**EDA + Visualization**

Zoom Recording: <https://berkeley.zoom.us/rec/share/8VLzqH0-B3pqE5xIedS8jqAch7UCbcX1efjj4xCv0fxip4aPEiiNOIHqoV7eOCTZ.AijMn57CaSa-hyDh>

This week’s main focus is on Exploratory Data Analysis (EDA). We want to be able to use data to tell us more about concise information, and to do this, we want to extract what we can to create logical conclusions as a result. The example I gave focuses a lot on whether the patients were predicted to have heart disease or chest pains, and the correlation these had with age, gender, etc. For this first week, we would like you to present your own take on what might be interesting about the data. Can you form connections between variables, and showcase them using visualization techniques?

The example contains various methods and functions to extrapolate data into a form we can easily see and understand. An important skill to develop throughout this project is working on how to learn new things on your own. This can come in the form of learning interesting techniques, or functions that pull useful information from our data. Please take some time to draw your own conclusions, and find an element that interests you! There are various topics to choose from, including:

* Exploring the relationship between other categorical variables (sex, fbs, restecg, exang). See if you can find something interesting, and how these relate to one another with meaningful visuals.
* You may also wish to investigate quantitative variables such as cholesterol (chol), and how they might relate to chest pain (cp).
* Anything else you can come up with on your own! Make this project unique!

Since most of the data might still seem very foreign, I would recommend reading a bit on the references and sources provided in the project spec to try and understand what a lot of the data actually means, especially the research paper from science direct. Additionally, spending time reading seaborn documentation may also be helpful, and exploring some of the potential plotting visualizations available to you. Ultimately, this first week is meant to get familiar with this dataset, as we move on to try new things and see how we can apply these techniques to other datasets as well.

**Note:** If you are unfamiliar with some of the packages we’ve used this week, or would like to find some more references to brush up on some basic functions and concepts, here is a link to a quick start guide as well as some documentation. I would recommend skimming through “10 minutes to pandas”, “Intro to data structures”, “Essential basic functionality” and “Indexing and selecting data”, which should be enough for our purposes: [10 minutes to pandas — pandas 1.2.2 documentation (pydata.org)](https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html)

Numpy is a bit more self explanatory. It will also be fine to just go through the absolute basics: [NumPy: the absolute basics for beginners — NumPy v1.20 Manual](https://numpy.org/doc/stable/user/absolute_beginners.html)

Matplotlib and Seaborn can be a lot more interesting, so you can also feel free to look a bit into cool visualizations that you might want to apply to your notebook, but don’t stress too much over knowing every single detail because it can get complicated in my opinion to tweak graphics to the “perfect” way.