**Stroke Prediction**

**Prediction Analysis**

***Author****:* George Ajayi, ***Contact:*** [g.ajayi86@gmail.com](mailto:g.ajayi86@gmail.com), 860-326-4284

**Business problem:**

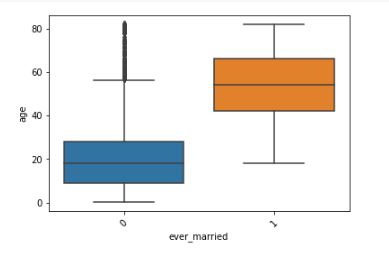
We need to resolve a classification issue which involves a prediction whether a patient will suffer stroke or not in the future or in the present using some contributing factors (data features).

**Data:**

The Data source is a csv flat file from World Health Organization, the Data is titled Stroke Prediction which consist of 12 columns and 5110 rows. The dataset consist of 3 different datatypes which are objects, floats and an integer. Both Numeric and Categorical data.

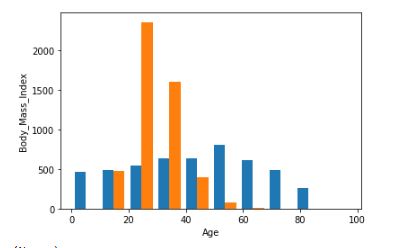
**Methods:**

* Import Packages
* Load Data
* Explore Data
* Perform a Validation Split
* Preprocess the data
* Analyze the data for correlating features
* Create & fit 2 separate Models
* Compare and Evaluate the score for the best choice
* Apply Feature engineering (PCA) for efficient runtime.

**Boxplot Visualization**

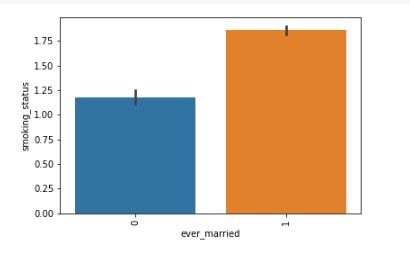
In the Boxplot above, the average mean of age of unmarried patient is under 20 while the average mean age of married patient in our dataset is 55. This box plot shows the age range among potential stroke patients based on marital status.

**Histogram**

****

In the Histogram above, it shows patient at age 24 and under, and those above the age of 75 have a lower Body Mass Index. Teenagers and young adult are a lot mobile and tend to engage in more outdoor activities than any other age groups which contributes to a lower Body Mass Index. Those in of age range of 75 and up tend to shrink down in size the older they get which also contributes to a lower Body Mass Index. The histogram shows other age groups with a higher Body Mass Index. Obesity and Excess weight does increase the risk of severe illness and other health problems.

**Bar-plot**

****

In this Bar-plot, we see that marital status correlates with smoking status. We also see that married patients have a higher smoking rate than unmarried patients. Marriage can cause lots of stress that turns a none-smoker into a smoker.

**Summary of the Model and the Metrics**

The 2 models were created to predict the outcome of stroke, the first model is a DecisionTreeClassifier and the 2nd model is KNeighborsClassifier. Evaluating both models using GridSearch CV and Classification Report, we end up with an identical test score for both first and second model.

***GridSearch CV test score:***

First Model = **0.9470684039087948,** 2nd Model = **0.9470684039087948**

***Classification Report test score:***

First Model = **0.95,** 2nd Model = **0.95**

**Model 2 was chosen for production because it has just the concise number of hyper parameter needed for tuning which also makes** KNeighborsClassifier **an ideal model to use when solving a classification problem such as the stroke prediction.**