Final Project

The time has come to show off what you've learned in this class. Your final project is worth 10 points, which is roughly 20% of your grade. If you have been paying attention, spending time on the homeworks, and following along with the class, it should not be too hard to get full (or even extra) credit. Below, I will outline what I want from each of the 5 sections. Each section is worth 2 points max. There are also 2 extra credit sections each worth 2 points max.

When I ask for justification it does not need to be long. Perhaps 1-2 sentences is really all I'm looking for. I want the project to be done in a Jupyter notebook, with heading seperating the different sections.

Due Date Monday, February 4th at 1 PM on Blackboard.

1 Finding and Cleaning Your Dataset

Find a dataset online that can be downloaded as a csv or excel file. Link to the website. Any dataset works, but DO NOT use one of the datasets built into scikit-learn.

You need to import your dataset into Pandas. The indicies may not be labeled exactly as you want them and the columns may also not be. Perhaps there are rows and columns that you don't want. Maybe there are NaN values that should be imputed. Fix these issues.

2 Elementary Data Analysis

Show me some cool summary stats about your dataset. The mean down each column may be an example. Justify this, if you are doing something looking at spread then you want to look at variance, not mean, for example. Don't just do 'data.describe()' as I only want the important stats. Feel free to show me more than 1 or to apply your own function but justify each.

3 Special Pandas

Use one of the special pandas functions (ex: groupby, pivot table, or if you have a time series something like rolling, or a multiindexed groupby).

4 Matplotlib

Plot something going on with the data. Remember to title, and label the plot and put legend if need be. Tell me what the plot shows and why its useful to have this plot. Describe as if I was your boss that didn't know too much about math. How can this make my fake company money?

5 Extra Credit: Plotting

Use seaborn or chartify to make a plot with your data that couldn't be made in matplot and tell me the utility of it.

6 Sklearn

Do some machine learning. Decide whether your task is supervised or unsupervised. Justify this. If unsupervised use kmeans with either elbow rule to decide amount of clusters or some justification of the number of clusters that should exist. If supervised decide whether you are doing classification or regression. Justify this. If classification use logistic regression, if regression use linear regression. In either case for supervised learning split into train and test, fit on train data and report scores on test data.

Again, act like I am your boss. What value does this analysis bring to my company?

7 Extra Credit: Post on github with a virtual environemnt

1 point for making a github repo and putting your project there with a nice readme.

1 point for using a virtual environment https://conda.io/docs/user-guide/tasks/manage-environments.html and sending that in with your project.

If you do both make sure have the environment file in your git repo.