# MATH 210 Project 2

*Data Science*

Data Science is a new field combining mathematics, statistics, and computer science to solve problems with data. Read more about data science:

`https://datascience.berkeley.edu/about/what-is-data-science/`

Your assignment is to find an **open** dataset online, use pandas to explore the dataset and write a Jupyter notebook to present your work. The subject of the dataset is completely up to you: economics, politics, sports, education, health, astronomy, transportation, etc. The requirement that the dataset must be open means that the readers of your notebook should be able to find the data online and recreate your analysis.

## INSTRUCTIONS

**Choose a dataset.** There are many online resources to explore open data (see the list below). Choose a subject that interests you and find a dataset you would like to explore. The only restrictions are: **The dataset should be . . .**

1. freely available online
2. at most 10MB
3. nontrivial (use your judgement and ask your peers for feedback)

**Ask questions about the dataset.** Why did you choose this data? What do you want to explore? What hypotheses do you want to test?

**Get feedback.** We will meet in groups in class on Wednesday April 6 to share our ideas and get feedback about the data we have chosen.

**Write a notebook exploring the dataset.** Use the pandas library to import, explore and visualize your data. See the project outline below.

## PROJECT OUTLINE

**Introduce the data.** What is the subject of your dataset? Why is it important/interesting? Where did you find the dataset? Where can the reader go to find the dataset for themselves?

**Ask questions.** What questions would you like to explore in the dataset? What hypotheses would you like to test?

**Explore the dataset.** Use pandas to import and explore the data. Make figures to present your results. Provide links to further resources.

**Summarize your findings.** Write a Jupyter notebook to clearly present your analysis with Python code, figures and text.

## SCHEDULE and DELIVERABLES

**Wednesday April 6 (in class)** – Peer feedback (3 points, submit responses to Connect)

**Wednesday April 13** – Submit Jupyter notebook (12 points, submit data and notebook to Connect)

## WHERE TO FIND DATA

City of Vancouver `http://vancouver.ca/your-government/open-data-catalogue.aspx`

Province of BC `http://www.data.gov.bc.ca`

Open Data Canda `http://open.canada.ca/data`

Statistics Canada `http://www.statcan.gc.ca`

Envionment and Climate Change Canada `http://climate.weather.gc.ca`

Quandl `http://www.quandl.com`

World Bank `http://data.worldbank.org`

NASA `http://data.nasa.gov`

`https://www.import.io` (scrapes data from a given webpage and converts it to .csv)

## GRADING RUBRIC

**Content (7 points)** Is the data introduced properly? Are questions well-posed? Is the data imported properly? Are quantities calculated properly? Is Python code used effectively to perform computations and plot figures? Are conclusions summarized clearly?

**Presentation (3 points)** Is the notebook organized with clear section headings? Are figures presented clearly to answer the questions? Is text written and presented clearly?

**Data (2 points)** Is the data[1] original? Is the data interesting or too simple?

[1] There is a HUGE amount of data available online. Choose your own data. Points will be deducted if you use the same data as someone else in the class.