CSC 371 - SPRING 2017
THE PRACTICE OF COMPUTER SCIENCE
DATA ANALYSIS PROJECT
UNIVERSITY OF VICTORIA

Due: Monday, April 10th, 2017 before 11:59pm. **Late assignments will not be accepted.**

1 Overview

So far, the assignments in this course have been rigidly structured and regimented. Moreover, the data analysis questions on assignments 2 and 3 were designed with a specific outcome in mind (for the obvious reason that an assignment question needs to have some kind of solution or marking is impossible). In practice, data analysis tends to be less structured and more exploratory. The goal of this project is to combine some of the applied aspects of the course with the design and exploration elements of data analysis that are hard to test with assignment questions.

For all projects, the final deliverable must include a written report documenting the project and its results. In addition to the written report, your submission must include all data and code used to generate your results. Depending on the project and topic, this may include source data, intermediate spreadsheets or database files used to clean up the data for analysis, SQL queries, material to generate visualizations (such as spreadsheet files or code in a language like R) and finished visualizations. You are permitted to use data and code from other sources (except other CSC 371 students) with proper citation. See the Section 'Submissions and Citations' below for more details.

2 Option 1: Exploratory

A typical data analysis project involves the following five stages (with the possible exception of phase 3 for some projects).

1. Data collection/harvesting (including gathering data from direct observations, web-based sources or academic literature).
2. Preprocessing (converting the data to a useful form for analysis).
3. Schema design (using E/R diagrams as appropriate).
4. Data analysis (using SQL or other methods, extract meaningful information or patterns from the data).
5. Presentation and Visualization (using Excel/LibreOffice/Numbers, Python/R or other tools).

Choose a data set of interest (or better yet, one relevant to your primary field of study) and develop an analysis project around it. Since this is a short project, most of the five phases above will be 'trivial', but you should focus on producing original results in at least one of the five phases, depending on your choice of topic. For example, the data collection might be easy if you're working with a pre-formatted dataset like our Ferry data, and no preprocessing might be needed for some datasets if they are already in a convenient form.

You could focus primarily on phases 1 and 2 by stitching together several datasets into a single, unified form, or you could focus exclusively on phase 3 by designing a data model and schema for a complex database. You could also focus exclusively on phase 5 by taking existing data and presenting it in an intuitive and interesting way.

However, you will likely need some original work in all five phases. Since the amount of work required depends on your choice of topic, talk to your instructor after choosing a topic to work out the expectations for the result. Since exploratory projects require more creativity and organization, you will not be expected to go into as much depth as projects in Option 2 below.

Some ideas:

- Primarily phase 1: Harvest data from an online source (e.g. the Canadian census or Elections Canada data over multiple years) and package it into an easily analyzable form. Construct at least one query to demonstrate that your dataset is convenient for analysis.
- Primarily phase 2: Cross-reference data from two sources to allow correlations to be investigated. For example, combine the census and election results data per riding (for a single year) to allow demographic information to be investigated. Construct at least one query to demonstrate that your dataset is convenient for analysis.
- Primarily phase 3: Choose a large-scale system (e.g. traffic management at an airport, logistics for a delivery system, etc.) and design a schema for it, consisting of E/R diagrams and SQL database construction statements. You should use some of the more advanced features of SQL table creation beyond those covered in class (talk to your instructor for details).
- Primarily phase 4: Using a dataset of your choice, produce a deliverable similar to that for Option 2 below. Since using your own dataset will add overhead to the project, the number of expected queries is fewer than an Option 2 project would entail. Depending on the data, you may also want to employ non-SQL-based analysis techniques (like text processing).
- Primarily phase 5: Using a dataset of your choice (or one of the datasets we used in class), produce interesting visualizations which illustrate salient aspects of the data. If you learn to use an advanced visualization toolkit (above the level of what we covered in class), the expectations of the result will be lowered accordingly (to compensate for the overhead of learning the new software).

3 Option 2: Analysis Only

A list of datasets has been posted to conneX. Choose one of them and, after familiarizing yourself with the schema, prepare a set of 8-10 analysis queries (of roughly the same level of depth as Assignment 3) and at least 2 visualizations (produced with your choice of software). It is expected that the visualizations will be above the level covered in class (that is, encapsulating more data or illustrating it more thoroughly).

4 Submissions and Citations

As mentioned above, your final submission must include a written report which summarizes your projects objectives and outcomes. It is expected that the report be written as a professional document and include a bibliography for all sources used (including the sources for data and code).

You may use any standardized citation style. The purpose of the report is simply to provide an overarching summary of the work done. Therefore, no 'minimum length' will be enforced, although the report is expected to be thorough and concise. For most projects, 2-5 single spaced pages (excluding any cover page, if present) should suffice. If you produce any plots or visualizations, they should be included in the body of the report and referred to in the text.

Your final submission must also include (electronic) copies of all queries and code used for your project. In general, all aspects of your queries and code that you want to be considered for marks should be referenced in some form in your report (that is, do not expect that we will go through your code looking for interesting things; anything interesting should be documented in the report).

If you used data which was not provided on conneX, you should also submit that. If the dataset is too large to submit, talk to your instructor to arrange an alternative submission.

Submission Instructions

All submissions will be accepted electronically. You are permitted to delete and resubmit as many times as you want before the due date, but no submissions or resubmissions will be accepted after the due date has passed. Ensure that each code file contains a comment with your name and student number.

After submitting, conneX will automatically send you a confirmation email. **If you do not receive such an email, your submission was not received.** If you have problems with the submission process, send an email to the instructor **before** the due date.