**Phase III: Inferential Statistics**

Experiment: down-sampled training set (class imbalance adjusted)

* As discussed above, to build a model which can classify on stroke patients rather than non-stroke patients, resampling was performed. Originally, there were 98% of non-stroke and only 2% of stroke cases represent in 42,512 of total observations. Thus, down-sampling with randomized replacement was performed to reduce the sample size of non-stroke cases and adjust proportion of classes being 1:1 ratio by matching exact sample size (2,142) of stroke cases to non-stroke cases.

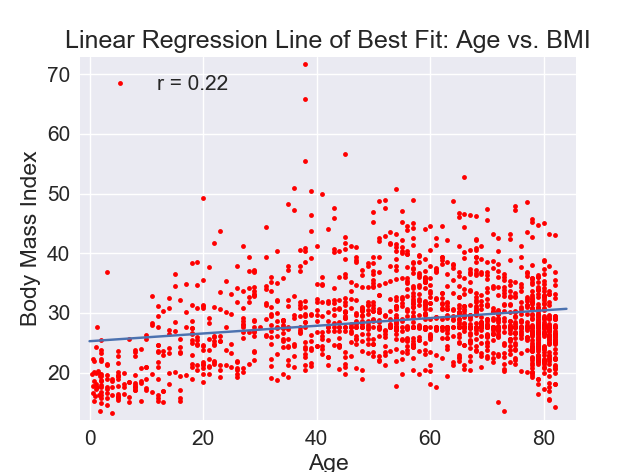
Correlation Matrix Plot: down-sampled training set (class imbalance adjusted)

* From independent variables to dependent variable interaction, the correlation of an independent variable “age” and a dependent variable “stroke” was reported as moderate positive with 0.59. Followed by a pair of avg. glucose level and stroke with a correlation value of 0.25 and a pair of bmi and stroke with a correlation value of 0.10.
* From independent variable to independent variable interaction, age and avg. glucose had a weak positive correlation of 0.25 followed by age and bmi with reported value of 0.22. Also avg. glucose level and bmi had a weak positive correlation of 0.21.

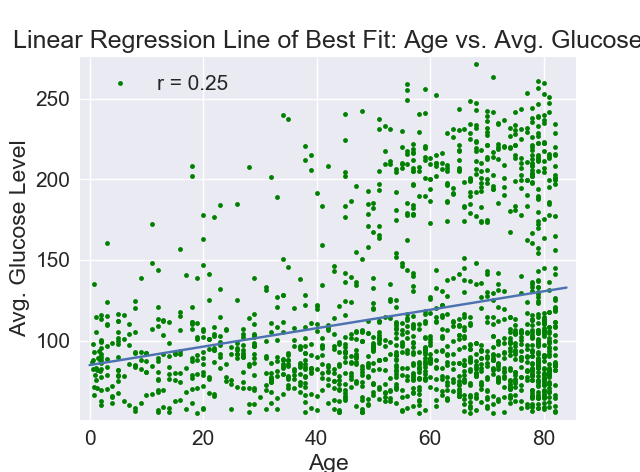
Linear Regression Analysis:

* Linear regression analyses were performed and plotted for following pairs of independent variables interaction: age vs. bmi, age vs. avg. glucose level and avg. glucose level vs. bmi.

1. Age vs. BMI:



1. Age vs. Average Glucose Level:



1. Average Glucose Level vs. BMI:

