CIS2250 Team Deliverable One

Team Toronto

Kanza Shams

Parth Patel

Israa Sinan

**Section One: Crime Data**

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. The data contains information from 1998 to 2015 and has information regarding all the different types of violations and their frequency. There are many different types of statistics available to interpret the data including actual incidents of each crime, rate per 100,000, percentage change in crime and also statistics specifically for crimes committed by adult and crimes committed by youth.

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 is the year that the information was collected for. The data is available for most geographical regions between the years 1998 – 2015 but it varies. The second column is for geographical location. 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 or capitals 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. For example, there is data for violent crime (general) and also first degree murder (specific) which also falls under all violent crime. The fourth column is the type of statistic there is information for. There are 14 different types of statistics. For each geographical location and violation there are 14 different statistics. For examples, one of the stats is “actual incidents” of the given violation at the geographical location but there is also statistics such as “rate per 100,000” and 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 in this data set.

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 rate committed per 100,000. Depending on the questions we want to answer and the relationships we want to identify, data for certain violations or certain cities might not be entirely necessary.

Data Reorganization

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.

Aggregate

We can aggregate the information based on type of violations. For example, if we wanted to look at violent crime across different geographical locations we can aggregate the data based on that. Also, it is possible to aggregate the data based on type of statistics. For example, the most accurate way to compare data across geographical locations would be to look at the statistic that is “Rate per 100,000 population” so we could aggregate all the data for the “Rate per 100,000 population” and ignore the 13 other statistics. Similarly, if we wanted to compare data for crimes committed by youth and crimes committed by adults we could aggregate the data based on that statistic to compare youth crime versus adult crime.

Encodings

The coordinate are already efficiently encoded for 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.

**Section Two: Population Centers**

Description of Data

Population Centers are described as an area with a population of at least 1,000 and no fewer than 400 persons per square kilometre. Population Centers are also called Urban Areas. All areas outside the population centres are called rural areas.The data set is organized by commas that separate the values that contain information regarding all the different characteristics in a household across Canada. There are numerous categories that encompass the characteristics of a household such as family structure, language spoken, type of housing and age characteristics. The data has information for the total number of people in each population center that fulfill a characteristic and goes further to divide the total number per category to total males and total females for appropriate categories.

Format and Organization

Population Centres in general is divided into three sections. The first section being small populations centres that have a population between 1000 to 29,999 people. The second section being the medium population centres that has a population of 30,000 to 99,999. Lastly, the third general section which is large urban population centres that contains 100,00 or more people. However, the data is organized in twelve columns. The first is the geographical code that represents the encoding of the geographical name. The second column gives the geographical name that was encoded in the first column. The third column is the population centeral that the data was collected from. The fourth column is the topic that the data carries. The topic may relate to age characteristics, population and dwelling counts, family characteristics, marital status, mother language, etc. The fifth column gives the specific characteristics or the subtopic, this includes the actual age groups, languages, number of people in a household, etc. The sixth column is the note section. In this column you are given a number that refers to an encoded message that is in another file which helps you understand more about the topic. The seventh column is the total amount of people that took part in the census. The eighth column is for any flags regarding the total count. The flagged columns are used to write about any quality control methods that might have affected the data. The ninth column is the amount of men that participated in the census. The tenth column is the number of flagged men. The eleventh column is the number of females that participated in the census. The twelfth column represents the number of flagged females.

Encodings

Encodings are within this data set many times. Firstly, the two digit geographical code that represents the geographical name of the province or territory in Canada. Secondly, the two digit code that represents the province or territory in Canada, can be equated to a four digit code that represents the population centres areas. Furthermore, some columns contain three dots [...]. This is meant to signify that there is no information available for that particular category since it is not applicable. Lastly, is the noted encodings that represents more information about the topic that the census data has been collected for.

Useful Fields

Most of the fields in the data set have inputs that help understand in which direction the country is going. However, there exists unused and unnecessary data sections in this data set. The eighth, tenth, and twelfth columns that represent the total flagged, males flagged, and female flagged are left without input. The unnecessary field in this data set is the population centers name. The reason for this is because the encode for the name has already been given to reduce the amount of columns and make the data set more understandable.

Data Re-Organization

Groupings

The data can be grouped in two ways. Firstly by age group. Age groups can show the cities that are more likely to be populated with crime. Age groups can be in different numbers. We may have ages 1-15 years olds, 16-18 years olds, and 19 years olds in different category age groups to show how the difference of each age group functions. Secondly by family characteristics. Family characteristics can include the size of the family, the type of house, the mother language spoken at home.

Aggregate

We can aggregate the information in the data set by looking at family characteristics. For example, you can group the household languages, types of housings and how many family members live in it. This data set is aggregated carefully from statistics canada and it can not be aggregated any further.

Encodings

The population centres could be encoded in a more strategic way to minimize the number of columns. For example, the data could be encoded with single digit numbers to represent the topic, and in this way the notes column could removed.

**Questions**

1. Is the crime rate higher in a larger populated location rather than a denser populated area?

The data that can be used to answer this question from the first data set (the crime data) includes statistics for the year 2011 (to match the census information) for all crimes and their rates committed per 100,000 people. We can look at this information for major cities in the crime data and compare it to the population density per square kilometre in the cities from the population center data set.

1. Is a single individual more likely to commit a crime than someone in a relationship?

To answer this question we can use the data from the second data set that has information on the marital status of people in many cities. There is data from the number of married or living with a common-law partner in each city as well as the number of not married and not living in a common law partner. We can gather data for each major city for this information and then compare it with the crime rates of the major city to see if there is any relationship between relationship status and crime.

1. Is it easier for someone living all alone to commit a crime than it is for someone with family?

To answer this, we can use the information in the census data regarding household and dwelling characteristics which tells you the number of one person households and two or more person households. We can look at the information for household characteristics for major cities and compare the crime rates in those major cities to see if there is correlation between how people live and crime rates.

1. Are more crimes committed by immigrants?

To see if there is a correlation between immigrants and crime we can look at how many people in selected cities speak non-native languages (languages other than French or English) in their household and compare that information to the crime rates of those cities.

1. Are youth more likely to commit a crime than adults?

We can compare the population of youth in each major city with the rate of youth crime in each city. To enhance this analysis we can compare the population of adults in each major city with the rate of adult crimes committed in the city. This will help us see if there is a stronger relationship between adult population and adult crime or stronger relationship with youth population and youth crime.

1. Is there a specific nationality that is more susceptible to crime than others?

To see if there is any relationship between the nationality of a person and their susceptibility to commit crime we can look at the data from the second set that provides information regarding “Detailed language spoken at home”. We can see if there is a particular language that is correlated with higher rates of crime across major cities.

1. Are people who live in apartment buildings more susceptible to being victims of crime?

To answer this question we can look at the dwelling characteristics for major cities provided in the second data set. It shows the number of people living in apartments or different types of houses (detached, semi-detached etc). We can compare how the number of people living in apartments is related to the number of crime within cities.

1. Does a lower population of children in a location cause a decrease in crime?

To answer this question we can look at the population of children aged 1-17 in particular cities which can be found in the second data set. We can compare the number of children in the cities with the rate of crime in those same cities. We can look at how population of children affects adult crime rates, youth crimes rates and crime rates in general.