**Machine Learning Portion of Website:**

Hypothesis: Number of deaths are dependent on obesity, tobacco, and income

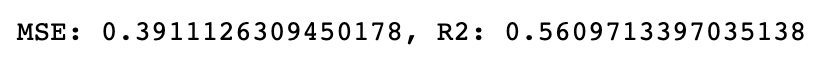
Testing and training data was scaled. Pandas get\_dummies and feature engineering was utilized to get a better fit model due to our dataset size and originally low R2

Artificially added new columns to make dataset more robust

* Sum
* Mean
* States & years transformed into categorical data

Analysis:

(This should pic of the MSE and R2 should be put under the two figures before the analysis)



* Training and testing data follows the relatively same pattern
* Testing data was used to make predictions and a separate plot reflecting that was created.
* The R2 of .56 indicates that the model explains about 56% of the variability of the data around the mean. Due to the number of data points clustered between .5 and 1.0 our prediction was probably a bit on the low side. Negative outliers that go extremely negative indicates the model’s predictions were a little too high.
* The mean scored error, which shows the averaged squared difference between the predicated and actual vales was .39, close to the desired 0 range.
* Overall, there is somewhat of a correlation between number of deaths and tobacco, obesity, and income.

(I thought this quote would be cool to include)

“Essentially, all models are wrong, but some are useful”

[George Box](http://en.wikipedia.org/wiki/George_E._P._Box)