# DATA 505 – Computational data science and analytics

## Project 2 – Exploratory Data Analysis

**Collaboration Policy:** Individual Assignment

**Total Points:** 100 (See rubric below)

**Project Topic:** In this project you'll select a publicly available dataset (from a data repository like Kaggle, https://www.kaggle.com) and perform the analysis/visualization items listed below. It is important to note, in this assignment you'll be making many of the decisions, and each decision must be described in your notebook. For instance, if your building a linear prediction model, what is your dependent variable and your dependent variable(s), and why did you select these, i.e. what was your hypothesis? Which software model did you choose, in which software library, and why? These are just a few examples. In general, assume you'll give your work to a colleague, and you're providing them descriptive information about your thought process. These "though processes" will be outlined in your markdown section or using code comments.

A. Design, develop, and evaluate the accuracy of a linear prediction model that has one dependent variable (i.e. $\hat{Y} = a + bX$). Your evaluation must include the correlation coefficient ($r$) **and one** of the following: 1) mean square error (MSE), 2) mean absolute error (MAE), or root mean square deviation (RMSD). Hint, you may use existing software models in the sckit-learn library.

B. Design, develop, and evaluate the accuracy of a linear prediction model that has two or more dependent variables (i.e. $\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$). Your evaluation must include the correlation coefficient ($r$) and **one** of the following: 1) mean square error (MSE), 2) mean absolute error (MAE), or root mean square deviation (RMSD). Hint, you may use existing software models in the scikit-learn library.

C. Visualize the prediction results in A or B.

D. Design, develop, and evaluate the accuracy of a two-class (i.e. only two labels, $y \in \{1, -1\}^n$) linear classification model using a support vector classifier (SVC). The training vectors $x_i \in \mathbb{R}^p, i = 1, \ldots n,$ are real-numbers ($\mathbb{R}$) with dimension ($p$). Your evaluation must include the correlation coefficient ($r$) and a 2x2 confusion matrix and **two** of the following: 1) positive predictive value (PPV), 2) negative predictive value (NPV), sensitivity, or specificity .

E. Visualize the classification results in D. Hint, you may use existing software models in the sckit-learn library.

**Approved software tools:** You may only use the Python programming language and the following technologies/software libraries: 1. Python notebook, 2. NumPy, 3. MatPlotLib, 4. SciPy, or 5. Scikit Learn

**Submission:** Submit your notebook and your data (in .csv format) to the Dropbox setup on OAKs before the due date. The name of your notebook must be, P2_<lastname>.ipynb, where <lastname> is your last

name ☺. For example, *P2_Munsell.ipynb* would be correct if I were to submit the assignment. Please, only use one notebook to complete this assignment, i.e. no other Python files (.py files) are allowed.

**Grading Rubric:**

| | |
|---|---|
| Python notebook compiles and runs with no syntax errors | 5 points |
| Your notebook must include a markdown section that describes what each cell is accomplishing. For instance, what is the cell computing or visualizing, and what mathematical, statistical, or visualization techniques is being applied. | 10 points |
| In each notebook cell, include comments that provide more in-depth information about how the computation is performed (i.e. step-by-step approach) or how the data is stored in a particular array or matrix. | 10 points |
| Items A, B, D (20 points each) | 60 points |
| Items C and E (7.5 points each) | 15 points |
| | 100 points |