**Machine Learning Portion of Website:**

Hypothesis: Number of deaths is correlated to obesity rates, tobacco use, and per capita income combined.

Pandas get\_dummies and feature engineering was utilized to get a better fit model due to our dataset size and originally low R2. Columns that were artificially added and calculated were sum and mean. States and years were transformed into categorical data. Testing and training data was preprocessed through scaling to normalize the data within a smaller range. We ran our analysis against our hypothesis to get our residual plot. Prior to that, we decided to run linear regressions to see the correlation between death and income, tobacco, and obesity individually. Actual and predicted data was plotted on the same scatterplot.

A screenshot of a cell phone

Description automatically generated

This plot shows the actual and predicted data follows the same pattern, showing a strong correlation between death and income. There is no clear trend between death and income alone.

A screenshot of a cell phone

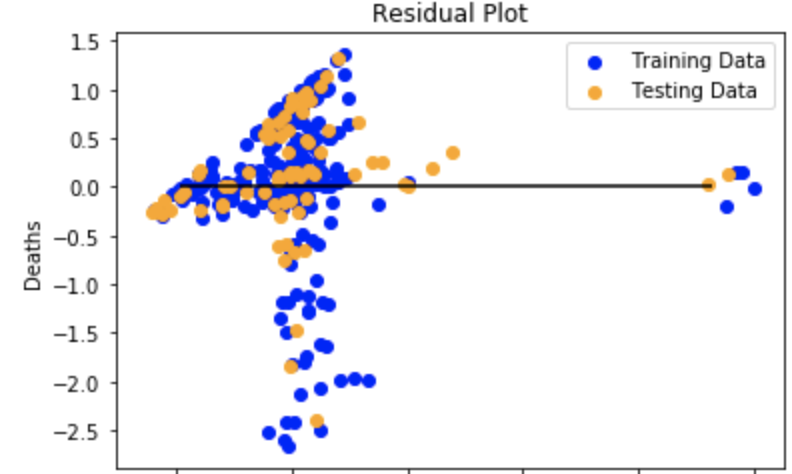
Description automatically generated

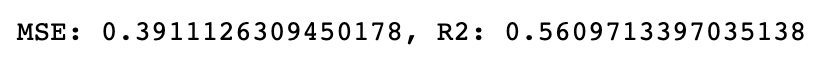
This plot shows a strong correlation between number of deaths and tobacco use as actual and predicted data follows the same pattern. There is a weak negative trend, as number of deaths decreases as tobacco use increases when examining directional impact.

A screenshot of a cell phone

Description automatically generated

This plot shows that there is a strong correlation between obesity and death as actual and predicted values follow the relatively same pattern. There are a couple outliers in the data that don’t make the correlation as strong. There is a weak positive trend when looking at the directional impact of the plot.





Training and testing data follows the relatively same pattern. Data points clustered between .5 and 1.0 indicates the model’s prediction was probably a bit on the low side. Negative outliers that are between -1.5 and -2.5 indicate the model’s predictions were a little too high. The R2 of .56 indicates that the model explains about 56% of the variability of the data around the mean. The differences between states are significant and can cause outliers and contribute to the skewed results of the analysis. The mean scored error, which shows the averaged squared difference between the predicted and actual values was .39, close to the desired 0 range. Overall, the probabilities of a positive correlation between number of deaths and tobacco use, obesity rates, and per capita income combined are relatively low. These factors alone are not enough to cause a strong correlation. Increasing the amount of data with more variables would have improved the results of the predicted model.

“Essentially, all models are wrong, but some are useful”- [George Box](http://en.wikipedia.org/wiki/George_E._P._Box)