Half-Time Special Topics Graduate Assistantship (10 hours/week).

Title:
World War I Woman's Defense Council Cards Data Cleaning Project

Project duration.
This project is for one-year as a part-time GA. (10 hours/week).

Special project description
This graduate assistant will be tasked with cleaning, or assisting with cleaning, of data previously entered from a WWI era survey into an electronic data base. The task includes checking data entries for accuracy, spelling, etc., and for free-response questions developing and/or implementing a protocol to use "like wording" with "like responses". Components of the task include developing code to "clean" said data, developing a code book of questions and responses, and preparing the data set so that others can more easily perform data analysis with the data set. The task includes identification of data entry problems and checking data entries against electronically scanned copies of the actual cards that were entered. The task may also include historical research related to the era and/or to the cards.

The project will emphasize data cleaning and preparation for later summary and analysis. The project will give the assistant substantial work with a survey data set with data from the WWI era. The data set will give substantial opportunity to learn and use in greater depth data cleaning techniques taught in STA 616 (Statistical Computing).

During World War I the Woman's Committee of the Council of National Defense conducted a voluntary nationwide survey of women to determine what skills U.S. women could contribute to the war effort. Various state and local chapters of this committee mobilized efforts to have women complete these cards.

Right after WWI most of the cards were probably thrown away because they were no longer considered to be useful. However, sets of cards have been found in Indiana, South Dakota, and Michigan. The largest such set, over 22,000 cards, are in the Grand Rapids Public Library (GRPL). Over the past 5 years the cards have been electronically scanned by staff and volunteers of the GRPL, and the cards were then transcribed into an electronic data base by the Western Michigan Genealogical Society (WMGS), GRPL staff, and other volunteers. The scanned cards can be found at
https://cdm16055.contentdm.oclc.org/digital/collection/p16055coll5

In February 2018 Dr. Gerald Shoultz received a tab-delimited copy of the data base. He has been working as time allows, in conjunction with staff at the GRPL, on data cleaning. Wikipedia defined data cleaning as "the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the **data** and then replacing, modifying, or deleting the dirty or coarse **data**." This includes such matters as (1) correcting spelling and typing errors, (2) correcting data that has clearly been entered in the wrong column,

(3) making sure that questions with categorical responses are answered as such, and (4) grouping similar responses so that they appear as such in the data set, especially for questions with an "other response". In many cases, one must go to the GRPL website to obtain the actual card that was transcribed into the data set and use information from that card to correct the database. There are 74 columns (variables) in this data set; many of the responses for the variables are not multiple choice and hence require more thought to clean. Cleaning such a data set is time-intensive and detail-intensive.

With this project comes many potential opportunities in addition to data cleaning experience. Students gain hands-on experience with historical data. Techniques used to clean this data can be sources of SAS Global papers and SAS conference presentations. Results of data analysis of cleaned columns can lead to co-authored presentations at statistics, demographic, and history conferences. Furthermore, for those wanting to more easily access the information found in the cards, either for personal genealogical information or for the set of women in its entirety, a cleaned data set would greatly facilitate related research.

Selection Process

The successful candidate will be limited to a graduate student accepted into the GVSU Biostatistics or Data Science Masters programs who has
(1) completed STA 318 (Statistical Computing, undergraduate at GVSU) and/or STA 616 (Statistical Computing, graduate, at GVSU), and/or
(2) is a SAS Certified Base Programmer for SAS 9, and/or
(3) has demonstrable professional experience equivalent to (1) and/or (2) that includes experience with SAS Character Functions and/or Perl Regular Functions.

Students will be required to submit resume's and cover letters to Dr. Gerald Shoultz. Applicants will be reviewed by Dr. Gerald Shoultz, Dr. Sango Otieno and Dr. Paul Stephenson. Other statistics department or computer science department tenure-line faculty members may be consulted as needed.

Work Station

Most computing will be done on servers at GVSU, so any GVSU computer lab or office would suffice for many of the required tasks.

Orientation

This full-time GA will be unique with respect to departmental GA orientation but initial meetings with G. Shoultz, the director of the student's graduate program (Bob Downer or Jerry Scripps), and the Department Chair (P. Stephenson) will outline all envisioned duties for this GA. In addition, the GA will be expected to meet with, and collaborate as needed, with Julie Tabberer and Will Miner of the GRPL. The first meeting will occur in August 2019. Another review meeting will occur in January 2020 to review the fall semester and plan for winter 2020.

Supervision
Dr. Shoultz plans to provide primary supervision of the GA through weekly meetings.

Contact Information:
Primary Contact:
Dr. Gerald Shoultz, Associate Professor
Dept. Statistics, 331-8689, shoultzg@gvsu.edu

Secondary contact:
Dr. Paul Stephenson, Professor and Chair
Dept. Statistics, 331-3355, downerr@gvsu.edu