**Umuzi assignment: Predicting salary differences**

**Due: Friday 15 June, 12pm**

**Submission:** Fork the ‘salary’ repository on the Umuzi-org Github page to your Github account. Create a notebook and commit your notebook changes to your git and Github repositories. Once you are done, create an html report from the notebook. The last GitHub commit done before 12pm will be marked.

**To complete this assignment, you should go through the following materials:**

https://github.com/justmarkham/DAT4/blob/master/notebooks/08\_linear\_regression.ipynb

http://scikit-learn.org/stable/auto\_examples/linear\_model/plot\_ols.html

Crash Course Statistics: Machine Learning & Artificial Intelligence (YouTube)

Khan Academy: Engageny Algebra Topic D, Lessons 14 – 18 (Modelling relationships with a line & Residuals)

https://www.khanacademy.org/math/engageny-alg-1/alg1-2/alg1-2d-modeling-relationships-line/v/fitting-a-line-to-data

In the next series of challenges, we will predict employee salaries from different employee characteristics (or features).

Import the data salary.csv to a Jupyter Notebook. A description of the variables is given on Sheet 2 of the csv file. Answer the following questions:

We are going to use a simple supervised learning technique: linear regression. We want to build a simple model to determine how well Years Worked predicts an employee’s salary.

1. Run a simple linear regression for Salary with one predictor variable: Years Worked
2. What percentage of the variance in employees’ salaries is accounted for by the number of years they have worked?
3. Does the model significantly predict the dependent variable? Report the amount of variance explained (R2) and significance value (p) to support your answer.
4. Report and interpret the standard error of the estimate. (Hint: Compare the standard error of the estimate to the standard deviation for Salary, which you would have calculated in Question 1.a)).
5. Examine the output from the Coefficients table:
6. What does the unstandardized coefficient (B) tell you about the relationship between Years Worked and Salary?
7. Compare the standardised coefficient (Beta) for Years Worked to the Pearson’s (zero-order) correlation coefficient from question 1.c). What do you notice? Explain why this is so.
8. Calculate the expected salary for someone who graduated 12 years ago.
9. Calculate the expected salary for someone who graduated 80 years ago. Are there any problems with this prediction? If so, what are they?
10. We have only looked at the number of years and employee has worked. What are employee characteristics might influence their salary?