

Assignment 10 - LA Crime Analytics

Data Source

Source Information

Attribute	Details
Source	Los Angeles Open Data Portal
Dataset	Crime Data from 2020 to Present
API Endpoint	https://data.lacity.org/resource/2nrs-mtv8.csv
Format	CSV
Update Frequency	Daily (live data)
Download Date	November 24, 2025
File Size	243.65 MB

Dataset Characteristics

Metric	Value
Total Records	1,004,991
Total Columns	28
Date Range	January 1, 2020 → May 29, 2025
Time Span	1,975 days (5.4 years)
Grain	One row per crime incident
Primary Key	<code>dr_no</code> (Division of Records Number)
Storage Location	Unity Catalog Volume: <code>workspace.la_crime_schema.datastore</code>

Databricks Environment Setup

Catalog: workspace

The screenshot shows the Databricks Catalog Explorer interface. On the left sidebar, under the Catalog section, the 'workspace' schema is selected. The main area displays a table of schemas with columns for Name, Owner, and Created at. There are 11 schemas listed:

Name	Owner	Created at
cdf_schema	vedant12mane@gmail.com	Nov 08, 2025, 07:08 PM
damg7370	vedant12mane@gmail.com	Oct 21, 2025, 06:49 PM
default	_workspace_admins_workspace_2803...	Sep 15, 2025, 08:48 PM
dlt_schema	vedant12mane@gmail.com	Nov 04, 2025, 09:18 PM
fi_dc_schema	vedant12mane@gmail.com	Nov 12, 2025, 10:31 PM
information_schema	System user	Sep 15, 2025, 08:48 PM
la_crime_schema	vedant12mane@gmail.com	Nov 24, 2025, 12:30 AM
medallion_architecture	vedant12mane@gmail.com	Nov 04, 2025, 05:11 PM
mmd_schema	vedant12mane@gmail.com	Nov 10, 2025, 06:07 PM
sd_schema	vedant12mane@gmail.com	Nov 17, 2025, 12:43 AM
spl_schema	vedant12mane@gmail.com	Nov 09, 2025, 10:51 PM

Schema: workspace.la_crime_schema

The screenshot shows the Databricks Catalog Explorer interface. On the left sidebar, under the Catalog section, the 'la_crime_schema' schema is selected. The main area displays the schema details. Under the 'Volumes' tab, there is one volume named 'datastore' owned by 'vedant12mane@gmail.com' created on Nov 24, 2025, 12:30 AM.

Name	Owner	Created at
datastore	vedant12mane@gmail.com	Nov 24, 2025, 12:30 AM

Volume: workspace.la_crime_schema.datastore

The screenshot shows the Databricks web interface. On the left, the sidebar includes sections like Workspace, Recents, Catalog (which is currently selected), Jobs & Pipelines, Compute, Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, AI/ML, Playground, Experiments, Features, Models, and Serving. The main area is titled 'Catalog Explorer > workspace > la_crime_schema > datastore'. It has tabs for Overview, Files, Details, and Permissions. Under Overview, there's a search bar, a 'Share' button, and a 'Upload to this volume' button. The 'About this volume' section shows it's owned by 'vedant12mane@gmail.com'. The 'Tags' section has a 'Add tags' button. The 'Files' section shows a table with columns Name, Size, and Last modified, which is currently empty. A 'Create directory' button is also present.

Profiling Methodology

Tools Used

- Platform:** Databricks Community Edition
- Compute:** Serverless
- Language:** PySpark (Python on Spark)
- Notebook:** [datap-profiling.ipynb](#)

Profiling Approach

We conducted **8 comprehensive profiling sections**:

1. Column-Level Analysis

- Null counts and percentages
- Distinct value counts (cardinality)
- Data type validation

2. Temporal Data Analysis

- Date range validation
- Date logic validation (report date \geq occurrence date)
- Reporting lag analysis
- Time format validation

3. Geographic Data Analysis

- Coordinate bounds validation (LA County)
- Area code validation (21 LAPD areas)
- Location field quality

4. Demographic Data Analysis

- Victim age validation
- Sex/gender distribution
- Descent/ethnicity validation

5. Crime Classification Analysis

- Crime code completeness
- Premise type distribution
- Weapon usage patterns
- Case status distribution

6. Data Quality Scoring

- Overall completeness calculation
- Quality metrics by category
- Severity classification

7. Issue Identification

- Critical issues requiring immediate action
- Medium priority issues
- Low priority issues

8. Recommendations

- Silver layer transformation rules
- Data cleaning strategies
- Data enrichment opportunities

Completeness Analysis

Null Analysis

Perfect Completeness (0% Null) - 18 Columns

Column	Null %	Completeness
dr_no	0%	100% ✓
date_rptd	0%	100% ✓
date_occ	0%	100% ✓
time_occ	0%	100% ✓
area	0%	100% ✓
area_name	0%	100% ✓
rpt_dist_no	0%	100% ✓
part_1_2	0%	100% ✓
crm_cd	0%	100% ✓
crm_cd_desc	0%	100% ✓
vict_age	0%	100% ✓
premis_cd	0%	100% ✓
status	0%	100% ✓
status_desc	0%	100% ✓
crm_cd_1	0%	100% ✓
location	0%	100% ✓
lat	0%	100% ✓

Column	Null %	Completeness
lon	0%	100% <input checked="" type="checkbox"/>

High Null Percentage - Expected (6 Columns)

Column	Null %	Reason	Action
crm_cd_4	99.99%	Most crimes have only 1-2 codes	<input checked="" type="checkbox"/> Expected - Keep as-is
crm_cd_3	99.77%	Most crimes have only 1-2 codes	<input checked="" type="checkbox"/> Expected - Keep as-is
crm_cd_2	93.12%	Most crimes have single classification	<input checked="" type="checkbox"/> Expected - Keep as-is
cross_street	84.65%	Not always recorded	<input checked="" type="checkbox"/> Expected - Keep as-is
weapon_used_cd	67.44%	Most crimes don't involve weapons	<input checked="" type="checkbox"/> Expected - Keep as-is
weapon_desc	67.44%	Most crimes don't involve weapons	<input checked="" type="checkbox"/> Expected - Keep as-is

Medium Null Percentage (3 Columns)

Column	Null %	Reason	Action
mocodes	15.09%	Modus operandi not always applicable	Handle as Unknown
vict_sex	14.39%	Institutional/business victims	Handle as Unknown
vict_descent	14.39%	Institutional/business victims	Handle as Unknown

Cardinality Analysis

High Cardinality (Identifiers)

Column	Distinct Values	Cardinality %	Category
dr_no	1,004,991	100%	Primary Key <input checked="" type="checkbox"/>
mocodes	310,941	30.94%	High cardinality dimension
location	66,566	6.62%	Medium-high (addresses)

Low Cardinality (Good for Dimensions)

Column	Distinct Values	Use Case
area	21	dim_location
area_name	21	dim_location
vict_descent	21	dim_victim_demographics
vict_sex	6	dim_victim_demographics
status	7	dim_status
status_desc	6	dim_status
part_1_2	2	Crime classification
crm_cd	140	dim_crime_type
premis_cd	315	dim_premise
weapon_used_cd	80	dim_weapon
vict_age	104	dim_victim_demographics

Temporal Data Quality

Date Ranges

Field	Earliest	Latest	Span
date_occ (Occurred)	2020-01-01	2025-05-29	1,975 days
date_rptd (Reported)	2020-01-01	2025-06-04	1,980 days

Date Logic Validation

Check	Result	Status
Report date ≥ Occurrence date	0 violations	✓ Pass
Future dates	0 records	✓ Pass
Dates before 2020	0 records	✓ Pass

Reporting Lag Analysis

Metric	Value
Minimum	0 days (same day)
Maximum	1,862 days (~5 years)
Average	12.18 days
Median	1 day
90th Percentile	13 days

Distribution

- **48% reported same day** (482,066 crimes)
- **22% reported next day** (222,401 crimes)
- **70% reported within 7 days**

Insight: Most crimes are reported quickly (median 1 day), but some cases like identity theft are reported much later (cold cases up to 5 years).

Time Validation Results

Check	Result	Status
Invalid times (outside 0000-2359)	0	✓ Pass
Invalid hours (>23)	0	✓ Pass
Invalid minutes (>59)	0	✓ Pass

Peak Crime Hours:

Hour	Crime Count	Period
12:00 PM	67,813	Afternoon Peak
6:00 PM	59,958	Evening Rush
5:00 PM	58,811	Evening Rush

Lowest Crime Hours:

Hour	Crime Count	Period
4:00 AM	18,757	Night
5:00 AM	17,290	Early Morning

Pattern: Crime follows a bell curve - lowest in early morning hours (4-5 AM), peaks at midday (12 PM) and evening rush (5-6 PM).

Geographic Data Quality

Expected LA County Bounds:

- Latitude: 33.7°N to 34.3°N
- Longitude: -118.7°W to -118.1°W

Validation Results

Issue	Count	Percentage	Severity
Zero coordinates (0, 0)	2,240	0.22%	Low
Out-of-bounds coordinates	11,610	1.16%	Low

Issue	Count	Percentage	Severity
Valid coordinates	991,141	98.62%	✓ Excellent

Geographic Distribution

LAPD Areas (21 total)

Top 5 High-Crime Areas:

Rank	Area	Crime Count	Percentage
1	Central	69,670	6.93%
2	77th Street	61,758	6.15%
3	Pacific	59,514	5.92%
4	Southwest	57,441	5.72%
5	Hollywood	52,429	5.22%

Bottom 5 Low-Crime Areas:

Rank	Area	Crime Count	Percentage
17	Foothill	33,133	3.30%
18	Hollenbeck	37,085	3.69%
19	Harbor	41,394	4.12%
20	Topanga	41,374	4.12%
21	Devonshire	41,756	4.15%

Reporting Districts

- Total Districts:** 1,210
- District Range:** 101 to 2199

Top 5 Reporting Districts by Crime:

District	Crime Count
162	5,403
1494	5,370
645	5,025
182	4,896
646	4,426

Demographic Data Quality

Victim Age Analysis

Age Statistics

Metric	Value
Minimum	-4 ⚠
Maximum	120
Average	28.92 years
Median	30 years

Age Quality Issues

Issue	Count	Percentage	Action Required
Negative ages	137	0.01%	Set to NULL in Silver
Zero ages (Unknown)	269,222	26.8%	Create "Unknown" age group

Issue	Count	Percentage	Action Required
Ages > 120	0	0%	None <input checked="" type="checkbox"/>

Age Group Distribution

Age Group	Count	Percentage
Unknown	269,222	26.8%
25-34	205,225	20.4%
35-44	162,147	16.1%
45-54	112,739	11.2%
18-24	94,952	9.4%
55-64	79,458	7.9%
65+ (Senior)	55,512	5.5%
0-17 (Juvenile)	25,735	2.6%

Insight: 26.8% unknown ages likely represent institutional victims (businesses, organizations) rather than individuals. Young adults (25-34) are the most common victims.

Victim Sex Distribution

Sex	Count	Percentage	Description
M (Male)	403,879	40.2%	Male victims
F (Female)	358,580	35.7%	Female victims
X (Unknown)	97,773	9.7%	Gender unknown
NULL	144,644	14.4%	Missing (institutional)
H (Non-binary)	114	0.01%	Non-binary
- (Dash)	1	0.00%	Unknown

Insight: Slight male victim majority (40% vs 36%). 14% missing sex data correlates with institutional victims.

Victim Descent/Ethnicity Distribution

Code	Descent	Count	Percentage
H	Hispanic/Latin/Mexican	296,404	29.5%
W	White	201,442	20.0%
B	Black	135,816	13.5%
X	Unknown	106,685	10.6%
NULL	Missing	144,656	14.4%
O	Other	78,005	7.8%
A	Other Asian	21,340	2.1%
K	Korean	5,990	0.6%
F	Filipino	4,838	0.5%
C	Chinese	4,631	0.5%

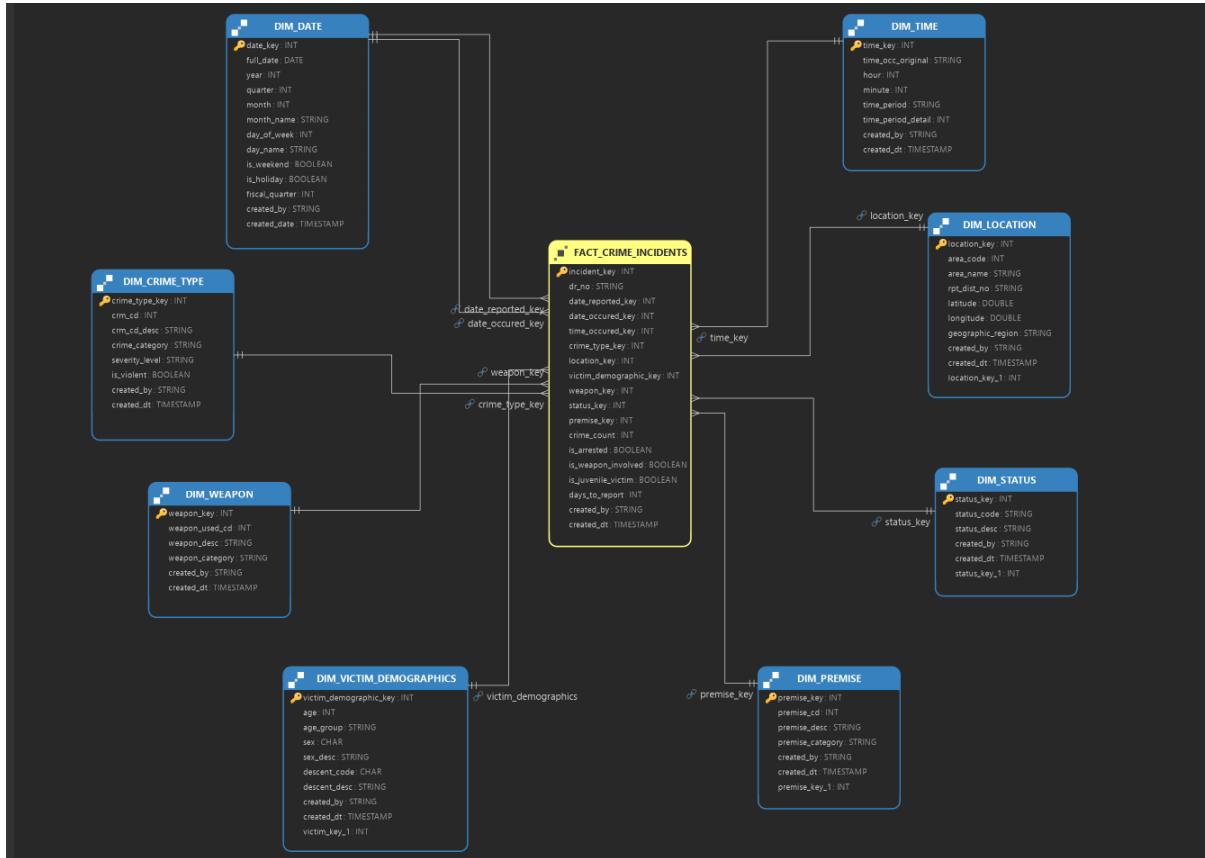
Insight: Distribution accurately reflects LA County demographics with Hispanic population being the largest group (29.5%).

Complete Descent Code Mapping:

A = Other Asian	K = Korean	U = Hawaiian
B = Black	L = Laotian	V = Vietnamese
C = Chinese	O = Other	W = White
D = Cambodian	P = Pacific Islander	X = Unknown
F = Filipino	S = Samoan	Z = Asian Indian

G = Guamanian I = American Indian - = Unknown
 H = Hispanic/Latin/Mexican J = Japanese

Dimensional Model



Databricks: ETL Pipeline

Bronze Layer - Raw Data Ingestion

Purpose

The Bronze layer serves as the **landing zone for raw data**, preserving the original data exactly as received from the source with minimal transformation.

Design Principles

- Immutable**: Raw data never modified, only appended
- Append-only**: Historical record of all data loads
- Minimal transformation**: Only add audit columns
- Schema preservation**: Keep original data types and structure
- Delta Lake**: Enable time travel and ACID transactions

Implementation Details

Table Name: `lacrime_incidents_bronze`

Source:

- Location:** Unity Catalog Volume `workspace.la_crime_schema.datastore`

- **File:** [crime_data_raw.csv](#) (243.65 MB)

- **Format:** CSV with header

Technology Stack:

- **Framework:** Delta Live Tables (DLT)
- **Ingestion Method:** Auto Loader (cloudFiles)
- **Table Type:** Streaming table
- **Storage Format:** Delta Lake

Schema - Bronze Layer

Column	Data Type	Description	Nullable
dr_no	Integer	Division of Records Number (Primary Key)	No
date_rptd	Timestamp	Date crime was reported	Yes
date_occ	Timestamp	Date crime occurred	Yes
time_occ	Integer	Time of occurrence (HHMM format)	Yes
area	Integer	LAPD Geographic Area (1-21)	Yes
area_name	String	Area name	Yes
rpt_dist_no	Integer	Reporting District	Yes
part_1_2	Integer	Crime classification (Part 1 or 2)	Yes
crm_cd	Integer	Crime code	Yes
crm_cd_desc	String	Crime description	Yes
mocodes	String	Modus Operandi codes	Yes
vict_age	Integer	Victim age	Yes
vict_sex	String	Victim sex (M/F/X/H/-)	Yes
vict_descent	String	Victim descent code (A-Z)	Yes
premis_cd	Double	Premise code	Yes
premis_desc	String	Premise description	Yes
weapon_used_cd	Double	Weapon code	Yes
weapon_desc	String	Weapon description	Yes
status	String	Case status code	Yes
status_desc	String	Case status description	Yes
crm_cd_1	Double	Crime code 1	Yes
crm_cd_2	Double	Crime code 2	Yes
crm_cd_3	Double	Crime code 3	Yes
crm_cd_4	Double	Crime code 4	Yes
location	String	Street address (rounded to 100 block)	Yes
cross_street	String	Cross street	Yes
lat	Double	Latitude	Yes
lon	Double	Longitude	Yes

Key Features

Auto Loader (cloudFiles)

Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage.

Benefits:

- **Incremental processing** - Only processes new/modified files
- **Schema inference** - Automatically detects column types

- **Schema evolution** - Handles schema changes gracefully
- **Scalability** - Efficiently processes billions of files
- **Exactly-once semantics** - No duplicates

Delta Lake Format:

All Bronze data is stored in **Delta Lake format**, providing:

- ACID transactions
- Time travel (view historical versions)
- Schema enforcement
- Efficient upserts and deletes
- Audit logging

```
# Bronze layer table
pl.create_streaming_table(
    name="lacrime_incidents_bronze",
    comment="Bronze layer: Raw LA crime data ingested from CSV files in Unity Catalog Volume")
```

dr_no	date_rptd	date_occ	time_occ	area	area_name	rpt_dist_no	
1	241514844	2024-10-24T00:00:00.000+00:00	2024-10-24T00:00:00.000+00:00	1245	16	N Hollywood	1526
2	240204007	2024-01-02T00:00:00.000+00:00	2024-01-01T00:00:00.000+00:00	2000	2	Rampart	218
3	241109524	2024-07-26T00:00:00.000+00:00	2024-07-25T00:00:00.000+00:00	2115	11	Northeast	1125
4	240605251	2024-01-21T00:00:00.000+00:00	2024-01-20T00:00:00.000+00:00	2315	6	Hollywood	627
5	241112706	2024-12-10T00:00:00.000+00:00	2024-12-08T00:00:00.000+00:00	1400	11	Northeast	1128
6	240317160	2024-12-21T00:00:00.000+00:00	2024-12-20T00:00:00.000+00:00	1756	3	Southwest	328
7	241105473	2024-01-28T00:00:00.000+00:00	2024-01-28T00:00:00.000+00:00	1455	11	Northeast	1162
8	240404593	2024-01-20T00:00:00.000+00:00	2024-01-19T00:00:00.000+00:00	900	4	Hollenbeck	407
9	241111589	2024-10-21T00:00:00.000+00:00	2024-10-17T00:00:00.000+00:00	1700	11	Northeast	1141
10	240211972	2024-07-16T00:00:00.000+00:00	2024-06-30T00:00:00.000+00:00	1900	2	Rampart	265

Bronze Layer Metrics

Metric	Value
Records Loaded	1,004,991
Load Duration	8 seconds
Data Volume	243.65 MB
Schema Version	1.0
Load Timestamp	2025-11-24 11:00:57
Data Retention	100% (no filtering)

Silver Layer - Data Cleaning & Enrichment

Purpose

The Silver layer provides a **validated, cleaned, and enriched** view of the data, serving as the foundation for analytics and ML workloads.

Key Principles:

- **Data quality** - Apply validation rules
- **Consistency** - Standardize formats
- **Enrichment** - Add derived fields
- **Performance** - Optimized for queries

Implementation Details

Table: lacrime_incidents_silver

Attribute	Value
Table Name	workspace.la_crime_schema.lacrime_incidents_silver
Table Type	Streaming Delta Live Table
Source	lacrime_incidents_bronze (streaming)
Records	1,004,900 (~99.99% retention)
Columns	48 (31 from Bronze + 17 enriched)
Quality Expectations	3 validation rules

Data Transformations

1. Data Cleaning (3 transformations)

Transformation 1.1: Clean Invalid Ages

```
vict_age_clean = CASE
    WHEN vict_age < 0 THEN NULL      # Fix negative ages (137 records)
    WHEN vict_age > 120 THEN NULL    # Fix impossible ages (0 records)
    ELSE vict_age
END
```

Impact: 137 records (0.01%) corrected

Transformation 1.2: Clean Invalid Coordinates

```
lat_clean = CASE
    WHEN lat = 0 AND lon = 0 THEN NULL  # Fix zero coordinates (2,240 records)
    ELSE lat
END

lon_clean = CASE
    WHEN lat = 0 AND lon = 0 THEN NULL
    ELSE lon
END
```

Impact: 2,240 records (0.22%) corrected

Transformation 1.3: Data Type Consistency

- All columns maintain correct data types from Bronze
- No type conversion issues detected

2. Data Enrichment (17 new columns)

Enrichment 2.1: Age Grouping

```
age_group = CASE
    WHEN vict_age_clean IS NULL OR vict_age_clean = 0 THEN "Unknown"
    WHEN vict_age_clean < 18 THEN "0-17 (Juvenile)"
    WHEN vict_age_clean < 25 THEN "18-24"
    WHEN vict_age_clean < 35 THEN "25-34"
```

```

WHEN vict_age_clean < 45 THEN "35-44"
WHEN vict_age_clean < 55 THEN "45-54"
WHEN vict_age_clean < 65 THEN "55-64"
ELSE "65+ (Senior)"
END

```

Business Value: Enables demographic segmentation for targeted crime prevention programs.

Enrichment 2.2: Time Period Classification

```

# Parse time_occ into hour and minute
hour = SUBSTRING(LPAD(time_occ, 4, '0'), 1, 2)::INT
minute = SUBSTRING(LPAD(time_occ, 4, '0'), 3, 2)::INT

# Create time periods
time_period = CASE
    WHEN hour BETWEEN 0 AND 5 THEN "Night"
    WHEN hour BETWEEN 6 AND 11 THEN "Morning"
    WHEN hour BETWEEN 12 AND 17 THEN "Afternoon"
    ELSE "Evening"
END

```

Business Value: Identifies temporal crime patterns for resource allocation (more patrols during peak times).

Enrichment 2.3: Reporting Lag

```
days_to_report = DATEDIFF(date_rptd, date_occ)
```

Business Value: Measures response time and identifies delayed crime reporting patterns.

Enrichment 2.4: Boolean Flags for Analysis

Flag	Definition	Business Value
is_weapon_involved	weapon_used_cd IS NOT NULL	Filter violent vs non-violent crimes
is_juvenile_victim	vict_age_clean < 18	Identify crimes against children
is_arrested	status IN ('AA', 'JA')	Calculate arrest rates

Complete List of Enriched Columns:

#	Column Name	Type	Description
1	vict_age_clean	int	Cleaned age (nulls for invalid)
2	lat_clean	double	Cleaned latitude
3	lon_clean	double	Cleaned longitude
4	age_group	string	Age grouping (Juvenile, 18-24, etc.)
5	time_str	string	Time as 4-digit string
6	hour	int	Hour of day (0-23)
7	minute	int	Minute of hour (0-59)
8	time_period	string	Night/Morning/Afternoon/Evening
9	days_to_report	int	Reporting lag in days
10	is_weapon_involved	boolean	Weapon used flag
11	is_juvenile_victim	boolean	Victim under 18 flag
12	is_arrested	boolean	Arrest made flag
13	silver_processed_datetime	timestamp	Processing timestamp
14	created_by	string	ETL flow identifier

Data Quality Expectations

DLT expectations are **declarative data quality rules** that automatically validate data.

Expectation 1: Valid Primary Key

```
expect_all_or_drop = {
    "valid_primary_key": "dr_no IS NOT NULL"
}
```

Expectation 2: Valid Date Logic

```
expect_all_or_drop = {
    "valid_date_logic": "date_rptd >= date_occ OR date_rptd IS NULL OR date_occ IS NULL"
}
```

Expectation 3: Valid Date Range

```
expect_all_or_drop = {
    "valid_date_range": "date_occ >= '2020-01-01'"
}
```

The screenshot shows the Databricks Pipeline Editor interface. The left sidebar displays the pipeline configuration with sections for Pipeline assets (data_transformation_flow, load_data_from_source, load_dimensions, load_fact) and Last runs. The main workspace shows a Python script titled 'Silver Layer Data Transformations' containing code for creating a streaming table and defining a DLT expectation. Below the script is a table view of the 'lacrime_incidents_silver' table, showing columns like dr_no, date_rptd, date_occ, time_occ, area, area_name, and rpt_dist_no, with 11 rows of sample data. A pipeline graph on the right visualizes the data flow between various stages.

Silver Layer Metrics

Metric	Value
Input Records	1,004,991 (from Bronze)
Output Records	1,004,900
Records Dropped	91 (0.01%) - failed expectations
Columns Added	17 enriched columns
Processing Duration	7 seconds
Data Retention	99.99%

Metric	Value
Quality Expectations Met	3/3 (100%) 

Gold Layer - Dimensional Model

Purpose

The Gold layer provides **business-ready, highly curated data** optimized for analytics, reporting, and machine learning.

Architecture: Star Schema

- 1 central **Fact Table** (measures and foreign keys)
- 8 **Dimension Tables** (descriptive attributes)

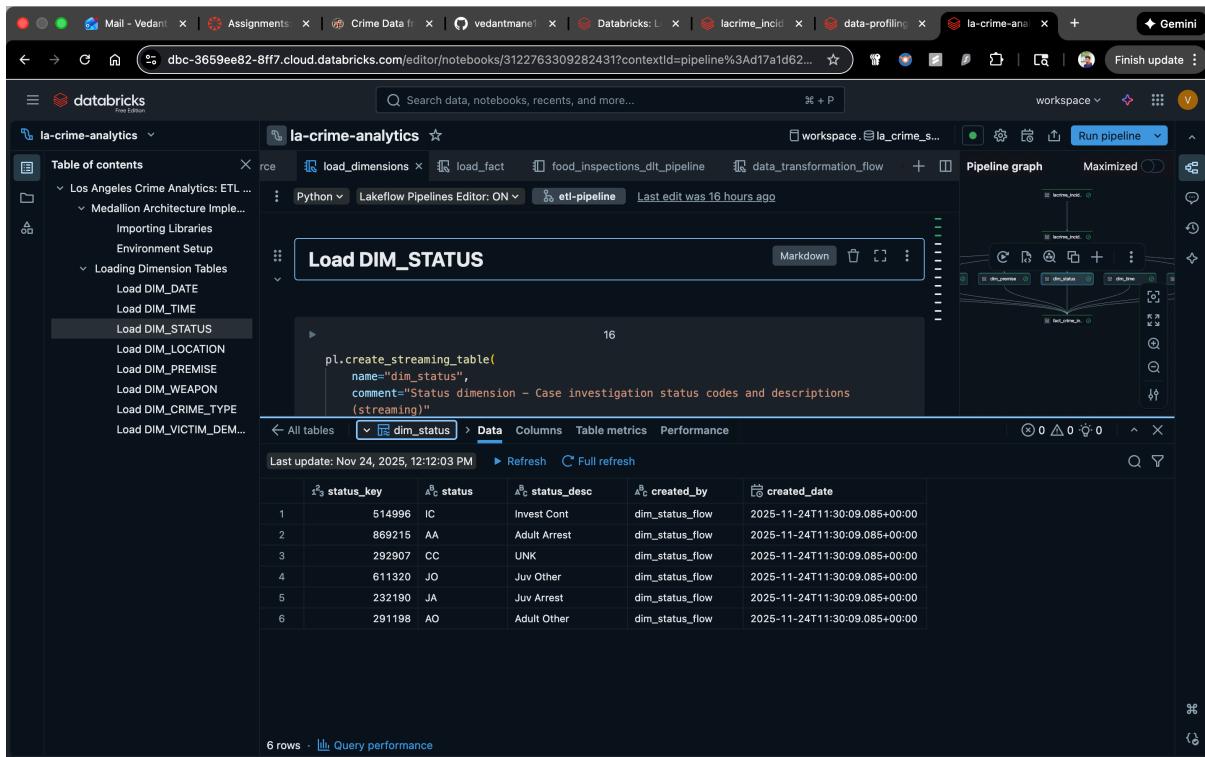
Benefits:

- **Query performance** - Optimized for BI tools
- **Business semantics** - Easy to understand
- **Denormalized** - Fast aggregations
- **Reusability** - Dimensions shared across facts

Dimensions

Dimension 1: dim_status

Purpose: Case investigation status lookup



```

pl.create_streaming_table(
    name="dim_status",
    comment="Status dimension - Case investigation status codes and descriptions (streaming)"
)
  
```

	status_key	status	status_desc	created_by	created_date
1	514996	IC	Invest Cont	dim_status_flow	2025-11-24T11:30:09.085+00:00
2	869215	AA	Adult Arrest	dim_status_flow	2025-11-24T11:30:09.085+00:00
3	292907	CC	UNK	dim_status_flow	2025-11-24T11:30:09.085+00:00
4	611320	JO	Juv Other	dim_status_flow	2025-11-24T11:30:09.085+00:00
5	232190	JA	Juv Arrest	dim_status_flow	2025-11-24T11:30:09.085+00:00
6	291198	AO	Adult Other	dim_status_flow	2025-11-24T11:30:09.085+00:00

Dimension 2: dim_location

Purpose: LAPD geographic areas and reporting districts

The screenshot shows a Databricks notebook titled "Load DIM_LOCATION". The code cell contains Python code to create a streaming table named "dim_location" with a comment about LAPD geographic areas and reporting districts. The resulting data table has columns: location_key, area, area_name, rpt_dist_no, geographic_region, created_by, and created_date. The data is truncated, showing rows 1 through 11.

location_key	area	area_name	rpt_dist_no	geographic_region	created_by	created_date
1	4470651	Rampart	292	Central	dim_location_flow	2025-11-24T11:30:09.816+00:00
2	458264	Hollenbeck	404	Central	dim_location_flow	2025-11-24T11:30:09.816+00:00
3	9421103	Wilshire	725	West	dim_location_flow	2025-11-24T11:30:09.816+00:00
4	139586	Southwest	374	South	dim_location_flow	2025-11-24T11:30:09.816+00:00
5	626326	Northeast	1122	Other	dim_location_flow	2025-11-24T11:30:09.816+00:00
6	7708985	Mission	1941	Other	dim_location_flow	2025-11-24T11:30:09.816+00:00
7	5091248	Olympic	2054	West	dim_location_flow	2025-11-24T11:30:09.816+00:00
8	6600082	Devonshire	1736	Valley	dim_location_flow	2025-11-24T11:30:09.816+00:00
9	5458339	Mission	1966	Other	dim_location_flow	2025-11-24T11:30:09.816+00:00
10	3700042	Rampart	242	Central	dim_location_flow	2025-11-24T11:30:09.816+00:00
11	9573225	Devonshire	1766	Valley	dim_location_flow	2025-11-24T11:30:09.816+00:00

Dimension 3: dim_time

Purpose: Time of day reference dimension

The screenshot shows a Databricks notebook titled "Load DIM_TIME". The code cell contains Python code to create a table named "dim_time" with a comment about it being a gold layer for time dimension. The resulting data table has columns: time_key, time_occ, hour, minute, minutes_since_midnight, time_24hr, time_12hr, and hour_12. The data is truncated, showing rows 1 through 11.

time_key	time_occ	hour	minute	minutes_since_midnight	time_24hr	time_12hr	hour_12
1	0	0	0	0	0 00:00	12:00 AM	
2	1	1	0	1	1 00:01	12:01 AM	
3	2	2	0	2	2 00:02	12:02 AM	
4	3	3	0	3	3 00:03	12:03 AM	
5	4	4	0	4	4 00:04	12:04 AM	
6	5	5	0	5	5 00:05	12:05 AM	
7	6	6	0	6	6 00:06	12:06 AM	
8	7	7	0	7	7 00:07	12:07 AM	
9	8	8	0	8	8 00:08	12:08 AM	
10	9	9	0	9	9 00:09	12:09 AM	
11	10	10	0	10	10 00:10	12:10 AM	

Dimension 4: dim_date

Purpose: Complete calendar dimension for temporal analysis

```

@dlt.table(
    name="dim_date",
    comment="Gold layer - Complete date dimension (2019-2029) with calendar attributes"
)
def dim_date():

```

date_key	full_date	year	year_name	quarter	quarter_name	year_quarter	month	m
1	20190101	2019-01-01	2019	1	Q1	2019-Q1	1	Janu
2	20190102	2019-01-02	2019	1	Q1	2019-Q1	1	Janu
3	20190103	2019-01-03	2019	1	Q1	2019-Q1	1	Janu
4	20190104	2019-01-04	2019	1	Q1	2019-Q1	1	Janu
5	20190105	2019-01-05	2019	1	Q1	2019-Q1	1	Janu
6	20190106	2019-01-06	2019	1	Q1	2019-Q1	1	Janu
7	20190107	2019-01-07	2019	1	Q1	2019-Q1	1	Janu
8	20190108	2019-01-08	2019	1	Q1	2019-Q1	1	Janu
9	20190109	2019-01-09	2019	1	Q1	2019-Q1	1	Janu
10	20190110	2019-01-10	2019	1	Q1	2019-Q1	1	Janu
11	20190111	2019-01-11	2019	1	Q1	2019-Q1	1	Janu

Dimension 5: dim_premise

Purpose: Location types where crimes occurred

```

pl.create_streaming_table(
    name="dim_premise",
    comment="Premise dimension - Location types where crimes occurred"
)

```

premise_key	premis_cd	premis_desc	premise_category	created_by	created_date
629586	124	BUS STOP	Other	dim_premise_flow	2025-11-24T11:30:07.
81582	253	MORTUARY	Other	dim_premise_flow	2025-11-24T11:30:07.
26492	140	BALCONY*	Other	dim_premise_flow	2025-11-24T11:30:07.
798112	106	TUNNEL	Other	dim_premise_flow	2025-11-24T11:30:07.
357876	247	CAR WASH	Vehicle	dim_premise_flow	2025-11-24T11:30:07.
393183	754	MUSEUM	Other	dim_premise_flow	2025-11-24T11:30:07.
542004	702	OFFICE BUILDING/OFFICE	Other	dim_premise_flow	2025-11-24T11:30:07.
541784	303	OIL REFINERY	Other	dim_premise_flow	2025-11-24T11:30:07.
407821	108	PARKING LOT	Street	dim_premise_flow	2025-11-24T11:30:07.
771770	301	GAS STATION	Other	dim_premise_flow	2025-11-24T11:30:07.
158839	144	GOLF COURSE*	Other	dim_premise_flow	2025-11-24T11:30:07.

Dimension 6: dim_victim_demographics

Purpose: Victim demographic profiles (age, sex, ethnicity)

Load DIM_VICTIM_DEMOGRAPHICS

```
pl.create_streaming_table(
    name="dim_victim_demographics",
    comment="Victim demographics dimension - Age groups, sex, and descent"
)
```

victim_demographic_key	vict_age_clean	age_group	vict_sex	sex_desc	vict_descent	descent_desc
1	527444	20	F	Female	B	Black
2	439898	41	M	Male	A	Other Asian
3	601241	35	M	Male	A	Other Asian
4	3806032	20	M	Male	J	Japanese
5	1509237	47	M	Male	X	Unknown
6	1724990	54	M	Male	C	Chinese
7	7913276	35	F	Female	C	Chinese
8	5695283	52	M	Male	X	Unknown
9	2366186	57	M	Male	F	Filipino
10	1892519	7	M	Male	O	Other
11	9546345	17	(Juvenile)	M	I	American Indian/Alaskan

Dimension 7: dim_weapon

Purpose: Weapons used in crimes (or no weapon)

Load DIM_WEAPON

```
pl.create_streaming_table(
    name="dim_weapon",
    comment="Weapon dimension - Weapons used in crimes (including No Weapon option)"
)
```

weapon_key	weapon_used_cd	weapon_desc	created_by	created_date
1	148597	311 HAMMER	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
2	58611	112 TOY GUN	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
3	485571	214 ICE PICK	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
4	927360	215 MACHETE	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
5	29750	513 STUN GUN	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
6	356858	104 SHOTGUN	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
7	29272	309 BOARD	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
8	328276	220 SYRINGE	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
9	190722	216 SCISSORS	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
10	26286	212 BOTTLE	dim_weapon_flow	2025-11-24T11:30:08.336+00:00
11	584873	307 VEHICLE	dim_weapon_flow	2025-11-24T11:30:08.336+00:00

Dimension 8: dim_crime_type

Purpose: Crime classification and categorization

Load DIM_CRIME_TYPE

```

pl.create_streaming_table(
    name="dim_crime_type",
    comment="Crime type dimension - Crime codes and classifications"
)

```

crime_type_key	crm_cd	crm_cd_desc	part_1_2	created_by	created_date
1	49732	763	STALKING	2	dim_crime_type_flow
2	481254	933	PROWLER	2	dim_crime_type_flow
3	208568	805	PIMPING	2	dim_crime_type_flow
4	102581	942	BRIBERY	2	dim_crime_type_flow
5	703599	623	BATTERY POLICE (SIMPLE)	2	dim_crime_type_flow
6	754215	762	LEWD CONDUCT	2	dim_crime_type_flow
7	943622	806	PANDERING	2	dim_crime_type_flow
8	56424	350	THEFT, PERSON	1	dim_crime_type_flow
9	52027	926	TRAIN WRECKING	1	dim_crime_type_flow
10	171230	330	BURGLARY FROM VEHICLE	1	dim_crime_type_flow
11	831140	944	CONSPIRACY	2	dim_crime_type_flow

Fact Table

Fact: fact_crime_incidents

Purpose: Central fact table connecting all dimensions with measures

fact_crime_incidents

```

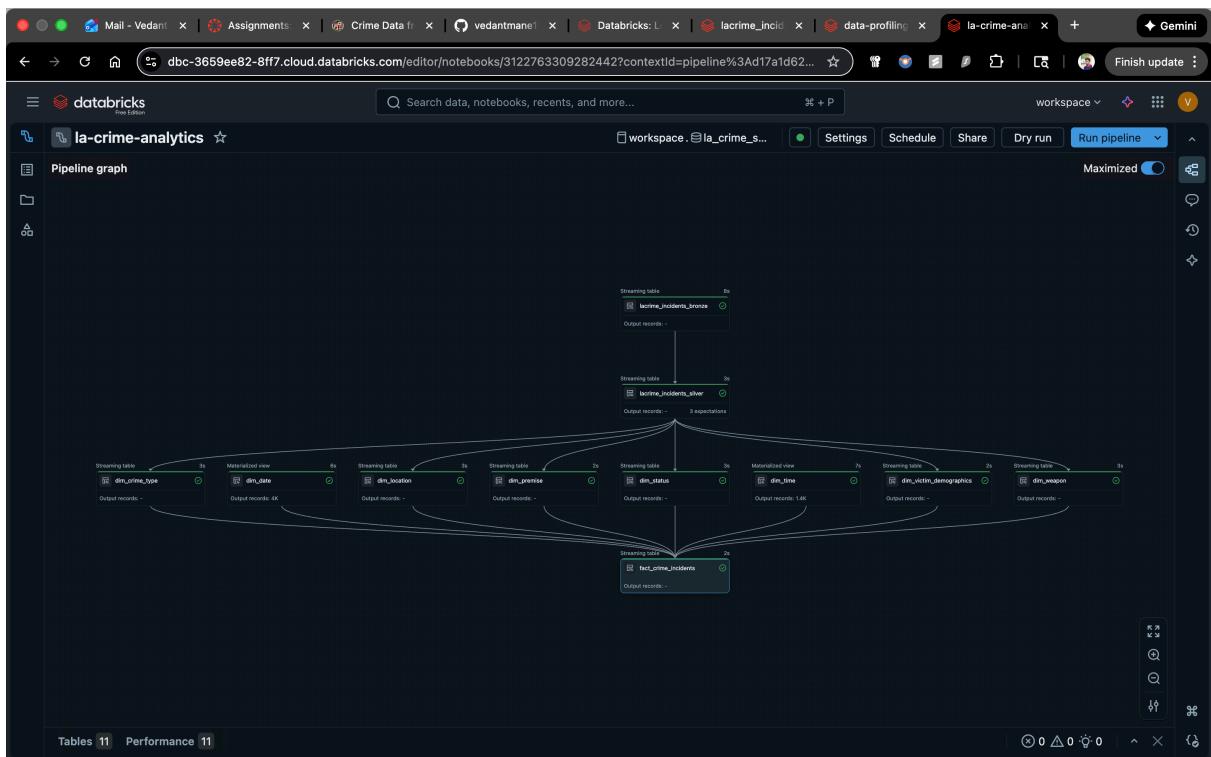
pl.create_streaming_table(
    name="fact_crime_incidents",
    comment="Fact table - Crime incidents with foreign keys to all dimensions and measures"
)

@pl.append_flow(
    target="fact_crime_incidents",
    name="fact_crime_incidents_flow",
    comment="Creates fact table by joining Silver with all dimension tables"
)

```

incident_key	dr_no	date_occurred_key	date_reported_key	time_occurred_key	location_key	crime_ty
1	1937455435	241514844	20241024	20241024	1245	3616091
2	1587023544	240204007	20240101	20240102	2000	4793673
3	489164771	241109524	20240726	20240726	2115	6050382
4	794410640	240605251	20240120	20240121	2315	5652167
5	1627758533	241112706	20241208	20241210	1400	2141804
6	1136274689	240317160	20241220	20241221	1756	5099954
7	1414105892	241105473	20240128	20240128	1455	4669274
8	1035859760	240404593	20240119	20240120	900	8088892
9	1454125872	241111589	20241017	20241021	1700	739328
10	1044251184	240211972	20240630	20240716	1900	9396098
11	954999042	240508062	20240417	20240418	800	3867699

ETL Pipeline Run



```
%sql
SELECT COUNT(*) FROM workspace.la_crime_schema.fact_crime_incidents

```

	COUNT(*)
1	1004991

1 row | 14.89s runtime

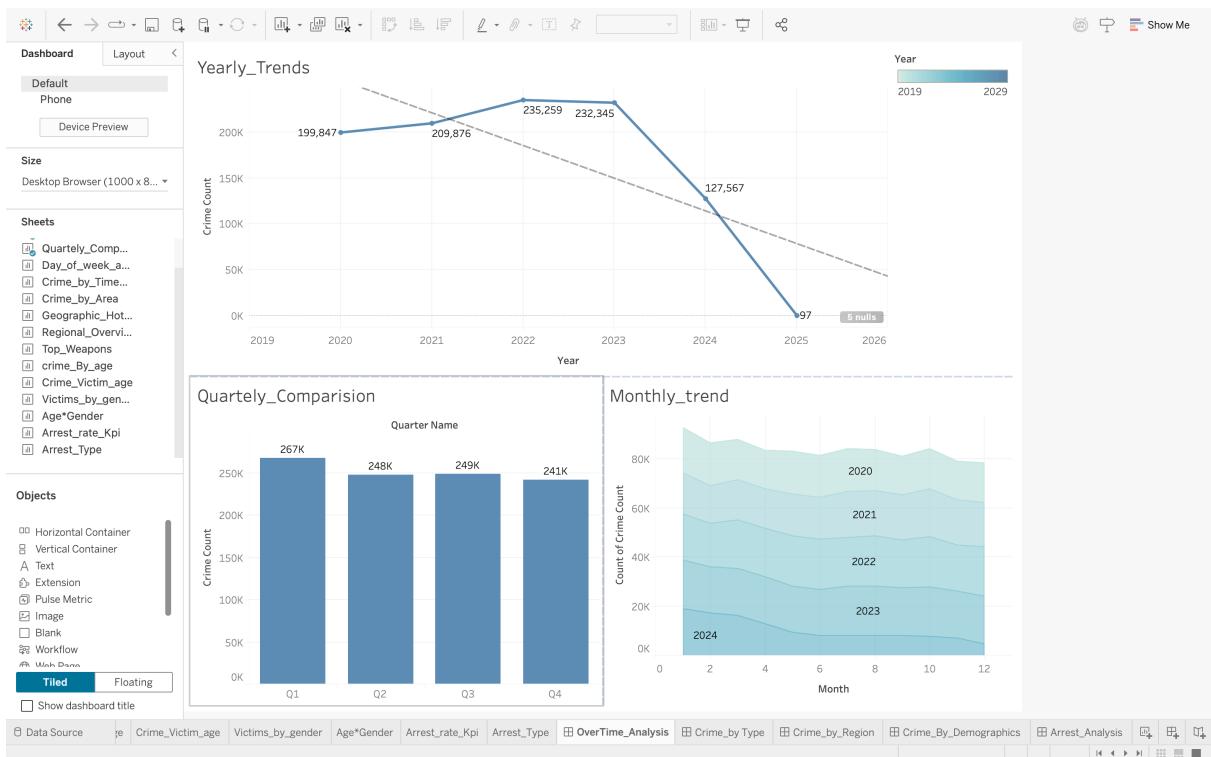
This result is stored as `_sqlid` and can be used in other Python and SQL cells.

Tableau Data Visualizations

Business Requirement 1

Crime Rates Over Time:

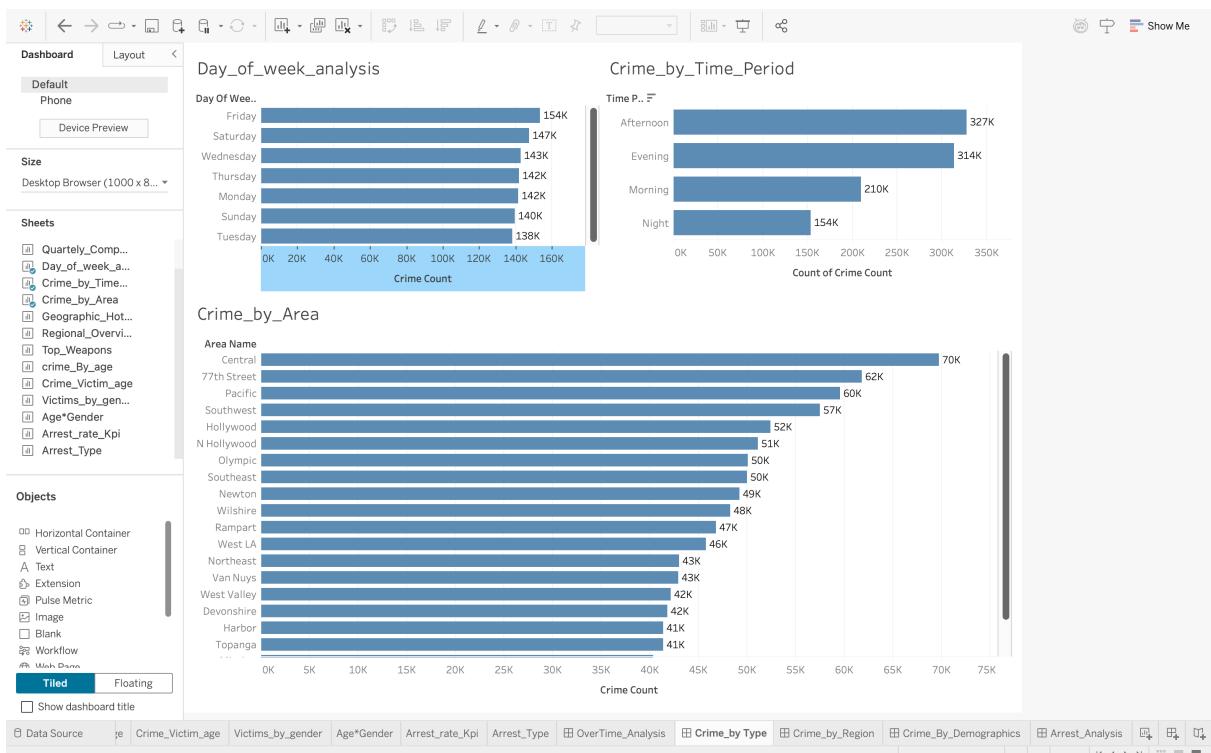
- What is the overall trend in crime rates over the years?
- How have crime rates changed on a monthly basis?
- How have crime rates changed on a quarterly basis?



Business Requirement 2

Day Time and Week Factors:

- Is there a correlation between the day of the week and the number of reported crimes?
- Do certain types of crimes tend to occur at specific times of the day?



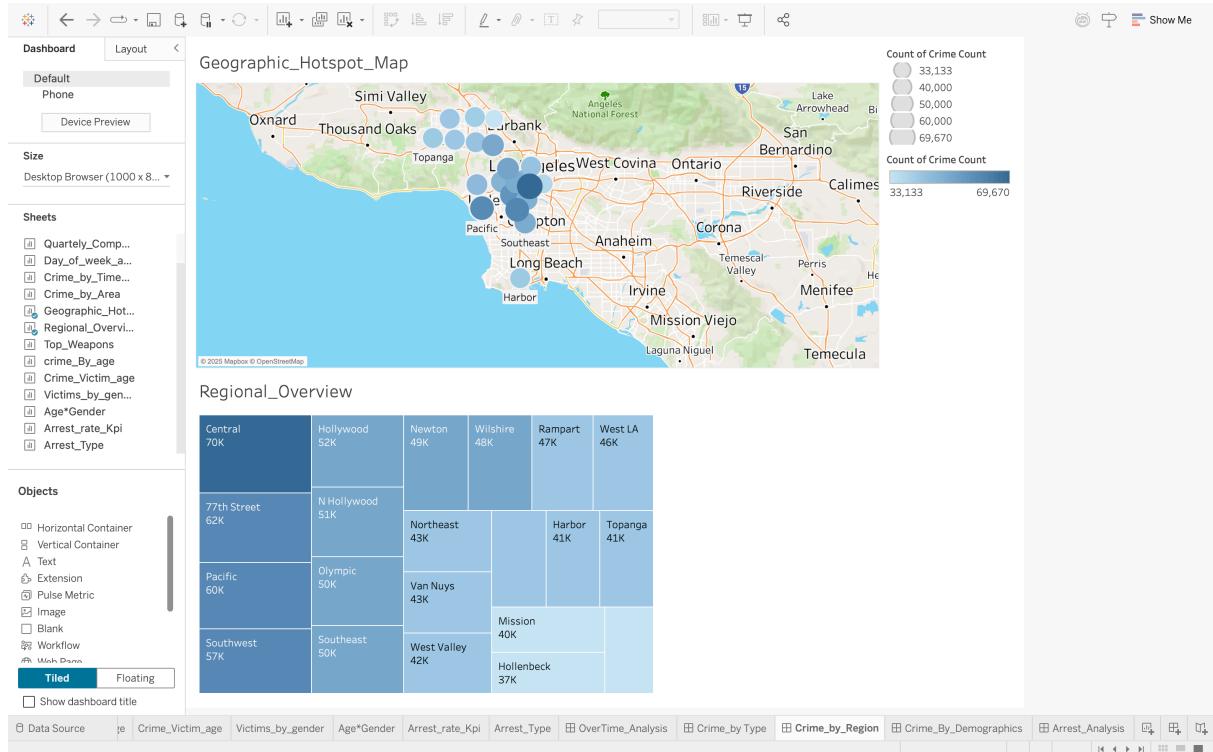
Business Requirement 3

Crime by Location:

- Where are the high-crime areas in Los Angeles?

We want to know (your inferences on):

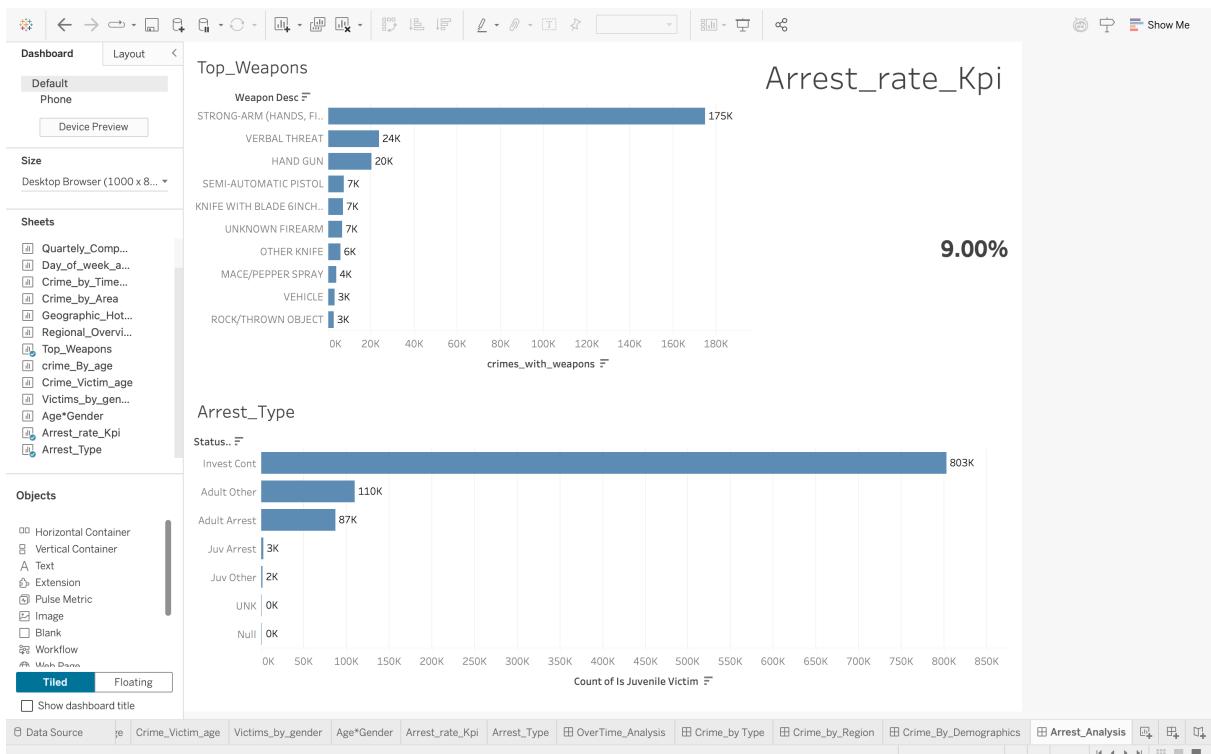
- By seeing the geo map, what are the crime hotspots?
- Can you identify areas where Los Angeles performs better or worse in terms of crime rates?



Business Requirement 4

Types of Crime:

- What are the most commonly used weapons in reported crimes?



Business Requirement 5

Demographic Analysis:

- Show in the visualization the age patterns in crime.
- Show in the visualization the gender-related patterns in crime.

Business Requirement 6

Arrests Ratio:

- What percentage of reported crimes result in Juvenile, Adult arrests?

