UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Integrated Analysis of Crime in Austin Using Multi-Source Data and SQL Storage

Data Management Course Report by:

- Samira Doshmankosh (946863)
- Bahareh Robaty Shirzad (946887)
- Setayesh Mohammadi Banadkooki (926072)

Abstract

This project investigates the relationship between crime distribution and population across different neighborhoods in the city of Austin, Texas. The analysis is based on the integration of multiple different data sources, including an open municipal court dataset provided by the City of Austin and several additional datasets collected through web scraping.

In particular, street-to-ZIP code mappings, neighborhood associations, and population statistics were extracted from public websites in order to enrich crime records with geographic and demographic information. A comprehensive data cleaning and normalization process was applied to address inconsistencies in street naming, missing values, and duplicate entries.

After performing correspondence investigation and schema integration, the final integrated dataset was stored in a relational SQLite database. Several SQL queries were executed to analyze crime frequency, crime rates per 1,000 inhabitants, and the distribution of offense types across neighborhoods.

The results highlight significant spatial differences in crime patterns and demonstrate how data integration techniques can enable meaningful analysis from initially disconnected datasets.

Keywords: Data Integration, Web Scraping, Crime Analysis, SQLite, Data Cleaning, Austin.

Table of Contents

Abstract
1. Group Components and Role
2. Introduction
    2.1 Research Questions
3. Data Sources
    3.1 Municipal Court Caseload Dataset
    3.2 Web Scraped Datasets
        3.2.1 Population by ZIP Code
        3.2.2 Street to ZIP Code Mapping
    3.3 Summary of Data Sources
4. Data Exploration and Cleaning
    4.1 Data Cleaning Procedures
    4.2 Correspondence Investigation
    4.3 Data Cleaning Procedures
    4.4 Impact of Cleaning on Integration Feasibility
5. Data Integration
    5.1 Integration Strategy
    5.2 Street-to-ZIP Integration
    5.3 ZIP-Based Enrichment with Neighborhood and Population
    5.4 Integration Results and Coverage
6. Storage and Querying
    6.1 Database Selection
    6.2 Database Schema
    6.3 Analysis Queries
        6.3.1 Total Crime Volume Q1
        6.3.2 Crime Distribution by Neighborhood Q2
        6.3.3 Population-Adjusted Crime Rates Q3
        6.3.4 High-Risk Streets Q4
        6.3.5 the most common types of offenses
        6.3.6 Temporal Crime Trends

7. Data Quality Evaluation
    7.1 Initial Quality Assessment (Before Cleaning and Integration)
    7.2 Identified Data Quality Issues
    7.3 Impact of First Cleaning Phase
    7.4 Impact of Second Cleaning and Integration Phase
    7.5 Data Quality Limitations
    7.6 Summary of Data Quality Evaluation
8. Conclusions and Future Work
    References :

1. Group Components and Role

This project, **Integrated Analysis of Crime in Austin Using Multi-Source Data**, is developed by **Bahareh Robaty Shirzad**, **Setayesh Mohammadi Banadkooki**, and **Samira Doshmankosh**, Master's students in Data Science at the University of Milano-Bicocca.

The aim was to fulfill the academic requirements of the Data Management (DM) course through a hands-on, end-to-end application of data engineering and analytical techniques using real-world data from the city of Austin. The group collectively defined the research objectives, acquired and profiled data from multiple sources (Open Data and Web Scraping), implemented data cleaning and normalization strategies, integrated datasets, and stored the final model in a SQLite database for SQL-based analysis. Each step reflects a shared effort and demonstrates strong collaborative planning across all phases of the project lifecycle.

2. Introduction

Urban crime represents a major social and economic concern for modern cities, and understanding how criminal activities are distributed across different urban areas is essential for effective policy making and resource allocation. Crime patterns are often influenced by several factors, including population density, socio-economic conditions, and neighborhood characteristics. Therefore, analyzing crime data together with demographic and geographic information can provide valuable insights into spatial disparities within a city.

However, urban data is typically fragmented across multiple sources and formats. In many cases, crime records, population statistics, and geographic mappings are maintained by different institutions or published on independent platforms. As a result, meaningful analysis requires the integration of different datasets that were not originally designed to work together. This makes data management techniques such as data acquisition, cleaning, correspondence investigation, and schema

integration crucial steps before any reliable analysis can be performed.

In this project, we focus on the city of Austin, Texas, combining municipal court crime records with demographic and geographic information collected through web scraping. Since the original crime dataset does not contain explicit geographic identifiers such as ZIP codes or neighborhood names, additional datasets were required to bridge this gap. Street-to-ZIP mappings, neighborhood boundaries, and population statistics were therefore collected from external web sources and integrated with the original crime records.

2.1 Research Questions

The study was driven by the following core research questions:

- How many crimes have occurred in Austin?
- Which neighborhoods are the most dangerous?
- How does crime vary when adjusted by population?
- What are the most common types of crime?
- How does crime vary across months within the same year?

By addressing these questions, the project aims to demonstrate how data integration and quality assessment techniques enable meaningful urban analysis from initially disconnected and various datasets.

3. Data Sources

This project integrates multiple datasets originating from independent sources. One dataset was obtained from an official open data portal, while additional datasets were collected through web scraping in order to enrich crime records with geographic and demographic information.

3.1 Municipal Court Caseload Dataset

The primary dataset used in this study is the **Municipal Court Caseload Information FY 2023**, published by the City of Austin through its open data portal (*data.austintexas.gov*). The dataset is provided in CSV format and contains detailed information about violations processed by the municipal court during the fiscal year 2023.

Relevant attributes used in this project include:

- **Offense Street Name:** Indicating the street where the violation occurred.
- **Offense Case Type:** Describing the category of the offense.
- **Race:** Reporting demographic information related to defendants.
- **Case Status:** Indicating whether the case is active or terminated.

Although the dataset provides detailed legal and offense-related information, it does not include structured geographic identifiers such as ZIP codes or neighborhood names. This limitation motivated the collection of additional datasets to enable spatial analysis.

3.2 Web Scraped Datasets

To enrich the crime records with geographic and demographic context, three additional datasets were collected using web scraping techniques.

All datasets were processed using Python. Web data were collected through automated scraping using the pandas library (read_html and CSV parsing), while data cleaning, normalization, and dataset integration were performed using pandas dataframes.
After preprocessing and integration, the final dataset was stored in a SQLite relational database in order to support structured SQL-based analytical querying.

3.2.1 Population by ZIP Code

Population statistics for the city of Austin were extracted from the website *zip-codes.com*, which provides demographic information aggregated by ZIP code. The data were collected using Python and the `pandas.read_html()` function to parse structured HTML tables directly into dataframes.

The resulting dataset includes:

- ZIP Code
- Population

Some ZIP codes were associated with missing or zero population values and were excluded from further analysis to avoid distortions in crime rate calculations.

3.2.2 Street to ZIP Code Mapping

Since the municipal court dataset reports only street names and does not include ZIP codes, a street-to-ZIP mapping dataset was required to associate each offense location with a geographic area. Street and ZIP code associations were collected from publicly available street directory websites through web scraping.

This dataset includes:

- Street Name
- ZIP Code

This mapping plays a central role in the integration pipeline by serving as a bridge between crime records and demographic information.

3.3 Summary of Data Sources

Dataset	Source	Acquisition Method	Key Attributes
Municipal Court Caseload FY2023	City of Austin Open Data	Direct CSV Download	Street, offense type, race
Population by ZIP Code	zip-codes.com	Web Scraping	ZIP, population
Street to ZIP Mapping	Street directory websites	Web Scraping	Street, ZIP
ZIP to Neighborhood Mapping	Real estate websites	Web Scraping	ZIP, neighborhood

4. Data Exploration and Cleaning

Before performing data integration, an exploratory analysis was conducted to assess dataset coverage and consistency. This step was essential to evaluate whether meaningful joins could be established across datasets.

4.1 Data Cleaning Procedures

Several cleaning steps were applied to improve data quality:

- **Uppercase Conversion:** All street names were converted to uppercase for case-insensitive matching.
- **Whitespace Removal:** Leading/trailing spaces and punctuation were removed.
- **Standardization:** Street suffixes were standardized (e.g., "STREET" to "ST").
- **Deduplication:** Duplicate street entries were removed.

The figure below illustrates the text cleaning pipeline applied to the raw data:

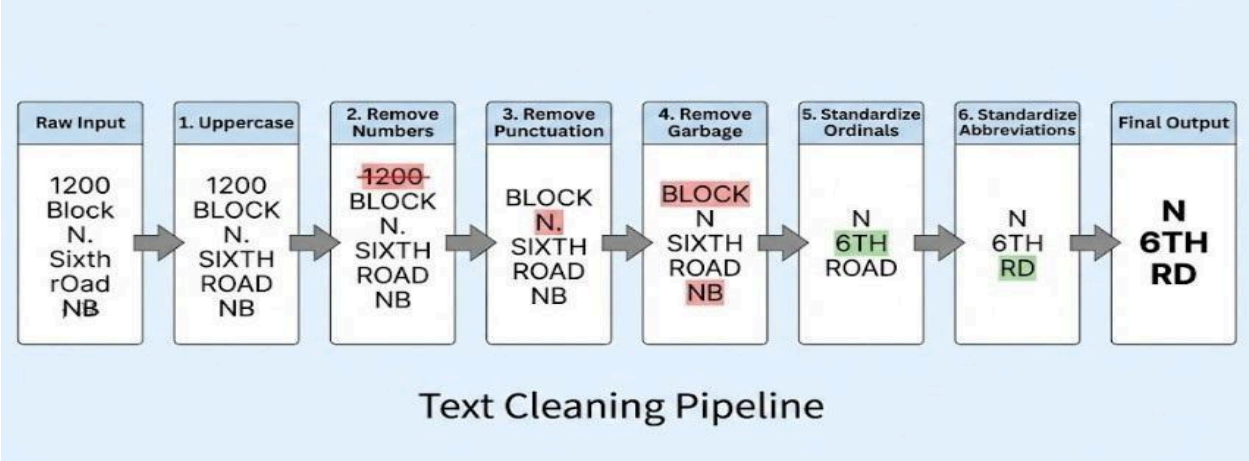


Figure 4.1: Text cleaning and normalization pipeline.

4.2 Correspondence Investigation

To evaluate integration feasibility, several overlap analyses were conducted prior to data integration. Intersections between ZIP codes from the population dataset, the neighborhood dataset, and the street-to-ZIP dataset were computed to identify a pool of valid ZIP codes common to all sources.

This analysis resulted in approximately 40 ZIP codes that were present across all geographic and demographic datasets, defining the valid geographic scope of the final analysis.

Additionally, overlap between cleaned street names in the court dataset and the street-to-ZIP mapping dataset was analyzed. Initial correspondence was limited due to inconsistencies in street name formatting, including variations in abbreviations, punctuation, and spacing.

After cleaning, a significant proportion of streets present in the crime dataset were found to correspond to streets in the ZIP mapping dataset and to fall within the valid ZIP code pool.

The figure below illustrates the correspondence investigation process, including intersection analysis across demographic datasets and string matching between street directories and crime records. The diagram highlights how a “golden set” of valid ZIP codes was identified prior to integration.

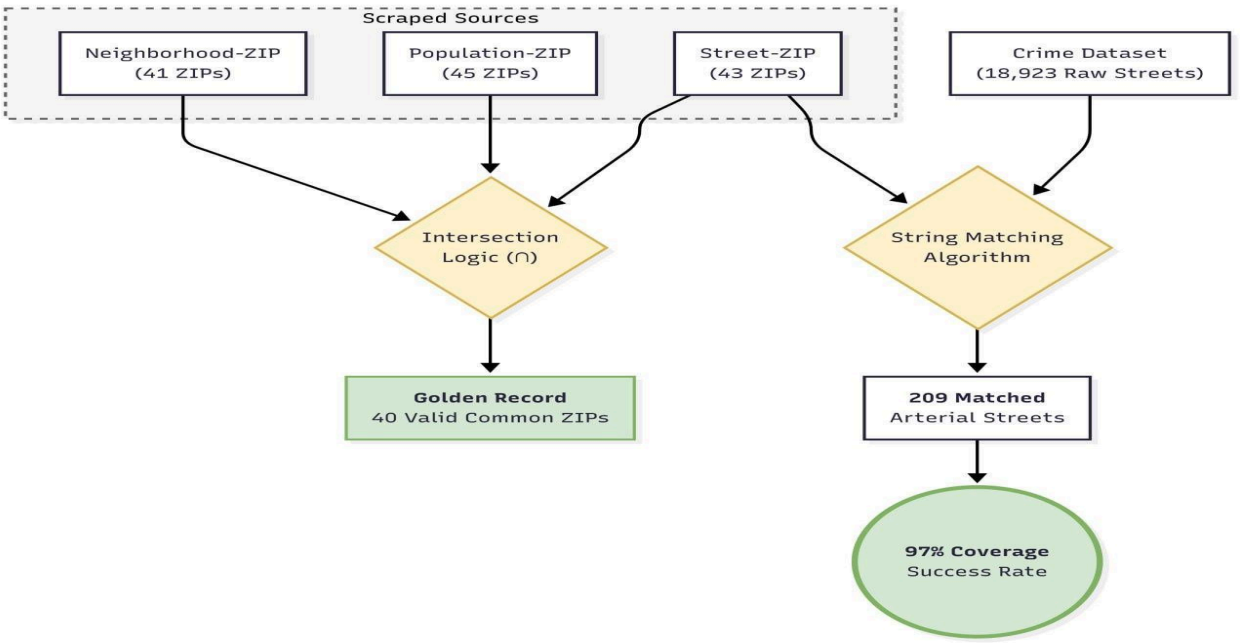


Figure 4.2: Correspondence investigation process, illustrating the intersection of ZIP codes and

string matching logic to identify the valid "golden set" of data.

4.4 Impact of Cleaning on Integration Feasibility

The cleaning process significantly increased the number of valid matches between datasets. In particular, street name normalization improved the overlap between the municipal court dataset and the street-to-ZIP mapping dataset, enabling reliable assignment of ZIP codes to a substantial subset of crime records.

Furthermore, restricting the analysis to ZIP codes present in all geographic and demographic datasets ensured that population-based crime rates could be computed consistently across neighborhoods.

Overall, the exploration and cleaning phase confirmed that meaningful integration was feasible and that sufficient geographic coverage existed to support neighborhood-level analysis.

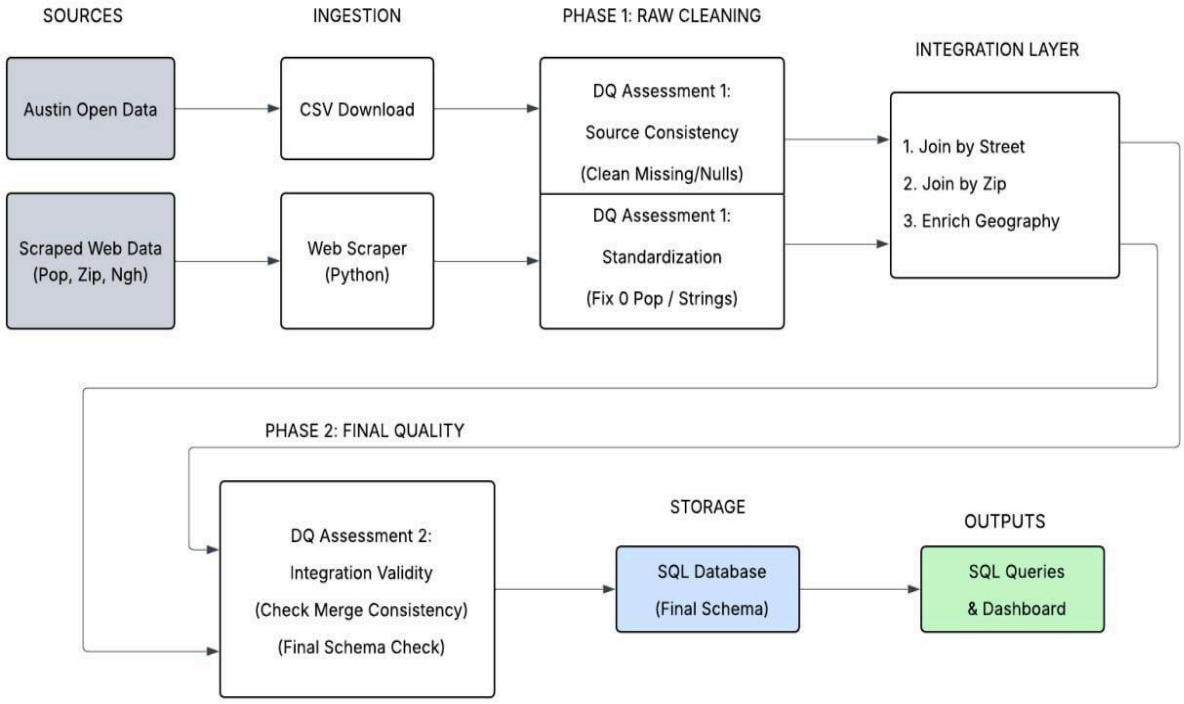


Figure 4.3: Complete data processing pipeline, illustrating the workflow from data ingestion and cleaning to integration and storage.

5. Data Integration

This Figure presents the complete data acquisition and integration pipeline adopted in this project, from raw data ingestion through cleaning, schema integration, database storage, and analytical querying.

After data cleaning and correspondence validation, the datasets were integrated in multiple stages in order to enrich municipal court crime records with geographic and demographic information. Since no single dataset contained all required attributes, integration was performed through a sequence of joins using intermediate mapping datasets.

5.1 Integration Strategy

The integration process followed a hierarchical approach:

1. Assign ZIP codes to crime records using street name matching.
2. Assign neighborhood names using ZIP codes.
3. Attach population values using ZIP codes.

This multi-step strategy was necessary because the primary crime dataset contained only street-level location information, while population and neighborhood data were available only at the ZIP code level.

5.2 Street-to-ZIP Integration

The first integration step consisted of assigning ZIP codes to crime records based on cleaned street names. The normalized street attribute from the municipal court dataset was joined with the street-to-ZIP mapping dataset using an equality match.

This join allowed each crime record to inherit one or more candidate ZIP codes depending on street coverage. Records for which no street match could be identified remained in the dataset but were excluded from geographic aggregation.

The outcome of this step produced an intermediate dataset containing:

- Original crime attributes
- Cleaned street names
- Associated ZIP codes

This intermediate dataset represented the geographic linking layer for subsequent integration.

5.3 ZIP-Based Enrichment with Neighborhood and Population

In the second integration stage, ZIP codes assigned to crime records were used as foreign keys to enrich data with neighborhood and population attributes.

Two additional joins were performed:

- Crime records were joined with the ZIP-to-neighborhood dataset to associate each ZIP code with a neighborhood name.
- Crime records were joined with the population dataset to associate each ZIP code with the corresponding population count.

Only ZIP codes present in all geographic datasets and associated with valid population values were retained for rate-based analysis. This ensured consistency across demographic indicators and avoided distortions caused by missing population data.

5.4 Integration Results and Coverage

After completing the integration pipeline, a final consolidated dataset was obtained containing:

- Offense characteristics
- Street-level location
- ZIP code
- Neighborhood name
- Population

Due to incomplete street mappings and missing ZIP-level demographic data, only a subset of the original crime records could be fully integrated into the final analytical dataset. However, correspondence analysis confirmed that the retained records were geographically consistent and covered multiple neighborhoods across the city.

The integration results demonstrate the challenges of combining different urban datasets and highlight the importance of intermediate mapping layers when direct geographic identifiers are not available.

6. Data Quality Evaluation

Data quality was evaluated throughout the project in order to assess dataset reliability, integration feasibility, and analytical validity. Since the integration process relied primarily on textual street names and ZIP codes, quality assessment focused on completeness, consistency, uniqueness, and joinability of these attributes across datasets.

Indexes were created on key attributes such as ZIP code and neighborhood name to improve query performance during aggregation operations.

6.1 Initial Quality Assessment (Before Cleaning and Integration)

Before any cleaning or integration was performed, exploratory analysis was conducted to evaluate basic data quality characteristics.

ZIP Code Coverage

The population dataset initially contained approximately 75 ZIP codes associated with Austin. However, 29 ZIP codes were associated with missing or zero population values and were therefore excluded from demographic analysis. This reduced the pool of usable ZIP codes to approximately 46.

The neighborhood dataset contained approximately 41 ZIP codes mapped to neighborhood names. The intersection of ZIP codes across the population, neighborhood, and street-to-ZIP datasets resulted in a common pool of approximately 40 ZIP codes that could be used for reliable geographic and demographic integration.

This restriction introduced a coverage limitation that had to be considered when interpreting final analytical results.

Street-Level Correspondence Issues

Initial correspondence checks revealed limited overlap between street names in the municipal court dataset and the street-to-ZIP mapping dataset. This was primarily due to inconsistencies in street name formatting, including variations in abbreviations, spacing, and punctuation.

As a result, direct matching of street names before cleaning produced low join success rates, indicating poor initial joinability.

6.2 Identified Data Quality Issues

Several data quality problems were identified during exploration:

- **Inconsistent street naming conventions**, causing semantically identical streets to be represented by different strings.
- **Duplicate records**, both at the street level and street–ZIP pair level, inflating apparent correspondence counts.
- **Ambiguous street-to-ZIP relationships**, where the same street appeared in multiple ZIP codes.
- **Missing or invalid demographic values**, especially in the population dataset.

These issues directly affected integration accuracy and required targeted cleaning strategies.

6.3 Impact of First Cleaning Phase

During the first cleaning phase, street names were normalized through uppercase conversion, whitespace trimming, punctuation removal, and suffix standardization (e.g., STREET → ST, ROAD → RD).

Duplicate street entries were removed in both the municipal court dataset and the street-to-ZIP dataset. After cleaning, overlap analysis showed that a significantly larger number of street names could be matched across datasets.

An important quality metric observed during this phase was that approximately **97% of streets that were common between the court and street-to-ZIP datasets were located within ZIP codes that were also present in both the population and neighborhood datasets**. This confirmed strong geographic consistency among the integrated sources within the valid ZIP pool.

6.4 Impact of Second Cleaning and Integration Phase

During the second phase, cleaning and integration were performed jointly. After normalization and deduplication, street-to-ZIP joins were applied to assign ZIP codes to crime records.

The primary integration quality metric computed was the **match rate**, defined as the proportion of crime records that successfully received a ZIP code after joining with the street-to-ZIP mapping dataset.

The observed match rate was approximately **40%**, meaning that ZIP codes could be assigned to roughly two out of five crime records.

This value reflects limitations in street coverage within the scraped mapping source rather than failures in cleaning. Many crime records refer to streets that were not present in the street directory dataset or were too ambiguously specified to allow reliable mapping.

However, within the restricted pool of ZIP codes that were common across all datasets, correspondence quality remained high. After filtering valid ZIP codes, matching performance exceeded **90%**, indicating that once geographic scope was constrained to consistent areas, integration reliability was strong.

It is important to note that the overall match rate computed on the full crime dataset is lower than correspondence rates computed during street-level overlap analysis, because only a subset of city streets was covered by the scraped street directory source. Therefore, integration coverage is constrained primarily by external data availability rather than internal data inconsistency.

6.5 Data Quality Limitations

Several limitations affecting data quality were identified:

- **Geographic coverage limitation**: analysis was restricted to ZIP codes with valid population and neighborhood data, reducing spatial completeness.
- **Scraping source reliability**: demographic and geographic datasets were obtained from third-party websites and may contain outdated or incomplete information.
- **Ambiguity of street-to-ZIP mapping**: streets spanning multiple ZIP codes introduce uncertainty in geographic assignment of crime records.
- **Text-based location dependency**: absence of coordinates or official geographic identifiers increases reliance on string matching, which is inherently error-prone.

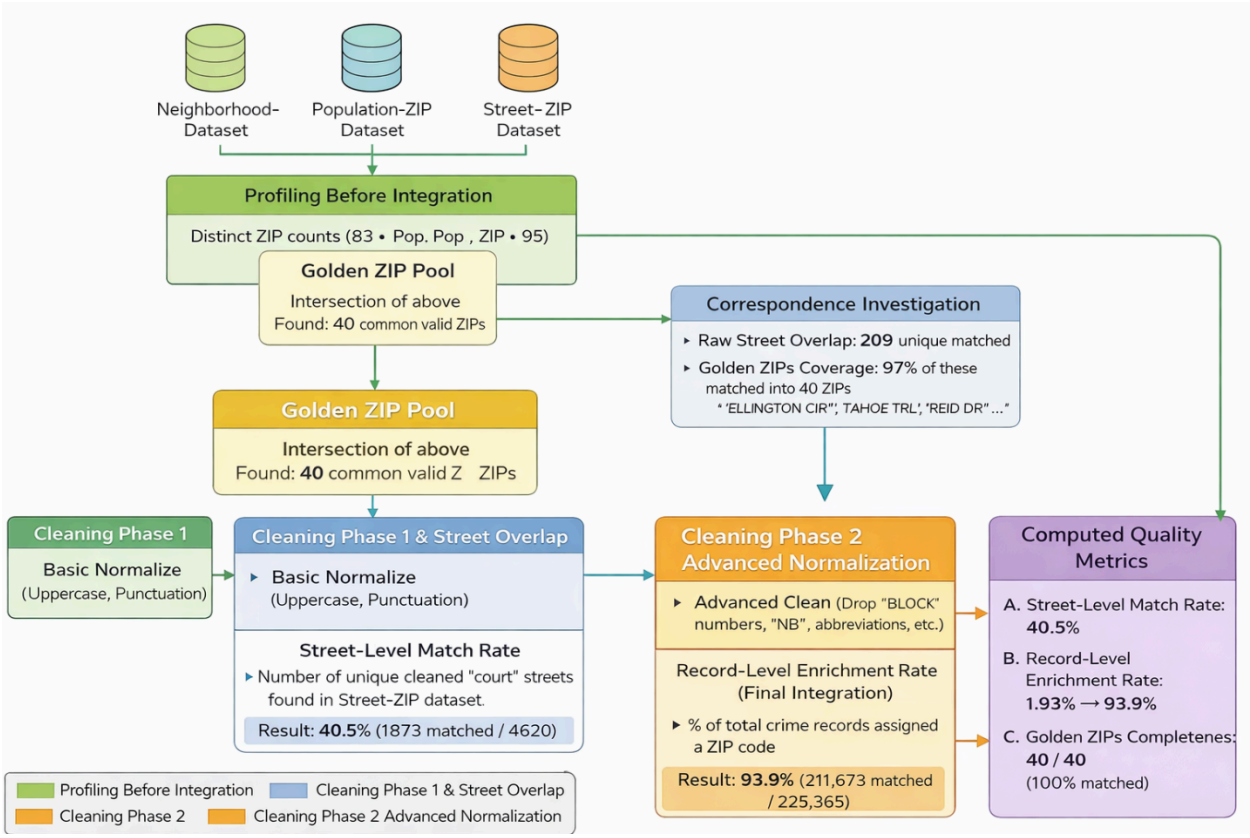
These limitations were mitigated through cleaning and correspondence filtering but could not be fully eliminated.

6.6 Summary of Data Quality Evaluation

Overall, data quality evaluation demonstrated that while raw datasets exhibited substantial inconsistencies and incomplete geographic information, systematic cleaning and correspondence validation enabled meaningful integration across independent sources.

Cleaning significantly improved joinability of street attributes, and geographic consistency across ZIP-based datasets was confirmed within the valid integration scope. Although full coverage of all crime records could not be achieved, the integrated subset of data provided reliable demographic and spatial context for neighborhood-level crime analysis.

This evaluation highlights the central role of data quality assessment in integration-driven analytical workflows and confirms that data management decisions directly affect the interpretability and reliability of final analytical results.



This figure illustrates the complete data preparation, cleaning, and integration workflow adopted in this project, together with the quality metrics computed at each stage.

First, three external scraped datasets (Neighborhood–ZIP, Population–ZIP, and Street–ZIP) are profiled to assess their consistency and coverage. By computing the intersection of ZIP codes across these sources, a Golden ZIP Pool of 40 common and valid ZIP codes is identified. This pool represents the most reliable geographic reference used during integration.

In Cleaning Phase 1, basic normalization is applied to street names in both the Court dataset and the Street–ZIP dataset, including uppercasing, removal of punctuation, and simple standardization of tokens. After this step, a preliminary street-level correspondence analysis is performed, resulting in a street-level match rate of approximately 40.5%, indicating that about 1,873 unique cleaned court streets are also found in the Street–ZIP dataset.

In Cleaning Phase 2, an advanced normalization pipeline is applied, removing block numbers, directional garbage tokens (e.g., NB, SB), highway expressions, and standardizing ordinals and abbreviations (e.g., “SIXTH” → “6TH”, “ROAD” → “RD”). This significantly improves string consistency across datasets and enables more accurate joins.

Using the refined street–ZIP associations and the Golden ZIP Pool, full record-level integration is then performed, enriching crime records with ZIP codes, neighborhood names, and population data. At this stage, the record-level enrichment rate increases from approximately 1.9% to 93.9%, meaning that almost all crime records are successfully assigned to valid geographic entities.

Finally, three quality indicators are reported:
(A) street-level match rate after cleaning,
(B) record-level enrichment rate after integration, and

(C) completeness of Golden ZIP coverage (40/40 ZIPs matched), confirming the consistency and robustness of the final integrated dataset.

7. Storage and Querying

After completing data integration, the consolidated dataset was stored in a relational database to support structured querying and efficient analytical processing. Using a database management system enables reproducible analysis and reflects realistic data engineering workflows.

7.1 Database Selection

A lightweight relational database system, SQLite, was selected for this project. SQLite provides full SQL support while requiring no server configuration, making it suitable for local experimentation and academic projects. The database was accessed directly from Python using standard database connectors.

The relational model was appropriate for this dataset because the integrated data are naturally structured in tabular form with clearly defined attributes and relationships.

7.2 Database Schema

The integrated dataset was stored in a main table containing the following attributes:

- Offense Case Type
- Cleaned Street Name
- ZIP Code
- Neighborhood
- Population
- Race
- Case Status

Additional auxiliary tables were optionally created for population and neighborhood reference data, but most analyses were conducted using the fully integrated crime table to simplify querying.

7.3 Analysis Queries

To address the research questions outlined at the beginning of this study, a series of analytical queries were executed against the integrated SQL database. The following results highlight key insights regarding crime distribution, geographic hotspots, and demographic correlations.

7.3.1 Total Crime Volume

Question:
How many crimes have occurred in Austin during the study period?

```sql
--  1.  How many crimes have occurred in Austin?

SELECT COUNT(*) AS total_crimes
FROM crimes_data;
```

Result:
The query returned a total of **752,902 crime records** in the municipal court dataset.

Interpretation:
This value represents the total number of recorded municipal court cases during the fiscal year 2023. It provides an overall baseline for the scale of urban crime addressed in this study and highlights the need for spatial and demographic breakdowns to understand how this volume is distributed across different areas of the city.

7.3.2 Crime Distribution by Neighborhood

Question:
Which neighborhoods report the highest number of recorded crimes?

```sql
--  2.  Which neighborhoods are the most dangerous?

SELECT Neighborhood, COUNT(*) AS total_crimes
FROM crimes_data
WHERE Neighborhood IS NOT NULL
GROUP BY Neighborhood
ORDER BY total_crimes DESC;
```

Results (Top Neighborhoods by Crime Volume):

Rank	Neighborhood	Total Crimes
1	Downtown Austin	106,361
2	North University, Hyde Park	95,018
3	East Austin	68,017
4	Hyde Park	62,293
5	South Lamar, Zilker, Bouldin Creek	37,575
6	Tarrytown, Clarksville	37,075
7	Mueller, Windsor Park	36,221

(Only the top neighborhoods are shown for brevity.)

Interpretation:
The results indicate that Downtown Austin reports the highest number of municipal court cases, followed by the North University–Hyde Park area and East Austin. These areas are characterized by high population density, commercial activity, and major transportation corridors, which may contribute to increased violation rates. However, absolute crime counts do not account for differences in population size, motivating further analysis using population-adjusted crime rates in the next section.

7.3.3 Population-Adjusted Crime Rates

Question:
How does crime vary across neighborhoods when adjusted by population size?

```
--  3.  How does crime vary when adjusted by population?

SELECT Neighborhood, COUNT(*) AS crime_count, Population,
       (CAST(COUNT(*) AS FLOAT) / Population) * 1000 AS crimes_per_1k
FROM crimes_data
WHERE Population > 0
GROUP BY Neighborhood
ORDER BY crimes_per_1k DESC;
```

Motivation:
While total crime counts provide useful information, they may be misleading when comparing neighborhoods with very different population sizes. To enable a fair comparison, crime rates were normalized by population and expressed as incidents per 1,000 residents.

Results:

Rank	Neighborhood	Crime Count	Population	Crimes per 1,000
1	Downtown Austin	106,361	11,625	9,149
2	Cherrywood	32,437	6,618	4,901
3	Hyde Park	62,293	17,071	3,649
4	Rosedale	30,070	8,426	3,569
5	East Austin	68,017	25,910	2,625
6	North University, Hyde Park	95,018	36,535	2,601

(Only neighborhoods with valid population data are included.)

Visualization:

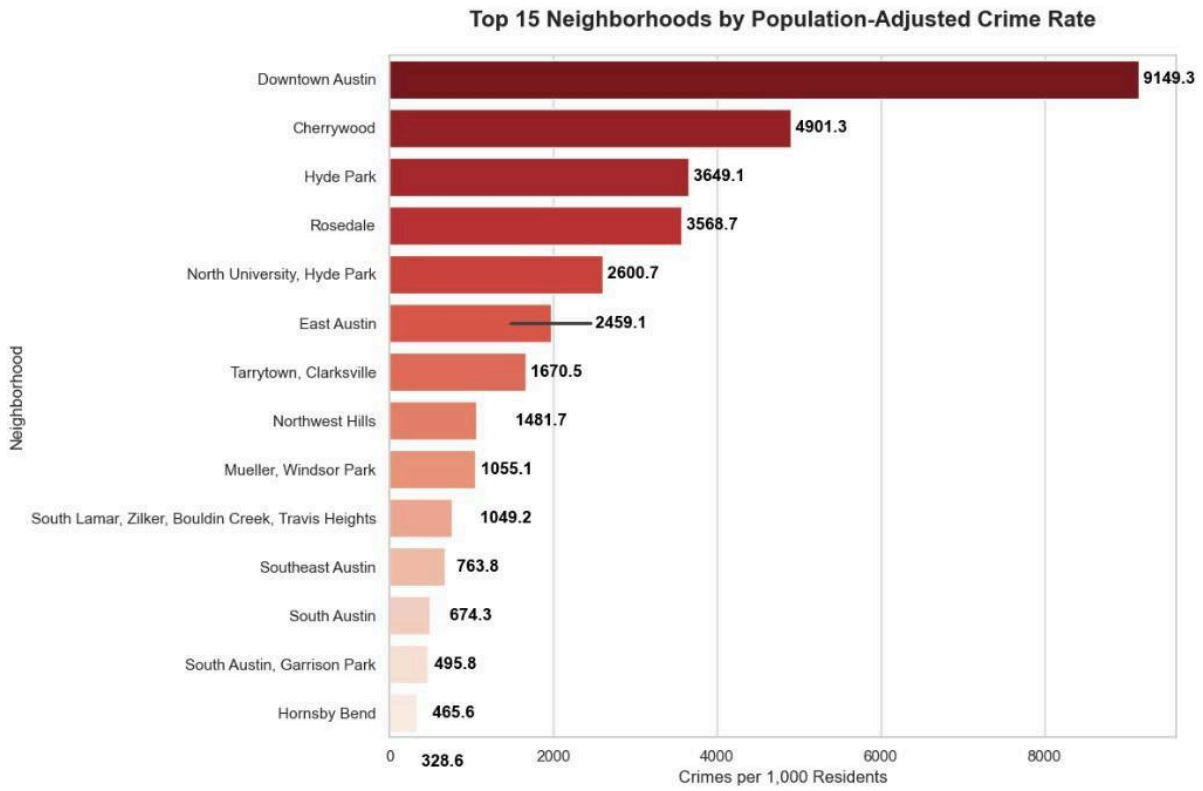


Figure 6.1: Top 15 neighborhoods (ZIP codes) with the highest number of recorded crimes

Interpretation:

This bar chart identifies the neighborhoods with the highest relative crime risk by standardizing the data to account for population size. Instead of relying on raw incident counts—which can misleadingly highlight highly populated areas—the analysis calculates the number of crimes per 1,000 residents. By sorting these values in descending order and isolating the top 15, the visualization highlights the specific communities where crime density is disproportionately high relative to the number of people living there. The use of a red gradient emphasizes intensity, providing a clear, risk-adjusted view of the most dangerous areas in the city.

7.3.5 the most common types of offenses

Question:
What are the most common types of offenses recorded in the Austin municipal court dataset?

```
--  4.  What are the most common types of crime?

SELECT "Offense Case Type", COUNT(*) AS total_incidents
FROM crimes_data
GROUP BY "Offense Case Type"
ORDER BY total_incidents DESC;
```

Results:

Rank	Code	Offense Type	Count
1	PK	Parking Violations	669,823
2	TR	Traffic Offenses	66,173
3	CP	Criminal Procedure	9,689
4	OR	Ordinance Related	5,257
5	NT	Non-Traffic	1,960

Interpretation:

The results indicate that **Parking-related violations (PK)** represent the vast majority of recorded cases, accounting for more than 85% of all incidents.
Traffic-related offenses (TR) constitute the second largest category, followed by significantly smaller volumes of criminal procedure (CP), ordinance-related (OR), and non-traffic (NT) cases.

This distribution confirms that the municipal court dataset is primarily dominated by administrative and regulatory violations rather than serious criminal offenses.
This characteristic is consistent with the institutional role of municipal courts, which mainly process parking tickets, traffic infractions, and minor city ordinance violations. Due to the dominance of parking violations, some spatial crime patterns may reflect parking enforcement policies rather than actual public safety risk.

7.3.6 Temporal Crime Trends

Question:
How does the number of reported crimes vary across different months of the year?

```
--  5.  How does crime vary across months within the same year?

SELECT substr("Offense Date", 1, 2) AS Month, COUNT(*) AS total_crimes
FROM crimes_data
GROUP BY Month
ORDER BY Month ASC;
```

Results:

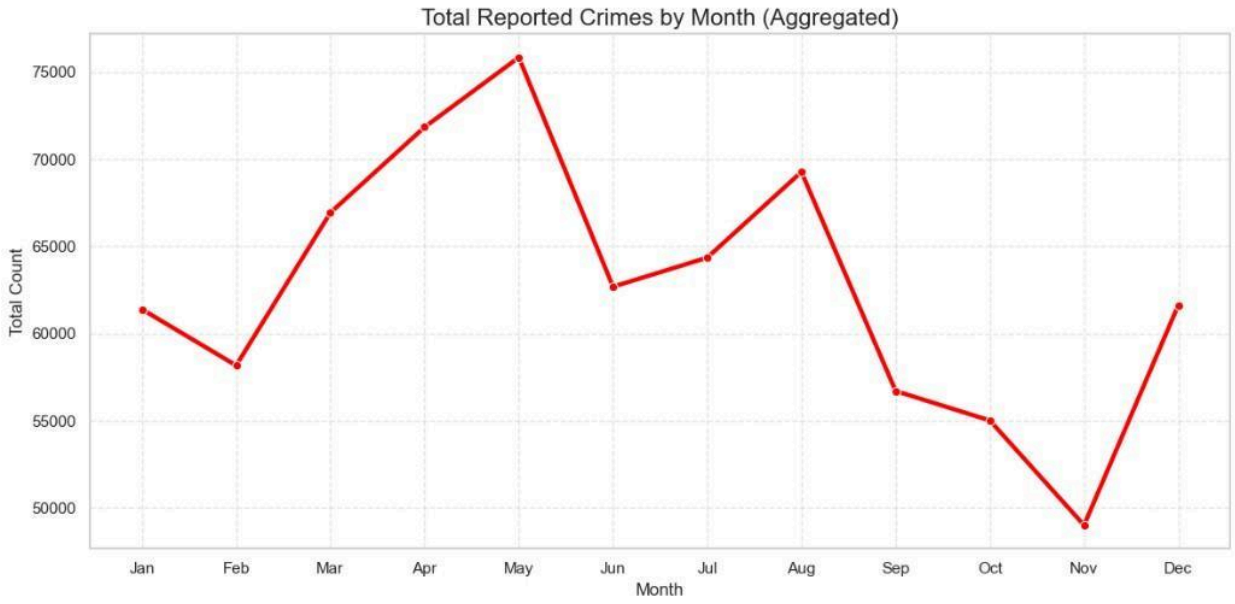


Figure 6.4: Monthly distribution of recorded crime incidents, highlighting seasonal variations.

Interpretation:

This line chart visualizes the distribution of criminal activity throughout the year by aggregating data from late 2023 and early 2024 into a single 12-month timeline. By plotting the months on the X-axis and total crime counts on the Y-axis, the visualization connects the data points to reveal the volume of reported crime for each specific month. This approach creates a continuous monthly profile, allowing for a clear comparison of crime levels from January through December without separating the data into distinct years.

Note: *It should be noted that this temporal trend reflects reported municipal court cases rather than actual crime occurrence time, and the analysis assumes that the offense date recorded in the dataset accurately represents the incident date.*

During data quality inspection, inconsistencies were observed in the formatting of the offense date field (e.g., some months were stored as "1/" instead of "01"). This issue was addressed during the cleaning phase to ensure consistent temporal grouping.

8. Conclusions and Future Work

This project demonstrated how different urban datasets can be combined to support meaningful spatial crime analysis through systematic data management techniques. By integrating municipal court records with demographic and geographic information collected via web scraping, it was possible to analyze crime patterns across neighborhoods in Austin, Texas.

The analysis showed clear variation in crime frequency and crime rates across different neighborhoods. Normalizing crime counts by population allowed more equitable comparison between areas with very different population sizes. Additionally, the distribution of offense types across neighborhoods suggested that certain categories of violations tend to concentrate in specific geographic areas.

From a data management perspective, the project highlighted several important challenges. The primary municipal dataset lacked structured geographic identifiers, requiring the use of intermediate mapping datasets to enable integration. Text-based street matching proved sensitive to formatting inconsistencies and required extensive cleaning and normalization. Even after cleaning, full geographic coverage could not be achieved due to incomplete street-to-ZIP mappings and missing demographic data for some ZIP codes.

Despite these limitations, correspondence analysis confirmed strong geographic consistency among the integrated datasets within the valid ZIP code pool. This indicates that the integration pipeline was reliable within its defined scope and that the resulting analytical dataset is suitable for neighborhood-level comparison.

Several directions for future work could significantly improve the quality and scope of the analysis. First, the use of official geographic identifiers such as census tract codes or latitude–longitude coordinates would reduce reliance on street name matching and improve spatial accuracy. Second, using official Census Bureau APIs could provide more reliable and complete demographic information. Third, fuzzy matching techniques or geocoding services could be employed to improve street-level matching and increase integration coverage.

In conclusion, this project illustrates how careful data acquisition, cleaning, integration, and quality evaluation are essential steps in transforming disconnected datasets into actionable analytical resources. The methodology developed in this study can be extended to other urban datasets and supports scalable data-driven investigations in smart city applications.

References :

1. Source: City of Austin Open Data
   URL: https://data.austintexas.gov/
2. Source:OpenAddressesProject
   URL: https://data.openaddresses.io/
3. Source:SpyglassRealty
   URL: https://www.spyglassrealty.com/zip-code-search.php

Tools and Libraries :

Python(pandas,sqlite3)
Used for: web scraping, cleaning, integration, and preprocessing
URL: https://pandas.pydata.org/

SQLiteDatabaseEngine
Usedfor:datastorageandanalyticalSQLqueries
URL: https://www.sqlite.org/