Spring 2022 **DATA 301:** Introduction to Data Science Alexander Dekhtyar

Mini Project 2: Final Project

Due date: Wednesday, June 9, 9:00pm

Assignment

Assignment Preparation

This is a **small team** project. Teams can be two, three, or four people in size.

Overview

The goal of this project is to get you test the skills you gained in this course ”out in the wild”.

Each team shall do the following:

1. **Select one or more datasets that are publicly available, and were not used in this class.** You can find datasets in various places, such as data science repositories (UCI Machine Learning Repository, Kaggle, etc), government organizations and international organizations (US Census Bureau, CDC, World Bank, etc...) and many other places.

2. **Ask analytical questions.** Come up with several analytical questions that would require you to apply the knowledge/methodology from this course to answer. You can ask any questions subject to the following conditions:

   (a) At least one question must involve classification.

   (b) At least one question must involve some form of clustering.

   The number of questions should be consistent with the number of people (and in some cases can exceed the number of people) in the team.

3. **Perform required data ingestion, cleaning, feature extraction and engineering, analysis, and visualization of the results.** You can work in multiple Jupyter notebooks (one per question), or you can create one Jupyter notebook for all your results.

4. **Explain your findings.** Add text to your notebook/notebooks, explaining the design of your analytical studies, your findings, and the conclusions you drew from them (i.e., how you answered your analytical questions).

Proposal.

By Thursday, May 26 each team shall submit a one-page document that contains the following information:

- Team name
- Team members (names, email addresses)
- Datasets you are planning to use (a brief description)
- Questions you are asking/analyses you want to perform

The document shall be short (1-2 pages tops).

Submit the document using the following `handin` command:

```
$ handin dekhtyar 301-proposal <file>
```

Submission

Submit three things:

- All your notebooks. The notebooks shall be clearly labelled, with the names of all team members at the top, the title of the notebook and the question/questions addressed in it present as well.
- All datasets you used for analysis (unless your analysis grabs data directly from web-accessible URLs). Basically, if I open your notebook, I should be able to run all of your cells right away, including any data ingestion you are doing.
- Project report. This can be a separate PDF document describing your work, and including the visualizations you constructed. Alternatively, this can be a curated (make all unnecessary code/output invisible, add appropriate text, collate all your notebooks into one) Jupyter notebook saved as a PDF document.

```
$ handin dekhtyar 301-project02 <file>
```

Good Luck!