**Instructions**

The goal of this problem is for you to explore how to properly analyze, visualize, split, clean and format data and perform linear regression, polynomial regression and regularization. You will use the happiness data (e.g. happiness data.csv) located [here](https://iu.instructure.com/courses/2201574/files/169393944?wrap=1)[Download here](https://iu.instructure.com/courses/2201574/files/169393944/download?download_frd=1). The data consists of the following attributes:

* Country name: name of the country
* Year: year data was collected
* Life ladder: information about how happy people are
* Log GDP per capita: market values of goods and services in a country
* Social support: how people feel they are supported by those around them
* Healthy life expectancy: rank of the country based on the happiness score
* Freedom to make life choices: how much freedom contributes to one’s feeling of happiness
* Generosity: have you donated money
* Perceptions of corruption: how do people perceive that there is corruption
* Positive affect: do you feel happiness, laughter and enjoyment?
* Negative affect: do you feel worry, anger or sadness?

Your task is to build a model that given attributes/features: Country name, Log GDP per capita, Social support, Freedom to make life choices, Generosity, Perceptions of corruption, Positive affect and Negative affect predicts Healthy life expectancy predicts (note that this means that you should ignore Year and Life ladder).

**Answer the questions below directly in your Jupyter Notebook, using Markdown cells**. Be sure to clearly indicate that your comment is an answer to a particular question.

1. Summarize the data. How much data is present? What attributes/features are continuous valued? Which attributes are categorical? **[5 points]**
2. Display the statistical values for each of the attributes, along with visualizations (e.g., histogram) of the distributions for each attribute. Explain noticeable traits for key attributes. Are there any attributes that might require special treatment? If so, what special treatment might they require? **[5 points]**
3. Analyze the relationships between the data attributes, and between the data attributes and label. This involves computing the Pearson Correlation Coefficient (PCC) and generating scatter plots. **[5 points]**
4. Select 20% of the data for testing. Describe how you did that and verify that your test portion of the data is representative of the entire dataset. **[5 points]**
5. Train a Linear Regression model using the training data with four-fold cross-validation using appropriate evaluation metric. Do this with a closed-form solution (using the Normal Equation or SVD) and with SGD. Perform Ridge, Lasso and Elastic Net regularization – try a few values of penalty term and describe its impact. Explore the impact of other hyperparameters, like batch size and learning rate (no need for grid search). Describe your findings. For SGD, display the training and validation loss as a function of training iteration. **[10 points]**
6. Repeat everything from part E with polynomial regression. Using validation loss, explore if your model overfits/underfits the data. **[10 points]**
7. Make predictions of the labels on the test data, using the trained model with chosen hyperparameters. Summarize performance using the appropriate evaluation metric. Discuss the results. Include thoughts about what further can be explored to increase performance. **[10 points]**