Final Project: Part 1

Dr. Ilkay Altintas

- Explain final project objectives and timeline
- List deliverables for Nov 20th

Final Project

- Culmination of everything learned in this course
- Expect to spend 24-32 total hours on the project
- Steps for the project:
  - Step 1: **Find a dataset or datasets**
  - Step 2: **Explore the datasets**
  - Step 3: **Identify 1-3 research questions and perform analysis**
  - Step 4: **Present your findings**

Schedule

- FINAL PROJECT PART 1: Due class on November 20, 2020
  - Deliverable: Presentation at class
  - 5 minutes, 3 slides
    - Dataset
    - EDA findings
    - 1-3 research questions to perform the analysis between November 20, 2020 and December 4, 2020.
- FINAL PROJECT PART 2: Due December 4, 2020
  - Deliverable: A Jupyter notebook with all the deliverables as asked for.
  - Jupyter notebook will be posted on November 20, 2020.

Step 1

- Based on your interest, identify a dataset which you will want to examine. You will find a starting point for where you can find open datasets at the end of this notebook, but feel free to use other datasets you have access to and can publicly share results about.
- This step may take some time, as you'll likely look at a number of datasets before you find one (or more) which holds promising data for the kinds of questions you want to ask. You are expected to use at least two interconnected datasets, e.g., two tables in one database or a combination of datasets which you can merge in some meaningful way.

Step 2

- In this step, you should explore what is present in the data and how the data are organized. You'll need to determine what common features allow you to merge the datasets.
- You are expected to answer the following questions using the *pandas* library and markdown cells to describe your actions:
  - Are there quality issues in the dataset (noisy, missing data, etc.)?
  - What will you need to do to clean and/or transform the raw data for analysis?
- You are also expected to use the *matplotlib* library to visually explore the datasets and explain your findings, specifically,
  - How are the data distributed?
  - What are some common trends?
  - What are the relationships between variables in your datasets?

Step 3 [Just pick the questions for Nov 20th!, do not complete the analysis!]

- Now that you have a better understanding of the data, you will want to form a research question which is interesting to you. The research question should be broad enough to be of interest to a reader but narrow enough that the question can be answered with the data. Some examples:
  - **Too Narrow:** What is the GDP of the U.S. for 2011? This is just asking for a fact or a single data point.
  - **Too Broad:** What is the primary reason for global poverty? This could be a Ph.D. thesis and would still be way too broad. What data will you use to answer this question? Even if a single dataset offered an answer, would it be defendable given the variety of datasets out there?
  - **Good:** Can you use simple sentiment analysis on comments about movies in a movie database to predict its box office earnings? If you have, or can obtain, data on a variety of movies and you have their box office earnings, this is a question which you can potentially answer well.

Remember, this course is for learning Python. You will not be graded on the complexity, accuracy or performance of your analytical methods. However, you are expected to use a Python library, e.g., *scikitlearn*, successfully to generate results and explain why you picked the methods you used.

Where to find project datasets?

- Web – some examples:
  - UCI's Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php
  - Kaggle: https://www.kaggle.com/datasets
  - KDnuggets Dataset (http://www.kdnuggets.com/datasets/index.html
    - government data (http://www.kdnuggets.com/datasets/government-local-public.html)
    - Data APIs (http://www.kdnuggets.com/datasets/api-hub-marketplace-platform.html
    - Data Mining Competitions (http://www.kdnuggets.com/competitions/index.html
  - US Government Data: https://www.data.gov/
  - UK Government Data: https://data.gov.uk/
  - Canada's Open Data Exchange: https://codx.ca/
  - World Health Organization: http://www.who.int/gho/en
  - World Bank: http://data.worldbank.org/
- **Remember!** Finding a dataset of interest and exploring it is most of your job for your final project. So expect this to take some time - and that's both perfectly normal and completely okay. Get started early and make sure to check if the dataset you find satisfies the project requirements.