# Assignment 3 – Linear Regression

*The purpose of this assignment is to use Python to learn how to perform regression analysis with python.*

This assignment provides you with an opportunity to demonstrate the achievement of the following course learning outcomes:

* Understand Linear regression
* Understand and apply the Python programming language
* Understand and apply feature selection principles

## Key Information

* **Type:** *Individual*
* **Weight:** 6.5%
* **Delivery:** Course website upload
* **Due Date:** 05/02/2019

## Expectations

You are expected to complete this assignment individually.

Respect for academic integrity is crucial to your success. Make sure you understand what constitutes acts of academic dishonesty in the page: [What is Academic Dishonesty?](http://mcmaster.ca/academicintegrity/students/whatis.html)

## Instructions

Find a data set that is suitable as a topic of a regression analysis; you may not use a data set that was discussed in class on January 30th, 2019. Briey describe your chosen data set and explain where it was sourced. Carry out a thorough analysis of your chosen data set using regression. Provide a clear and concise description of the results and state what conclusions can be drawn from your analysis.

Note: If you are doing the following optional assignment, you can use wine data set as well.

**Use recursive feature elimination to select features (optional):**

Recursive feature elimination is based on the idea to repeatedly construct a model (for example an SVM or a regression model) and choose either the best or worst performing feature (for example based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimization for finding the best performing subset of features.

The stability of RFE depends heavily on the type of model that is used for feature ranking at each iteration. Just as non-regularized regression can be unstable, so can RFE when utilizing it, while using ridge regression can provide more stable results.

Sklearn provides RFE for recursive feature elimination and RFECV for finding the ranks together with optimal number of features via a cross validation loop.

* + Import: RFE from sklearn.feature\_selection
  + Import: LinearREgression from sklearn.linear\_model
  + Define features and target as X and Y
  + Use linear regression as the model
    - lr = LinearRegression()
  + Rank all features, i.e continue the elimination until the last one
    - rfe = RFE(lr, n\_features\_to\_select=1)
    - rfe.fit(X,Y)

Print Features sorted by their rank

What observations can you make based on obtained ranking?

## Submission

Please submit your .py or .ipynb files to avenue before 05/02/2019.