COMP4432 - Assignment 1

Learning objective: This assignment will offer the opportunity to construct a proof-of-concept model (a model built with a minimal demand of time and effort that demonstrates that more investment would be worthwhile). A linear regressor will be built using a single input feature that has been identified as the best predictor of the target variable from exploratory data analysis.

Part 1: Data Importing, Exploration, and Preparation

a) From the diabetes bunch located in SKLearn Datasets, load the diabetes dataset and target set into a single Pandas DataFrame.

b) Calculate and present the descriptive statistics and skewness for the features in the DataFrame.

c) Show a histogram of each feature in the DataFrame.

d) In a Markdown cell, share some thoughts on the preprocessing and distributions of the data (ie, centers, variances, and scales of the features, should further preparation or transformations be considered?)
Hint: The *DESCR* key within the bunch shares details on the data preparation.

e) Using correlation analysis, identify the individual feature that best predicts the target (disease progression). Document the single feature selected in a Markdown cell.
Hint: This is potentially a single line of code using built-in Pandas functionality.

f) Using *train_test_split* from SKLearn, partition the data into training and testing data sets. Set aside 20% of the data to be used as the test set. When implementing the splitting function, set the random state so the work can be reproduced. The training and testing input data should consist of only the single feature identified in Part 1d.

Part 2: Model Training

a) Instantiate a linear regressor from SKLearn and train it with the single feature identified from Part 1d and targets from the **training** data.

Part 3: Prediction and Evaluation

a) Print the linear regressor's feature coefficients (slope and intercept)

b) Print the root mean squared error (RMSE) of the model's predictions on the **training** data

c) Print the standard deviation of the **training** data target values

d) In a Markdown cell, discuss the implication of the difference between the RMSE value in Part 3b and the standard deviation in Part 3c.

e) Print the root mean squared error (RMSE) of the model's predictions on the **testing** data

Part 4: Visualization

a) Using either Matplotlib or Seaborn, construct a scatter plot with the single feature training and testing data identified in Part 1e on the x-axis, and the target (disease progression) on the y-axis. Use different color codes to distinguish the training data from the testing data. Label the x-axis with the single feature name, and the y-axis with target attribute name. Plot the regression line from Part 2 on this same graph. Include a legend to identify the training and testing data, and the regression line.