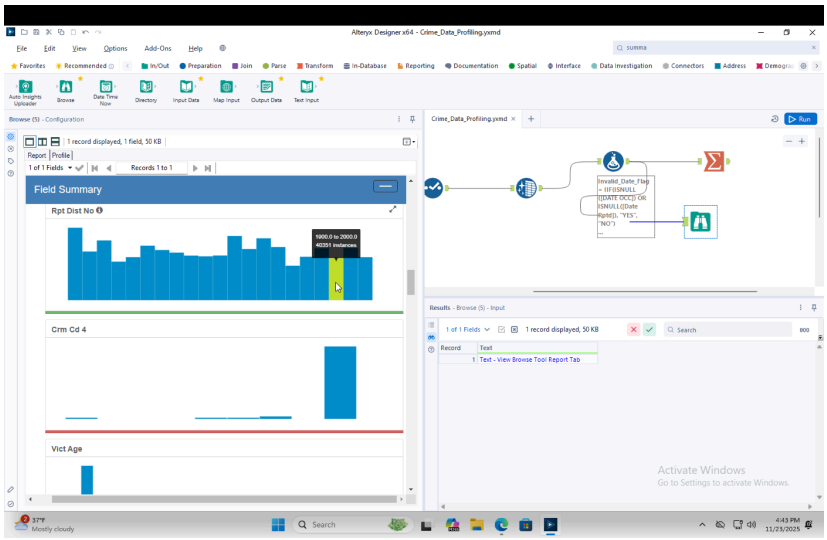


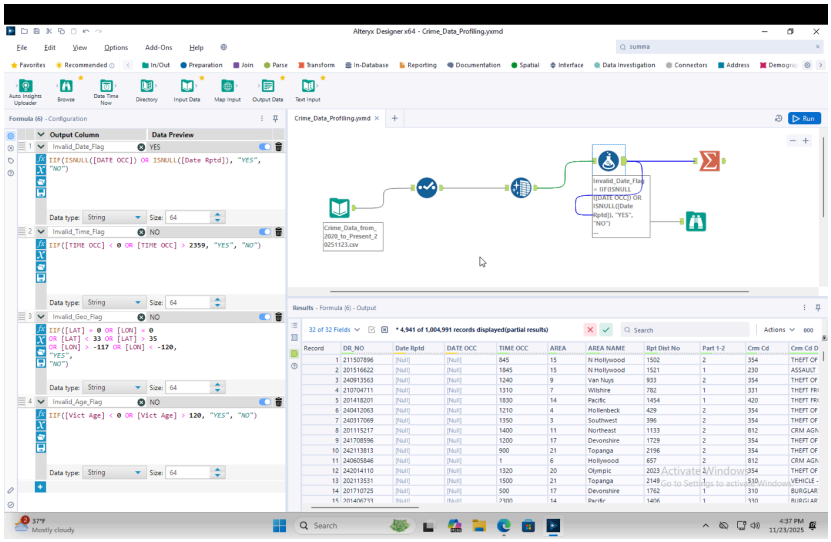

Alteryx Designer x64 - *New Workflow1
Field Summary (4) - Configuration
Configuration
Select the fields to produce summary info
All
None
OBJECTID
Date Rptd
DATE OCC
TIME OCC
AREA
AREA NAME
Rpt Dist No
Sample input data.
Note: Date fields are not sampled in order to correctly calculate intervals.
Random N Records
Number of Records
5000
Random N% of Records
Percent of Records
10
*New Workflow1
Run
Crime_Data_from_2020_to_Present_2025112...csv
Results - Field Summary (4) - Out - Output
22 of 22 Fields
28 records displayed
Search
Data
Metadata
Actions
Record
Name
Field Category
Min
Max
Median
Std. Dev.
Percent Missing
Unique Values
1 Crm Cd 3 Numeric 310 999 998 52.350982 99.769749 39
2 Crm Cd 2 Numeric 210 999 998 110.354348 93.110346 127
3 Part 1-2 Numeric 1 2 1 0.489969 0 2
4 Crm Cd 4 Numeric 821 999 998 27.06985 99.993632 7
5 Weapon Used Cd Numeric 101 516 400 123.734528 67.437818 80
6 LAT Numeric 0 34.3343 34.0599 1.610713 0 5426
7 LON Numeric -118.6676 0 -118.3225 5.582386 0 4982
8 Crm Cd 1 Numeric 110 956 442 205.073602 0.001095 143
9 Crm Cd Numeric 110 956 442 205.273553 0 140
10 DR NO Numeric 817 252104146 220915885 131971182.559925 0 1004991
11 Premis Cd Numeric 101 976 203 219.302137 0.001592 315
12 Rpt Dist No Numeric 101 2199 1139 611.160487 0 1210
13 Vict Age Numeric -4 120 30 21.992719 0 104
14 TIME OCC Numeric 1 2359 1420 651.06127 0 1439
15 AREA Numeric 1 21 11 6.110255 0 21
16 Status Desc String [Null] [Null] [Null] [Null] 0 6
17 Cross Street String [Null] [Null] [Null] [Null] 84.652997 10414
18 Mocodes String [Null] [Null] [Null] [Null] 15.086603 310941
19 Vict Descent String [Null] [Null] [Null] [Null] 14.393761 21
20 LOCATION String [Null] [Null] [Null] [Null] 0 66566
40°F
Cloudy
Search
9:54 AM
11/23/2025

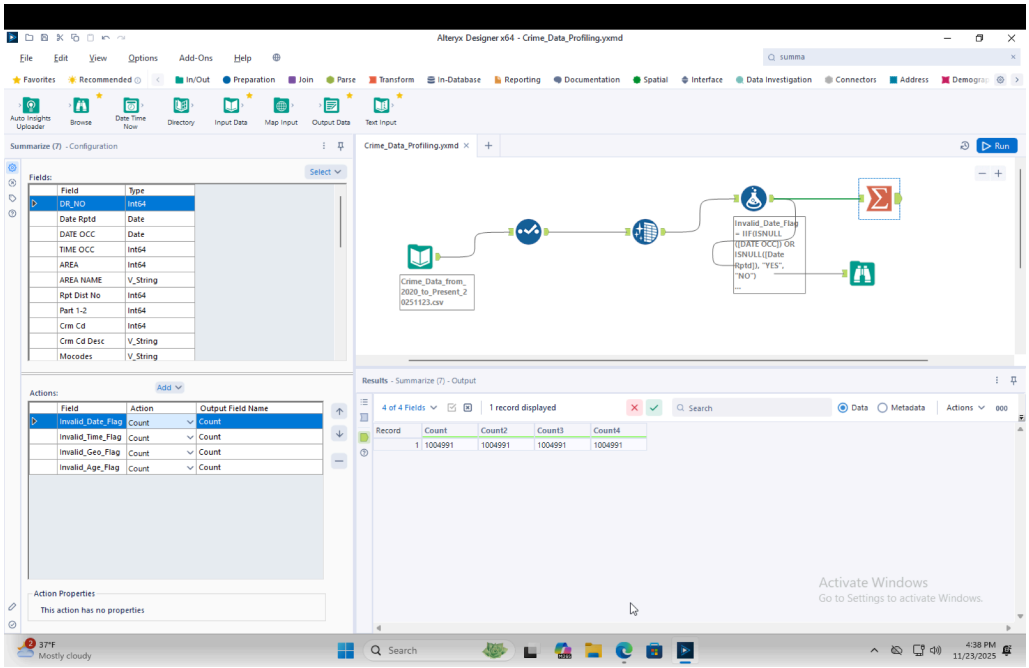

Alteryx Designer x64 - *New Workflow1
File
Edit
View
Options
Add-Ons
Help
Search for tools, help, and resources...
Favorites
Recommended
In/Out
Preparation
Join
Parse
Transform
In-Database
Reporting
Documentation
Spatial
Interface
Data Investigation
Connectors
Address
Auto Field
Create Samples
Data Cleanse Pro
Data Cleansing
Filter
Formula
Generate Rows
Imputation
Multi-Field Binning
Multi-Field Formula
Multi-Row Formula
Oversample Field
Random % Sample
Rank
Record ID
Sample
Select
Select Records
Sort
Tile
Data Cleansing (3) - Configuration
Options
Remove Null Data
Remove null rows
Remove null columns
Select Fields to Cleanse
All
None
DR_NO
Date Rptd
DATE OCC
TIME OCC
AREA
AREA NAME
Rpt Dist No
Replace Nulls
Replace with Blanks (String Fields)
Replace with 0 (Numeric Fields)
Remove Unwanted Characters
Leading and Trailing Whitespace
Tabs, Line Breaks, and Duplicate Whitespace
All Whitespace
Letters
Numbers
Punctuation
Modify Case
Upper Case
*New Workflow1
Run
Crime_Data_from_2020_to_Present_2025112...csv
Results - Data Cleansing (3) - Input2
0 of 0 Fields
Cell Viewer
Data
Metadata
Actions
No data available. Use Ctrl+R to run the workflow.
40°F
Cloudy
Search
9:47 AM
11/23/2025


Alteryx Designer x64 - Crime_Data_Profiling.yxmd
Field Summary
Crm Cd 3
LON
Crime_Data_Profiling.yxmd
Run
Results - Browse (5) - Input
Text - View Browse Tool Report Tab
Activate Windows
Go to Settings to activate Windows.

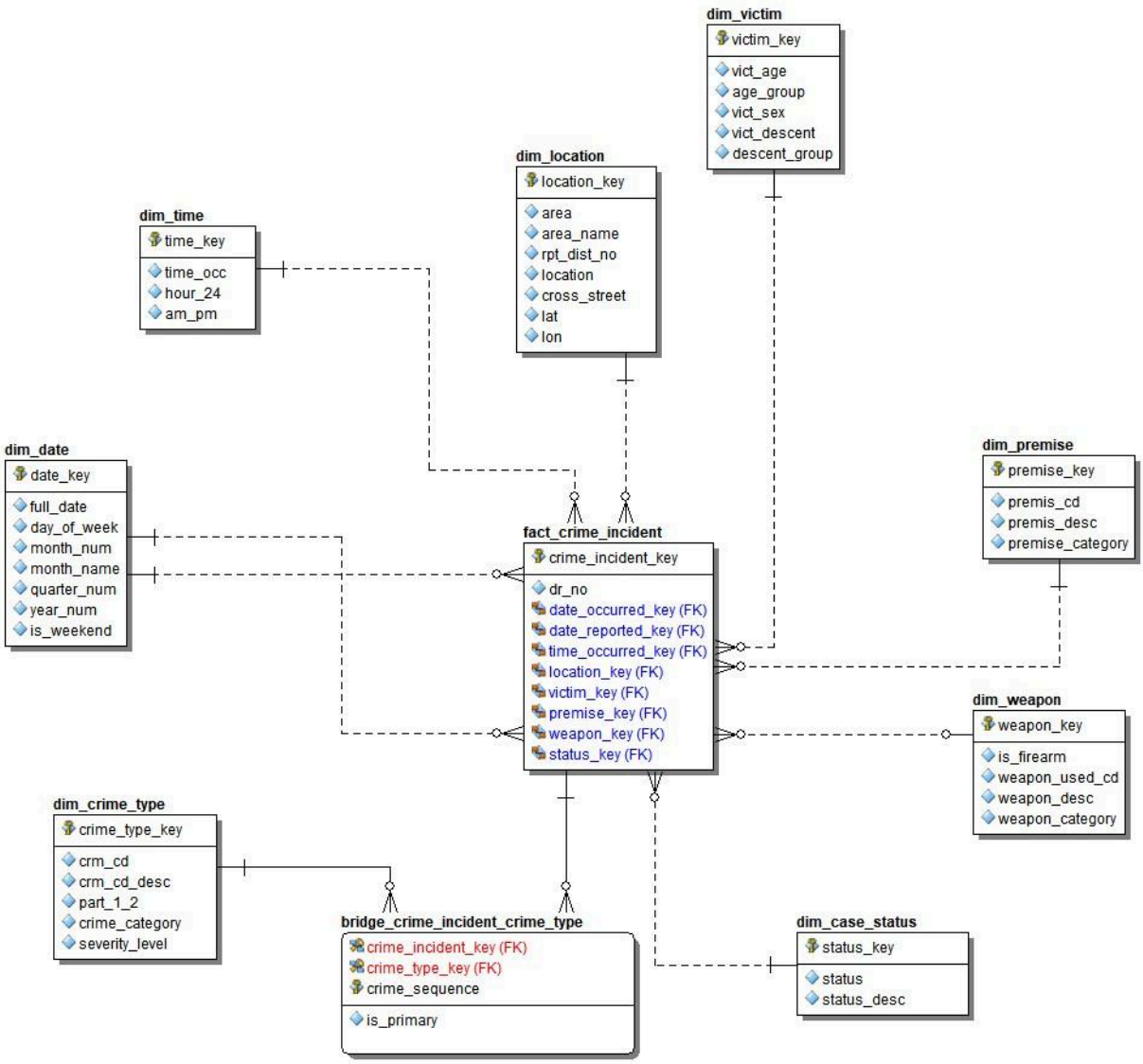

Alteryx Designer x64 - Crime_Data_Profiling.yxmd
Field Summary
Crm Cd
DR_NO
Crime_Data_Profiling.yxmd
Run
Results - Browse (5) - Input
Text - View Browse Tool Report Tab
Activate Windows
Go to Settings to activate Windows.


Alteryx Designer x64 - Crime_Data_Profiling.yxmd
Field Summary
Rpt Dist No
Crm Cd 4
Vict Age
Crime_Data_Profiling.yxmd
Run
Results - Browse (5) - Input
Text - View Browse Tool Report Tab
Activate Windows
Go to Settings to activate Windows.


Alteryx Designer x64 - Crime_Data_Profiling.yxmd
Crime_Data_Profiling.yxmd
Results - Formula (6) - Output


Alteryx Designer x64 - Crime_Data_Profiling.yxmd
Summarize (7) - Configuration
Crime_Data_Profiling.yxmd
Results - Summarize (7) - Output
Activate Windows
Go to Settings to activate Windows.

Data Modelling



Professor Feedback & Model Revisions

1. Consolidate Unnecessary Dimensions

Professor Feedback: "The reporting district dimension is unnecessary complexity. It should be consolidated into the location dimension since reporting districts are subsets of geographic areas."

Change Made:

- **Removed:** Separate DIM_REPORTING_DISTRICT table
- **Consolidated into:** DIM_LOCATION which now includes area, area_name, rpt_dist_no, location, cross_street, lat, lon

Rationale: Reporting districts are always part of a larger geographic area. Having a separate dimension adds unnecessary joins without providing analytical value. Users naturally think of crime locations hierarchically (Area → Reporting District → Specific Location).

---

2. Simplify Date Hierarchy

Professor Feedback: "You don't need separate dimension tables for date components. A single date dimension with attributes is the standard approach and sufficient for all temporal analysis."

Change Made:

- **Removed:** Multiple date hierarchy tables
- **Created:** Single DIM_DATE with all temporal attributes: full_date, day_of_week, month_num, month_name, quarter_num, year_num, is_weekend

Rationale: A single date dimension is industry standard (Kimball methodology). All date-related queries can use this one dimension. Separate tables for month, quarter, year create unnecessary complexity and slower queries.

---

3. Add Bridge Table for Multiple Crimes

Professor Feedback: "Your fact table has multiple crime code columns (CRM_CD_1 through CRM_CD_4) which violates normalization principles. Use a bridge table to properly model the many-to-many relationship between incidents and crime types."

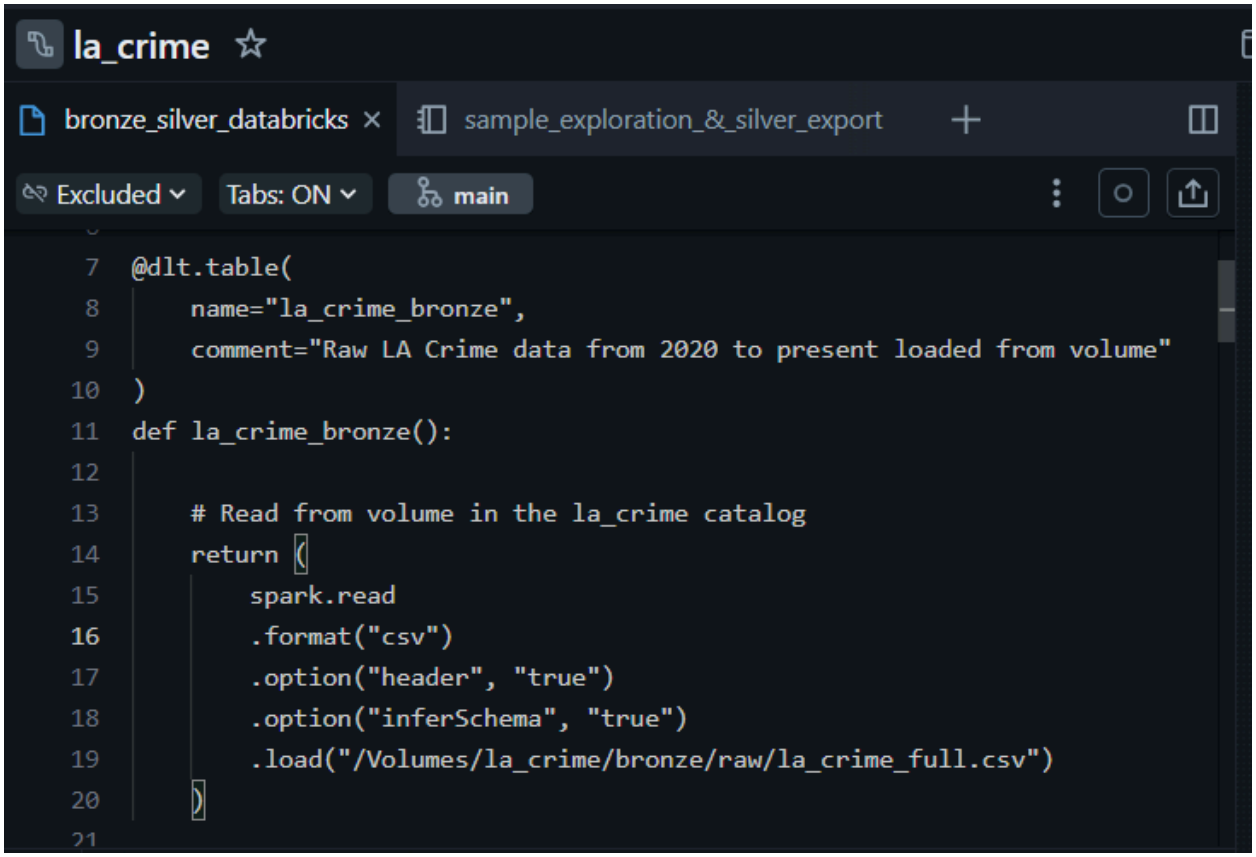
Change Made:

- **Added:** BRIDGE_CRIME_INCIDENT_CRIME_TYPE table
- **Attributes:** crime_incident_key (FK), crime_type_key (FK), crime_sequence, is_primary
- **Removed:** Multiple crime code columns from fact table (kept only primary crime reference)

Rationale: Bridge tables are the proper way to handle many-to-many relationships in dimensional models. This allows an incident to have multiple associated crimes while maintaining referential integrity. The crime_sequence indicates order, and is_primary flags the main crime.

Medallion Architecture

Bronze

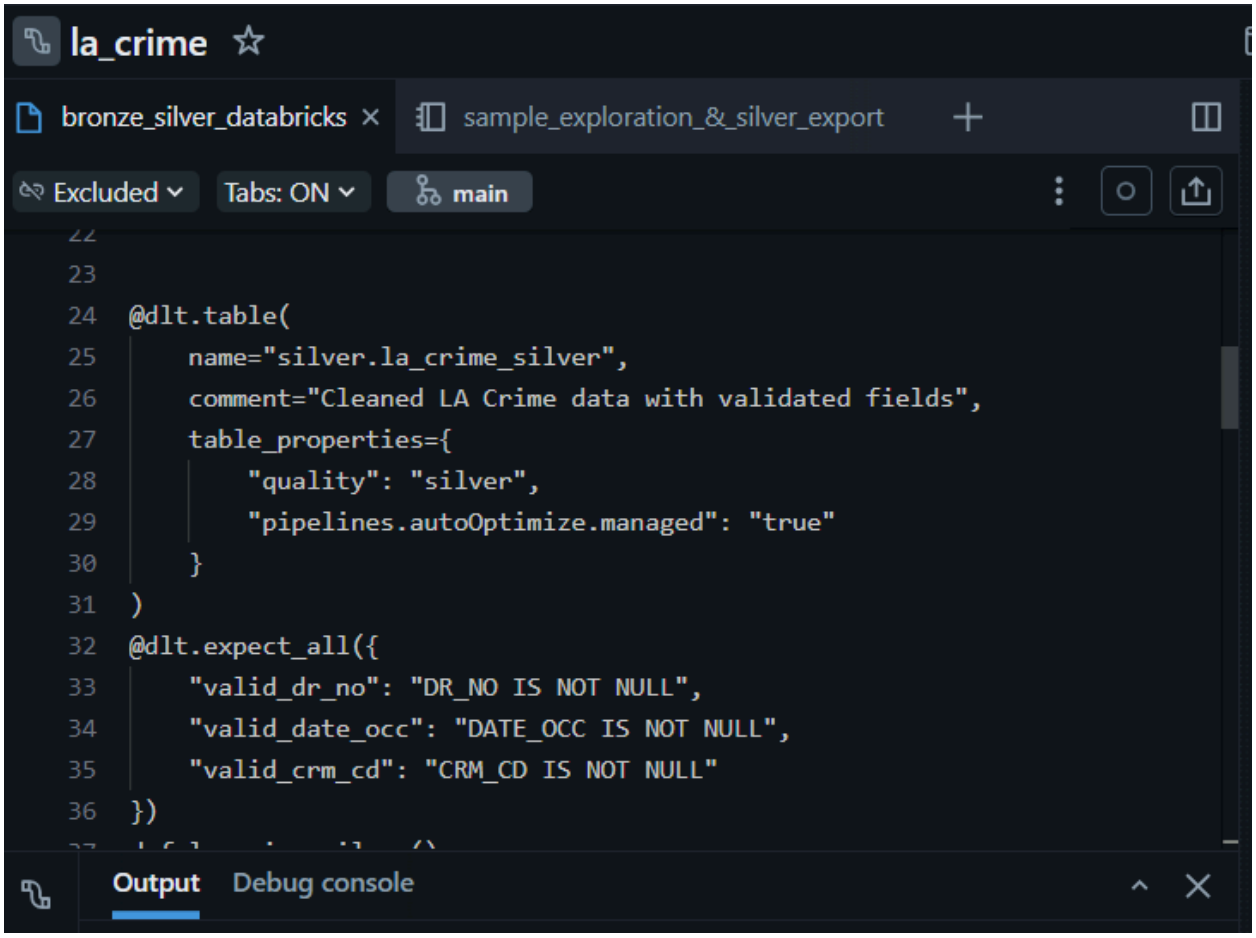


Purpose:
The bronze layer is designed to ingest the **raw LA crime dataset** from 2020 to the present, preserving the source data exactly as it exists. This layer acts as the **foundation for all downstream transformations** and analytics.

Process / ETL Steps:

1. **Ingestion:**
    - The CSV files are loaded from `/Volumes/la_crime/bronze/raw/la_crime_full.csv` using Spark.
    - Schema is inferred automatically (`.option("inferSchema", "true")`).
    - Headers are preserved to maintain column names.
2. **No transformations applied:**
    - All data is retained in its original form, including whitespaces, inconsistent casing, and invalid or missing values.
    - This ensures **traceability** and a historical record of all ingested raw data.
3. **Usage:**
    - Serves as a source for the **silver layer**, where data cleaning and validation occurs.

SILVER



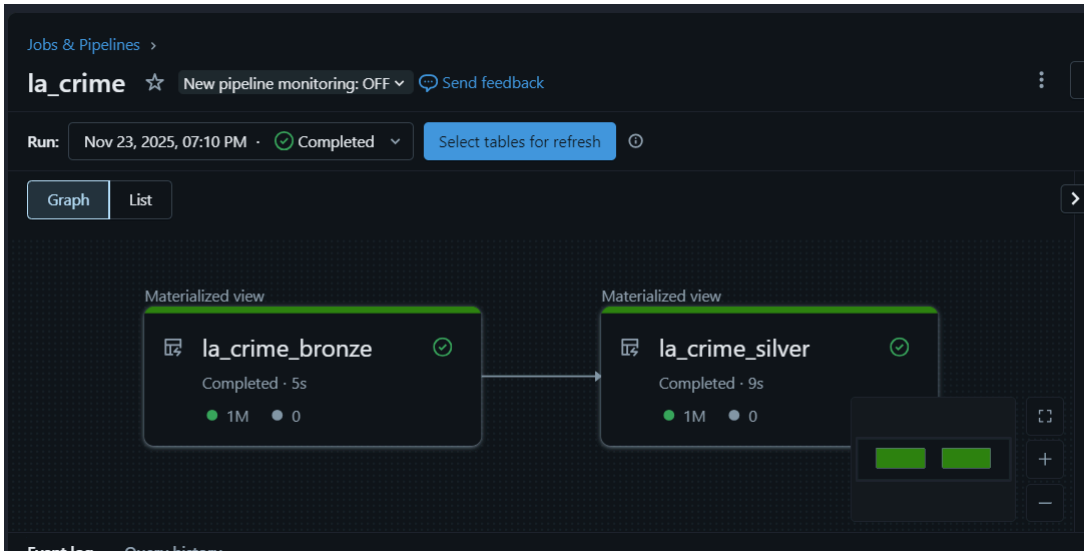
Purpose:
The silver layer transforms raw data into a **clean, validated, and standardized dataset** that can be reliably used for reporting and analytics. It ensures **data quality** while preserving the integrity of the original data.

Process / ETL Steps:

1. **Read Bronze Data:**
    - The silver table reads directly from `la_crime_bronze`.
2. **Field Cleaning and Standardization:**
    - **Trim** leading/trailing spaces from string fields (`DR_NO`, `CRM_CD_DESC`, `MOCODES`, `PREMIS_DESC`, `WEAPON_DESC`, `LOCATION`, `CROSS_STREET`).

- **Uppercase** key categorical fields (`AREA_NAME`, `VICT_SEX`, `VICT_DESCENT`, `STATUS`).
- **Normalize spacing** in the `LOCATION` field using regex to replace multiple spaces with a single space.

3. **Validation of Critical Fields:**
   - Ensure `DR_NO`, `DATE_OCC`, and `CRM_CD` are not null (`@dlt.expect_all`).
   - Remove records missing these essential identifiers.
4. **Data Type and Value Corrections:**
   - `VICT_AGE` is set to NULL if the value is non-positive.
   - Latitude and longitude are set to NULL if both are zero, preventing invalid coordinates.
5. **Additional Quality Controls:**
   - Automatically managed table optimization (`pipelines.autoOptimize.managed=true`) ensures **performance improvements** for queries.
   - Silver layer acts as a **trusted dataset**, ready for consumption by gold-layer aggregations and analytics.



GOLD

The CSV exported from Databricks Silver is uploaded to a Snowflake table stage using the `PUT` command (`PUT file:///path/to/la_crime_silver.csv @%LA_CRIME_SILVER;`) before loading it into the `SILVER.LA_CRIME_SILVER` table for Gold-layer processing.

The **Gold layer** represents the **curated, analytics-ready dimensional model** of LA crime data, derived from the Silver layer. It follows a **star schema** design with fact and dimension tables, suitable for **BI reporting, dashboards, and analytics**.

This Python Snowflake notebook builds **simplified dimension and fact tables** to support streamlined analysis while maintaining traceability to the Silver data.

Process / ETL Steps

1. Read Silver Layer

- Data is read from `SILVER.LA_CRIME_SILVER` using Snowpark.
- Key validation checks and initial record counts are performed.

2. Dimension Tables Creation

- **DIM_DATE**: Contains unique dates for `DATE_OCC` and `DATE_RPTD`. Derived columns include:
    - `DATE_KEY` (YYYYMMDD as integer)
    - `FULL_DATE`
    - `DAY_OF_WEEK`, `MONTH_NUM`, `QUARTER_NUM`, `YEAR_NUM`
    - Duplicate dates from reporting and occurrence dates are merged.
- **DIM_TIME**: Simplifies `TIME_OCC` into:
    - `TIME_KEY` (integer)
    - `HOUR_24`, `MINUTE`, `AM_PM`
- **DIM_LOCATION**: Aggregates by `AREA` and `RPT_DIST_NO`:
    - Location name, cross street, latitude, longitude
    - Surrogate `LOCATION_KEY` assigned with `row_number()`
- **DIM_CRIME_TYPE**: Maps `CRM_CD` and description, and categorizes crimes:
    - `CRIME_CATEGORY` (e.g., Theft, Burglary)
    - `SEVERITY_LEVEL` (High, Medium, Low)
    - Surrogate `CRIME_TYPE_KEY`
- **DIM_VICTIM**: Standardizes victim attributes:
    - `AGE_GROUP` buckets
    - `DESCENT_GROUP` mapping (e.g., Asian, Hispanic)
    - Surrogate `VICTIM_KEY`

- **DIM_PREMISE**: Standardizes premises descriptions and categories:
    - Surrogate `PREMISE_KEY`
- **DIM_WEAPON**: Categorizes weapons and flags firearms:
    - Surrogate `WEAPON_KEY`
- **DIM_CASE_STATUS**: Unique case statuses with `STATUS_KEY`.

---

3. Fact Table Creation

- **FACT_CRIME_INCIDENT**: Core fact table linking:
    - `DR_NO`, `DATE_OCC`, `DATE_RPTD`, `TIME_OCC`, `AREA`, `VICTIM`, `PREMISE`, `WEAPON`, `STATUS`
- Joins with dimension tables to replace raw attributes with **surrogate keys**.
- Fact table uses `row_number()` to generate `CRIME_INCIDENT_KEY`.

---

4. Bridge Table

- **BRIDGE_CRIME_INCIDENT_CRIME_TYPE** handles **multiple crimes per incident**:
    - Maps `DR_NO` to multiple `CRM_CD` fields (`CRM_CD`, `CRM_CD_1…4`)
    - Assigns sequence and primary/secondary flags
    - Surrogate keys from `DIM_CRIME_TYPE`

```
39
40  show tables;
```

Results | Chart

	created_on	name	database_name	schema_name	rows	bytes
1	2025-11-23 21:16:36.190 -080(	BRIDGE_CRIME_INCIDENT_CRIME_TYPE	LA_CRIME	GOLD	2023753	7951872
2	2025-11-23 21:16:24.997 -080(	DIM_CASE_STATUS	LA_CRIME	GOLD	6	9216
3	2025-11-23 21:16:21.814 -080(	DIM_CRIME_TYPE	LA_CRIME	GOLD	140	24064
4	2025-11-23 21:16:19.187 -080(	DIM_DATE	LA_CRIME	GOLD	1901	29184
5	2025-11-23 21:16:20.958 -080(	DIM_LOCATION	LA_CRIME	GOLD	1210	53248
6	2025-11-23 21:16:23.474 -080(	DIM_PREMISE	LA_CRIME	GOLD	314	21504
7	2025-11-23 21:16:20.164 -080(	DIM_TIME	LA_CRIME	GOLD	1439	24576
8	2025-11-23 21:16:22.631 -080(	DIM_VICTIM	LA_CRIME	GOLD	2937	33280
9	2025-11-23 21:16:24.225 -080(	DIM_WEAPON	LA_CRIME	GOLD	80	18432