Foundations of Data Science

Semester 1 Mini-project

In the mini-project, you will get to choose one of the tools you liked most in S1 and apply it on a dataset of your choice (under some constraints). Please work with your WS collaboration group on the tasks below. This is not graded; the aim is for you experiment with one of the tools you learned in S1 on a dataset of your choice. You will present your results in the week 1 workshop of S2.

Please divide up tasks between yourselves, e.g. after an initial discussion, one or two of you might go in search of data, another of you might do data cleaning, and another the coding and another the presentation. If your collaboration group isn't active, please try to join up with another collaboration group.

1. Choose a dataset

You can choose any dataset you like, provided that it has the following characteristics:

- Multivariate (at least 3 variables)
- Available for public download.

Some sources for datasets:

- Kaggle: https://www.kaggle.com/datasets
- UCI: https://archive.ics.uci.edu/ml/datasets.php
- Data is plural: https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit
- Google: https://datasetsearch.research.google.com/
- The FDS Learn Wiki -> "Sources of Data".

We do not suggest that you do web-scraping for this task.

2. Choose a problem

- Choose a few representative samples of instances in your data and look at them closely to get some ideas for potential questions. Come up with an interesting problem, such as predicting the value of a variable, classifying the instances to one of several labels, or understanding how different groups of instances in the data behave.
- Select one (or more, if needs be) of the tools you learned in the course, e.g., Linear regression, PCA, K-means or k-NNs, to solve the problem. You will probably need to make some choices about which variables to include and whether to do some pre-processing (e.g., addressing missing values, generating new variables), etc
- Use the tools to attack the problem and analyze the results.

3. Prepare a joint presentation with your group

Your presentation should include the following:

- DATASET: Describe the dataset you use; Explain why it is appropriate for answering these questions.
- QUESTIONS: What are the questions you wanted to explore? Why are they interesting to you?
- TOOL: Which tool are you using to solve the question? For example, if this is a prediction problem, clearly state what is the target variable (class) you are trying to predict, which variables (features) you are using to predict the class, and why you chose them. If this is a clustering task, clearly state what you want to understand through the clustering, and which features you are using to define groups.
- ANALYSIS & FINDINGS: What analyses did you conduct to answer your questions? What did you find? (support with plots, but no code here)
- LIMITATIONS: What are some limitations of your analyses and potential biases of the data you used?
- FUTURE DIRECTIONS: What new questions came up following your exploration of this data? Describe at least one question that could not be answered using your data alone, and specify what additional data you would collect to address it.