POLITECNICO
MILANO 1863

Project ID: 50

Assigned Dataset: crime and offense_Codes
DATASET Number: 14

Authors:
Bahman Amirsardary
Milad Ramezani Ziarani

GitHub: https://github.com/Bahman75/Data_Quality

Student ID:	10900660, 10930504
Advisor:	CINZIA CAPPIELLO
Academic Year:	2024-2025

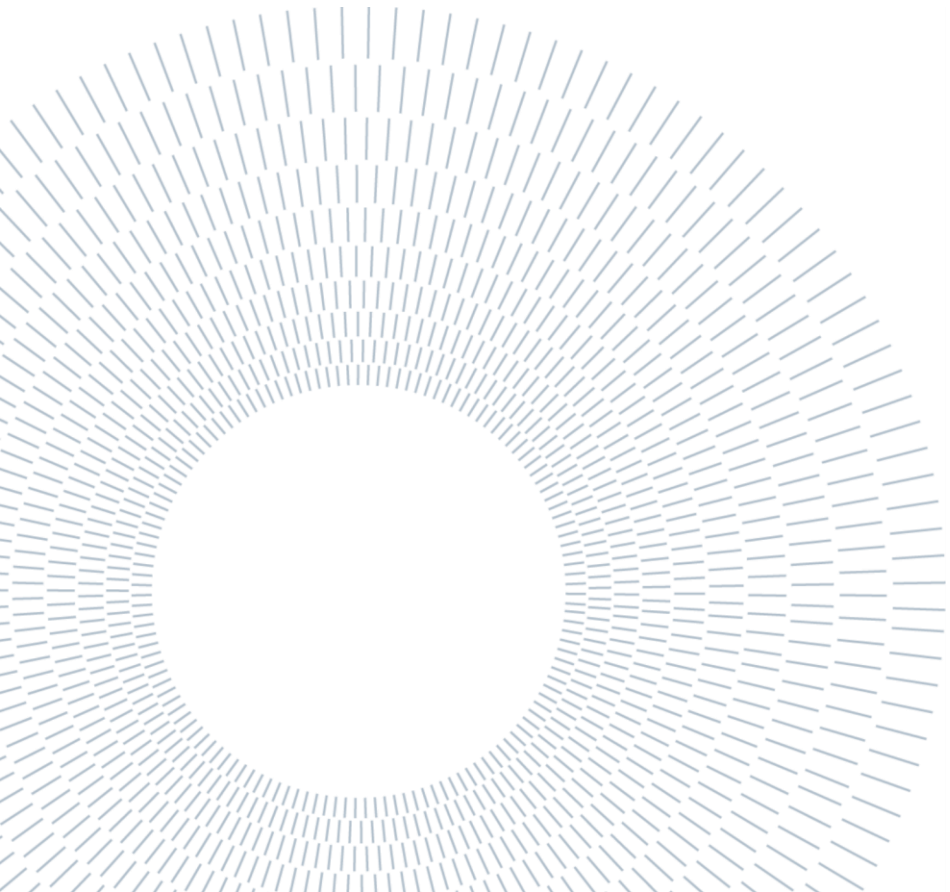


Table of Contents

1. Introduction

Importance of Data Quality in Analytical Systems

Data quality is a cornerstone of any analytical system. The accuracy, reliability, and usefulness of insights derived from data depend on its quality. High-quality data ensures that decisions are based on accurate, complete, and consistent information, which is vital for sound decision-making. In contrast, poor data quality can result in incorrect conclusions and biased outcomes. This is especially critical when working with large datasets, such as crime data, where the quality of the data influences the performance of predictive models, data visualizations, and overall system efficiency. Therefore, improving data quality through processes like imputation, cleaning, and validation is crucial for any data-driven analytical system.

Objectives of the Project

This project aims to assess and enhance the quality of crime data by addressing missing values, identifying and managing outliers, and ensuring consistency across variables. The main objective is to clean and preprocess crime datasets to prepare them for reliable analysis. This involves imputing missing values, removing duplicates, and handling inconsistent entries. Additionally, the project seeks to apply various data wrangling techniques to improve the completeness and accuracy of the data, making it suitable for advanced analytical tasks, such as crime trend analysis and predictive modeling. Ultimately, the goal is to create a high-quality dataset that can offer actionable insights for law enforcement agencies and other stakeholders involved in crime prevention and analysis.

2. Data Description

Overview of Data Sources

Crime Dataset

This dataset provides detailed information about individual crime incidents, including unique incident numbers, offense codes, descriptions, locations, and timestamps. It includes key attributes like district, reporting area, shooting information, and geographic coordinates, enabling geographic mapping of the crime incidents.

Offense Codes Dataset

This dataset links numeric codes to specific types of crimes, like larceny, vandalism, or motor vehicle accidents. It standardizes the classification of offenses, making it easier to analyze and compare crime data across regions and time periods.

Schema of the Data Sources

Crime Dataset

- INCIDENT_NUMBER: Unique ID for each crime.
- OFFENSE_CODE: Numeric code for the offense type.
- OFFENSE_CODE_GROUP: Broad category for the offense.
- OFFENSE_DESCRIPTION: Description of the offense.
- DISTRICT: Police district of the crime.

- REPORTING_AREA: Subdivision within the district.
- SHOOTING: Indicates if a shooting occurred (Y/N).
- OCCURRED_ON_DATE: Date and time of the incident.
- YEAR, MONTH, DAY_OF_WEEK, HOUR: Time details of the crime.
- UCR_PART: Classification based on crime severity.
- STREET: Street where the crime occurred.
- Lat & Long: Geographic coordinates of the location.
- Location: Combined latitude and longitude.

Offense Codes Dataset

- CODE: Numeric offense code
- NAME: The description of the offense (e.g., "LARCENY")

These schemas structure the data to allow for efficient analysis, integration, and querying of crime-related information.

3. Data Profiling and Exploration

Data profiling helps to understand the structure and quality of the dataset. This step involves identifying patterns, assessing the completeness of the data, and highlighting potential issues.

Display of Unique and Distinct Values

Crime Dataset

- Incident Numbers: Unique identifiers for each crime
- Offense Codes: Distinct offense codes (e.g., Larceny, Vandalism)
- District: Various districts where crimes occur
- Day of Week: Days when crimes occurred
- Hour: Hour of the day when crimes were reported
- Location: Latitude and Longitude values for mapping crimes

Offense Codes Dataset

- Offense Codes: Numerical representations of different crime types
- Offense Names: Descriptions of each offense, such as "LARCENY" or "MANSLAUGHTER"

4. Data Cleaning and Transformation

Data cleaning is essential to ensure that the dataset is accurate and ready for analysis. Several techniques were applied in this process:

Handling Missing Values

Completeness

Missing data was identified by calculating the proportion of missing values at each attribute. This helped in deciding which columns needed imputation or removal.

Imputation Techniques

- **MICE (Multiple Imputation by Chained Equations)**

This method generates multiple imputed datasets based on relationships between variables, effectively handling missing data in continuous and categorical attributes.

- **Mode-Based Imputation**
  For categorical variables with missing values, mode-based imputation was applied, where the missing entries were filled with the most frequent value (mode) within their respective columns. For instance, street names such as "Washington Street," the most common district associated with this street, identified as "B2" was used to fill the missing data.

Managing Outliers and Duplicates

Outliers
Statistical measures such as IQR and z-scores were used to identify outliers, which were either removed or transformed based on their impact on analysis.

Duplicates
Duplicate records were removed by identifying repeated values across key attributes to ensure data integrity.

Column Splitting and Renaming

Splitting Columns
The NAME column in the offense code table was split into two parts based on the - separator, creating NAME and DESCRIPTION columns for better clarity.

Renaming Columns
Columns were renamed to enhance clarity and standardization (e.g., changing "Lat" and "Long" to "Latitude" and "Longitude").

Sorting Data by Time

The dataset was sorted by the "Occurred on Date" column to arrange the records chronologically. This sorting helps identify crime trends over time and detect patterns related to specific periods.

Merging data set

Duplicates were removed, and the offense code table was merged with the crime dataset using OFFENSE_CODE. The NAME column was excluded, keeping only DESCRIPTION for clarity

5. Data Quality Assessment

This section outlines the methods used to evaluate and ensure the quality of the crime datasets.

Completeness Analysis

The completeness of each column was assessed by calculating the percentage of non-missing entries. Critical attributes such as Offense Code, District, and Latitude/Longitude were checked for missing values, and appropriate imputation or removal was applied.

Consistency Checks

Validating Offense Codes
Ensuring the Offense Code matched entries in the Offense Codes dataset.

Date and Time Format
Verifying consistent formats in date and time attributes.

Geospatial Consistency
Ensuring that Latitude and Longitude coordinates were within valid geographical bounds.

Deduplication

Duplicate records were identified by checking for repeated values across key attributes and removed to ensure the dataset's integrity.

6. Geospatial and Temporal Analysis

Understanding the geographical and temporal distribution of crimes is vital for deriving actionable insights.

Handling Time Series Data

The "Occurred on Date" field was converted into a time series format, enabling the identification of crime patterns across different time intervals. This analysis focused on:

- Daily Patterns

Geospatial Data Transformation and Enrichment

Mapping Crime Incidents
Visualizing crime incidents on a map to identify high-crime areas and hotspots.

7. Data Visualization and Reporting

Data visualization and reporting play a vital role in presenting insights and ensuring that the data is understandable and actionable. For the crime dataset, summary reports and visualizations were generated using Sweetviz, a Python library designed to create comprehensive visualizations and exploratory data analysis reports. The following steps outline how Sweetviz was used to enhance the data understanding:

Summary Reports using Sweetviz

Sweetviz is an open-source Python library that automates the process of generating detailed visualizations and reports for exploratory data analysis. It provides a quick and easy way to explore the dataset and communicate its key characteristics to both technical and non-technical stakeholders.

Automatic Report Generation
Sweetviz generates a detailed report summarizing the dataset's structure, distributions, and relationships between features. It displays the count of missing values, unique values, and the distribution of numerical features.

Comparison of Datasets
The library also enables the comparison of different subsets or different versions of datasets, which can be particularly useful when examining pre- and post-imputation datasets, or when comparing different temporal or spatial segments of the data.

Visualization
Sweetviz produces various visualizations, such as bar charts, histograms, and heatmaps, to provide deeper insights into the dataset. These visualizations help in identifying trends, outliers, and patterns in the crime data, such as the frequency of different offenses or the geographical distribution of incidents. By using Sweetviz, key data characteristics such as missing values, distribution of offenses across districts, and relationships between various features (like time of day, district, and offense type) were quickly summarized, leading to a more efficient data analysis process.

8. Code Implementation

In this project, Python was used to carry out data cleaning, transformation, geospatial analysis, and data visualization. Below is an overview of the key Python scripts and libraries utilized, as well as some key code snippets that highlight the core functionalities of the data analysis process.

Overview of the Python Scripts and Libraries Used

Several Python libraries were employed to manipulate, clean, and analyze the data. The primary libraries used include:

Pandas
Used for data manipulation and analysis, including handling missing values, cleaning, and transforming the dataset.

NumPy
Utilized for numerical operations, particularly in handling arrays and performing mathematical computations.

Matplotlib and Seaborn
Libraries for creating static, animated, and interactive visualizations, used to generate graphs such as histograms, bar charts, and heatmaps.

Sweetviz
A library used for automated exploratory data analysis and report generation, providing valuable insights into the dataset.

Geopandas
For spatial data analysis, enabling the processing and visualization of geospatial data.

MICE (Multiple Imputation by Chained Equations)
Used to handle missing data through multiple imputation techniques.

Scikit-learn
Employed for data preprocessing and machine learning tasks.

Key Code Snippets

Here are some key code snippets that illustrate the main tasks carried out in the analysis:

Data Loading and Initial Exploration

```
import pandas as pd
# Load the crime dataset
crime_data = pd.read_csv('crime_data.csv')
# Preview the first few rows of the dataset
crime_data.head()
```

Handling Missing Values using MICE

```
from sklearn.impute import SimpleImputer
# Using MICE to impute missing values
from miceforest import KernelDataSet
# Create the MICE imputation model
dataset = KernelDataSet(crime_data)
# Impute the missing values
dataset.impute()
imputed_data = dataset.complete()
```

Geospatial Data Visualization

```
import geopandas as gpd
import matplotlib.pyplot as plt
# Convert the crime data to GeoDataFrame
gdf = gpd.GeoDataFrame(crime_data, geometry=gpd.points_from_xy(crime_data['Long'],
crime_data['Lat']))
# Plot the geospatial data
gdf.plot(marker='o', color='red', markersize=5)
plt.title('Crime Locations')
plt.show()
```

Summary Report Generation using Sweetviz

```
import sweetviz as sv
# Generate a report of the crime dataset
report = sv.analyze(crime_data)
# Display the report
report.show_html('crime_report.html')
```

Handling Time Series Data

```
# Convert the 'OCCURRED_ON_DATE' column to datetime format
crime_data['OCCURRED_ON_DATE'] = pd.to_datetime(crime_data['OCCURRED_ON_DATE'])
# Sort the dataset by date
crime_data_sorted = crime_data.sort_values(by='OCCURRED_ON_DATE')
# Display the first few rows of the sorted data
crime_data_sorted.head()
```

These code snippets represent the fundamental steps carried out in this project, including data loading, handling missing values, geospatial analysis, and reporting. They showcase the powerful capabilities of Python in managing and analyzing complex crime data.

9. Results and Observations

This section outlines the key findings and outcomes from the data cleaning and transformation process, focusing on improvements in data quality metrics, the resolution of data inconsistencies, and overall dataset readiness for advanced analysis.

Improvements in Data Quality Metrics

The data cleaning process led to significant enhancements in key data quality dimensions, such as completeness, consistency, and accuracy. These improvements are summarized below:

Completeness

Reduction in Missing Values

Missing data across critical columns, including *OCCURRED_ON_DATE*, *LAT*, *and LONG*, were imputed using advanced techniques such as Multiple Imputation by Chained Equations (MICE) and Mode-Based Imputation. This ensured a substantial reduction in missing entries, making the dataset more suitable for analysis.

Overall Dataset Completeness

The completeness of the dataset improved. Columns such as *SHOOTING*, *Lat*, *Long*, and *STREET* saw significant reductions in null values.

Consistency

Standardization

Date formats were standardized, and categorical variables like *DAY_OF_WEEK* and *MONTH* were aligned with correct data types and values.

Column Transformation

The *NAME* column in the *OFFENCE_CODES* dataset was split into *NAME* and *DESCRIPTION* based on a hyphen (-) separator. Null values in the new *DESCRIPTION* column were replaced with empty strings for clarity.

Data Integrity

Duplicate Records

The *CRIME* dataset contained **duplicate records**, identified and removed to ensure that each entry corresponded to a unique crime event. Repeated entries with the same *INCIDENT_NUMBER* were eliminated to improve accuracy.

Additionally, 151 duplicate rows were found and deleted from the offense code table.

Outlier Management

Outliers in numerical columns such as geospatial coordinates (*LAT* and *LONG*) were identified and either corrected or excluded, ensuring reliable analysis.

Summary of Data Cleaning Outcomes

The data cleaning process included multiple steps, each contributing to enhanced dataset quality:

Handling Missing Values

- Advanced imputation methods were used to fill missing values in critical columns, such as geospatial coordinates (*LAT* and *LONG*) and temporal data (*OCCURRED_ON_DATE*).
- Null values in categorical columns, such as *SHOOTING*, were replaced with consistent placeholders (e.g., *'N'*).

Data Transformation

- The *NAME* column was transformed into two separate columns, *NAME* and *DESCRIPTION*, to improve clarity and usability.
- The *OCCURRED_ON_DATE* column was split into *YEAR*, *MONTH*, *DAY*, and *TIME*, facilitating detailed time-based analysis.

Sorting and Ordering

- The dataset was sorted chronologically based on the *OCCURRED_ON_DATE* column, which had been converted to datetime format. This enabled effective time-series analysis and trend identification.

Geospatial Enrichment

- Geographic data was enriched by converting *LAT* and *LONG* into geospatial points, preparing the dataset for spatial analysis and visualization.

Sweetviz Analysis

- A *Sweetviz* report provided an overview of the dataset:
  - **Total Rows**: 285008
  - **Total Features**: 14

Observations and Insights

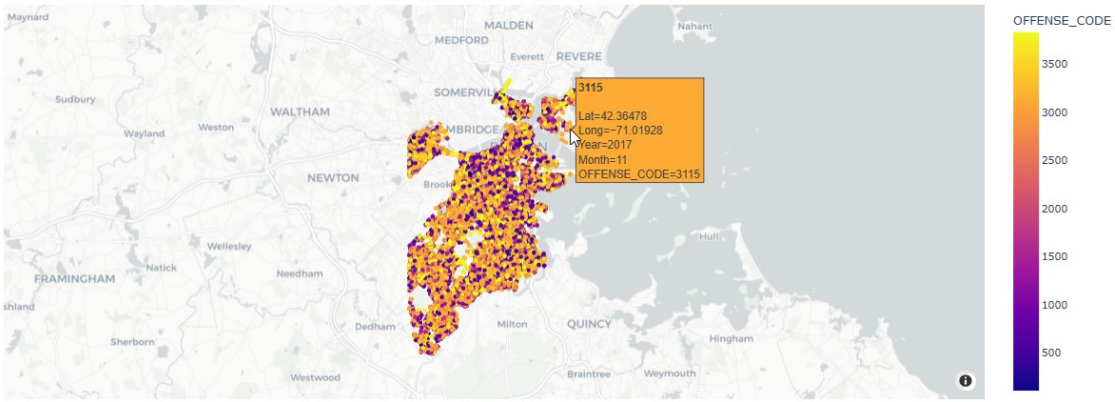
Unique Offense Codes

The *OFFENCE_CODES* dataset contained **425 unique offense codes**, demonstrating consistency with the *codes_1* DataFrame.

Geospatial Visualization of Crime Data Over Time

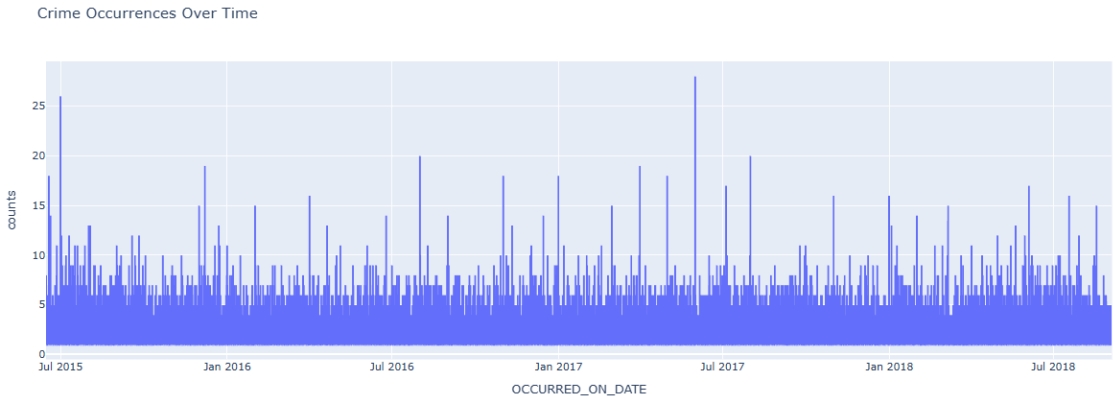
In this report, a geospatial timeseries map is generated using Plotly Express to visualize crime data. The map is created by plotting latitude and longitude coordinates from the *CRIME_no_duplicates* DataFrame, with each point representing an offense. The map's hover feature displays details such as the offense code, year, and month. Different offense codes are color-coded to allow easy differentiation between them. The map is zoomed in at a level of 10, and the "carto-positron" style is applied for a clean and readable

display. This interactive map provides a clear geospatial representation of crime patterns over time.



Trend of Crime Occurrences Over Time

In this report, a time series plot is created to show the trend of crime occurrences over time. The data is grouped by the date of occurrence (*OCCURRED_ON_DATE*), and the number of crimes on each date is counted. This aggregated data is then visualized using a line plot, which illustrates fluctuations in the number of crimes throughout the period under study. The plot provides valuable insights into crime patterns, helping to identify periods of higher or lower activity. The chart's title, "Crime Occurrences Over Time," clearly reflects the trend being analyzed.



10. Conclusion

In this project, a comprehensive data cleaning and transformation process was undertaken to improve the quality and usability of the crime dataset. By applying various data quality enhancement techniques, the dataset was made more suitable for analytical tasks, ensuring that key issues related to completeness, consistency, and accuracy were addressed.

Key Takeaways

Data Quality Improvement

The dataset saw substantial improvements in completeness through advanced imputation techniques such as Multiple Imputation by Chained Equations (MICE) and mode-based imputation. Missing values were effectively addressed, contributing to a more robust dataset for analysis.

Effective Handling of Outliers and Duplicates

Outliers and duplicates were identified and managed appropriately, ensuring that the dataset accurately represented unique crime incidents and their corresponding attributes.

Geospatial and Temporal Enrichment

Geospatial data was transformed to enable spatial analysis, and the temporal data was cleaned and organized for effective time series analysis, allowing for insights into crime trends over time.

Improved Data Usability

By splitting columns, renaming variables, and performing various transformations, the dataset was made more user-friendly, facilitating further analysis and visualization.

Future Enhancements

While the data quality improvements achieved in this project are significant, there are several opportunities for further enhancement:

Advanced Imputation Techniques

Future work could explore additional imputation techniques, such as deep learning-based methods, to further improve the accuracy of missing data imputation, especially for complex or multivariate missing data patterns.

Integration with External Data Sources

Incorporating additional data sources, such as socio-economic factors, weather conditions, or neighborhood characteristics, could enrich the analysis and provide deeper insights into crime patterns.

Predictive Modeling

With the cleaned and enriched dataset, the next logical step would be to build predictive models to forecast crime occurrences or identify high-risk areas based on historical data.

Real-Time Data Integration

Integrating real-time crime data could improve the timeliness and relevance of the analysis, allowing for more dynamic and up-to-date insights into crime trends and patterns.