Unit 6: Regression Project

- **Select a dataset from Kaggle (not too hard, not too easy). It must contain at least one categorical explanatory variable and the variable that you are trying to predict must be numerical.**

- **You will download and analyze your dataset using Pandas and scikit-learn and create a readable Jupyter that explains and shows your analysis and summarizes your findings and conclusions**

Some things to include in your Jupyter Notebook write-up:

1. Did your data need cleaning? If so, what was your process? Justify any decisions.

2. Use a one-hot matrix to transform your categorical variable and add it to your input.

3. Use .corr and .sort_values to sort your variables according to correlations with the response (y) variable. What explanatory variables were most positively correlated with your response variable? What variables were most negatively correlated?

4. Is your response variable skewed? If so, how did you transform it? Did that help your predictions?

5. Use sns.pairplot to plot scatterplots of your variables versus each other. Choose an explanatory variable that does not appear to be linearly related to your response variable. Try to come up with an optimal degree polynomial that fits the data better. Include a plot of your degree versus test error and clearly show the U-shape. If there is a polynomial that fits the data better than the linear model, add the polynomial features of this variable to your input matrix.

6. Use a pipeline to scale your data before applying RidgeCV.

7. What alpha was best?

8. What is the equation of your model? Use variable names instead of y,x1,x2,x3…. If there are many variables, you can list the first three and then do "…".

9. What is the R^2, adjusted R^2, and MSE (Mean Squared Error) on your train and testing data? They should be close, or else you are overfitting.

10. Explain what the words "bias" and "variance" mean in the context of your model.

Make sure that your code is clean enough for me to read and prefaced by markdown cells which describe the work the code isdoing in each step. Discuss what is interesting or unexpected in the context of your data.

Some possible DataSet Candidates:

Stroke Prediction: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Will it rain tomorrow? https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

Bankruptcy Prediction: https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

Australian Housing Prices: https://www.kaggle.com/dansbecker/melbourne-housing-snapshot

Wine Price Predictor: https://www.kaggle.com/zynicide/wine-reviews

World Happiness: https://www.kaggle.com/unsdsn/world-happiness

Student Exam Performance: https://www.kaggle.com/spscientist/students-performance-in-exams

KAGGLE COMPETITION:

Housing Price Predictor Competition
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Spaceship Titanic
https://www.kaggle.com/c/spaceship-titanic