**IBM Machine Learning Course – EDA Project**

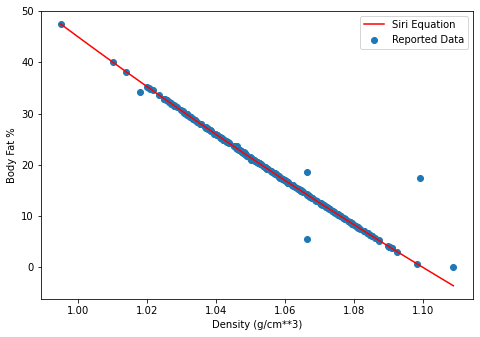
**Background:**

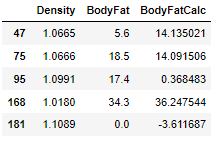
The dataset selected for this project is found here: <https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset>. Features of the data include body measurements (e.g. height, weight, abdomen circumference) of 252 males. The first two columns of the dataset are the hydrostatically measured body density and the corresponding body fat percentage calculated by the Siri equation. The goal of the model is body fat percentage (BFP) prediction from measurements taken at home, without the need for a hydrostatically measured body density.

**Exploratory Data Analysis (EDA) Outline:**

Data cleaning will begin by corroborating the tabulated BFP data with the Siri equation. Discrepancies between calculated and reported BFP will be removed and attributed to human errors such as typos or data wrangling issues. It is noted that these data can later be evaluated with a model residual analysis once a model is selected. A Body Mass Index (BMI) feature will be added to the dataset to uncover incorrect height and weight records, and to be used in the predictive models. Later, BMI may be used to stratify the model by weight category (healthy, overweight, obese, etc) to increase the precision of the results. Pair plots will be constructed to identify features that are not strongly correlated with the body fat percentage target. Variables with low to no correlation will be removed from the model. Early iterations of the model will be based on a small number of features. Additional features will be added in subsequent iterations. The production model will, naturally, be selected from a tradeoff between the number of features and their benefit to prediction results.

**EDA Results:**

 Comparing the Siri Equation and the reported data allowed 5 observations to be removed from the data set. Reimplementing the Siri equation using the densities provided led to this plot:

Where the red line is the Siri equation predictions for body fat % and the blue points are the measurements in the dataset. Considering the reported data was directly calculated from the Siri equation, points that do not directly fall on this line are deemed erroneous and removed from further analyses. A tabular view of the outlying points is shown here:

The addition of a BMI feature revealed one observation with a BMI of 165! The entire row was removed from analysis.

Pair plots illustrate certain observations to be more strongly correlated with the target variable than others. The BMI, Abdomen, and Chest measurements will be the critical features used to predict the target; while Age, Ankle, and Height likely can be ignored without detriment to model performance. Further exploration of features relationships to the target variables will be conducted upon implementation of a model.

**Hypothesis Testing:**

Three hypotheses about these data are that the three best predictors of body fat percentage will be measurements of BMI, Abdomen, and Chest. An easily verifiable hypothesis is that Height is not correlated with body fat percentage. The alternative hypothesis is patient height does have a positive influence on body fat percentage. A correlation coefficient hypothesis test with alpha = 0.05 was done and the resulting p-value was 0.15. Since p>alpha, we do not reject the null hypothesis.