3250 TERM PROJECT

The Project

The 3250 Term Project is your opportunity to showcase what you've learned in the course. You will be working in assigned groups. The objective is to select a data set, transform and clean it, conduct a data analysis, an report key statistical features and insights of the data. If you are already familiar with regression and/or classification and would like to try out one of the predictive modelling algorithms in scikit-learn it is fine to do so but this is not required. Alternatively, you can propose a custom project. Some examples of custom projects/reports that would be appropriate:

- Strengths and weaknesses of using Blaze (http://blaze.pydata.org/)
- Functional programming features of Python
- Naïve Bayes
- Bootstrap Sampling
- Strengths and weaknesses of using PyPy (https://pypy.org/)
- Parallel programming in Python

Data

The data you use for your analysis must be real (i.e. not randomly generated). The following are links to a variety of interesting datasets for your consideration. The dataset you use does not need to be from this list. If you plan to use a dataset from work you must have written approval from your employer to discuss the dataset with the class.

https://www.analyticsvidhya.com/blog/2016/11/25-websites-to-find-datasets-for-data-science-projects/
http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=9e56e03bb8d1e310VgnVCM10000071d60f89RCRD
https://www.ontario.ca/search/data-catalogue?sort=asc
http://open.canada.ca/en/open-data
http://koaning.io/fun-datasets.html

Topic Approval

Prior instructor approval of your choice of project is required. If you would like to undertake a different topic please submit your idea to the instructor via email for feedback and approval by Week 8 (and preferably earlier).

Requirements & Due Dates

Each group must submit two documents, a report and a presentation which are due the second-last class of the term. If your project is a data analysis, your Jupyter notebook(s) with the code must be submitted together with the report.

Report

The report should *be of a quality that is suitable for a management briefing*. If it is a data analysis it should follow this outline:

1. **Data Preparation**: What was your data source (e.g. web scraping, corporate data, a standard machine learning data set, open data, etc.)? How good was the data quality? What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis? What challenges did you face?
2. **Analysis**: What trends, correlations and/or patterns do you see in the data?
3. **Conclusions**: What did you learn about your data set?

If you're doing a custom project the outline should include Objectives, Analysis and Conclusions. The report should be about 15 pages in length. If you would like to include any code and samples of data, in addition to the Jupyter notebook, these should be included as an appendix, not in the main body of the report. Data samples should be limited to less than a page.

If you refer to any external sources in your report (articles, blog posts, the Statistics Canada website, etc.), you must include the references (link, book title) to the sources of the information that you reference. This includes a URL for the webpage where your dataset came from if from the Internet.

Presentation

The presentation should be no more than fifteen PowerPoint slides long. Each group will have 15-20 minutes to present.

Marking Scheme

Marks will be allocated as follows for a total out of 40:

- Suitability as a report and presentation to management – 10 marks
  - Spelling
  - Explanation of any technical terms used
  - Formatting
  - Easy to follow
  - Follows structure as described above
- Correctness and thoroughness of analysis – 15 marks
  - Use of appropriate techniques
  - Checking that the data meets any assumptions the model or test requires
  - Correct interpretation
- Presentation quality – 10 marks
  - Spelling
  - Explanation of any technical terms used
  - Formatting

    - Easy to follow
- Novelty – 5 marks
    - Is this an interesting and different analysis?