Lab 2 - Linear Regression

CS 207

Due: Monday, February 15th at 11:59pm

Your task for this lab is to apply linear regression analysis to a given data set and to your project data. You should look in the `lab2.linreg` directory in the course repository. In that project, you'll find the dataset specific to this lab. You should add any files you create here. Added files may include additional websites and should also include any scripts used to create your analysis. The work for this lab that is towards your large project should be done within the `project` subdirectory.

Before diving in to the lab work below, you should work your way through chapters 4-6 in the d3 textbook, making sure to actually do the examples they lead you through.

Overall, your goal is to analyze the data as if it were assumed to adhere to a linear model (appropriately challenging this assumption as needed based on what you find through the analysis). You should be sure to consider multiple slices of the data for different independent/dependent variable pairs, trying to choose the pair that best fits a linear model while still answering an interesting/relevant question about the data. Your final websites should be like short (1-2 page) research papers investigating this data. You’ll hand in two of these (one for the given data set and one specific to your project data set).

When determining your simple linear regression, you are expected to use existing packages that calculate the linear regression for you. You may find it easiest to do this using python, exporting your resulting data to a csv file to be imported into your d3 visualization. Be sure to read the included files when you look at the `lab2.linreg` directory, as these have useful tips about the data and about good python packages to use for linear regression. These files also include some pre-written python functions that you may find useful.

The main pieces of your lab work to be handed in are:

1. Analysis of the given data set using simple linear regression. (This will be done in the `lab2.linreg` directory.)
2. Analysis of your project data set using simple linear regression. (This will be done in the `lab2_linreg/project` directory.)

3. A modified version of your project data removing any features you think should be removed based on this linear regression analysis.

These should be handed in by submitting a website through the appropriate project directory containing the analysis. For each part of this lab, you should be sure to include the following:

a) Give a written description that includes the question you were considering, the answer you found, and your analysis of what those results mean.

b) Include the hard data used in your analysis, i.e., the raw error values found.

c) Describe how you handled missing data and what this might have meant for your results.

d) A visualization showing the line of best fit you calculated and the individual data points for at least one graph described in your written analysis.