**Project Data Design**

**Team Brockville**

Ahmed Mahmoud - 0966966

Joshua Prpic - 0971107

Alex Montague - 0959687

**Table of Contents**

**1. Data Set One (Crime Stats)**

**1.1………………………….Brief description of the data**

**1.2………………………….Data formats and organization**

**1.3………………………….Data encodings**

**1.4.………………………....Useful fields**

**1.5………………………….Data re-organization**

***1.5.1………………..Groupings***

***1.5.2………………..Aggregations***

***1.5.3………………...Encodings***

**2. Data Set Two (2001 Census Profile)**

**2.1………………………….Brief description of the data**

**2.2………………………….Data formats and organization**

**2.3………………………….Data encodings**

**2.4.………………………....Useful fields**

**2.5………………………….Data re-organization**

***2.5.1………………..Groupings***

***2.5.2………………..Aggregations***

***2.5.3………………...Encodings***

**3. Possible Questions based on the Data Sets**

**1. Data Set One (Crime stats)**

**1.1 Brief description of the data**

This data is the collection of information regarding crimes committed in Canada over a period of time. It is a sorted record of all crimes committed with specific statistics on them like rate per 100,000. It only contains the information for the years 1998 to the year 2015.

**1.2 Data formats and organization**

The data is formated in a organized in a manner described at the start of the document.

It is formatted into 7 columns:

* + Ref\_Date: The reference year
  + GEO: The location the statistic is referring to
  + VIOLATIONS: The crime committed
  + STA: The statistic on the crime
  + VECTOR: The documents arbitrary reference number to a specific statistic in that specific city
  + Coordinate: The coordinate is an encoded version of the location, crime and statistic column (Share the same function as the vector but is not arbitrary)
  + Value: The actual data point

As far as the subgrouping go it is first organized by the the city, then for that city they list all the crimes and their respective statistics that go along with it. These statistics are then organized by the year so the earlier year would be higher than the years following it. So for every statistic, there would be a list of years with the information relating to that stat in the given year.

**1.3 Data encodings**

The data utilizes some forms of encoding. It uses a coordinate system for encoding. However, it leaves much of the information in plain text storage. The coordinate encoding is simply representing the field information (location, crime and statistic) in a short hand numerical representation. Where the first number is the location, the second is the crime and the third is the specific statistic relating to the crime.

**1.4 Useful fields**

There are many useful fields within the crime data. The data format has various fields as shown above. The most useful fields are the geographic region, violation, statistic, value, and coordinate fields. The geographic region field is an indicator of where the data encompasses. This allows the data to be concise based on where and also can be used to determine which locations carry certain trends. Next, the violation field simply represents which violation the is associated with the record. This is also a very important field as it is in essence what describes the crime associated with the value. Then there is the statistic field which describes what statistic the value will represent. This is also very useful as many of the commonly used statistics that we may need to calculate are already in the statistic field. Moreover, the value field is arguably the most important and useful field as it is the data point itself. The value is what we will use in accordance with these other fields in order to make statements or answer different questions about crimes in Canada. Finally, the coordinate field contains the coordinate encoding which can be used, as later explained, to save storage space so it is also an important field.

**1.5 Data re-organization**

**1.5.1 Groupings**

The groupings are quite sufficient when it come to organizing the information, it clearly groups all of the information given in a logical manner. The only grouping that could be removed is the vector column as it serves the same function as the coordinate but instead of a number relative to the data like coordinate uses, it uses an arbitrary number that has no relation to the data. Therefore it is useless and can be removed.

**1.5.2 Aggregations**

The aggregation of the data could be set up in many different ways depending of what is needed from the data. For example depending of the purpose of the data it could be aggregated to be shown by crime instead of by city, this would require a fair chunk of processing as there is a fairly large chunk of data.

**1.5.3 Encodings**

Seeing as the coordinate contains all the information in the location, crime and statistic column, you could remove these three columns entirely in favour of the coordinate and a table to show what the number in the coordinate represent. This in turn would save a large amount of storage space as in the crime stats documents every crime is restated as well as the coordinate.

**2. Data Set Two (2001 Census Profile)**

**2.1 Brief description of the data**

This is the 2001 Canadian Census profile that contains a variety of information throughout the census collection period. There is data on topics from the “Highest Level of Schooling” all the way till “Health Status”. Many topics are split up into sections like age-groups or levels called characteristics and is then further split into counts based on gender which gives further information about the topic at hand.

**2.2 Data formats and organization**

The data is formatted well and is generally quite organized. The data’s format is listed at the top of the CSV file as:

1. Geo\_Code: An encoding for the region that is associated with the statistic
2. Prov\_Name (Region\_Name more broadly): The province or region
3. Topic: Simply the topic that the statistic represents
4. Characteristic: The different sections of the topic (i.e. age groups)
5. Note: Any notes about the data is given here
6. Total: This is the total value of the statistic
7. Male: The male count value of the statistic
8. Female: The female count value of the statistic
9. Flag: A flag used to give information about the statistic

The data is organized is a very straightforward way. Each column is filled in accordance with its column heading and any unknown or non-applicable values are then left empty. Additionally, the data is chunked based on the Geo\_Code, and thus, based on the region that the value is associated with. Every region has the same topics with the same characteristics in the same order. This allows for extraction of exact information based on simple arithmetic to get the location of the data that is wanted.

**2.3 Data encodings**

In the 2001 Census there are two types of encoding present in the data. The Geo\_Code column has a specific value for each region and by taking in the values and using an accompanying table with all the associated regions the Geo\_Code can be used to encode the region and would save storage space. They also encode the notes column, each note has a specific number assigned to is so in the actual data there is only a number provided when referencing notes, this saves storage space as it prevents the repetition of long generic notes about the information.

**2.4 Useful fields**

For the most part all the fields within the 2001 Census Profile is useful in some way. The Geo\_Code encoding in combination with the Prov\_Name (or Region\_Name) represents the region from which the data represents. This is vital to determine where the values are from and how location affects these values. The topic, characteristic and note fields are also extremely important in the organization of the data both in the CSV and when it is utilized. The topic is the matter that the value represents, the characteristic is the subgroup or section of the topic that the value belongs to, and the note field is simply any extra information about the value that may be of importance. These fields are what divide the values into discrete domains so they can be used to derive statements about very specific subjects and characteristics. Finally the total, male, and female fields are the actual data points based on gender. These values are of course useful as they are the count based on the gender, topic and characteristic. All these fields are vital in order to answer questions about the census.

**2.5 Data re-organization**

**2.5.1 Groupings**

The grouping are sufficient in their organization of the data. There are no real significant changes that need to be made to the organization of the grouping that were provided as everything grouped based on the topic and related topics come in subsequent.

**2.5.2 Aggregations**

Changes in the aggregations could be done simply with a program. There would only need to be changes in the aggregation of the data depending on what is being done with the data. It would require a fair amount of processing power to change the aggregation of the data as it is a fairly large amount of data.

**2.5.3 Encodings**

The only real encoding in the 2001 Census data is the Geo\_Code. While this is useful to find the region that the value pertains to it would be more useful to implement an encoding system like in the crime data (data set one). A more useful encoding would take the coordinate form of: “Geo\_Code.Topic” similarly like the crime statistics. This would allow the data to be stored more compactly and instead of having the Geo\_Code and the topic as separate the Geo\_Code and the topic can be stored as one coordinate saving storage space similarly again to how the coordinate encoding in the crime data can be used in order to save storage space.

**3. Possible Questions based on the Data Sets**

**3.1. Does level of education have an impact on severity of crime?**

* + Data on the highest level of education received for that specific area with relation to the severity of crime for the same area. Using the Geo\_Codes and the GEO fields in the Census and Crime data respectively data about the general population’s education can be compared with the number of severe crimes.
  + Ex. “Highest level of Schooling” and “School Attendance” versus the number of high-tier crime values (i.e. Murder of all forms, robbery etc.)

**3.2 Does annual income reflect the frequency of crimes?**

* + Annual salaries for each area and the number of crimes for the same area. Again, using the Geo\_Codes and the GEO fields will allow for finding information based on the specific region. Then by comparing the frequency of crimes to the income of the individuals in the same region.
  + Ex. Province specific “Median Total Income” versus province specific “Total Violations”.

**3.3. Are youth with single/divorced parents more likely to commit a crime?**

* + This would look at the number of youths charged for a crime in a specific province and then compare that to the number of children with single parents for a specific province. You could then look at the same statistic, just for children with married parents.
  + Ex. “Total, youth charged” versus “Children with Lone Parent”

**3.4. Are wealthy persons more likely to commit a tax fraud/evasion crime?**

* + Annual salaries for an area and the number of tax related crimes in that area. Using the Geo\_Codes and GEO fields like the previous questions, we would be able to match the data up based on the specific location. Then you can compare the frequency between tax crimes and income.
  + Ex. Province specific “Median Total Income” versus province specific “Income Tax Act” or “Tax Fraud” charges.

**3.5. How have crimes changed over time (which crimes remained the same)?**

* + This is simply what is considered a violation in every year. The crimes that are new, no longer considered crimes, or have been renamed will have no value and thus using this information the different violations can be tracked over time.
  + Ex. VIOLATION field in the crime data, the value associated with the field (no applicable or not) and the REF\_YEAR field.

**3.6 Which violations are the most common in each province/region?**

* + This is simply what is the violation with the largest incident value for each province over time in the crime data.
  + Ex. The largest crime incident value and its associated violation and the GEO field for each province/region

**3.7 What level of schooling is correlated with what occupation?**

* + By looking at the level of schooling of a population and the number of individuals in each occupation type a conclusion can be drawn about generally what different levels of schooling lead to.
  + Ex. “Highest level of schooling” and the different occupation topics (i.e. industry, unemployment etc.)

**3.8 How does immigration affect the crime rates of a region?**

* + Every region has specific values about immigration and the different crime values. By looking at the immigration values as well as various crimes links can be made between immigration and the decrease or increase of crimes in a region. This information can also be used to find trends about specific crimes and their links to immigration within that region.
  + Ex. The total number of immigrants in a region and any specific crime such as “Murder, first degree” over a period of time

**3.9 Are there differences between crimes in rural vs urban areas?**

* + By looking at the 2001 Census data as it pertains to more specific regions that are split up depending on whether they are rural or urban the data can be compared to the different violations and the values associated with them to see what crimes are more common in rural vs urban areas
  + Ex. The “VIOLATIONS” field in the crime data and the values associated with it compared over time between rural and urban areas.

**3.10 Are there any changes in crime due to cultural shifts?**

* + Crime may evolve based on what is happening in region at the time. By looking at the values of different crimes during big cultural shifts in a specific region a link could be drawn between the cultural shift and the change in crime values.
  + Ex. The “VIOLATIONS” field and large known cultural shifts in the region