The Life and Death Team Project

Part I. (Week 6 Lab)

- Introduction
- Where Does the Data Come From?
- Processing the Mortality Data
- Before 2003...There are Differences!
- Birth Data
- Starting the Project

Introduction

The project for your team is a challenge to use the deaths and births data provided by the CDC in the United States to answer questions about the population in our neighbour to the south. In your last lab assignment you saw some of the data and now you will start to explore this big and rich data set. Below you will find links to the data sources for the Team Project and some processed data files.

Where Does the Data Come From?

It comes from the Centers for Disease Control and Prevention Vital Statistics Data Available Online website (http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm). This contains two sources of information that you will need for your project:

- Mortality Multiple Cause-of-Death
- Birth Data Files

The documentation for the Birth Data Files is on this website but the documentation for the Mortality Data Files is located http://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm.

Processing the Mortality Data

The data that you worked on in your last lab was pre-processed for you - the data on the CDC site has many fields and is not in CSV format. The following instructions will show you how to download and preprocess the files for the years 2003 to 2014. Instructions on what to do for earlier years will follow.

On the CDC Vital Statistics Data Available Online site, do the following:

- Go to the Mortality Multiple Cause Files.
- In the left column (U.S. Data) download 2014 (83 MB) - the file mort2014us.zip will be downloaded.
- Move mort2014us.zip to the directory where you will be working on this data.
- Unzip mort2014us.zip - you will get the file VS14MORT.DUSMCPUB.

- Process the data with the readMORT.pl script: perl readMORT.pl VS14MORT.DUSMCPUB > 2014.txt
- Remove "blanks" in the data with the noblanks.pl script: perl noblanks.pl 2014.txt > 2014MortUSA.txt.

This data file contains 2,631,171 records with the following fields:

- Year
- Month of Death
- Day of the Week
- Age
- Sex
- Race
- Marital Status
- Education (89 format)
- Education (03 format)
- Resident Status
- Place of Death
- Injury at Work?
- Manner of Death
- Method of Disposition
- Autopsy?
- ICD (cause of death code)

All of these fields are explained in the documentation you received in the previous lab. It is also on the documentation site previously mentioned in this document.

Before 2003...There are Differences!

You will have to look at the documentation for every year to see if the locations of the data fields of interest are in the same places in each year. Look at both the documentation and the readMORT.pl script to see the location in the original file of the data fields that we are extracting. For the years previous to 2003 we have differences such as the lack of the fields "Method of Disposition" and "Autopsy?" and a different set of Education fields. The following Perl script (readMORT2002.pl) will read the year 2002 (as some others - yours to discovery!) file and create a preprocessed file that is similar to the one that we created before (in CSV format but missing some fields).

Once you have downloaded and unzipped the 2002 Mortality data file to get Mort00us.dat, do the following:

- perl readMORT2002.pl Mort00us.dat > 2000.txt
- perl noblanks.pl 2000.txt > 2000MortUSA.txt

This file will have 2,407,193 records.

One of your project tasks should be to research the different formats and information in the Mortality files given the year. The Mortality files go back to 1968 - can you use all of this data?

Birth Data

The Mortality files are a bit depressing so let's look at the Birth Data Files. Once again we have data from 1968 to 2014. There are a few things that you will have to do to use this data. The Zip files on the CDC site cannot be "unzipped" on a Mac using unzip. You need to use the program 7zX. I will get this program installed soon on the Macs in Reynolds 114. But until then you can still research the format of the files (the documentation is on the CDC site) and I have downloaded and processed the 2014 data for you. The script to process the raw data is readBirth2014.pl and the fields captured are

- Year
- Resident Status
- Month
- Time
- Day of the Week
- Place of Delivery
- Method of Delivery
- Attendant
- Child's Sex
- Child's Weight
- Child's Number
- Mother's Age
- Mother's Race
- Mother's Marital Status
- Mother's Education
- Mother's Total Birth Order (number of children)
- Mother's Birth Interval (time since previous birth)
- Father's Age
- Father's Race
- Father's Education

The resultant processed data file has 3,998,175 records. It can be found on http://ontology.socs.uoguelph.ca/~dastacey/LifeandDeath/

Starting the Project

- Go to the CDC sites mentioned.
- Collect some data.
- Create test data files - the files are so big that while you are developing your system you should work with just a fraction of each file.
- Review some of the questions about the data that you talked about in Week 1's lab. Look at the documentation on the Birth and Mortality data - what would be interesting to know?
- Can you answer these questions with just the data from the CDC? Do you need other data sources, e.g. what about census data?
- Go back to your Trello site and start recording some of your thoughts.

Part II. (Week 7 Lab)

The Query Engine

In Week 1's lab we discussed questions that could be asked and answered by the type of data you will be using in your project. In Lab Assignment 4, you developed a program that allowed you to find data that gave the distribution of "Manner of Death" (*e.g.* suicide, homicide, *etc*.) by sex and by month in a particular year.   The data was put in a form that could be visualized.

Let us use the homicides in 2003 as an example. Using your Lab Assignment 4 code, run the following:

```
$ perl getstats.pl 2003MortUSA.txt 3 "Homicide" > hom2003.txt
```

Your output file (hom2003.txt) should look like the following:

```
"Sex","Month","Homicide"
Female,01/January,338
Female,02/February,291
Female,03/March,351
Female,04/April,331
Female,05/May,374
Female,06/June,326
Female,07/July,369
Female,08/August,364
Female,09/September,377
Female,10/October,326
Female,11/November,309
Female,12/December,330
Male,01/January,1134
Male,02/February,1012
Male,03/March,1203
Male,04/April,1201
Male,05/May,1256
Male,06/June,1265
Male,07/July,1404
Male,08/August,1348
Male,09/September,1244
Male,10/October,1234
Male,11/November,1152
Male,12/December,1248
```

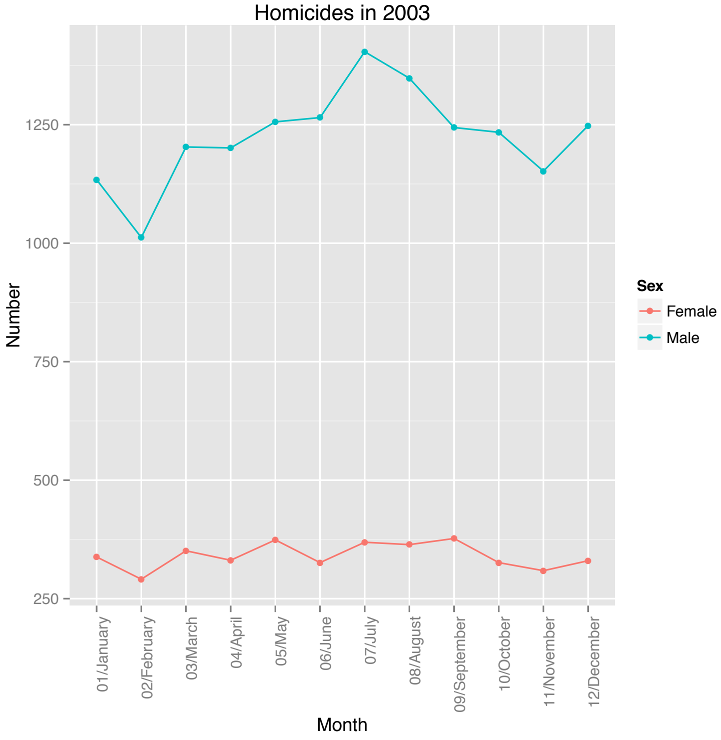
Download the program *plotMe.pl* and run the following:

```
$ perl plotMe.pl hom2003.txt Homicide2003.pdf   "Homicide"  "Homicides in 2003"
```

This will produce a PDF file and if you do the following:

```
$ open Homicide2003.pdf
```

you will see:



This is a visualization of the homicides in the year 2003 by sex and by month. Make sure that your team understands how this program works, how you can change and adapt it and how you can produce other visualizations using R. More information about using R visualizations will be posted on the CourseLink site this week.

The query engine that you will develop for your project will start from this script to collect data that will answer a question (*i.e.* How many men as opposed to women were killed by month in 2003?) and this script to visualize the answer. You must develop a set of Perl scripts that allows you to answer questions that relate to the deaths and births in the USA from 1968 to 2014.

Design Task

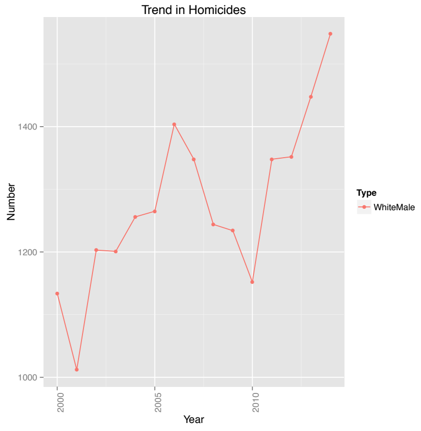
Your first design task is to:

- Look through the data (deaths and births) and see what kind of questions you can ask. For example, in the death data, you can look to see if there has been an increase in suicides in the years 2000 to 2014. Within this time period, what group has had an increasing number of suicides? Men, women, blacks, white, white men, black men, white middle-aged men?
- As you develop questions, consider how you are going to find the answers. If we look at our suicide example and you are going to see what groups have an increasing number of suicides, then you will need to develop a script that looks through many years of data (*e.g.* 2000 to 2014) and gets the number of suicides for a particular group (or groups). For example, you might develop a script that allows you to select the Manner of Death, Race, Sex and Age and then outputs that for a range of years.

```
$ perl manner.pl “Suicide” “White” “Male” 30 55 2000 2014
```

where manner.pl has 7 command line parameters: manner of death, race, sex, start of age range, end of age range, start year, end year

- This might produce a file that looks like:



- This is only an example but you should be able to see that you can ask many questions of this data. But you have to find out what is in the data and then develop a set of interesting questions to ask.

Project Deliverable #1

On Trello, you will place a document that answers the following questions:

1. What data fields will you be using in the death and births data files? Remember that some data fields might not be available in the same form or at all in every year.

2. What years of data will your system cover? How much data does that represent after you preprocess the data? How long will it take to process 1 year of data, 10 years of data, all of the data (1968 to 2014)?

3. How will you be organizing your team to look at the data? Who will be checking out the various types of data (death, birth) and which years?

4. List the types of questions that you will *initially* be trying to answer. Organize them into various categories based on type of answer, amount of data needed to find the answer, how the answer will best be presented, *etc*.

5. Estimate the number of Perl scripts you will need to do the following:
   a. Collect the data from the death and birth files
   b. Organize the data into an "answer"
      i. Tables or numbers or answers such as "Yes", "No", "Increasing", *etc*.
      ii. Visualizations (line graphs, bar graphs, pie charts, *etc*.)

   The smaller the number of scripts the better – look for commonality in the processing and look for opportunities to **aggregate** the data, *i.e.* you might create files that contain summarized/aggregated data for some fields for some time periods. For example, you might create a file that contains the suicide data for a decade (*e.g.* 1980 to 1989) where each line represents the number of female suicides per month in a number of age groups.

This is due on **Sunday, March 13 at midnight**. Please make the document easy to find on your Trello site. The document should be downloadable as a **Word** or **PDF** document.