Team Ottawa

Deliverable 1 - Data Design

Contributors:

Alex Thiele
Mark Waters
Anirudh Sodhi

Section: Data Set 1 (Crime CSV file)

Brief Description of the Data:

The CSV file has seven columns of information that it is separated into. The first column is labeled Ref_Date and shows the date at which the crime was either recorded or committed. The second column is labeled GEO and is the geographical location at which the crime occurred. Most of these are general, such as Canada, but some of them are very specific, such as St. John's, Newfoundland and Labrador. The third column is labeled VIOLATIONS, and it states the general category for what each statistic falls into. An example of what would appear in this column would be Total, all Criminal Code violations (excluding traffic). The fourth column is labeled STA and is the specific statistic of the general crime that was committed. An example would be Rate, youth charged per 100,000 population aged 12 to 17 years. The fifth column is labeled Vector, and it consists of values that start with a v followed by a series of 8 numbers. These numbers change based on the column STA so it is assumed that it is a reference to that statistic in that violation in that location in that year. The sixth column is labeled Coordinate and it consists of three numbers, each separated by a period. The first number corresponds to the geographical location, the second number corresponds to the type of violation, and the third number corresponds to the type of statistic. The seventh column is labeled Value and it is the information that is related to each specific statistic and each specific crime.

The CSV file starts with Canada, but there does not appear to be a way that the geographical locations are organized, with the exception being that cities in the same province are consecutive to each other. An example is in the file, New Brunswick comes, first, then immediately after that set of geographical data is finished, Moncton, New Brunswick is the next geographical location.

As mentioned above, the coordinate value has three numbers, each separated by a period. The first number changes with the geographical location, and changes the least frequently. The second number changes along with the violation, and changes the second most frequently. The third number changes with the statistic, and changes the most frequently.

Another piece of information that was discovered in this CSV file was the fact that no year went below 1998, and no year went above 2015.

Data Formats and Organization:

The data is organized into seven separate columns, each with its own specific value. It is formatted in a csv (comma separated version) file.

Data Encodings:

The vector column is an encoded version of the specific violation and statistic. The coordinate column is an encoded version of the location of a specific piece of data in the file.

Useful Fields:

Ref_Date field is useful in narrowing down crime data results to a specific year. Vector field can be used to find a specific statistic value for the type of violation. The Coordinate field can be used to find a specific record in the entire dataset as each digit of the coordinate corresponds to a column in the dataset. The value fields is useful in finding records that have values within a specific range or higher/lower than a specific value.

Data Reorganization:

Grouping and Aggregations:

The way the data is organized means that in order to process any type of aggregation or grouping would require going through each record in the data set exactly once.

The data can be aggregated or grouped by the following fields (or a combination of the following fields). The first one would be the year, and for this you would print out the entire dataset but only for that specific year. The second field would by statistic, and for this you would print out a specific statistic for every type of violation.

The coordinate values are useful in aggregating data by violation and statistic as the last two digits of the coordinate correspond to these two specific fields.

Encodings:

A way to encode the data better would be to add an extra digit to the end or the beginning of the coordinate value that would correspond to the year.

The vector encoding process could be described at the beginning of the CSV file, to enhance readability.

Section: Data Set 2 (Census CSV)

Brief Description of the Data:

The Census data, available for years 2011, 2006, and 2001, shows population data for different geographic locations in Canada. The data is available for the entire country, individual provinces, cities, economic regions, census tracts, and several other categories.

Each file contains total population numbers, age groups, marital status, family status, mother tongue, and language spoken at home. There are around 90 non-Aboriginal and 9 Aboriginal languages listed. According to StatsCan, the number of languages included was intentionally limited to only the most popular ones.

Three CSV files in a folder are downloaded each time a location is specified. The first file is a short index file that enumerates the geographical codes for each location. Other flags are

described, including a data quality flag, and error flags. The second file is the main CSV containing the data. The third file includes supplementary notes, citing information, and symbol descriptions.

Data Formats and Organization:

The data is organized into 11 comma-separated columns in a CSV file. A single geographical area is covered at a time, with the data being further sectioned off into Topics and Characteristics. These sections are consistently organized from left to right.

Data Encodings:

Each unique geographical location is assigned its own encoding, which are described in the first "index" CSV and used in the main CSV file. The method of encoding differs from file to file. For example, when downloading census data for all of Canada or for "Economic Regions", geographical locations are listed as integers. However, the geographical codes for "Forward Sortation Areas" are alphanumeric.

Where applicable, notes are encoded by integers ranging from 1 to 23. These notes are explained in detail within the third (metadata) CSV file.

Useful Fields:

In the main CSV file, encoding the notes with an integer saves space and avoids needless repetition of the sentence describing each note.

Descriptive fields such as language (English, French, etc) are repeated for each geographical location. This takes up more space, but makes the document more readable than if each category were encoded with an integer. At a glance, no matter where the reader is in the file, they can see the pertinent information without having to refer to an encoding table.

Characteristics further organized by sub-characteristics are tabbed, so that a reader can quickly determine the level of detail by looking at the indentation.

Data Reorganization:

Grouping and Aggregations:

- Processing would require going through each record in the dataset once.
- Data can be aggregated/grouped by the following fields (or combination of the following fields):
    - Topic: print out all records matching a specific topic for each location.
    - Total: aggregate records for specific Topic that have a Total value within a specific range.
    - Languages could be aggregated for multiple locations (for example, to see language differences in the east coast vs west coast)

Encodings:

Each unique Topic and language could be assigned its own integer, similar to how each unique location is assigned a unique integer in the Geo_Code field. This would be useful if compressing the data to as small a size as possible were a priority.

Section: Questions

1. What is the concentration of a specific type of violation in a particular area?
    - This question can be answered by using:
        i. Geo field to match the location
        ii. Violations field to match the crime
        iii. Statistic value of "Rate per 100,000" to find the concentration

2. Is the crime rate rising faster than the population growth rate in a specific range of years?
    - Data for "population and dwelling counts" from the census dataset and the data for the statistic "percentage change in rate" from the crime dataset can be used to answer this question.

3. What percentage of youth (or adult) offenders actually get charged for the crime?
    - Variations to this question could include:
        i. Is this number higher (or lower) for a certain type of violation?
        ii. Is this number been increasing/decreasing over the years?
        iii. Is there significant variation in this statistic among multiple geo locations?
    - This question can be answered by using the values for the statistics: "Total, youth charged", "Total, youth not charged", "Total, adult charged", "Total, adult not charged"

4. Which violations are the most common among youth (or adult) offenders?
    - Same data used in question 3 can be used to answer this question as well.

5. Is there a correlation between population decrease and crime increase in any particular area?
    - Same data used in question 2 can be used to answer this question as well.

6. Is there a correlation between population density and the rate of crime in an area?
    - This question can be answered by using:
        i. Geo field to match the location
        ii. Total violations field
        iii. Additional data would be required for the size of an area in sq km, in order to calculate population density

7. Which kinds of crimes correlate with certain age groups in the population?
    - This question can be answered by using:
        i. Violations field to match the crime. Aggregating crimes into groups (violent, theft, fraud, drug, etc) may be useful.
        ii. “Total, youth charged” and “Total, adult charged”

8. What is the demographic of the population in areas with the highest (or lowest) crime rate?
    - For each location, the data for “Age characteristics”, “Marital status”, “Family characteristics”, and “Household and dwelling characteristics” from the census dataset can be compared to the data for “Total, all violations” and “Actual incidents” from the crime dataset.