We have taken our first steps into Machine Learning. We have learned three different cost functions, and we have learned how to use gradient descent to optimize them to approach the optima(minimum). We have also learned how to set up and maintain a database, as well as how to build out the appropriate helper functions(api) to communicate with our database.

This project will be quite similar to the previous project. There will be a few more requirements in the Code, and that will also then write-up, as well as your presentation.

You will not need to submit a dataset for approval. You will however be expected to source your own data either through an API, or through webscraping via beautifulsoup or selenium. Remember, you know yourself better than I do. If you want the challenge, take it, but if you need the time, find a nice, easy to work with API, and a nice, relatively clean data set.

I will outline the deliverables that will be expected of you.

The raw code is going to be graded separately from the clarity of your notebook.

Many people did not properly comment code, or made no comments.

In order to fix this, the raw code will be graded out of 20 and now the comments and structure will be graded separately out of 10. You will be much more heavily penalized for not adhering to the standards expected both on the syntax of your code, and the structure of your codebase. To achieve high marks in the new category, ensure that you are using clean segmented code with comments that explain your thought process, and what the code is doing.

Jupyter Notebook Code

Your code should be bug free, and appropriate blocks of code should be encapsulated and reused by making them into functions. You will be working with a database, and so your helper functions should be extensible, reusable, and multifaceted.

The expectations of what code you should contain are outlined.

Code that sources your data.

Code that stores your data into a database.

Code that accesses your data from your database, and transforms it as necessary.(helper functions)

Code that cleans your data, and makes it easier to work with.

Code that checks for a linear relationship between the columns you want to perform regression analysis on.

Code that performs Regression Analysis on the given columns (Ensure that you are tuning your hyper parameters for a true best fitting line)

Code that visualizes the regressor line(s) that were created.

Code that visualizes aspects of your data set as a whole.

Jupyter Notebook Styling

Your code should be clearly segmented, and clearly commented. Get rid of any code blocks that you ran one ofs to explore EDA(A common problem).

The expectations of how your code is outlined below:

Ensure that each line of logic is commented. Not every line of code needs to be commented on, but if there is some step of logic being applied that should be explained, ensure that the reader can easily understand what your code is doing.

Ensure that code is being segmented. Code that runs together should be unified under one code block, not spread across multiple cells. It is one thing to test and experiment, but the final product should be clean and easy to read.

Ensure that you are using appropriate variable names and placeholder variable names. No more xs,ys,zs. If you need to create a simple loop iterating over a range, that is ok, but if you need to store the resultant call of a function, ensure that the variable that it is being stored in has a name that represents the data stored inside.

Ensure that you are cluing in the reader(s) as to the overall reasoning behind certain methodologies and functions. It is one thing to explain what a function is doing, but explaining the roadmap as to why that function needs to be built, and what that function plans to do further down the line is different.

Written Project Wrap Up.

This should be an ESSAY. We are changing the requirements here as well. 5 paragraphs now, following the typical structure of opening paragraphs, 3 body, and 1 closing paragraph.

This should capture one of two things.

Either your process, what methods you used, how you sourced your data, the functions you created and their use cases within your projects, and your next steps OR it should encapsulate the findings you were able to make regarding your data and next steps. (A mix of both is fine)

It is inappropriate to write an essay explaining the data set. That does not show me what you learned.

Written Script

This was another area people struggled with. Writing a script is not writing down just 3 bullet points, nor is it just copy and pasting your essay. A Script is supposed to be a tool that you can use to read from in the case that you get lost while explaining your project. If you don’t have enough, then it won’t be a tool to get you back on track. If your script contains every single thing about your project, you would just be reading from it and it would be dry. Use this time to be creative and explore yourself. Explain what you liked about working with the dataset, or what you learned while working with the dataset, or what you thought you should have done/approached differently.

In class presentations

This is scheduled for the 18th.