Crime Data Chronicles: A Comprehensive Project for Cleaning, Uncleaning and Analysis

**Guide:** Prof. Kamellia Reshadi, Prof. Theodoros Soldatos

**Subject:** Data Management: Data Acquisition and Data Cleaning

**Date:** 08/03/2024

**Project by:**

Ashwith Anand Poojary(11038008)

Muskaan Chhabra(11038196)  
   Rakesh Hadne Sreenath(11038138)  
        Aadishree Borkar(11037522)

A person and person in police vests

Description automatically generated

Image Link: https://ichef.bbci.co.uk/news/976/cpsprodpb/C428/production/\_101661205\_polgerafp30jul17.jpg.webp

## Introduction

In our group project, we carefully assigned tasks to make the most of each member's skills. Aadishree focused on running queries and cleaning data, laying a foundation for our analysis. Rakesh and Ashwith, who excel in data profiling, paid attention to cleaning duties, contributing to the overall data quality. Aadishree and Muskaan took on the task of handling datasets. Muskaan ensured the integrity of the cleaned dataset, while Ashwith skillfully crafted queries to navigate through the uncleaning process.

Initially struggling to find a dataset that met our requirements, we decided to simultaneously search for one, which improved our efficiency. By using Tableau Prep Builder for uncleaning and GCP DataPrep for cleaning, we streamlined our processes and maintained workflow.

Our weekly team meetings, held every Thursday in the library, were crucial for communication and collaboration. These gatherings provided a platform to review progress, address challenges faced by team members, and collectively brainstorm solutions. The regular meeting schedule not only strengthened our team dynamics but also created an atmosphere of mutual support and shared learning. Our primary goal was not just project completion but collaboration and promoting knowledge exchange among team members.

The emphasis on working to learn not only improved the standard of our output but also helped foster a supportive and united team environment, making our project experience both fruitful and rewarding.

Dataset Used

* Dataset reflects incidents of crime in the City of Los Angeles.
* Link: <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>
* Data Last Updated: Last updated February 22, 2024

A screenshot of a computer

Description automatically generated

### Why couldn’t we use Open Refine Application?

The dataset we chose was too large and it could not be opened in Open Refine as it has its limitations. It was only able to work on smaller datasets. Hence, we did not move forward with this application.

A screenshot of a computer

Description automatically generated

## User story: The Crime Chronicles of Los Angeles

As a data analyst for law enforcement agencies:

* I want to retrieve crime counts for each police area name, so that we can identify crime hotspots and allocate resources efficiently.

As a law enforcement officer:

* I want to know the most common type of crime in each area, so that we can focus our efforts on addressing prevalent criminal activities.

As a community safety advocate:

* I want to determine the average victim age for different types of crimes, so that we can tailor prevention and support initiatives to better protect vulnerable demographics.

## Data Profiling

In the data profiling process, Rakesh and Ashwith along with the team engaged in several crucial tasks aimed at gaining a thorough understanding of the dataset. Initially, we conducted an assessment to determine the number of unique values present for each attribute, offering insight into the dataset's diversity and breadth. This step allowed us to grasp the range of data variations, identifying potential anomalies or redundancies.

Subsequently, we meticulously categorized the types of data within the dataset, distinguishing between numeric, textual, date-based, and categorical data. This classification was instrumental in devising appropriate strategies for data manipulation and analysis, ensuring that he could effectively interpret and utilize the information contained within.

Following this, we closely examined the distribution of values across the dataset, aiming to uncover any discernible patterns, outliers, or irregularities. This scrutiny enabled us to discern underlying trends and tendencies within the data, providing valuable context for subsequent analysis. Lastly, we endeavored to establish connections and dependencies between different parts of the dataset, elucidating how various attributes interacted with each other. By discerning these relationships, we could prioritize key factors for analysis and gain a holistic understanding of the dataset's structure and dynamics. Ultimately, through these systematic profiling efforts, we laid a solid foundation for informed decision-making and analysis, ensuring that we could derive meaningful insights and make informed choices based on a comprehensive understanding of the data.

When we profiled the dataset, we essentially examined all the different features of information to figure out which ones were useful and which ones weren't. It's like sorting through a pile of stuff to decide what to keep and what to throw away. We kept the things that were important for our analysis and got rid of the things that didn't really matter. This helped us clean up the data more efficiently. By doing this, we ensured that our profiling analysis would be based on the most relevant and reliable information.

## Steps taken in Data Profiling

**Single-Column Analysis-**

1. Unique Values:

Profiling involves determining the number of unique values for each attribute. This helps in understanding the variety and distribution of data within the dataset.

A screenshot of a computer

Description automatically generated

1. **Data Types:**

Profiling also involves Identifying the types of data present in the dataset, such as numeric, text, date, or categorical. Understanding data types is crucial for ensuring appropriate handling and analysis of the data.

A screenshot of a computer

Description automatically generated

1. **Value Patterns and Distribution:**

Identify patterns and distributions of values within each attribute to gain insights into data characteristics and behavior, so that we can consider which features to consider.

A screenshot of a computer

Description automatically generated

1. **Accuracy:**

Data inaccuracies can lead to misleading analysis results and incorrect conclusions. So, we need to know how well data values reflect the real-world entities or phenomena they represent.

A screenshot of a computer

Description automatically generated

1. **Structure discovery:**

Data profiling is focused on understanding dependencies that delineate relationships among different columns within a dataset.

A diagram of a crime fact table

Description automatically generated

## Data Cleaning – In Google Cloud DataPrep

Data cleaning stands as a cornerstone in the data analysis process, serving to refine raw datasets by addressing errors, inconsistencies, and inaccuracies. In our endeavor, Aadishree and Rakesh with the team navigated this critical phase utilizing Google Cloud DataPrep, a cloud-based data preparation tool renowned for its capacity to simplify and automate such processes, thereby enabling users to efficiently manage and transform data at scale.

Our initial step involved importing the raw crime dataset, sourced from local files in .csv format, into the DataPrep environment, priming it for examination and refinement. Within this framework, our focus was on rectifying quality issues prevalent within the dataset. This included a comprehensive audit to identify and rectify missing values, duplicate records, inconsistent formatting, and outliers. Notably, we encountered several columns that, upon closer inspection, revealed limited relevance to our analytical objectives while being riddled with missing data. Consequently, we opted to remove these columns, streamlining the dataset for enhanced clarity and utility.

Additionally, our efforts extended to standardizing data formats to ensure consistency and usability. For instance, the 'Date Occurred' column underwent a transformation to convert it into a more accessible datetime format, facilitating easier interpretation and analysis. However, despite our concerted efforts, we faced challenges in handling the 'Time Occurred' column, which originally presented data in a 24-hour military time format (for example 1800, 2200, etc.). Attempts to convert this format into the conventional HH:mm:ss format proved to be arduous and impractical for querying purposes, ultimately leading us to make the decision to remove this column from the dataset. While this decision required careful consideration, it was deemed necessary to preserve data integrity and usability for subsequent analytical tasks.

Throughout this meticulous process, Aadishree and Rakesh leveraged the capabilities of Google Cloud DataPrep to streamline and expedite data cleaning tasks, ensuring that the resulting dataset was refined, structured, and primed for further analysis. By meticulously addressing quality issues and standardizing data formats, we laid a solid foundation for subsequent analytical endeavors, facilitating more accurate insights into the underlying data.

## Steps taken in Data Cleaning

Recipe: (Using Google Cloud DataPrep)

A screenshot of a computer

Description automatically generated

1. **Consistency:**

Standardize formats and resolve conflicting information.

A screenshot of a phone

Description automatically generated

1. **Completeness:**

Incomplete data may contain missing values or mismatch values and lack certain necessary attributes.

A screenshot of a computer

Description automatically generated

1. **Accuracy:**

Accuracy in data cleaning refers to the degree to which the data reflects the true values or information it is intended to represent.

A screenshot of a graph

Description automatically generated

1. **Cardinality**:

* Number of unique values in a column. Analyzing cardinality helps in understanding the diversity and distribution of data within a column.
* High cardinality columns may contain mostly unique values and might be suitable for indexing, while low cardinality columns may have a limited set of values and could potentially be candidates for categorization.

A screenshot of a graph

Description automatically generated

## Queries:

Application used: BigQuery

Worked by Aadishree and Muskaan

Steps taken:

1. Retrieving the count of crimes for police area name

A screenshot of a computer

Description automatically generated

1. The Most Common Type of Crime Committed

A screenshot of a computer

Description automatically generated

1. Number Of Crimes in Every Month Of The Year 2021

A screenshot of a computer

Description automatically generated

1. Retrieve The Number of Crimes Reported In Each Location Type

A screenshot of a computer

Description automatically generated

1. Retrieving the count of crimes grouped by their status descriptions.

A screenshot of a computer

Description automatically generated

1. Average age of victims in crimes

A screenshot of a computer

Description automatically generated

1. Total count of crimes involving female victims, grouped by each location.

A screenshot of a computer

Description automatically generated

## Data Uncleaning:

## As part of a unique project aimed at testing data cleaning methodologies, Ashwith and Muskaan were tasked with intentionally introducing errors and inconsistencies into a clean dataset. This exercise aimed to provide a hands-on learning experience for the team, allowing them to better grasp the challenges associated with data cleaning and the significance of maintaining data quality. Leveraging Tableau Prep's diverse range of data manipulation capabilities, the team embarked on systematically degrading the quality of the dataset across various dimensions. Drawing upon their prior experience with Tableau Prep from class assignments and tutorials, the team felt confident in their ability to execute this task effectively. The approach adopted involved deliberately undermining several key dimensions of data quality, including consistency, validity, accuracy, completeness, conformity, and uniqueness. To achieve this, a series of intentional manipulations were performed on the clean dataset. These manipulations included mismatching column values, inserting null values, adding duplicate entries, introducing negative values, altering column data types, removing essential columns, and incorporating extraneous ones. Furthermore, to exacerbate discrepancies, the team established a data quality rule, only to intentionally violate it. Through this deliberate introduction of errors and inconsistencies, the team aimed to simulate real-world scenarios where datasets are often plagued by various quality issues. By experiencing firsthand the process of degrading data quality, the team gained valuable insights into the complexities involved in data cleaning and the importance of maintaining high standards of data quality. This exercise not only enhanced their understanding of data cleaning methodologies but also underscored the critical role of data quality in driving reliable and meaningful insights from datasets.

## Steps in Data Uncleaning:

Recipe: (Using TableauPrep)

A diagram of a diagram

Description automatically generated with medium confidence

1. Affecting Completeness:

-Adding the null values to different fields

A screenshot of a group

Description automatically generated

1. Affecting Uniqueness:

-Adding a duplicate field to reduce the uniqueness

A screenshot of a computer

Description automatically generated

1. Affecting Validity:

-Adding negative value in age field

A screenshot of a computer

Description automatically generated

1. Affecting Accuracy:

-Making sure the column name does not accurately describe the data present under its rows.

-making sure the data present in the row does not accurately represent as to what comes under its Column

A screenshot of a computer

Description automatically generated

1. Affecting Consistency:

-by changing data types.

A screenshot of a computer

Description automatically generated

## What we cannot implement in Data Uncleaning–

What is not valid according to the given dataset.

* Currency – the degree to which the data is current in the column.
* Timeliness – when the data is readily available for use.

## 

## Conclusion:

The cleaning process provides critical insights that can guide the allocation of resources and focus law enforcement efforts.

Conversely, the uncleaning process underscored how data quality issues could obscure these insights.

This exercise underscores the importance of using tools like GCP DataPrep, BigQuery, and Tableau Prep for rigorous data management.

## 

## References:

<https://catalog.data.gov/dataset/crime-data-from-2020-to-present>

https://ichef.bbci.co.uk/news/976/cpsprodpb/C428/production/\_101661205\_polgerafp30jul17.jpg.webp