CS513: Theory & Practice of Data Cleaning
Project Phase - I

Sim Bhamra (sbhamra2@illinois.edu)
Albert Yeh (ayeh2@illinois.edu)

The goal of this project is to apply the data cleaning concepts and practices we learned during class to crime data from the city of Denver, Colorado. Both team members are excited about using this dataset as Sim pursued a bachelor's degree in criminal justice and Albert is interested in how crime changes over time. Also, this data set as it is a great candidate for data cleaning activities is substantial in size and can provide us with useful crime insights which our team is interested in.

Dataset:

The Denver City crime data is part of an open data catalog provided by the city government and collected by the Denver Police Department. The data can be found at this location: https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime.

For this project we will be using two of the tables provided, the first is the crime table which contains all criminal offenses in the city and county of Denver from January 2017- June 2022. The second is the offense code table, this can be used to cross-reference the offense code given in table 1 to find the actual crime name according to the National Incident-Based Reporting System or NIBRS. Finally, we will be using the NIBRS crime type definitions to help further categorize the crimes.

Use Cases:

Given this dataset and our interest in understanding, crime trends our main use case will be to show the change in each type of crime according to the NIBRS definition in the last 5 years in the city and county of Denver. In order to do this, after cleaning the data, we will perform various analytic tasks such as averaging and comparing change according to population across the years and crime types.

A minor use case that can be completed with no data cleaning at all is to determine the number of traffic crimes committed in this time period. Another use case would be to determine the most and least common crimes in the time period provided.

On the other hand, a minor use case that is unrealistic is a detailed analysis of temporal data relating to the crime, this includes crime duration and the amount of time between the incident and the report. This is due to the fact that there is a large amount of missing data for the “last_occurrence” column.

Description of Dataset:

The Denver City Crime data includes criminal offenses in the city and county of Denver for January 2017- June 2022. The crime data is based on the NIBRS and withholds all crimes identifying sexual assault and child abuse victims and juvenile perpetrators, victims, and witnesses of certain crimes. The offense data includes the offense code used by the police department of Denver and characteristic details for that offense code.

Using pandas in python we were able to explore the data in some detail, the notebook and code for this can be found at:
https://github.com/simranBhamra/CS-513-Final-Project/blob/main/Dirty_Data/Data-Exploration.ipynb

Crime Dataset:

This dataset includes a total of 20 columns and 376812 rows. The table below shows the type and description for the columns of the crime data.

Column	**Data Type**	**Description**
INCIDENT_ID	integer	Identifier for the incident, key
OFFENSE_ID	integer	Identifier for the offense, key
OFFENSE_CODE	integer	Identifer for the offense type
OFFENSE_CODE_EXTENSION	integer	extension assigned to each offense, from the police department
OFFENSE_TYPE_ID	object	type of offense
OFFENSE_CATEGORY_ID	object	category of the offense
FIRST_OCCURRENCE_DATE	object	date and time of the start of the offense
LAST_OCCURRENCE_DATE	object	date and time of the end of the offense
REPORTED_DATE	object	date and time the offense was reported to the police
INCIDENT_ADDRESS	object	street address of the incident
GEO_X	float	x coordinate of the location of the offense
GEO_Y	float	y coordinate of the location of the offense
GEO_LON	float	longitude of the location of the offense
GEO_LAT	float	latitude of the location of the offense
DISTRICT_ID	float	district where the offense occurred
PRECINCT_ID	float	precinct id where the offense occurred
NEIGHBORHOOD_ID	object	neighborhood id where the offense occurred
IS_CRIME	integer	0 or 1 indicating if the offense was a crime, 1 meaning yes

IS_TRAFFIC	integer	0 or 1 indicating if the offense was traffic-related, 1 meaning yes
VICTIM_COUNT	integer	the number of victims involved

In addition to this, we looked at the amount of missing data. The following table shows the number of missing values for each column in the dataset.

Column	**Number of Missing Values**	**Percentage of Missing Values**
INCIDENT_ID	0	0.00%
OFFENSE_ID	0	0.00%
OFFENSE_CODE	0	0.00%
OFFENSE_CODE_EXTENSION	0	0.00%
OFFENSE_TYPE_ID	0	0.00%
OFFENSE_CATEGORY_ID	0	0.00%
FIRST_OCCURRENCE_DATE	0	0.00%
LAST_OCCURRENCE_DATE	183996	48.80%
REPORTED_DATE	0	0.00%
INCIDENT_ADDRESS	5267	1.00%
GEO_X	5267	1.40%
GEO_Y	5267	1.40%
GEO_LON	5267	1.40%
GEO_LAT	5267	1.40%
DISTRICT_ID	584	0.20%
PRECINCT_ID	584	0.20%
NEIGHBORHOOD_ID	584	0.20%
IS_CRIME	0	0.00%
IS_TRAFFIC	0	0.00%
VICTIM_COUNT	0	0.00%

Offense Dataset:

The offense dataset has no missing data and includes 9 columns and 300 rows. We will be using this data as a reference so we can give the crime data the appropriate names for the crime. The table below shows the type and description for the columns of the offense data.

Column	**Data Type**	**Description**
OBJECTID	integer	identifying the offense
OFFENSE_CODE	integer	code identifying the type of offense
OFFENSE_CODE_EXTENSION	integer	extension assigned to each offense, from the police department
OFFENSE_TYPE_ID	object	defines the type of offense
OFFENSE_TYPE_NAME	object	the full name of the offense
OFFENSE_CATEGORY_ID	object	the full name of the offense type
OFFENSE_CATEGORY_NAME	object	the full name of the offense category
IS_CRIME	integer	0 or 1 indicating if the offense was a crime, 1 meaning yes
IS_TRAFFIC	integer	0 or 1 indicating if the offense was traffic-related, 1 meaning yes

Data Quality Issues:

As mentioned above the offense dataset will be used as a reference to add the offense name and standarized the offense to the NIBRS. There are no data quality issues present in the offense dataset. There are a couple of data quality issues present in the crime data such as missing values, incorrect types, formatting issues, incorrect addition of special characters, and columns being combined. The following is the general description of the data quality issues that exist.

- First Occurence Date
    - Type is incorrect
    - The column includes both the date and time as an object
    - ex) 1/4/2022 11:30:00 AM
    - Not broken up by year but instead all one long list
- Last Occurence Date
    - Type is incorrect
    - The column includes both the date and time as an object
    - ex) 1/4/2022 12:00:00 PM
    - Almost half of the data is missing
    - Not broken up by year but instead all one long list
- Reported Date

    - Type is incorrect
    - The column includes both the date and time as an object
    - ex) 1/3/2022 11:01:00 AM
    - Not broken up by year but instead all one long list
- Offense Type ID, Offense Category ID
    - The crimes are not standardized to the NIBRS
    - There is no full name given to the crime, so it isn't clear what the actual offense is
- Incident Address
    - This only includes the street name not the full address
    - Column name should be adjusted or the rest of the address such as city, state, and zip should be added
    - Some entries have special characters included such as /
- Neighborhood ID
    - Some of the names could be the same for example university vs university hill
    - Some entries look like they could be garbage values, such as cbd
- Offense Type ID , Offense Category ID, Incident Address, Neighborhood ID
    - Type should be a string instead of object
- District ID, Precinct ID
    - Type should be integer instead of float

Methodology:

To achieve our goal we intended to use primarily Python, OpenRefine, and YesWorkflow. The final product should include seven tables, the first six tables will be the crime data cleaned and split by year whereas the seventh table will include the percentages of change and summaries of all 5 years together. Our initial plan is as follows.

1. Complete data exploration using Python.
2. Using OpenRefine and Regex or Python and Regex split the date and time values in the first occurence date column, last occurence date column, and reported date column.
3. Standardize the new split columns to date and time ISO types respectively.
4. Remove the two new last occurence date columns as over half the data is missing.
5. Convert offense type ID , offense category ID, incident address, and neighborhood ID to strings and remove leading and trailing whitespaces.
6. Remove special characters from incident address column.
7. Adjust address column name or add city, state,and zip to make a full address.
8. Consolidate neighborhood ID column and remove garbage values.
9. Join offense table to crimes table using the offense code.
10. Remove redundant columns that were added during the join.
11. Check integrity constraint violations using Python.
12. Split table into separate tables according to year of first occurence date column.
13. Calculate offense type totals, and percentages for each year using Python and store in new table.
14. Create final YesWorkflow workflow diagram.

Assignment of Tasks:

Sim	**Albert**
Data exploration	Fixing data types for offense type ID , offense category ID, incident address, and neighborhood ID
Data cleaning for first occurence date column, last occurence date column, and reported date column	Data cleaning for address column
Joining tables and removing redundant columns	Consolidation for neighborhood ID
Table splitting and formatting	Integrity constraints
Calcuate summaries for final table	Calcuate summaries for final table
YesWorkflow diagram	YesWorkflow diagram
Final report	Final report