Description of Data

This data is a compilation of comma separated values that contain information regarding all the different types of violations committed across Canada, the provinces and territories and certain major cities.

Format and Organization

The data is organized with 7 columns and multiple rows that correspond to different years for numerous cities, provinces and Canada as a whole.

The first column in the data is the reference date for which the value is a certain year. The data is available for the years 1998 – 2015.

The second column is for geography. The geography varies beginning with statistics for the whole country, entire provinces and certain major cities across Canada. Every province is mentioned as a whole at least once but not all cities are represented. Throughout the document there are 48 different geographical regions represented.

The third column contains information regarding the type of violation. There are 255 categories for the different types of violation. This includes specific and general types of violations.

The fourth column is the type of statistic there is information for. There are 14 different types of statistics.

The fifth column is the vector number. A vector number is a unique 8 digit number preceded by a ‘v’ that identifies the particular combination of geography, violation and statistic.

The sixth column contains a coordinate. This coordinate is 3 numbers separated by periods such as 1.1.1. The first number in the coordinate contains an encoding for the geography. The second number in the coordinate contains an encoding for the type of violation. The third number in the coordinate contains encoded number for the type of statistic. Each row has a unique coordinate associated with it’s geography, violation and statistic.

The seventh column in the data is a value. The value is the number associated with the information within the row.

Encodings

The sixth column of the data is a coordinate that is encoded to contain information for each combination of geography, violations and statistics. The coordinate is formatted like this: a.b.c where a, b and c are numerical digits. The number in place of a is representative of the geographical location. For example, the number 1 represents Canada. There are 48 possible geographical locations. The number in place of b is representative of the type of violation. There are 255 possible violations that can be represented here and each number corresponds with a different type of violation. The last part of the coordinate, c, would represent the type of statistic being used. There are 14 different types of statistics referenced.

Useful Fields

Not all the fields are particularly useful for the purpose of inferencing the data. For examples, the fifth column, which contains the vector does not provide any valuable information and is only used to identify the particular combination of geography, violation and statistics. Therefore, it is not necessary to use the vector value. Furthermore, since the sixth column, the coordinate, contains encoded values for the geography, violations and statistics (the second, third and fourth columns) we can do without those as well. Simply using the coordinate values gives us the information provided by the second, third and fourth columns.

Ultimately, the columns that will be most useful for the purpose of inferencing this data would be the first, sixth and seventh column which contains the year, coordinate and value respectively.

With regards to the statistics, it would be most accurate to look at only

Data Re-Organization

Groupings

The data can be grouped in various manners. Firstly, we can group the data by year to compare different statistics across different geographical locations per year. Secondly, we can group the data by provinces and compare data across different provinces as a whole and ignore the data for major cities. Thirdly we can group the data based on the type of violation we want to look at such as violent crimes versus sex crimes versus theft/fraud etc.

Aggregations

Encodings

Since the coordinate already encodes the geography, violations and type of statistic and the only other useful information is the year and value. It is not useful to encode the value since there is a wide variety of numbers the value can take on. However, it is possible to encode the year. Since the information in the data is for the years within 1998 to 2015, there are 17 possible values the year can take on. To reduce the amount of space taken up, the year can be encoded as a single digit representing a year. For example, the year 1998 can be encoded as the value 1, 1999 can be encoded as 2 and so on.