# Sean McLean

# ALY 6020

# Module 6

# Final Exam

## Introduction

The final project looks at two datasets from a Boston hospital that has partnered with the healthcare company that I am currently employed with. The datasets pertain to demographics and health information of stroke patients and the demographics of cancer patients in every county in the United States. The goal of the project is to understand through data analysis and research what variables contribute to strokes, cancer, and deaths linked to cancer. The analysis conducted consists of the implementation of several models that will help solve these problems.

## Data Preparation/Analysis

After uploading the necessary packages and libraries, the two datasets were imported into the Python notebook. The stroke dataset contains 12 variables total and 5,111 cells, with the 'stroke' variable as the dependent variable being examined. The target variable is a binary variable because it only has values of 0 or 1, indicating whether a patient has had a stroke or not. The cancer dataset features 34 variables and 3,048 rows with the dependent variable being identified as the 'target death rate' variable. These values in the column pertain to the rate of deaths from cancer per capita in each county for an unspecified time period.

A conclusion of the summary statistics shows missing values in a few of the columns, outliers in several columns of the cancer dataset, and a few variables that are deemed unnecessary before the modeling process begins. Of the columns with missing values, only one of them ('PctSomeCol18_24' = percentage of people in the 18-24 age range that attended some college) contained a considerable number. And because it was regarded as an irrelevant variable to the study because there are similar variables in the cancer dataset, it was entirely removed.

During this execution, the ‘binnedInc’ variable was also removed because it was redundant when compared to other income variables. The other rows in both datasets that had some missing values were replaced with the mean value of the column so no important data was left out of the study.

The stroke dataset contained a combination of data types that are both categorical and numerical, so for preparation before modeling the categorical variables were converted to binary variables. This process was effective because the variables in the dataset each had only a few possible values. While analyzing the cancer dataset there were several rows in the ‘median_age’ variable that contained values that were too high for any individual to live. Those values were erased and updated with the mean age of the variable.

The only column in the cancer dataset that was categorical was the ‘geography’ variable, but converting this column to numerical was difficult in that each value was different in the dataset since it a dataset of every county in the country. To simply the variable the counties were eliminated and the remaining states in each cell were converted to numerical values ranging from 0-49. Each value pertains to the state, and it is in alphabetical order, so as an example 0 is Alabama, 1 is Alaska and so on until 49 which is Wyoming. After these data preparation revisions, the summary statistics were evaluated again to verify all the updates in both datasets have successfully executed.

## Modeling/Results/Insights

The models used for the stroke dataset were logistic regression modeling as well as random forest modeling. The logistic regression model was used because of its ability to provide input on the relationships between the dependent and independent variables. The random forest model was also incorporated for how it handles non-linear attributes and for identifying the most impactful independent variables. Both models seek to understand the accuracy of the dataset and its relationships between variables (Anaconda, 2024).

After dividing the dataset into training and testing sets, a classification report for both models were executed to show its accuracy. The key metric in the report is looking at the F1 score which is calculated when combining the precision and recall values. The precision value will tell us how many of the predicted positives are actually true positives while the recall will provide how many of the actual positives were correctly identified by the model (Shung, 2018). The reports for each model are similar in their results, with both F1 scores for the rows in the dataset that had a value of 0 or no stroke at 0.97 or 97 percent, and 0.00 or 0 percent for rows in the dataset that had a 1 or had a stroke. This was based off a testing set that had 1,444 rows of 0 or no stroke patients and only 89 patients with a 1 or that had a stroke.

The results of the classification report were verified by creating a confusion matrix that showed that the logistic regression model was 100 percent accurate and that the random forest model was 99 percent accurate (See Appendix A). To better understand the model performance in imbalanced scenarios, an ROC plot was built. The ROC curve displays how the true positive rate (sensitivity) and false positive rate (1 - specificity) change as the classification threshold is adjusted. The AUC (Area Under the Curve) quantifies the model's ability to differentiate between positive and negative classes, with a value close to one indicating strong performance and a value near 0.5 suggesting that the model performs no better than random guessing (Hoo, 2017). Both models perform very well with the logistic regression model performing slightly better at 0.84 AUC (See Appendix B).

The imbalance of the dataset is the main issue because there are very few people that have had strokes compared to not. If the question is what factors can contribute to having a stroke, then there is not enough data to provide enough evidence that certain variables have an effect or not. This dataset could have some value if the question is adjusted to what factors can lead to decreased chances of having a stroke since most people in the dataset have not. Some other issues are that there are not enough variables about the individual's health that can be looked at when determining who is more or less at risk. Different modeling methods could then be utilized to show these linear relationships to the target variable and could potentially make up for the imbalance in the dataset.

The models used for the cancer dataset were generalized linear regression modeling as well as gradient boosting modeling. These two models were selected because of their ability to handle non-linear relationships and how they can interpret the data. The dataset was divided into a training and testing set that was fitted into a GLM model where the results were visualized in a residual plot (See Appendix C). These results are compared with plots that show the residuals versus the fitted plot (See Appendix D) and the actual values versus the predicted values in a generalized linear model (See Appendix E). The data points in the residual plots show the residuals exhibit a random dispersion around zero with no noticeable patterns. This randomness suggests that the model's assumptions are valid and that it appropriately fits the data (Frost, 2024). The majority of the data points in the linear regression model are within range of the regression line with a few outliers, so it indicates that the model's predictions are accurate.

The test data is fit into a gradient boosting model that returned a root mean square error of 18.26 units which suggests that the model is working well because of the largest values in the target variable. A plot is created to show the feature importance of the model, and the variables that have the most effect on the target death rates are ‘incidence rate,’ ‘pct bach deg 25 and over’ and ‘pct hs 25 and over’ (See Appendix F). An evaluation of the generalized linear regression model results reveals that several variables contain p-values below 0.05 and positive coefficient values. These variables ('povertyPercent', 'PercentMarried', 'PctHS18_24', 'PctHS25_Over', 'PctEmpPrivCoverage') were each placed into scatterplots to be compared and analyzed with the dependent variable (See Appendix G). Of the five variables, three of them ('povertyPercent,' 'PctHS18_24', 'PctHS25_Over') showed positive linear relationships that suggest they impact the target death rates in their communities. These findings were verified with a correlation matrix that showed the three variables having a positive correlation to the dependent variable with the other two variables ('PercentMarried', 'PctEmpPrivCoverage') showing a negative correlation (See Appendix H).

After the adjustment of the column ‘geography’ to look at the states of each cell instead of the county, a bar plot was created to see if any trends and patterns are noticeable (See Appendix I). The states with the highest target death rates are Kentucky, Mississippi, Arizona, Tennessee, Louisiana, and Alabama, and the states with the lowest rates are Utah, Hawaii, California, and New Mexico. These findings indicate that the data can be looked at by region in future analysis since most of the states with the highest rates and lowest rates are in the same regions. Overall, the findings show that the demographics with the highest target death rates are in states in the south region with stronger relationships toward people that are living in poverty and highest level of education are in age groups of 18-24 and 25 and over with a high school diploma.

Some of the short comings with the cancer dataset model are that there are many variables that have a negative relationship with the dependent variable. This could mean that these factors are not influencing the cancer and death rates and that other variables need to be considered and added to the dataset for more analysis. Another concern when looking at the data is whether there could be biases toward certain groups of the population that are contributing to higher death rates. The data could be used against them if it connotes that it is only a certain area of the country that have a higher risk of cancer and death and could exacerbate the situation instead of improving the rates. This is also why more attributes that could pertain to cancer and death should be added so other factors instead of just demographics are looked at.

## Recommendations

I would recommend both datasets to county health departments across the country to provide them the demographic information and models to show the relationships between the variables and the target variables. The revisions in data can also show which states have the highest death rates among cancer patients so county health departments can react and respond quicker to these problems. These findings can be essential to identify early risk factors for health issues like strokes and cancer and figure out ways to mitigate them. This would also allow them to place funding and resources in the necessary places that can help decrease occurrences like strokes and decrease cancer and death rates.

The dataset I would look at first for further research is the cancer dataset due to more variables and the findings from previous modeling that was conducted. From the analysis there were positive linear relationships that can be explored further that could detect key demographics in looking at the dependent variable. A revision of the dataset by looking at certain variables while removing irrelevant ones could also boost the coefficient of determination which is currently above average at 0.67 but could be improved. Overall, the cancer dataset is more balanced and contains more insight than the stroke dataset and can be vital on a national level with the data it contains.

## Conclusion

The two datasets offered the opportunity to use multiple modeling techniques that gave insights into how key variables play a role in medical conditions like strokes and cancer. Through data cleaning and preparation, each was set up for proper modeling and data visualizations that showed their relationships. While both had their flaws, the modeling displayed the accuracy of each dataset and what the next steps should be toward making them more effective.

## References

Anaconda Distribution. (2024). Anaconda (Version 4.12.3). Retrieved August 16th, 2024, from https://www.anaconda.com

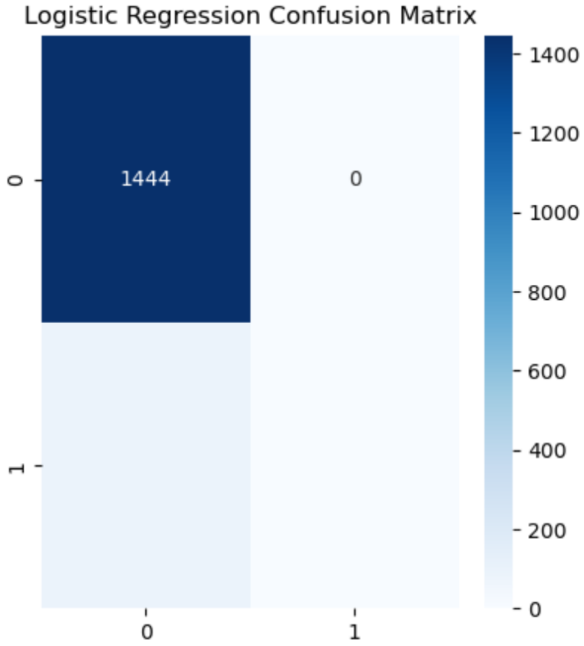
Frost, Jim (2024, August 16$^{th}$). *Check Your Residual Plots to Ensure Trustworthy Regression Results*! Statistics By Jim. https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

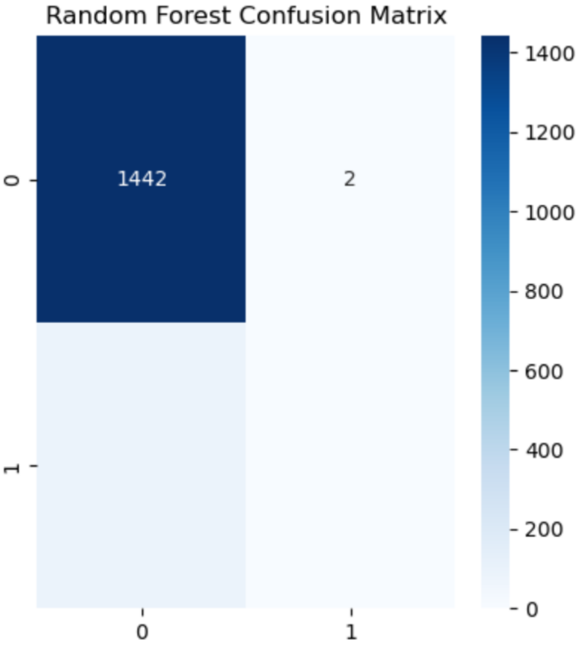
Hoo, Z., Candlish, J., & Teare, D. (2017). *What is an ROC Curve?* Emergency Medicine Journal, 34(6), 357-362. https://doi.org/10.1136/emj-2023-123456

Shung, K. (2018, Mar. 5th). *Accuracy, precision, recall, or F1?* Towards Data Science. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

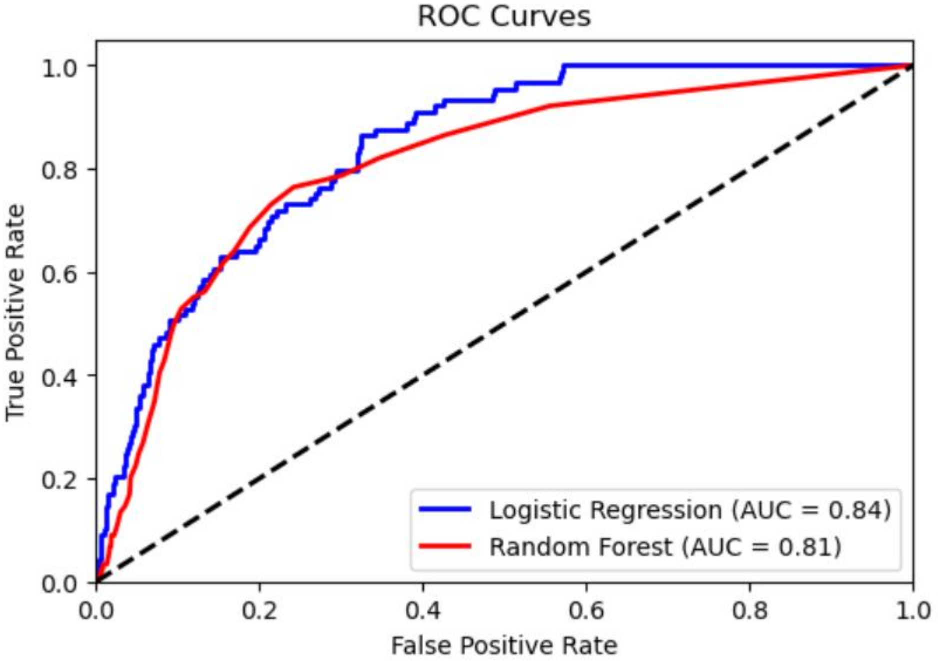
# Appendix

## Appendix A

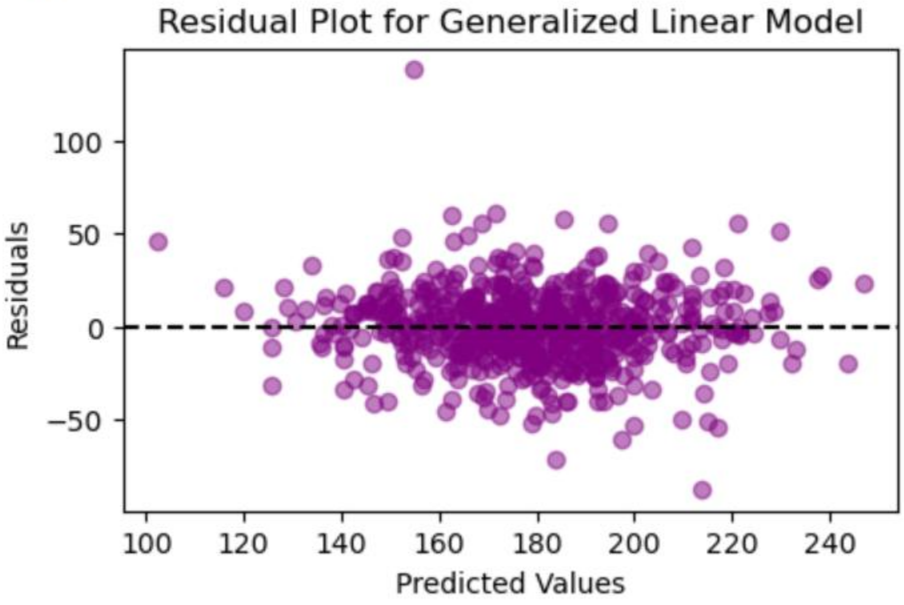




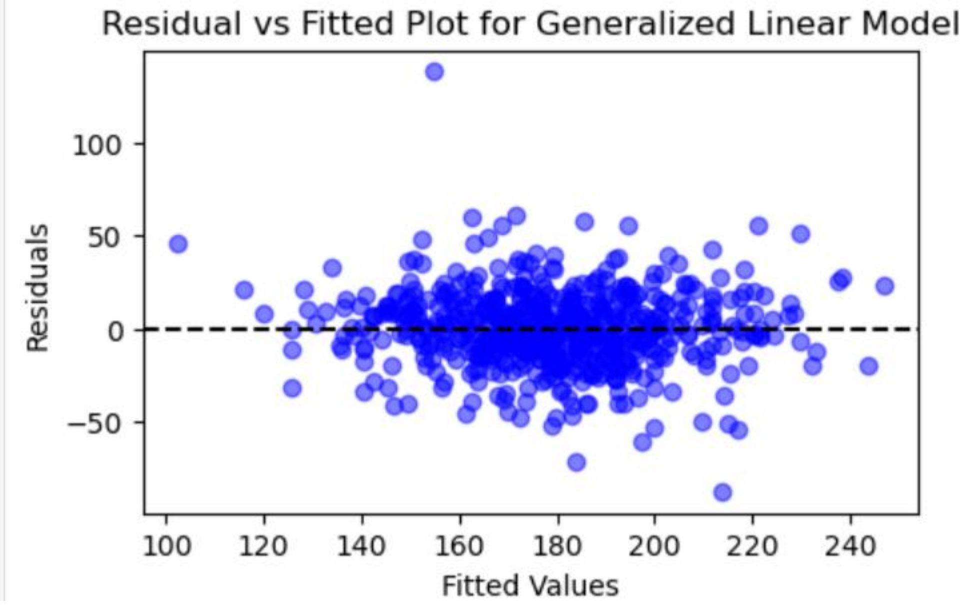
## Appendix B



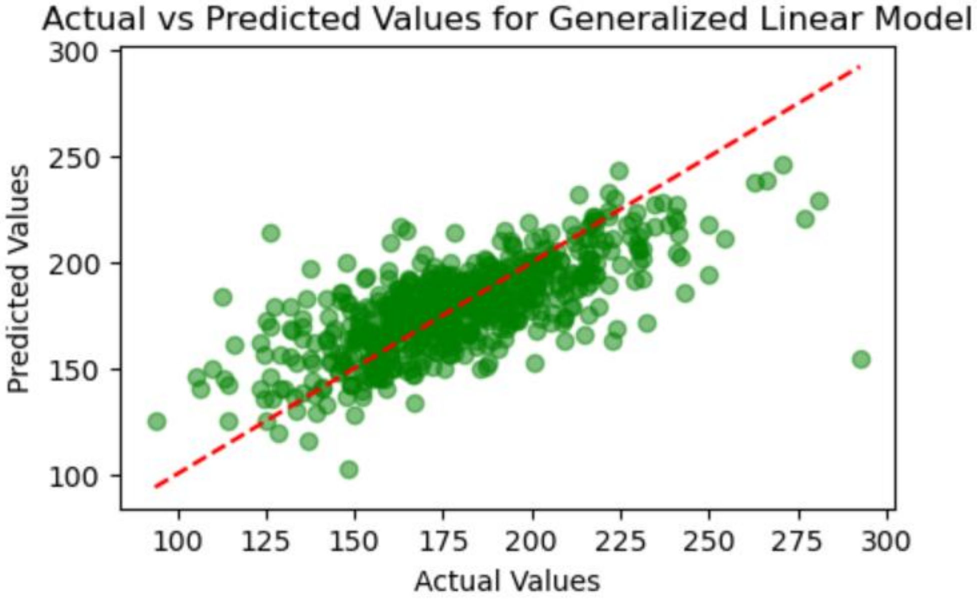
## Appendix C



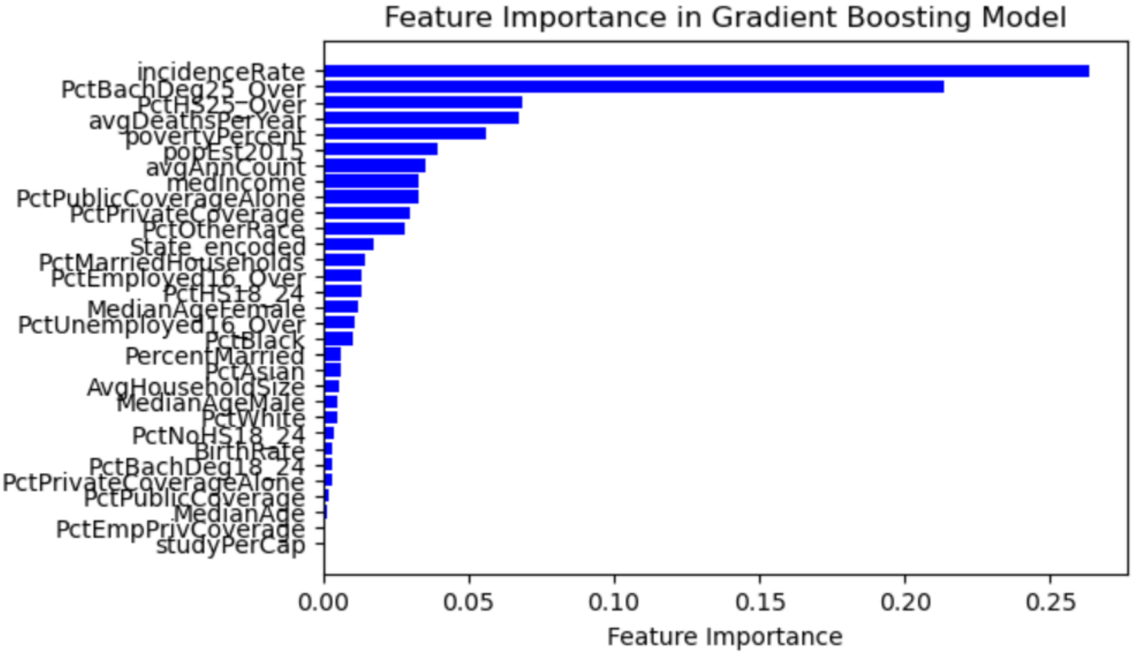
## Appendix D



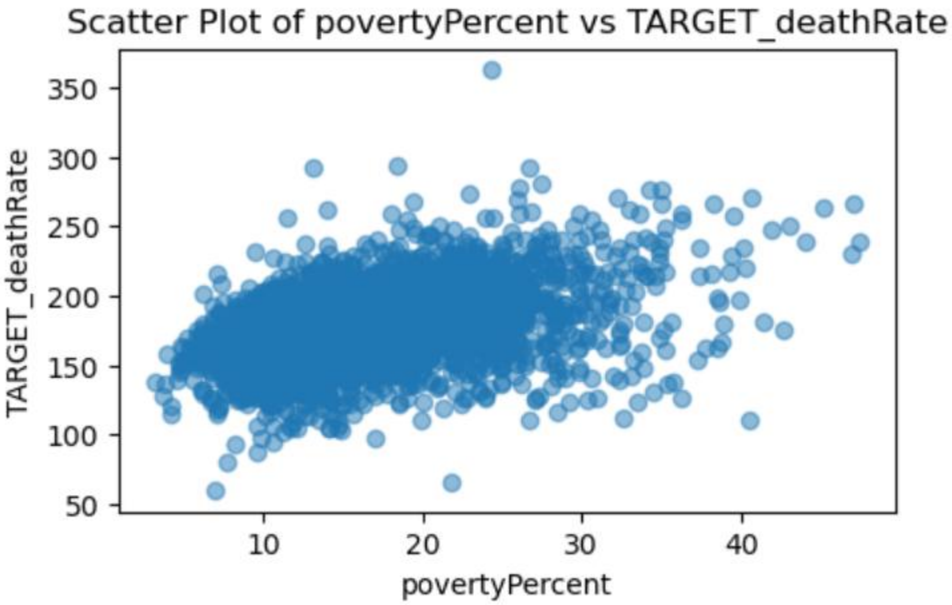
## Appendix E

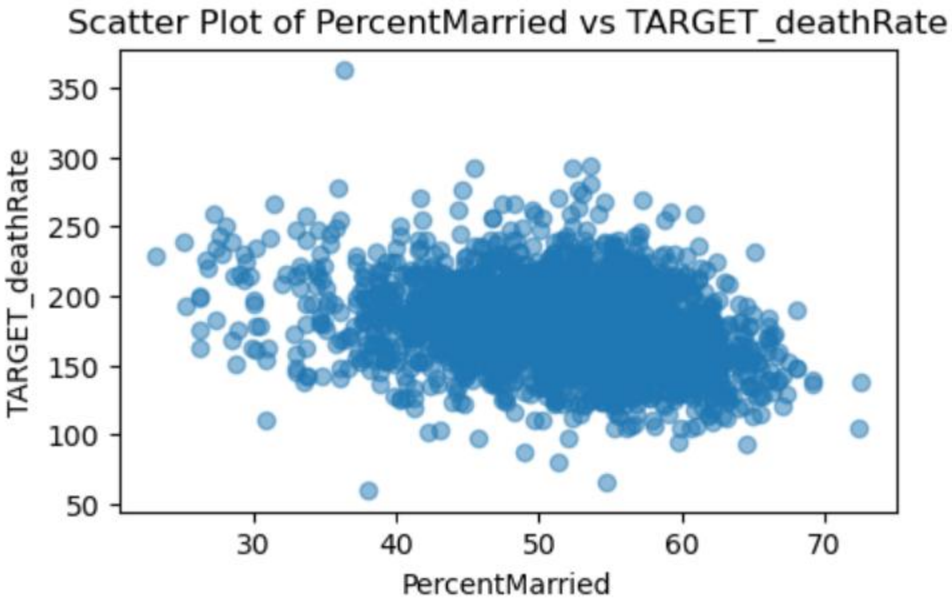


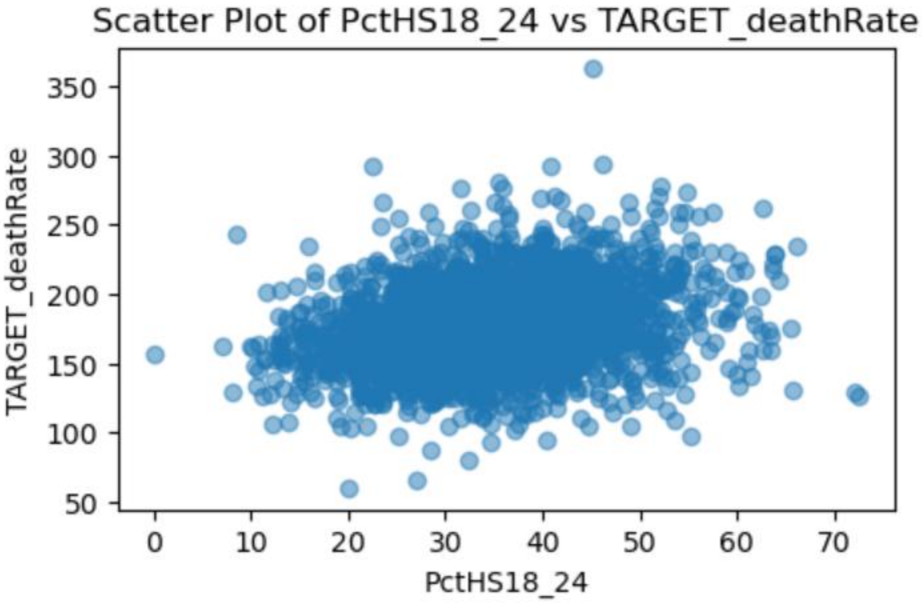
**Appendix F**



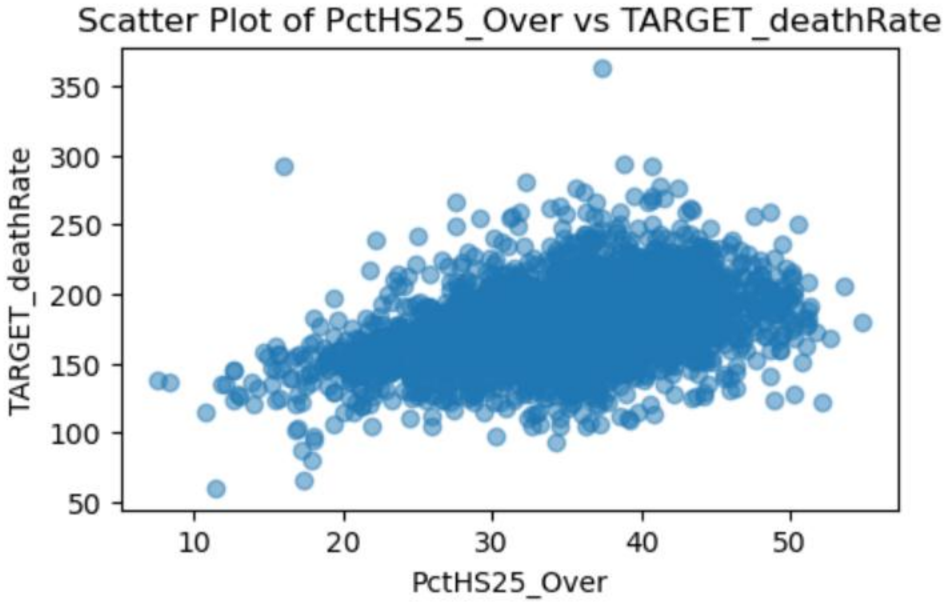
**Appendix G**

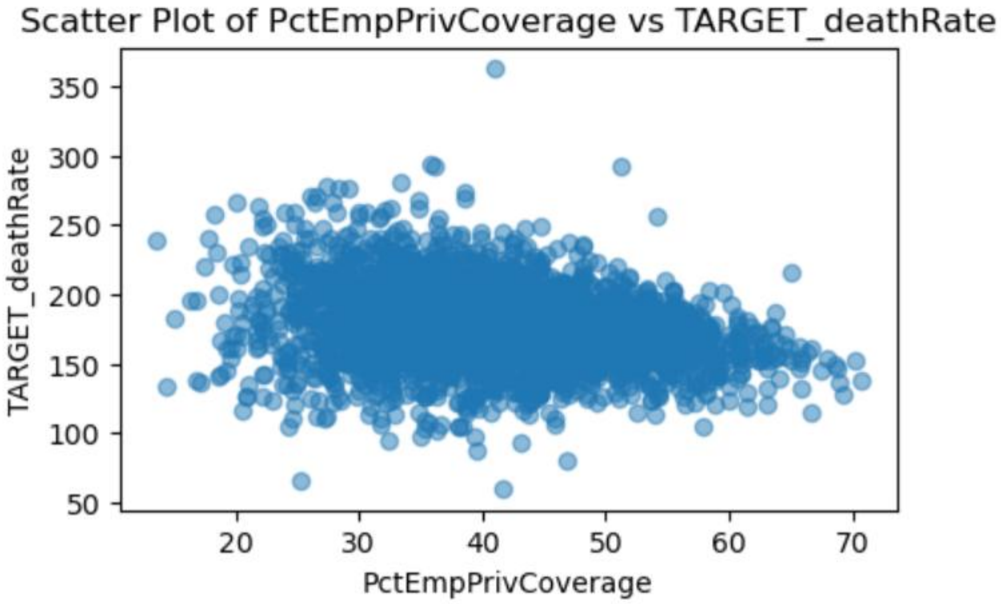

Scatter Plot of povertyPercent vs TARGET_deathRate
TARGET_deathRate
350
300
250
200
150
100
50
10
20
30
40
povertyPercent


Scatter Plot of PercentMarried vs TARGET_deathRate
TARGET_deathRate
350
300
250
200
150
100
50
30
40
50
60
70
PercentMarried

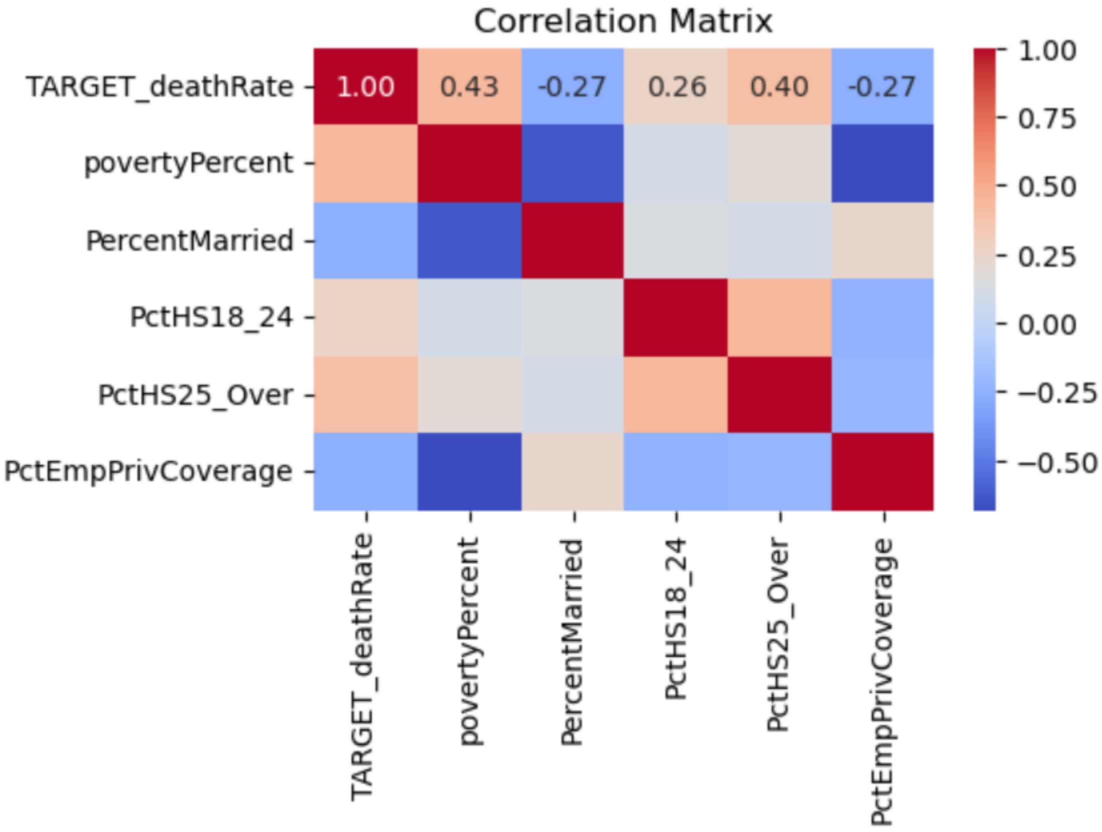

Scatter Plot of PctHS18_24 vs TARGET_deathRate
TARGET_deathRate
350
300
250
200
150
100
50
0
10
20
30
40
50
60
70
PctHS18_24

_

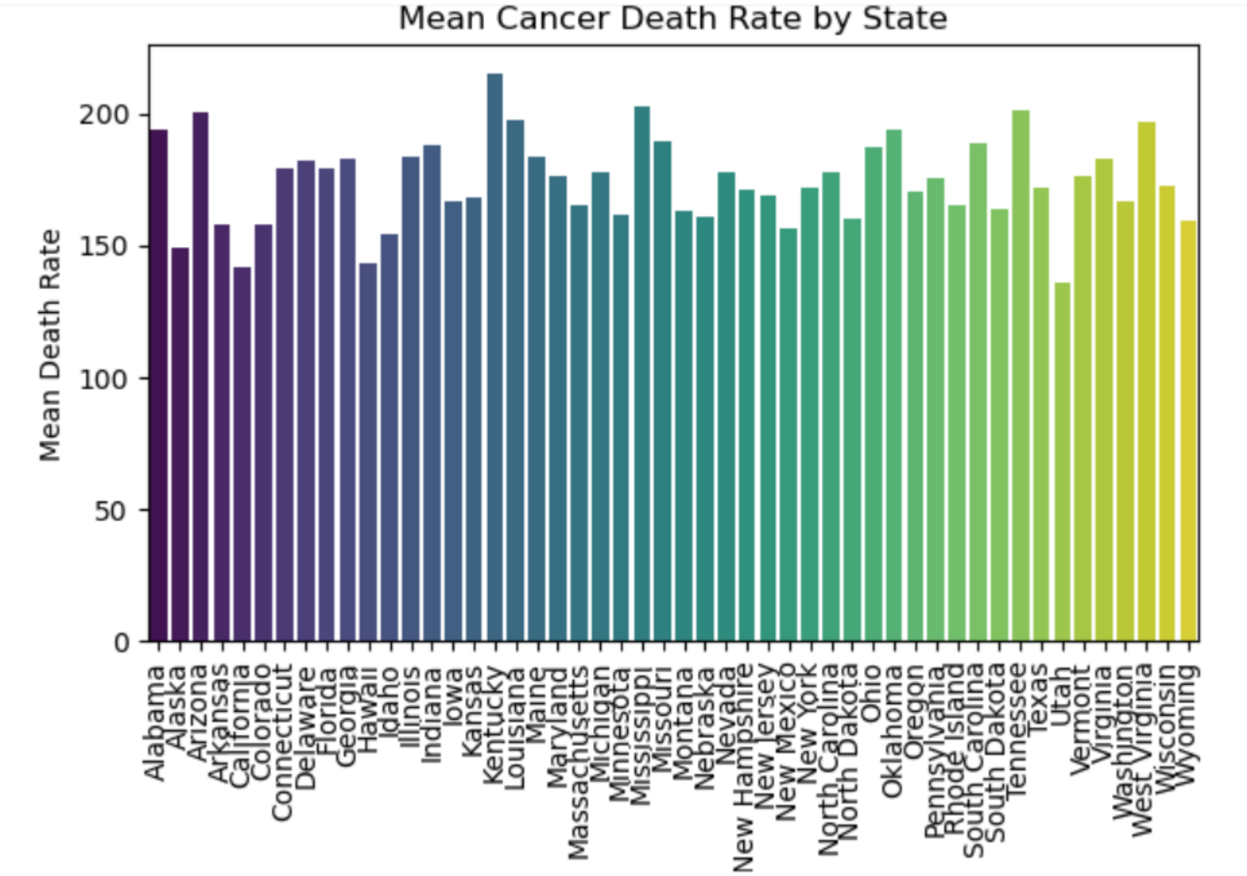

Scatter Plot of PctHS25_Over vs TARGET_deathRate
TARGET_deathRate
350
300
250
200
150
100
50
10
20
30
40
50
PctHS25_Over


Scatter Plot of PctEmpPrivCoverage vs TARGET_deathRate
TARGET_deathRate
350
300
250
200
150
100
50
20
30
40
50
60
70
PctEmpPrivCoverage

**Appendix H**


Correlation Matrix
TARGET_deathRate
povertyPercent
PercentMarried
PctHS18_24
PctHS25_Over
PctEmpPrivCoverage
1.00
0.43
-0.27
0.26
0.40
-0.27
TARGET_deathRate
povertyPercent
PercentMarried
PctHS18_24
PctHS25_Over
PctEmpPrivCoverage
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50

## Appendix I


Mean Cancer Death Rate by State
Mean Death Rate
0
50
100
150
200
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming