International Population Data Linkage Conference 2018
The LIDIC Hackathon: LInked Data Innovation Challenge

Information for Participants

Date: September 11, 2018

Sponsors: We are grateful for sponsorship of this workshop by Statistics Canada and IBM.

Description: Participants will engage in a team-based analysis of a complex, linked, synthesized dataset provided by Statistics Canada. This synthesized data base links socioeconomic and mortality data representing the Canadian population. The data base was derived from existing linked data available from Statistics Canada.

Objectives:
- To encourage innovative thinking about complex linked databases
- To stimulate interdisciplinary and inter-jurisdictional data collaborations
- To facilitate an environment for creative thinking about data
- To promote networking amongst participants

Format: You will work in teams of 3-4 workshop participants. Each team will conduct an innovative and policy relevant analysis of the simulated dataset. You will receive an e-mail from a Statistics Canada representative with a link to a Google Drive file that contains the synthesized dataset, codebook, and other relevant information about the data. You will also receive a separate e-mail from me with your team assignment. Your time on-site at the workshop will be spent in the analysis and presentation preparation, in anticipation of team judging later in the conference. You should plan to connect with your team electronically prior to the conference, to become familiar with your team's composition, strengths, and interests.

Analysis Tools: There are no restrictions on the analysis tools that can be applied to the synthesis data. IBM, as a sponsor of this workshop, has information on tools available through the IBM® Academic Initiative program. This program provides professors and students access to the latest technologies in open source and IBM software, hardware, course materials, training, technical support, and other resources. Faculty members, research professionals, and students at accredited institutions all over the globe can join the IBM Academic Initiative at no cost. There is no limit on the number of members from an institution that can join.
All academic software is available at: https://onthehub.com/download/free-software/ibm-ai
A list of all the products available for use is available at:
https://ibm.onthehub.com/WebStore/ProductsByMajorVersionList.aspx?cmi_mnuMain_child=43ffb823-db5f-e611-9420-b8ca3a5db7a1&cmi_mnuMain=445a8b3f-5765-e611-9420-b8ca3a5db7a1
This includes IBM Cloud storage.

Suggested Analysis Topics: This linked database contains rich demographic, socioeconomic (income, education, employment), identity (immigration, ethnicity) and geographic (urban rural)

information linked to mortality data (including cause of death). Here are some example research questions that could be applied to the synthesized dataset:

1) What socioeconomic factors contribute to the greatest inequalities in mortality?
2) Does the healthy immigrant advantage extend to the second generation?
3) Does risk of cardiovascular mortality vary among immigrants from different source countries?

Note however, that you are not limited to one of these questions. Choosing a question with your team members will be an important component of your team's success with this Challenge.

Judging and Prizes: An expert panel of judges will evaluate each team's analysis in a conference plenary session on September 14 from 8:30 am – 9:30 am. Each team will have five minutes to present one infographic that describes their analysis and its significance/relevance to policy. All team members should plan to participate. There will be time for questions from the judges (and audience) following your presentation.

Prizes will be awarded for:
1. Best Use of Data
2. Best Visualization
3. Best Overall Analysis

In addition, Statistics Canada will work with the winning teams to disseminate and promote their analyses. This is a really exciting opportunity to showcase the outcomes of this event.

Evaluative Criteria: The judges will evaluate each team's analysis on the following criteria:
- Innovate use of the linked, synthesized data in a policy relevant analysis.
- Appropriate and novel analyses of the data.
- Creative and informative presentation of findings.
- Evidence of teamwork in the analysis and presentation of results.

Name and Contact Information for Workshop Facilitator: Lisa Lix, University of Manitoba, Canada; e-mail: lisa.lix@umanitoba.ca

Synthetic File Information
This synthesized data base links socioeconomic and mortality data representing the Canadian population. This synthetic data base was derived from existing linked data available at Statistics Canada.

The data provided for this Hackathon is synthetic. It has been created using the R package *Synthpop*. The overall distributions of the variables resembles that of the original dataset, allowing the user to perform descriptive statistics and complex analysis. However, please note that this synthetic file produces synthetic results.

Documentation about the synthetic file is available at this link:
https://drive.google.com/open?id=1zWT9jSFYQoPsSRaYPhe7FHf3OLfBQ8fm

Conditions of Use:
By participating in this IPDLN Hackathon and accessing the Socioeconomic-Mortality linked data set, the participant agrees to:
- use the Socioeconomic-Mortality linked data set for the purpose of IPDLN Hackathon only;
- not share or disclose the Socioeconomic-Mortality linked data set to any other party not involved in the IPDLN Hackathon;
- destroy the Socioeconomic-Mortality linked data set at the end of IPDLN Hackathon without retaining any copy thereof.

Number of records: 4,346,649
Number of variables: 34

Contact Information:
For information related to the file content or its use, please contact
Julie Bernier, Director
Health Analysis Division, Statistics Canada
julie.bernier@canada.ca