Scientific Data Analysis
Final Project
– 2018 –

Deadline: December 11, 11:59PM

This course is all about data and drawing conclusions from data. In the final project, you are expected to apply the tools and methods introduced in this course on a *real-life* dataset. You are encouraged to apply additional tools and methods beyond the scope covered in the course.

You need to *perform* the analysis, and *report* the conclusion of your analysis.

Kaggle

There is a variety of datasets that are publicly available online.

Kaggle.com is the authoritative platform for competitive analytics. It hosts over 10,000 datasets over many topics. Here are just a few datasets that are good choices for analysis.

1. San Francisco Police Reports from 2016-2018
   https://www.kaggle.com/san-francisco/sf-police-calls-for-service-and-incidents
2. Google Play Apps
   https://www.kaggle.com/lava18/google-play-store-apps
3. Madrid daily weather reports from 1997-2015
   https://www.kaggle.com/juliansimon/weather_madrid_lemd_1997_2015.csv

See https://www.kaggle.com/datasets for more datasets.

Canadian Open Data

As a response to the new digital age, many federal governments have been pushing the open data initiative where governments make their operational data available to the public for analysis.

The Canadian Open Data portal contains some very interesting datasets.

1. Agriculture activities across Canada from 1971-2006.
   https://open.canada.ca/data/en/dataset/ba79d499-2af3-41fb-b087-bf84c6087de8
2. National Science and Engineering Research Council Awards data from 1991-2017.
   https://open.canada.ca/data/en/dataset/c1b0f627-8c29-427c-ab73-33968ad9176e
3. Food in Canada
   https://open.canada.ca/data/en/dataset/a683c640-b5fd-48f8-a0f1-d619b8f7e04c

See https://open.canada.ca/en/open-data for more datasets.

FAQ

Q: How much data do I need?

> A: You need to work with at least **one** tabular dataset (.csv file). You can work with multiple datasets from different sources. Finding multiple data sources that can be combined can be challenging, but it's an exciting opportunity. You want to work with **large** datasets with at least several hundred rows of data.

Q: How should I choose the data to work with?

> A: Kaggle typically hosts cleaner datasets, and offers a user friendly interface to browse the available datasets. The Canadian Open Data portal, on the other hand, offers some very socially relevant datasets to examine this country's state of existence. Take some time to download multiple datasets. Use **Jupyter notebook** (in conjunction with Pandas) to examine the data. See if the data "speaks" to you early on. Don't worry too much about jumping to conclusions. The most important factors in picking a *good* dataset or datasets are: (1) make sure there is **enough** data (multiple columns and over hundreds of rows), and (2) you have sufficient understanding of the underlying data.

Q: Can I work team members?

A: Yes. You can choose to form a team *upto three members*. I do **not** recommend working in a team at this point of your career/education. It's actually very easy to do a good job on this project without any additional help. All members of the team will receive the **same** grade. Working with others will invariantly create **more** work for yourself, and unfairness in work/reward ratio.

Q: What do I need to submit?

A: You are required to work with Jupyter notebook. This means that you need to submit (1) the .ipynb file, (2) either the HTML file or PDF (recommended) file that contains **the executed results** of all the cells of your notebook. You are *not* required to submit the raw data.

Q: What do I need to include in my notebook?

A: You need to include as much text in Markdown as Python code. You need to clearly indicate the data source or data sources. Before each cell, you need to clearly explain your thought process, and the analytical method the code is implementing. You need to render the summary of your analysis, either as Pandas dataframe or matplotlib figures. It is absolutely crucial for you clearly state the **conclusion** of your analysis and the **evidences** that supports your conclusion.

Submission

You need to submit:

- clearly indicate the members of the project
- the original .ipynb
- the .html or .pdf output with all the cells executed.

Grading Rubrics

	None 0%	Little 20%	Some 50%	Lots 80%	Awesome 100%	Total
Data volume (10%): you should clearly indicate the number of rows and columns of your dataset.						
Data quality (10%): discuss why the data source is interesting and relevant to data science.						
Feature coverage (40%): in your worksheet, clearly indicate the features you are applying during the data analysis.						
Graphical and tabular summarization (20%): your worksheet should generate intermediate and final results in the appropriately designed text and graphical output.						
Well written English comments (20%): your worksheet should make good usage of Markdown to explain the intention of the code, explain the significance of the output, and the conclusions. We really care about the English. Pay attention to the writing.						
Making the world a better place (bonus 10%): if your data analysis can unveil an opportunity to improve social values, quality of life, or any other aspect of Ontario or Canada, you will be awarded some bonus.						