## **Final Project: Proposal**

This is a **group assignment**. You can find your group on Canvas/People/Project Group.

## **Goals**

1. Choose a dataset of interest
2. Report on the dataset’s *potential* for data analysis
3. List of initial research questions related to the dataset.

# **Introduction**

For the project proposal, you need to choose a dataset according to your interest and perform an initial data analysis to get to know your data. After this initial analysis you will outline your research questions. During at this step, you will not answer the research questions, but you will list them and indicate the data attributes that will help you answer them according to your initial analysis.

**Datasets**

This document lists a few websites that publish open datasets, these are suggested sources. However, you are not limited to choosing only from these, you can propose a dataset from a different source too. The only requirements are that (1) the format of the dataset is supported or can be converted into a format supported on Python/Jupyter for data analysis; and (2) this dataset is open/free for non-commercial or academic usage, preferably dataset with real-world data avoiding toy datasets.

*Also avoid datasets posted in the course for assignments.*

* Kaggle datasets: <https://www.kaggle.com/datasets?datasetsOnly=true>
* Open Data by the Government of Canada: <https://open.canada.ca/en/open-data>
* Open Data Toronto: <https://open.toronto.ca/>
* Open data Seattle: <https://data.seattle.gov/>
* Open data USA: <https://www.data.gov/>
* Specialized for Text Analysis (e.g., frequency of words)
  + eBooks and Texts: <https://archive.org/details/texts>
  + Gutenberg free eBooks: <https://www.gutenberg.org/>

# **Submission**

You will be submitting your notebook report via Canvas. More details:

* 1 submission per group
  + 1 person submits in behalf of the group.
  + Named <project-groupID>-proposal
* Each submission contains 2 versions of your notebook:
  + The ipynb notebook file
  + The pdf OR the HTML notebook file executed
* In the notebook file, the group will
  + Include the group information as displayed on canvas (e.g., Project Group 1).
  + Include the list of group members’ names
  + The name and source of the chosen dataset (URL).
  + A brief description of the chosen dataset
    - What is it about? How was is collected?
  + Description of each of the data attributes
  + The code and the analysis of basic characteristics of your dataset
    - Obs1. Similar to the “get to know your data” part on assignment 1.
  + Your proposal and motivation
    - Why did you choose this dataset over others?
    - Was there any other dataset you also found interesting?
    - What are the data analysis questions you want to use this dataset for?
      * At least 5 questions for proposal, you may include more questions in the actual final project report
      * What methods do you expect to apply when you process the data to answer these questions?
    - Do you see any *potential* for data science applications using this data?
      * Examples: models for prediction, recommendation, data visualization, etc.

# **Feedback & Expectations**

You will receive feedback on this proposal which will guide to develop your final project report.

At this point, you are only expected to show genuine interest in a dataset and present its potential for further data analysis. You are not expected to perform thorough data analysis at this step. In your final project, you will be analysis the chosen dataset (if approved by the TA) in details. Otherwise, the TA and instructor will help you find an alternative dataset to fulfill your interests.

**Extra**

* Tips for formatting markdown cells: <https://www.tutorialspoint.com/jupyter/jupyter_notebook_markdown_cells.htm>