Predicting Cancer Mortality in U.S. Counties: 2015

By Dan Loew

https://www.linkedin.com/in/danielloew/

The Problem: Understanding Socioeconomic Influences on Cancer Mortality

Cancer is an ever-present threat to our health, productivity, and potential, and the effects that socioeconomic, geographic and dietary factors have on cancer mortality are an important area of study.

Understanding which of these factors have the greatest influence on cancer mortality can be a difficult area of study.

The Solution: A Machine Learning Model

This project aims to construct a robust machine learning model using Ordinary Least Squares Linear Regression (OLS-LR) and other regression algorithms to predict cancer mortality rates at the county level using 2015 data from the Census and other sources. Through the construction of this model, the relative influence of socioeconomic, geographic, and food environment variables (i.e. the predictive feature set) can be elucidated by examining the Linear Regression algorithms' coefficients.

An innovative technique of using logarithmic and exponential transformations of the predictive feature set will be used to capture nonlinear relationships between predictive features and cancer mortality within the linear context of the OLS-LR algorithm.

The Data Cleaning Process: Part I

- A pre-assembled dataset was downloaded from https://data.world/exercises/linear-regression-exercise-1 which was comprised of a feature set of socioeconomic indicators and a target variable of the cancer mortality rate for 2015 for 3,047 out of 3,141 total U.S. counties (97% of all U.S. counties and county equivalents).
- This dataset was pre-assembled from three sources: the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov.
- This dataset was cleaned by filling the median value of a feature if that feature had missing or obviously erroneous data, removing a feature because it was a source of machine learning leakage containing the information being predicted in the target variable, and transforming a string feature with categorical answers into a set of binary features.

The Data Cleaning Process: Part II

- The latitude/longitude centroids, water and land square mileage, and state for the 3,047 counties were downloaded from the Census and added to the original DataFrame.
- The latitude/longitude location of the nation's top 10 oncology hospitals and eight regional urban centers were also added, and L1 and L2 distances from each these locations to each county were calculated and added as predictive features. The L1 and L2 distances to the closest top 10 oncology hospital and regional urban center were also calculated and added as predictive features.
- The latitude/longitude locations of the nation's 1,313 EPA-designated Superfund Cleanup sites were also used to calculate the L1 and L2 distances from each county to the closest of these sites, which were added as predictive features.

The Data Cleaning Process: Part III

- A large set of food environment variables such as the percentage of the county's populace with low access to a grocery store, or the number of farmers markets in a county, were downloaded from the U.S. Department of Agriculture's Food Environment Atlas. A set of 75 features of interest from this Atlas were added to the core feature set.
- In order to capture nonlinear relationships between the target variable and the predictive feature set, versions of most of the features with logarithmic and exponential transformations of their values were computed, plotted, and tested for their contribution to the OLS-LR model's accuracy.
- The data cleaning process improved this accuracy from 49.2% to 64.1%.

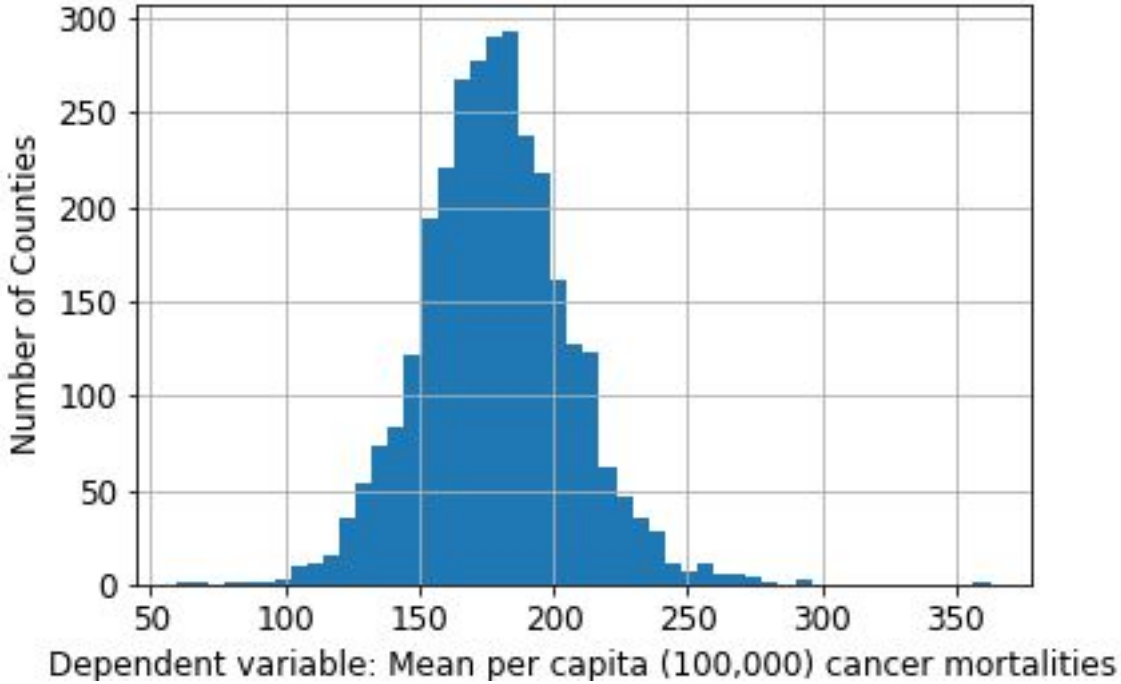
Visual Exploratory Data Analysis (EDA)

A comprehensive exploratory data analysis (EDA) of the cleaned and expanded cancer mortality DataFrame was carried out utilizing:

1. descriptive statistics and distribution plots of the per capita cancer mortality target variable and each feature in a selected subset of the overall feature set used for machine learning later in the project,
2. the correlation and covariance between the target variable and each feature in this selected subset,
3. plots of the logarithmic and exponential versions of the features that added to the OLS linear regression model's accuracy, with supporting details, and
4. geographic distributions of differing rates of cancer mortality across the U.S.A.

Target Variable: Cancer Mortality Rates

The distribution of cancer mortality rates per 100,000 people is visualized below.



Salient Predictor Features

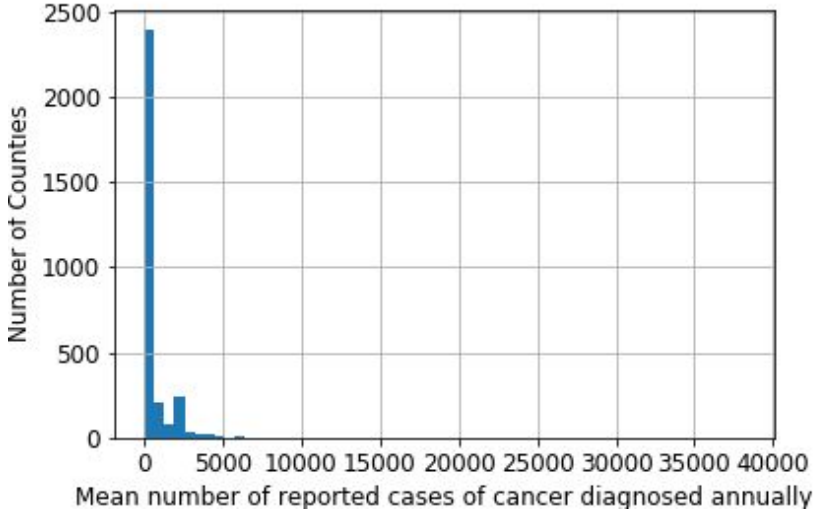
Several of the predictor features stood out as having interesting distributions, strong negative or positive correlations with the target variable of per capita cancer mortality, or had intriguing non-linear relationships with the cancer mortality that were uncovered through logarithmic and/or exponential transformations of specific predictive features' values.

These salient predictor features are further explored through the next series of slides.

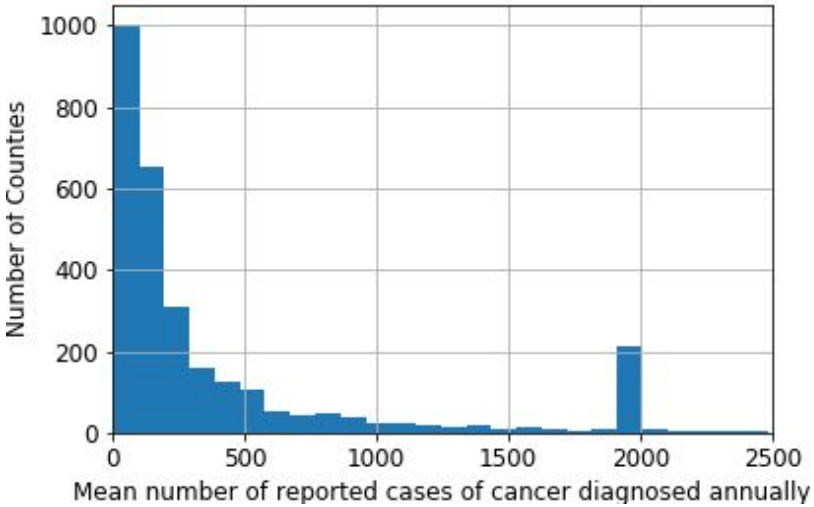
Mean Number of Annual Cancer Diagnoses

The first feature of interest is ‘avgAnnCount’, the mean number of reported cases of cancer diagnosed annually. As can be seen, the majority of the distribution lies between 6 and approximately 2,500 diagnosed cases. There are several outliers, due to several counties with very large populations.

Full Distribution



Majority of the Distribution



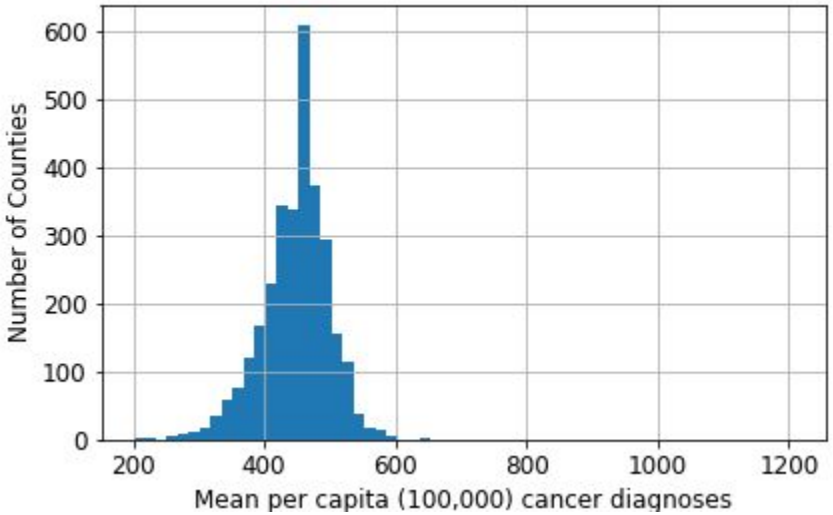
Mean Number of Annual Cancer Diagnoses

There is a slightly negative correlation of -0.14 between the number of diagnosed cancer cases and the number of cancer mortalities. This is a confusing correlation, as one would think that an increase of actual diagnosed cases would result in an increase in cancer mortality. This result could be due to the 'avgAnnCount' and 'TARGET_deathRate' variables being on different scales (i.e., actual count versus per capita), but future research is recommended for this correlation.

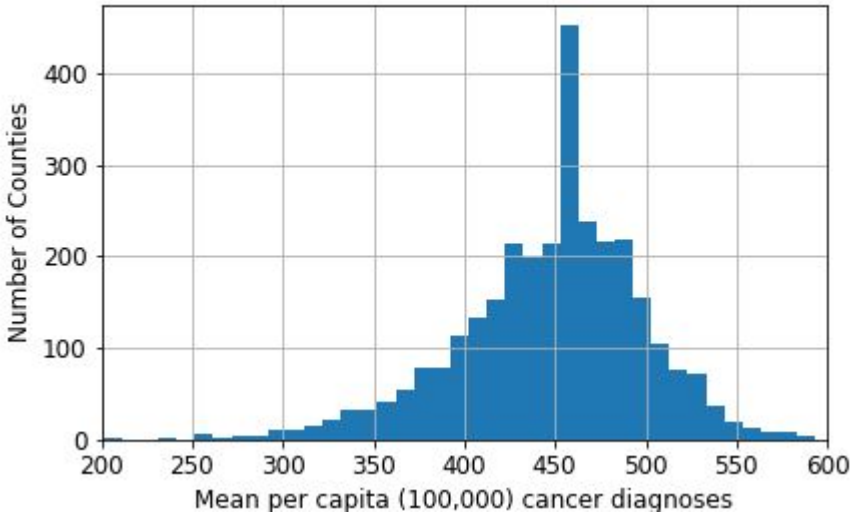
Mean Per Capita (100,000) Cancer Diagnoses

The 'incidenceRate' feature provides the mean per capita (100,000) cancer diagnoses. The bulk of the distribution lies between 200 and 600. The feature had a moderately strong correlation of 0.45 with the cancer mortality rate.

Full Distribution

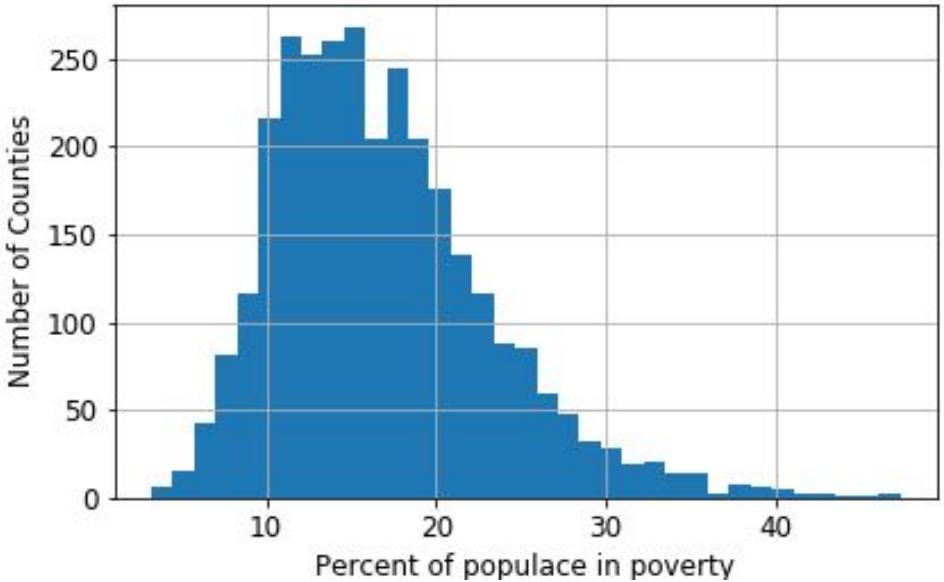


Majority of the Distribution



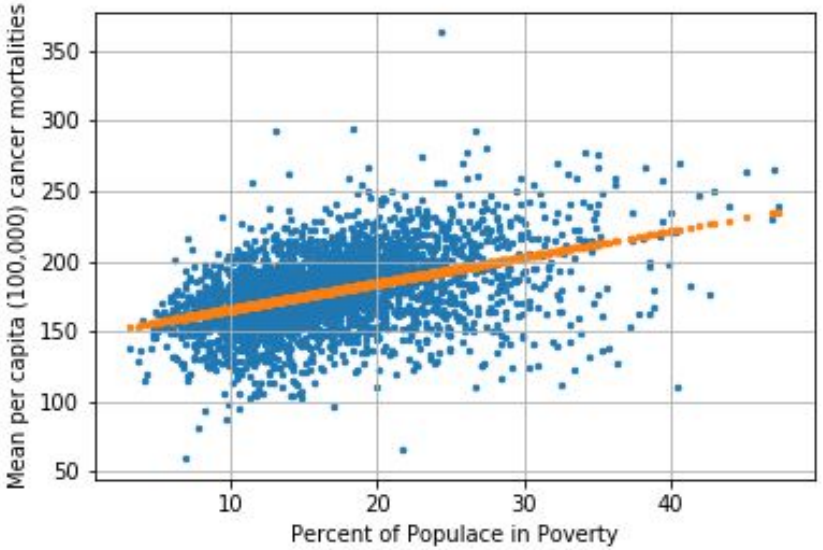
Poverty and Cancer Mortality

The percent of the populace in U.S. counties who live at or under the poverty line ranges from 3% to 47%, and had a correlation of 0.43 with cancer mortality, showing poverty moderately increased the risk of dying from a cancer diagnosis.



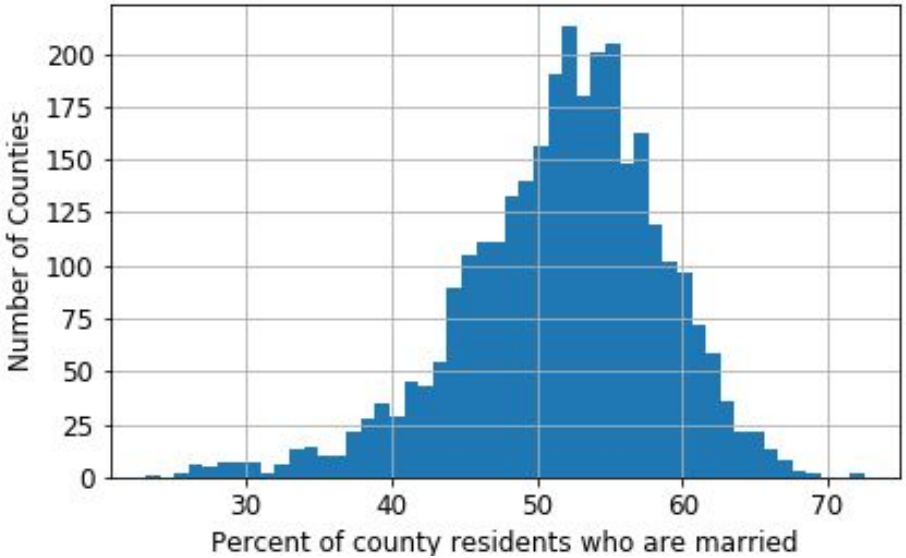
Poverty and Cancer Mortality

The moderately positive correlation between poverty and cancer mortality is seen below. Adding poverty's logarithmic and exponential transformations increased the linear regression model's accuracy by 0.0008.



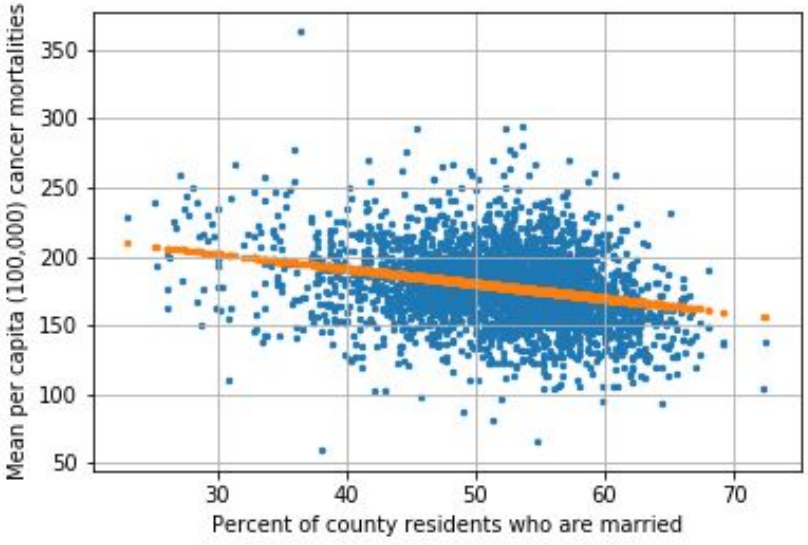
Percent of County Residents Who Are Married

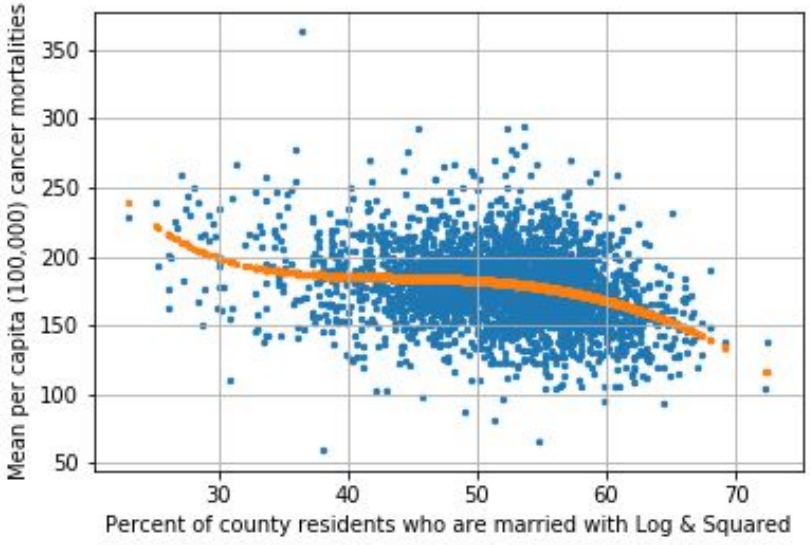
The percentage of county residents who were married ranges from 23% to 73%, and had a correlation of -0.27 with cancer mortality. This suggests a relationship between being married and being less likely to die from cancer, but of course no causative relationship can be posited here.



Percent of County Residents Who Are Married

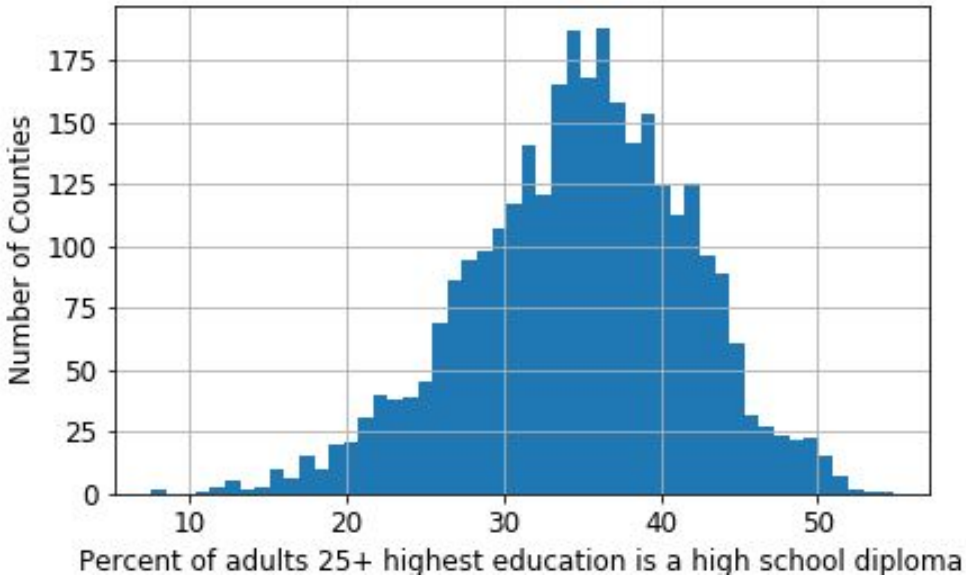
The weak negative correlation between marriage and cancer mortality is seen below. Adding its logarithmic and exponential transformations increased the linear regression model's accuracy by 0.003.





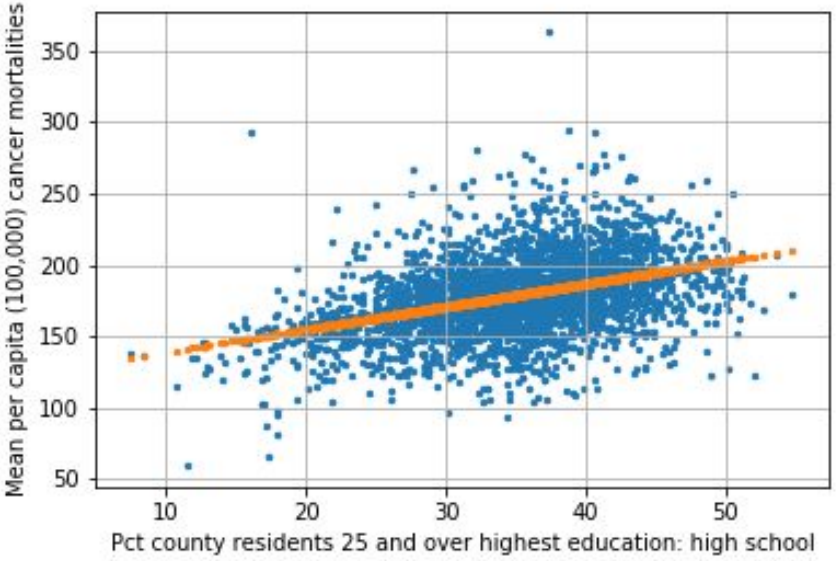
Percentage Adults With Only High School Diploma

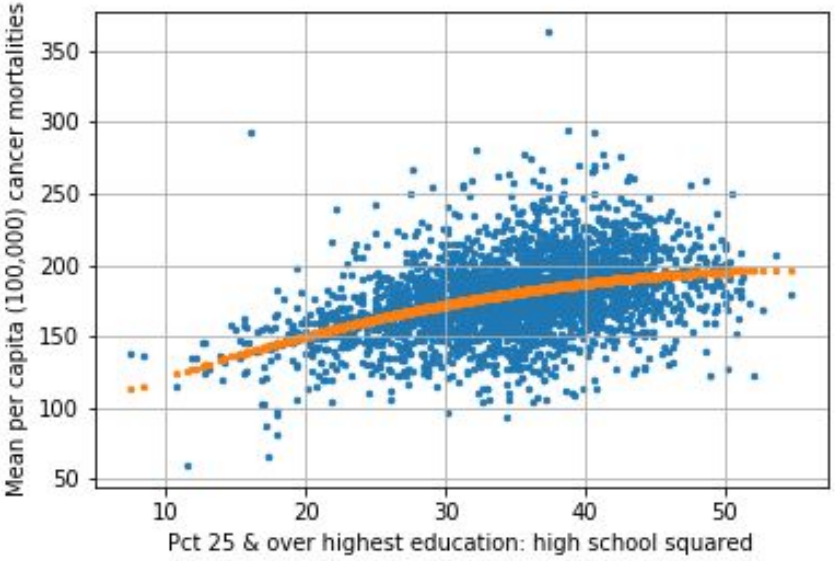
The percentage of adults whose highest education is a high school degree ranges from 7% to 54.8% and had a correlation of 0.41 with cancer mortality, showing a relationship between lower education and an increased rate of cancer mortality.



Percentage Adults With Only High School Diploma

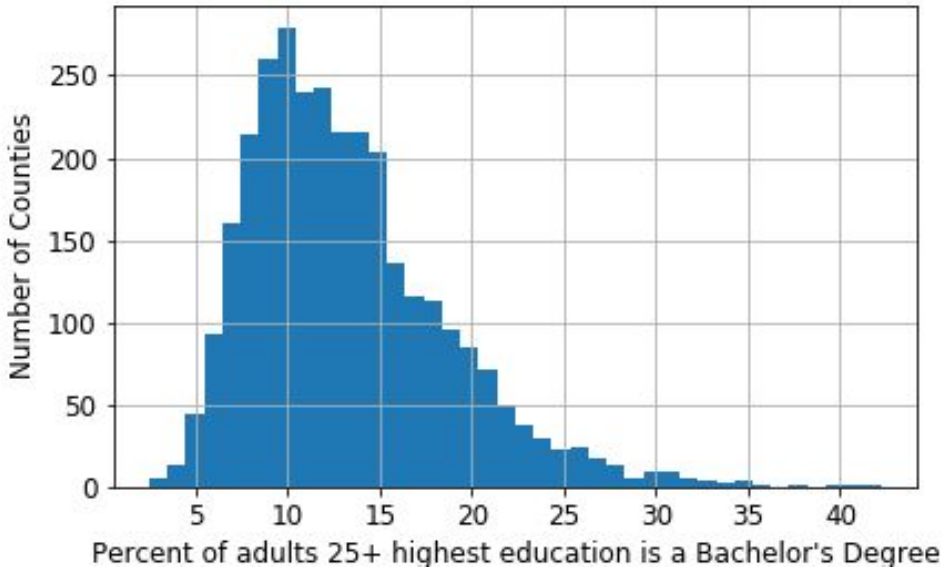
The moderately positive correlation between the percentage of adults with only a high school diploma and cancer mortality is seen below. Adding the exponential transformation increased the linear regression model's accuracy by 0.0002.





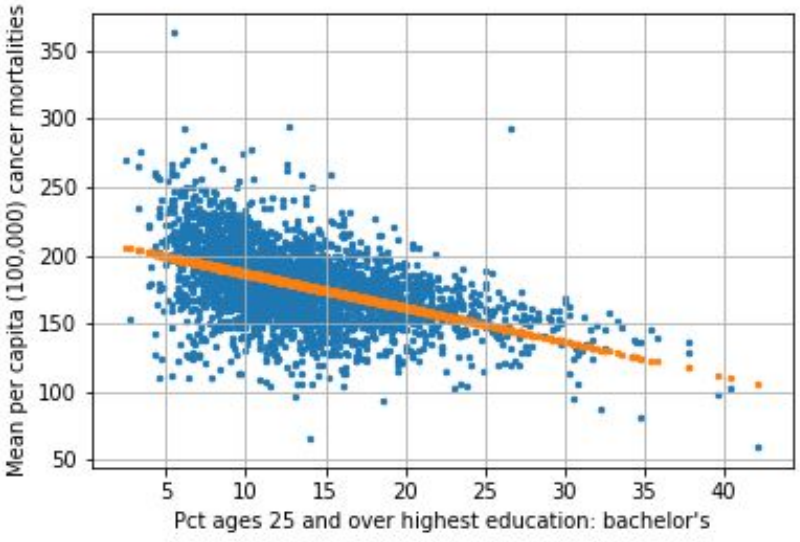
Percentage Adults With Bachelor’s Degrees

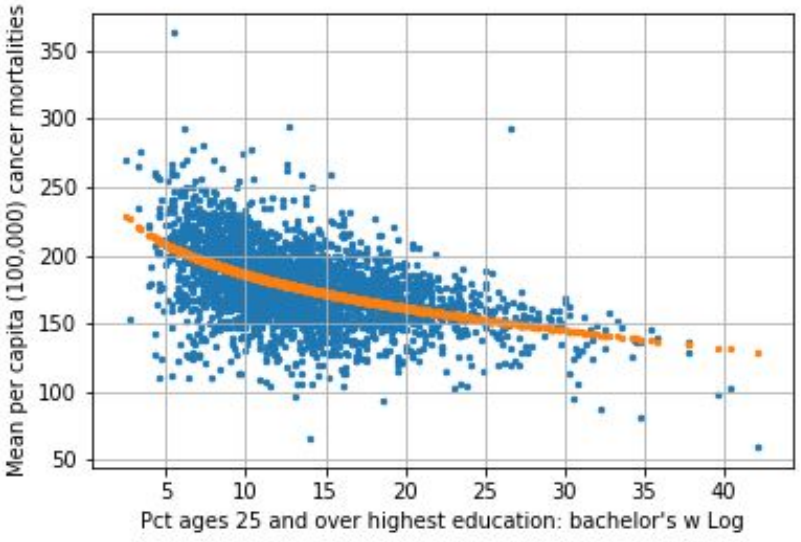
The percentage of adults whose highest education is a Bachelor’s degree ranges from 2.5% to 42.2% and had a correlation of -0.49 with cancer mortality, showing a relationship between higher education and a decreased rate of cancer mortality.



Percentage Adults With Bachelor’s Degrees

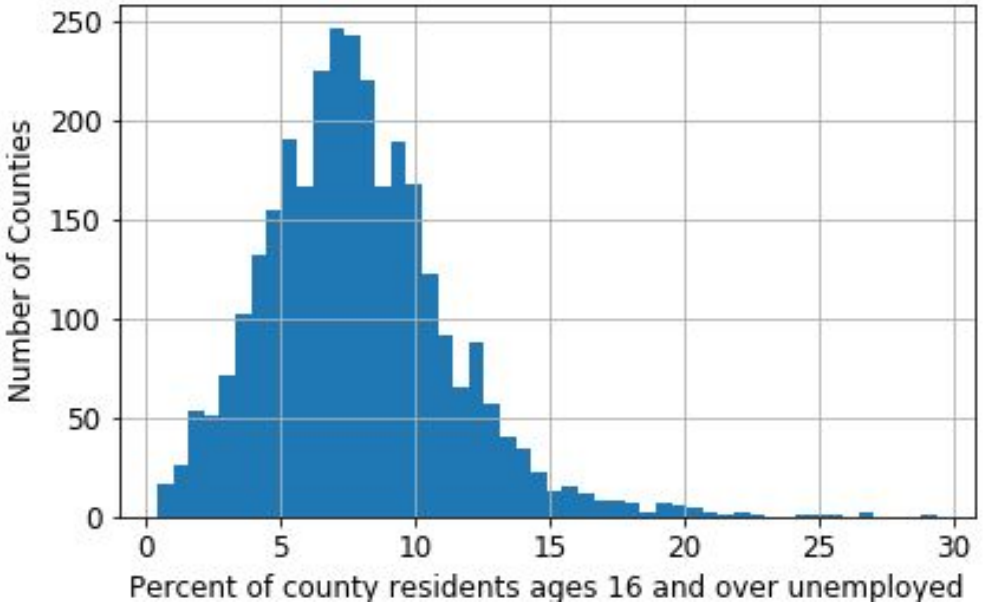
The moderately negative correlation between the percentage of adults with Bachelor’s degrees and cancer mortality is seen below. Adding the logarithmic transformation increased the linear regression model's accuracy by 0.0004.





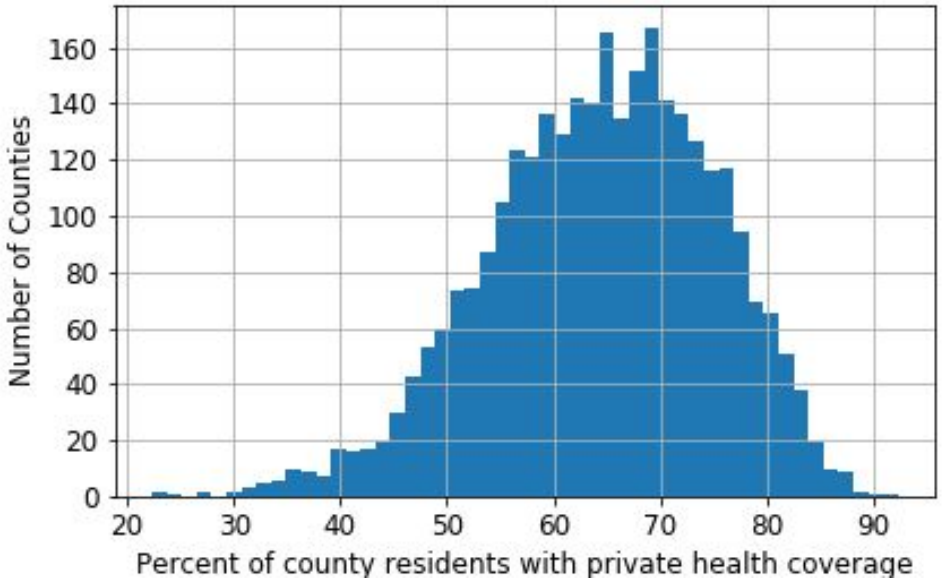
Percent Unemployed (ages 16+)

The percentage of county residents that are 16 years and over who are unemployed ranged from 0.4% to 29.4%, with a moderately strong correlation of 0.38 with cancer mortality.



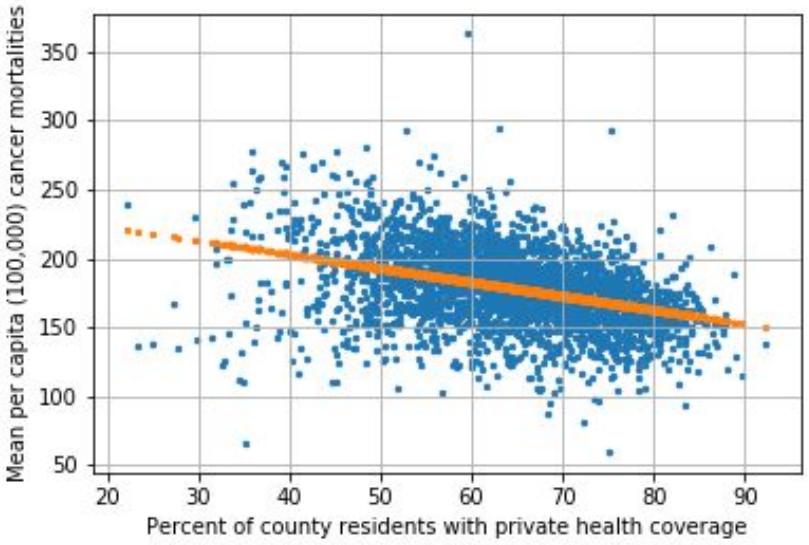
Percent with Private Health Coverage

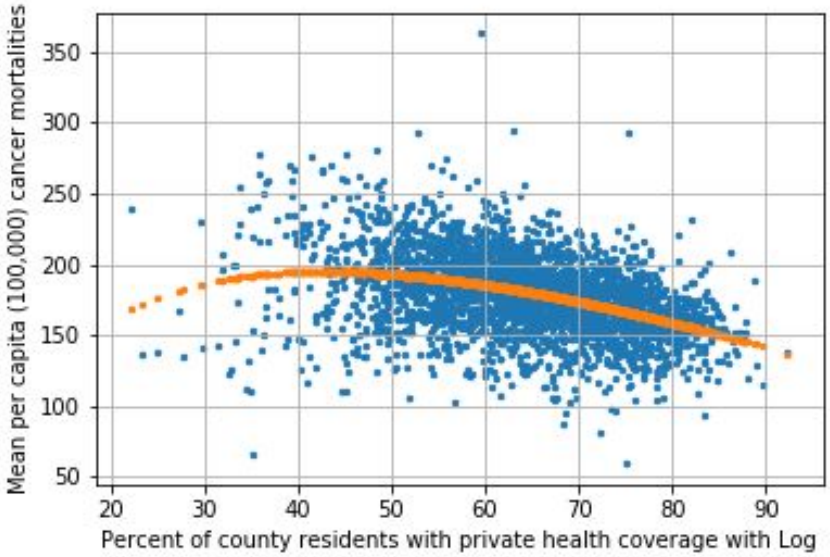
The percentage of county residents with private health coverage ranges from 22.3% to 92.3% with a negative correlation of -0.39, showing a relationship between private health coverage and a reduced risk of cancer mortality.



Percent with Private Health Coverage

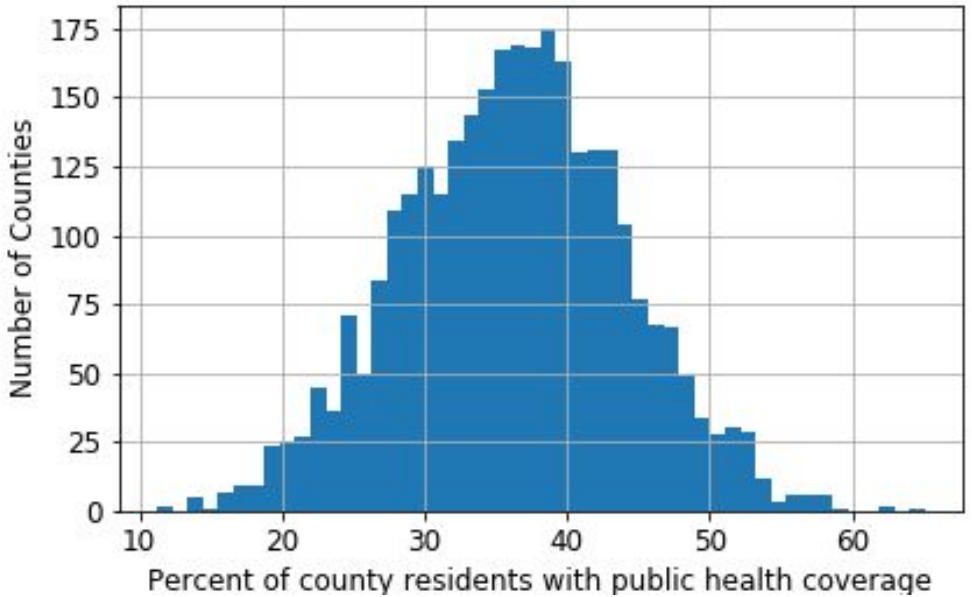
The moderately negative correlation between the percentage of adults with private health coverage and cancer mortality is seen below. Adding the logarithmic transformation increased the linear regression model's accuracy by 0.008.





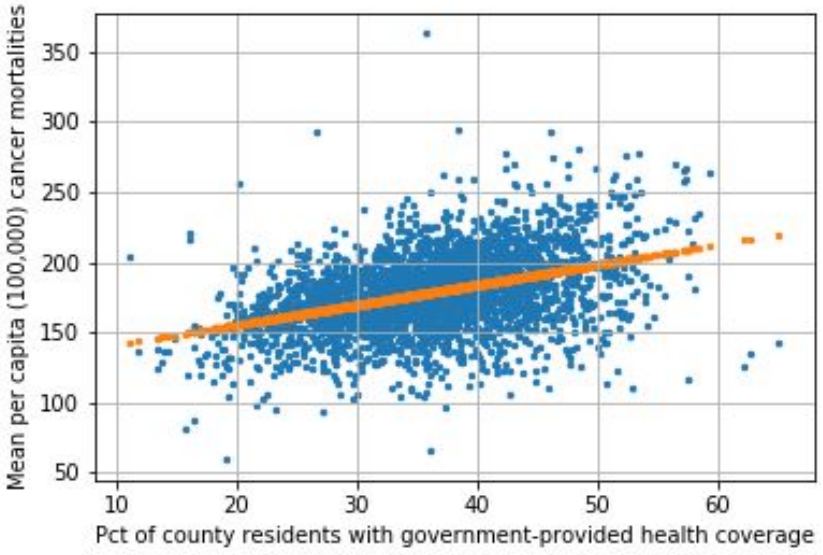
Percent with Public Health Coverage

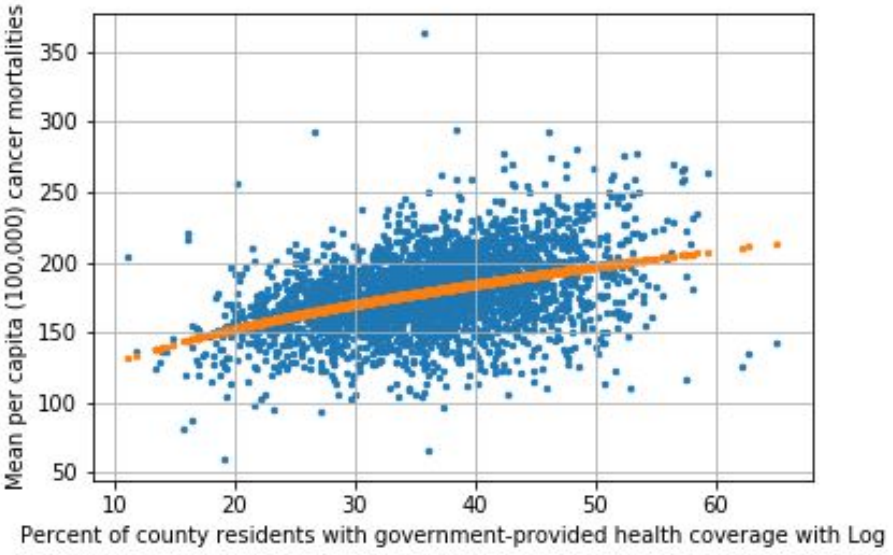
The percentage of county residents with public health coverage ranged from 11.2% to 65.1% with a moderate correlation of 0.41, showing a relationship between public health coverage and an increased risk of cancer mortality.



Percent with Public Health Coverage

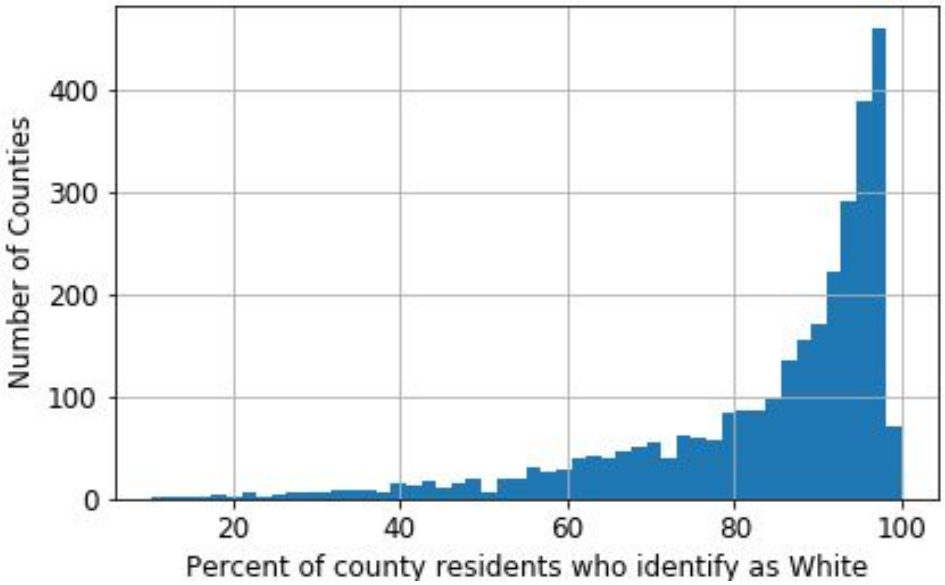
The moderately positive correlation between the percentage of county residents with public health coverage and cancer mortality is seen below. Adding the logarithmic form increased the linear regression model's accuracy by 0.008.





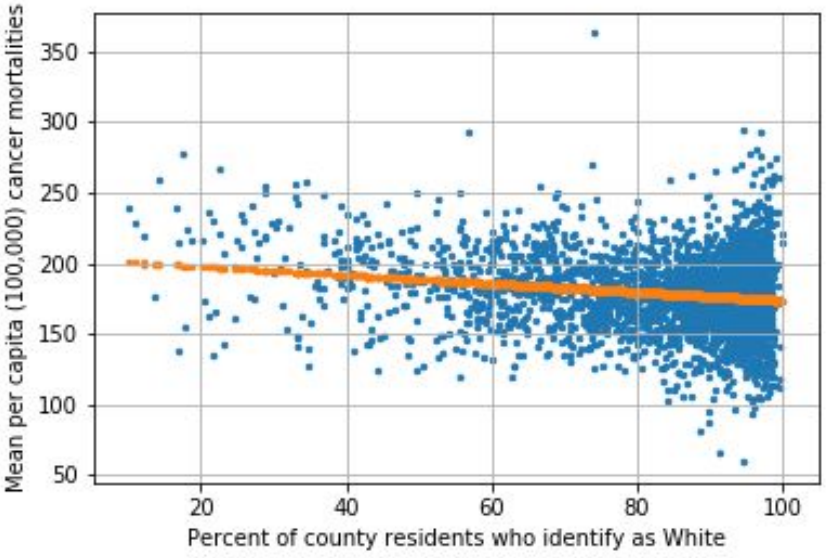
Percent of County Residents Who Identify as White

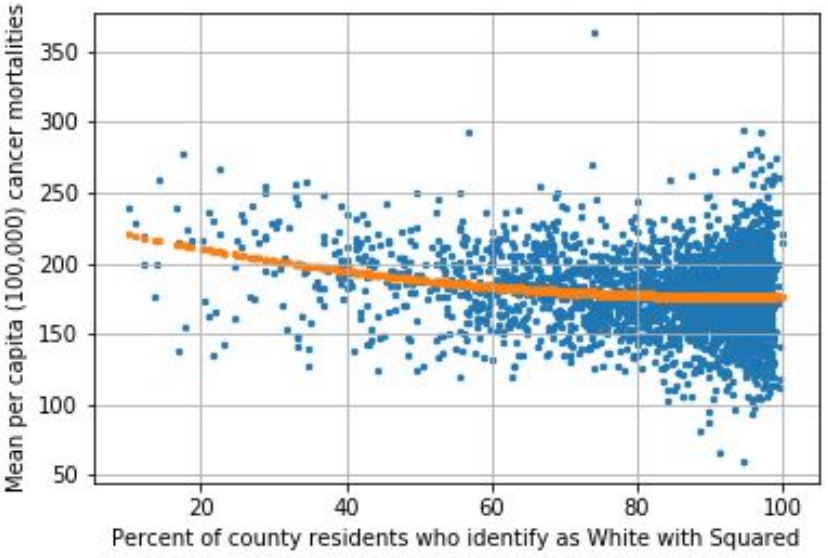
The percentage of county residents who identify as White ranges from 10.2% to 100% with a weak negative correlation of -0.18, showing a relationship between being White and a reduced risk of cancer mortality.



Percent of County Residents Who Identify as White

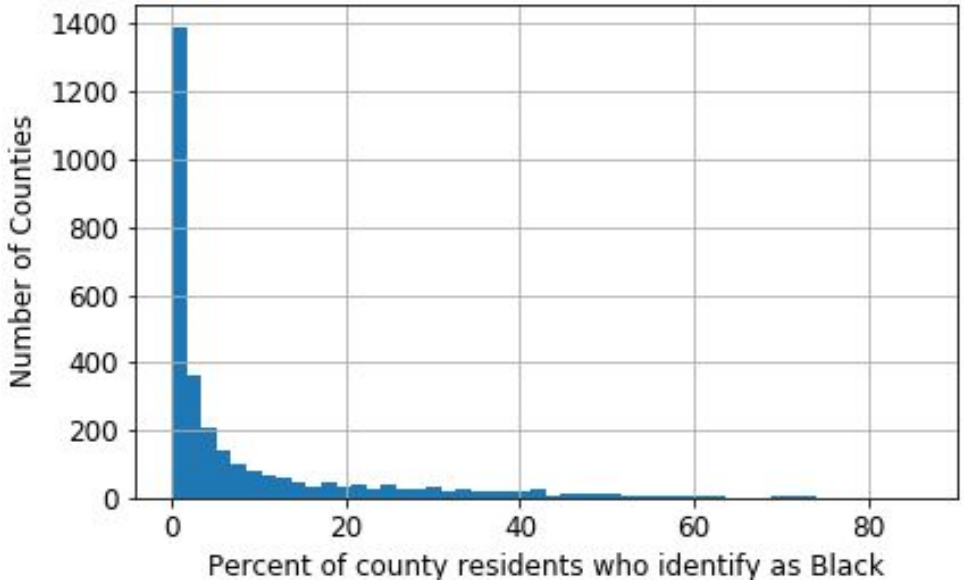
The weak negative correlation between the percentage of county residents who identify as White and cancer mortality is seen below. Adding the exponential transformation increased the linear regression model's accuracy by 0.0003.





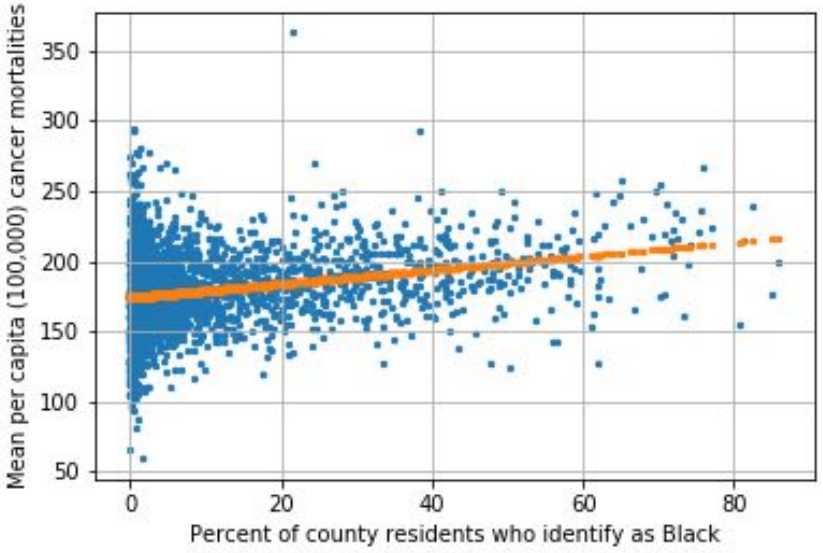
Percent of County Residents Who Identify as Black

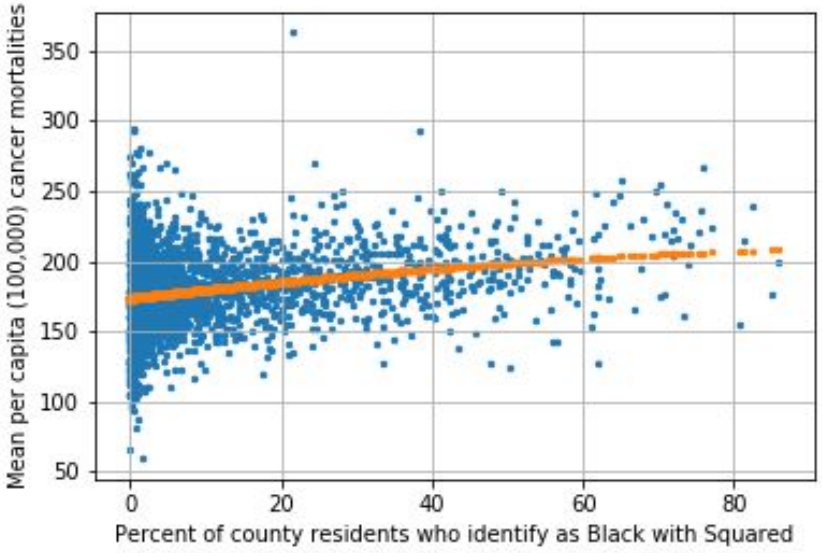
The percent of county residents who identify as Black ranges from zero to 86% with a positive correlation of 0.26, showing a relationship between being Black and an increased risk of cancer mortality.



Percent of County Residents Who Identify as Black

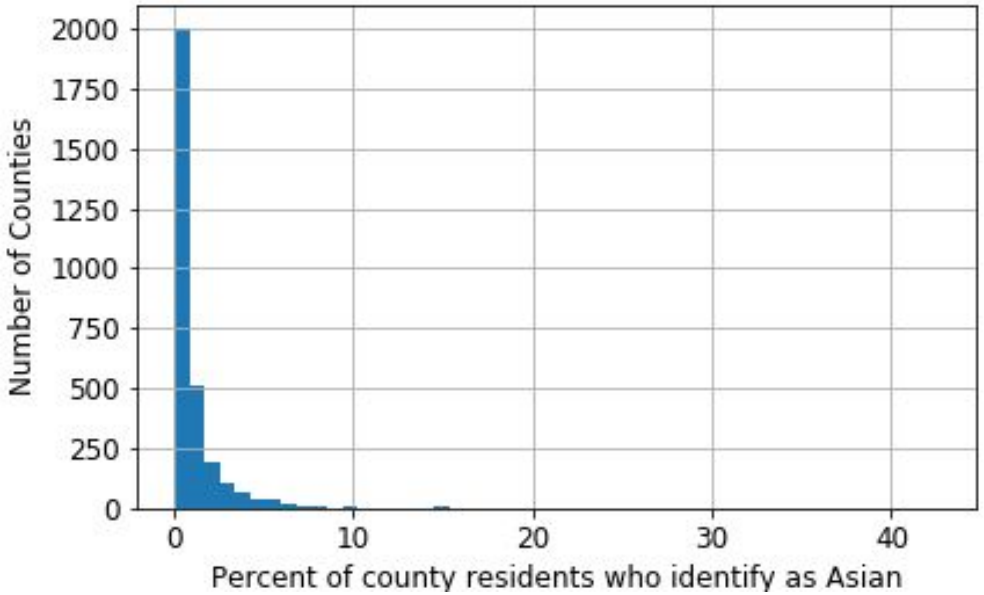
The positive correlation between the percentage of county residents who identify as Black and cancer mortality is seen below. Adding the exponential transformation increased the linear regression model's accuracy by 0.004.





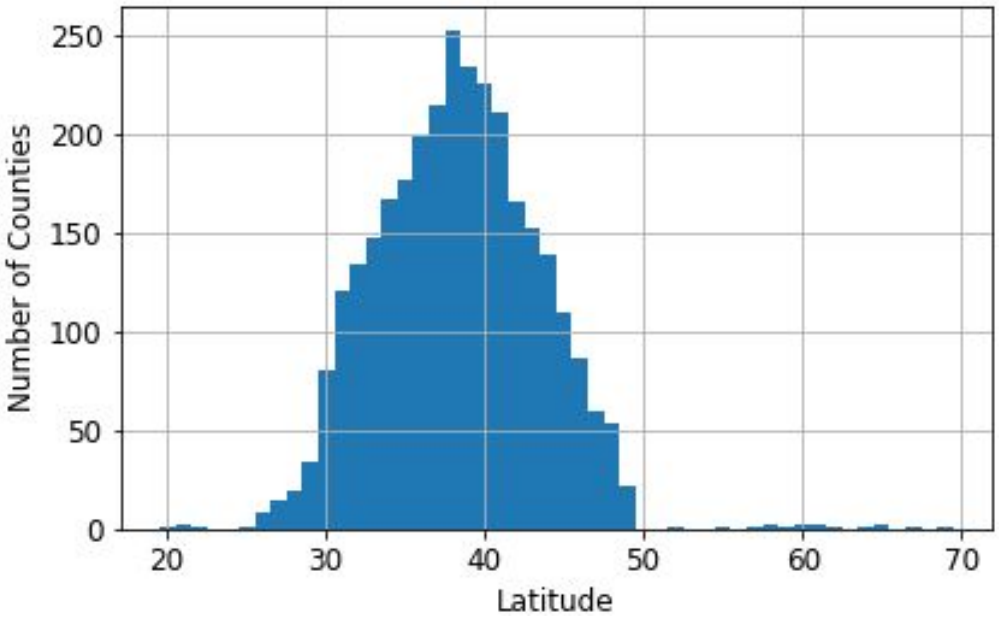
Percent of County Residents Who Identify as Asian

The percentage of county residents who identify as Asian ranges from zero to 43% with a weak negative correlation of -0.19, showing a relationship between being Asian and a reduced risk of cancer mortality.



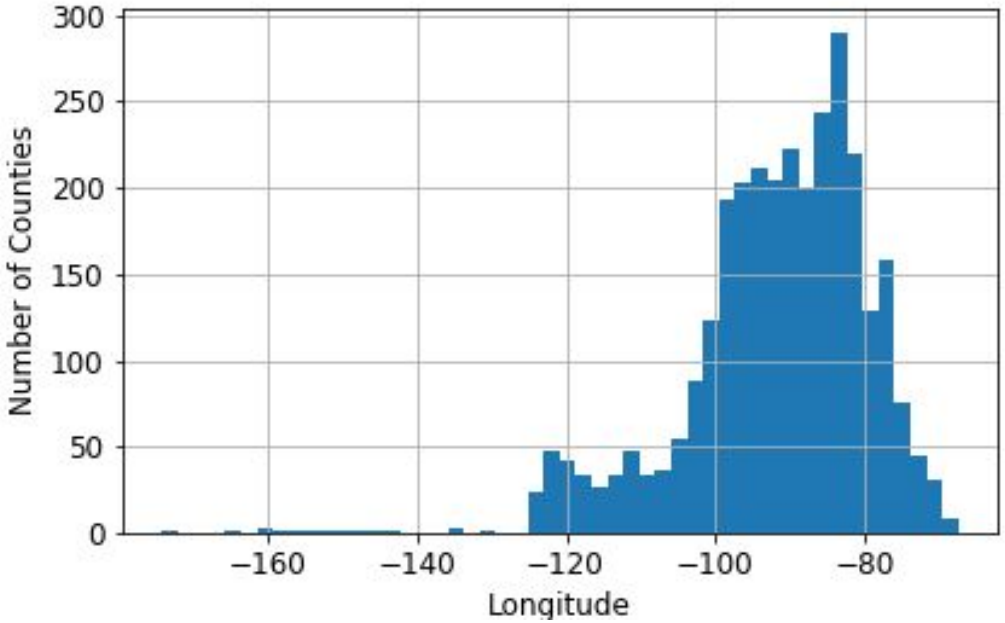
Latitude of Counties

The United States' latitude ranges from 19.6 to 70 with a correlation of -0.18 with cancer mortality, suggesting that generally the further north a county is the lower its cancer mortality rate is.



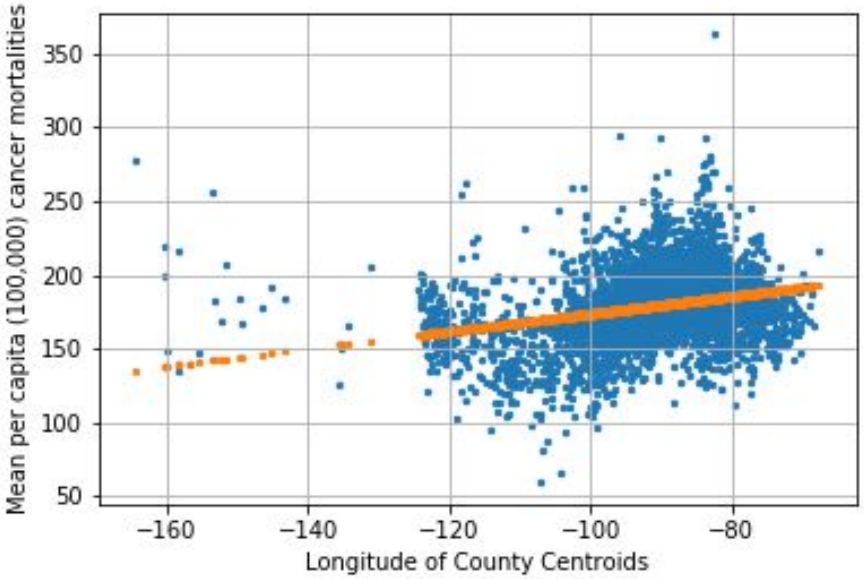
Longitude of Counties

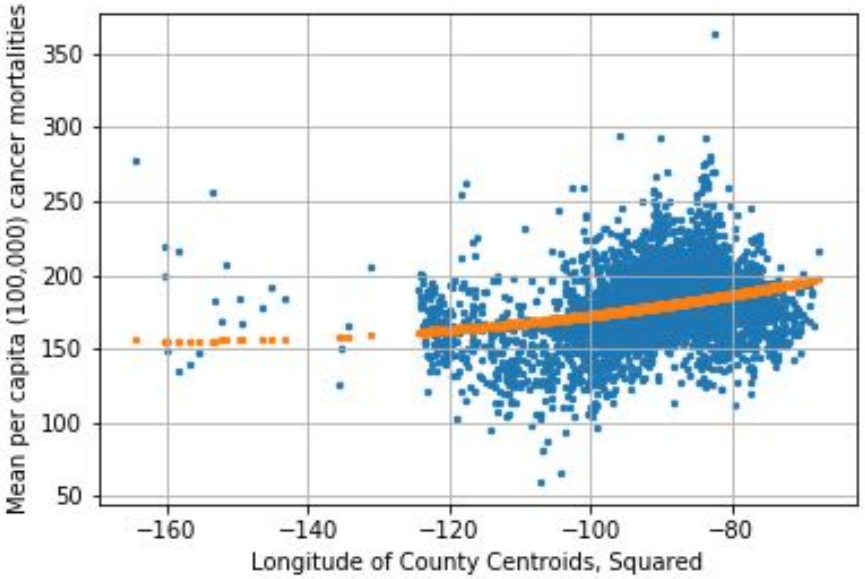
The United States' longitude ranges from -164.2 to -67.6 with a correlation of 0.26 with cancer mortality, suggesting that generally the further east a county is the higher its cancer mortality rate is.



Longitude of Counties

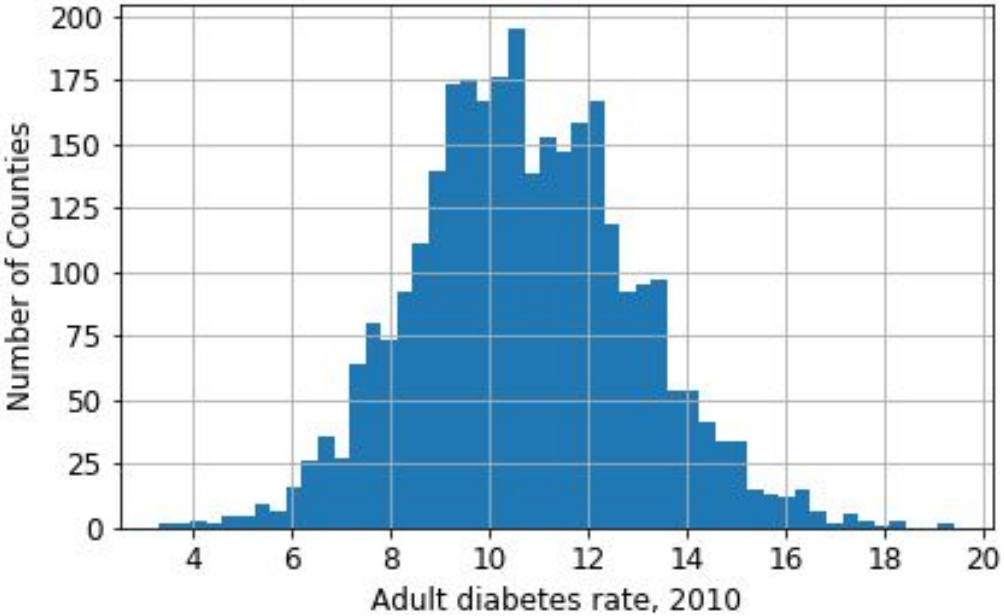
The positive correlation between longitude and cancer mortality is seen below. Adding the exponential transformation increased the linear regression model's accuracy by 0.0005.





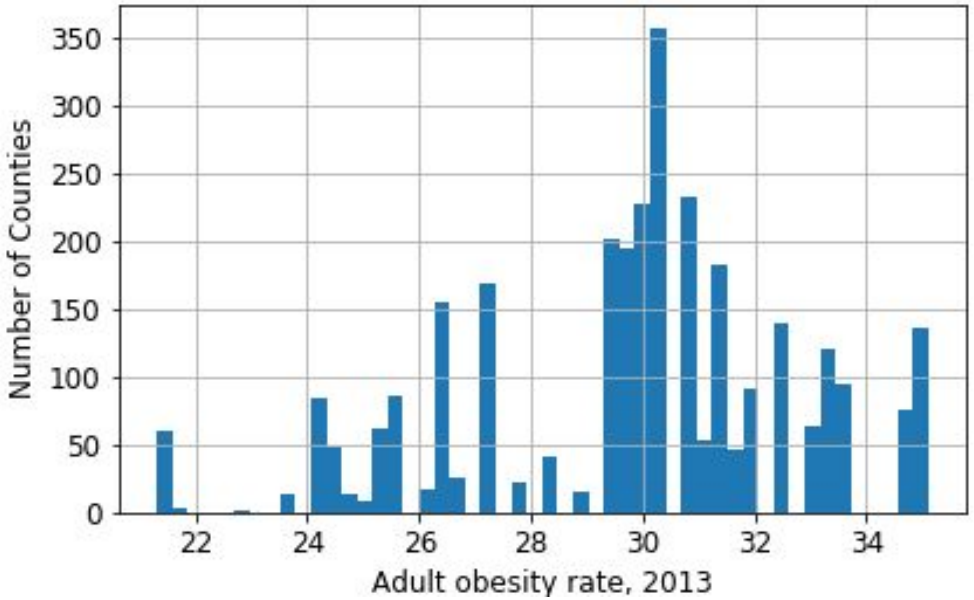
Diabetes

The percentage of adults in each county with diabetes (in 2010) ranges from 3.3% to 19.4% with a correlation of 0.53 with cancer mortality, showing a strong relationship between the two health conditions.



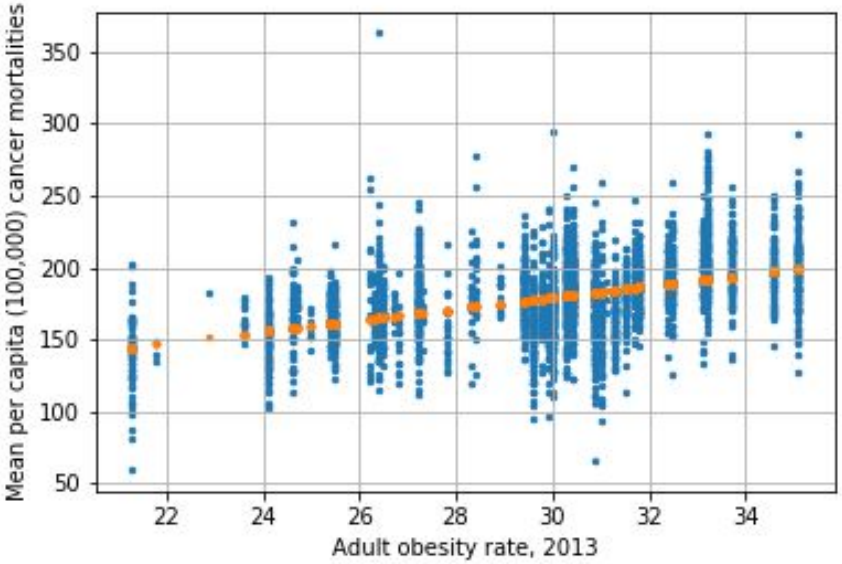
Obesity

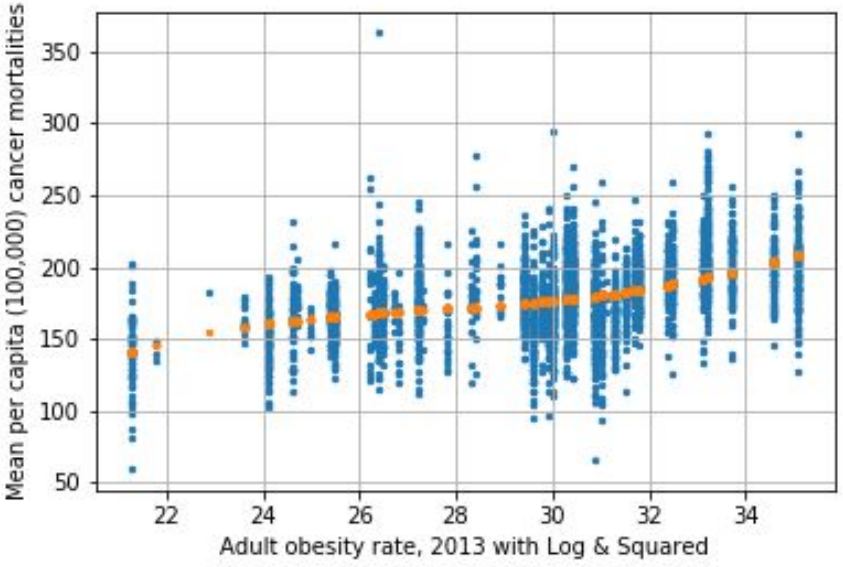
The percentage of adults in each county with obesity in 2013 ranges from 21.3% to 35.1% with a correlation of 0.43, showing a strong relationship between the two health conditions.



Obesity

The positive correlation between obesity and cancer mortality is seen below. Adding the logarithmic and exponential transformations increased the linear regression model's accuracy by 0.000005.





Geographic Distribution of Cancer Mortality

The geographic distribution of cancer mortality rates across the continental U.S. is visualized in the latitude/longitude scatter plot maps on the next two slides. The distribution was split into quintiles and each point on the scatter plot maps represents one of the 3,047 counties in the DataFrame.

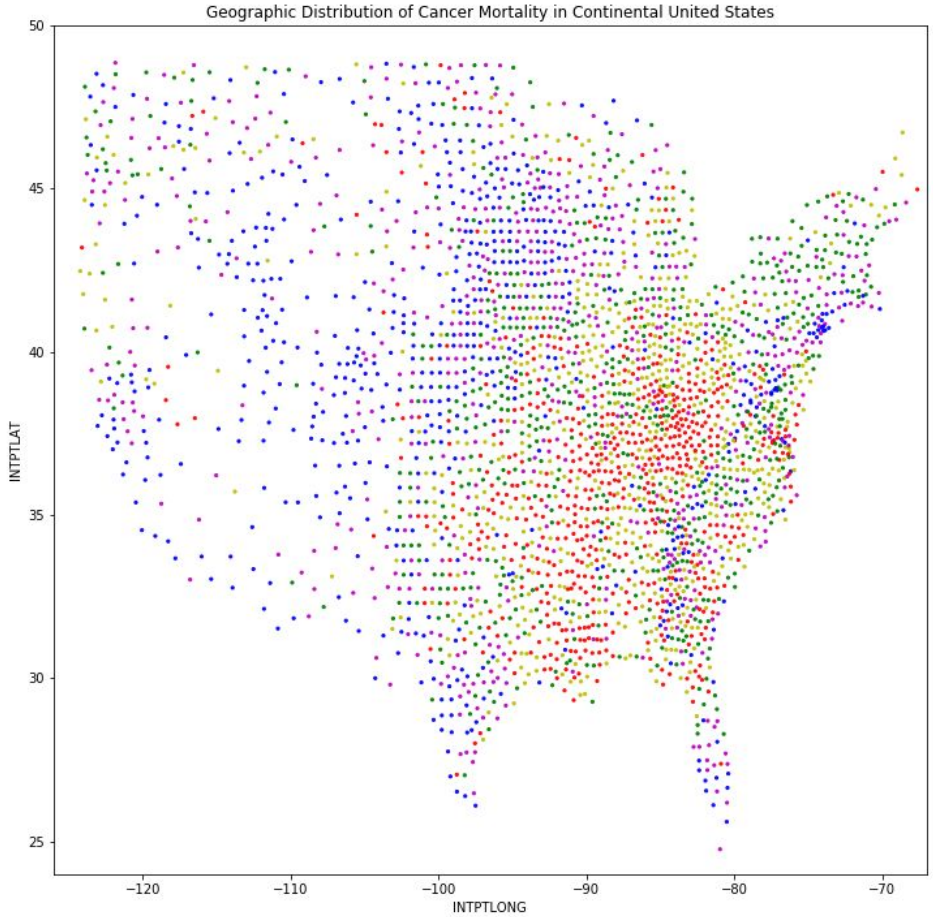
The first scatter plot map just shows the raw latitude/longitude coordinates and the second scatter plot shows these coordinates overlaid on a map of the continental United States. Although the second scatter plot map was overlaid using the official latitude/longitude extent of the continental U.S., the county data points and the map are offset due to the curvature of the earth.

Although high cancer mortality rate counties are sprinkled throughout the continental United States, one can easily see a particularly high concentration in the American South and eastern Midwest regions.

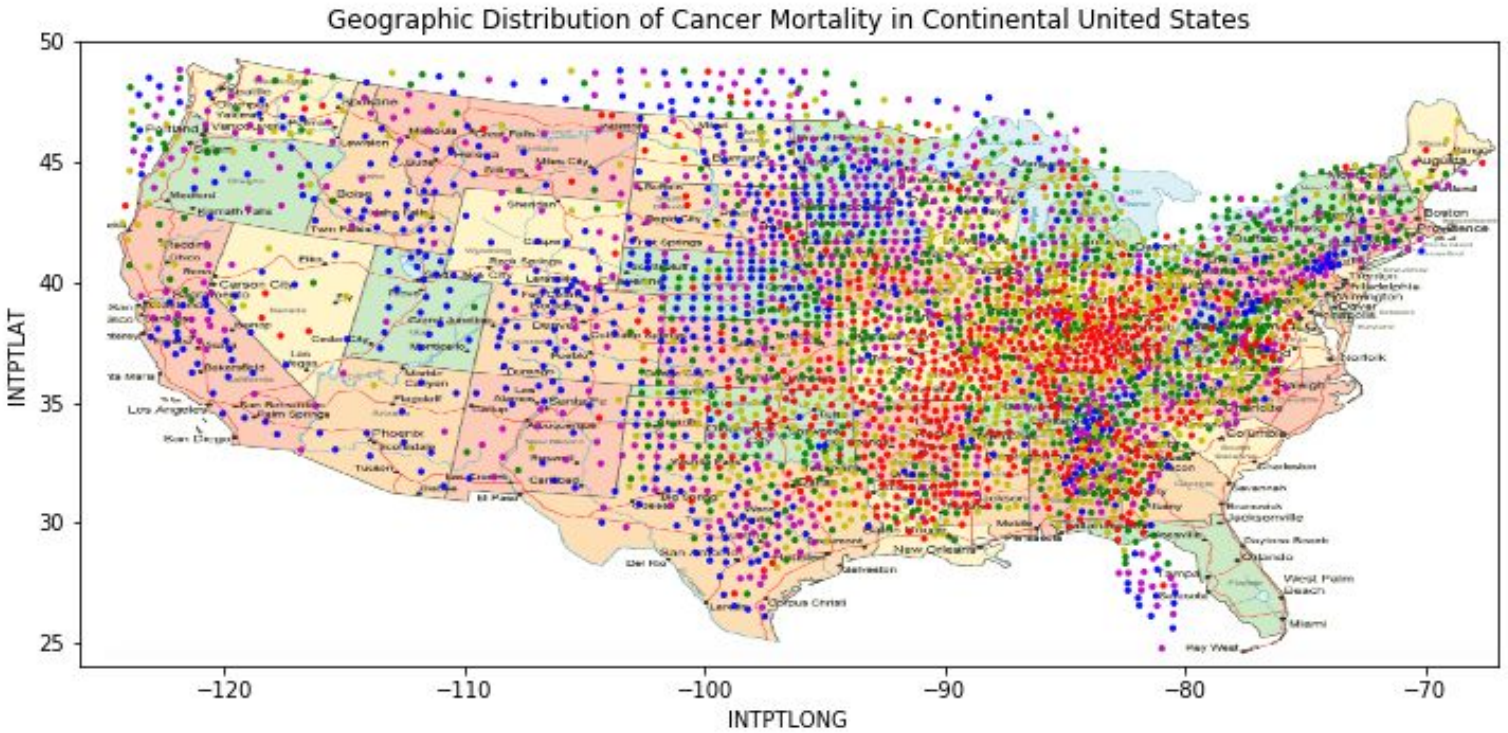
Scatter Plot Map of Cancer Mortality

LEGEND

- Lowest Mortality
- Low Mortality
- Medium Mortality
- High Mortality
- Highest Mortality



Scatter Plot Map of Cancer Mortality Overlaid on U.S.



Hypothesis Testing Using the T-Test

Five hypothesis tests were run on a series of null hypotheses about whether the cancer mortality rates seen in different types of counties is due to chance. The null hypotheses were evaluated using the t-test.

The t-test was used for the reason that although the majority of U.S. counties and U.S. counties were included in the DataFrame (97%), it is still a sample of the overall U.S. and t-tests generally work better when population parameters are not fully known.

Hypothesis Testing Using the T-Test: 1st Hypothesis

The first null hypothesis tested was that the differing cancer mortality rates seen in majority White counties and majority Black counties was due to chance. The median cancer mortality across the U.S. in 2015 was 179 per 100,000 people. For majority White counties,, the cancer mortality rate was also 179 per 100,000 people with a standard deviation of 27.2. For majority Black counties, the cancer mortality rate was 202 per 100,000 people with a standard deviation of 28.2.

The t-score for this first hypothesis test was approximately 8.69 and the p-value was 6e-18. Therefore, the null hypothesis was rejected. Although causality and the identification of confounding variables is outside the scope of this analysis, one cannot say that the difference in cancer mortality rates seen in majority Black and majority White counties is due to random chance.

Hypothesis Testing Using the T-Test: 2nd Hypothesis

The second null hypothesis tested was that the differing cancer mortality rates seen in counties where the majority of the populace had private health insurance and counties where the majority of the populace had public health insurance is due to chance. The cancer mortality rate for majority private health insurance counties was 177 per 100,000 people with a standard deviation of 25.8, while the cancer mortality rate for majority public health insurance counties was 199 per 100,000 people with a standard deviation of 38.1.

The t-score for this second hypothesis test was approximately 9.1 and the p-value was 1.6e-19. Therefore, the null hypothesis was rejected and it cannot be said that the differences between these two groups of counties was due to random chance.

Hypothesis Testing Using the T-Test: 3rd Hypothesis

The third null hypothesis tested was that the differing cancer mortality rates seen in counties with a high rate of unemployment and counties with a low rate of unemployment was due to random chance. A "high rate" was defined as being above the median of nationwide unemployment. The counties with a high unemployment rate had a cancer mortality rate of 188 out of 100,000 people with a standard deviation of 27.8, while the counties with a low unemployment rate had a cancer mortality rate of 169 out of 100,000 people with a standard deviation of 24.6.

The t-score for this third hypothesis test was approximately 19.6 and the p-value was 2.7e-80. The third null hypothesis was therefore rejected and the difference in cancer mortality rates seen in low unemployment and high unemployment counties cannot be said to be due to random chance.

Hypothesis Testing Using the T-Test: 4th Hypothesis

The fourth null hypothesis tested was that the differing cancer mortality rates seen in counties where the median income of the populace was below the national median and counties where the median income of the populace was above the national median was due to chance. The counties with a median income below the national median had a cancer mortality rate of 189 per 100,000 people with a standard deviation of 28.6, while the counties with a median income above the national median had a cancer mortality rate of 168 per 100,000 people with a standard deviation of 22.6.

The t-score for this fourth hypothesis test was approximately 22 and the p-value was 5.2e-100. The fourth hypothesis test was therefore rejected and the difference in cancer mortality rates in low income counties and high income counties cannot be said to be due to random chance.

Hypothesis Testing Using the T-Test: 5th Hypothesis

The fifth null hypothesis tested was that the differing cancer mortality rates seen in counties where a high percentage of the adult populace's highest level of education was a high school diploma (“high school counties”) and counties where a high percentage of the adult populace's highest level of education was a college degree (“college degree counties”) is due to chance. A "high percentage" was defined as being above the median. The “high school counties” had a cancer mortality rate of 188 out of 100,000 people with a standard deviation of 27, while the “college degree counties” had a cancer mortality rate of 168 out of 100,000 people with a standard deviation of 23.6.

The t-score for this fifth hypothesis test was approximately 22.2 and the p-value was 3.6e-100. Therefore, the fifth hypothesis test was rejected and the difference in cancer mortality rates between “high school counties” and “college degree counties” cannot be said to be due to chance.

Midway Summary

The features with the most salient correlations to cancer mortality involve financial income of county residents, the poverty level of each county (including persistent poverty, child poverty, and persistent child poverty), the level of education among the county residents, the levels of employment and unemployment in each county, the levels of private vs. public insurance, race, latitude/longitude, diabetes, and obesity. These features along with the entire feature set will be further explored through the analysis of their linear regression coefficients in the machine learning section of this project.

In-Depth Analysis Using Machine Learning

This project created machine learning regression models using Ordinary Least Squares (OLS) Regression, Ridge Regression, LASSO, ElasticNet, Stochastic Gradient Descent (SGD) Regressor, Kernel Ridge Regression, and Random Forest algorithms to try and predict population cancer mortality rates in 97% of U.S. counties in the year 2015.

These models were created not only to predict cancer mortality, but to also identify the most salient predictors of cancer mortality by looking at the coefficients of the best performing regression algorithm.

Evaluating Regression Algorithms (Part I)

The best performing regression algorithm for the model was identified by evaluating the accuracy score and root mean squared error (RMSE) of a set of regression algorithms. These regression algorithms were run on unscaled and scaled data, and utilized different values for:

- the regularization hyperparameter ‘alpha’ (for all algorithms except for simple OLS linear regression),
- the L1 ratio (for ElasticNet and SGD Regressor),
- the penalty (L1, L2, or ElasticNet for SGD Regressor),
- Epsilon (SGD Regressor)
- and the number of estimators (for Random Forest).

Evaluating Regression Algorithms (Part I)

The LASSO and ElasticNet algorithms used their internal normalization setting to scale the data, as they would not converge otherwise. The MinMax scaler was used for scaling data on the other algorithms. These regression algorithms' accuracy and RMSE scores were stored in a hyperparameter tuning table. The best performing scoring regression algorithm was then identified for the model.

The best performing regression algorithm was Ridge Regression using the 'auto' solver and an Alpha of 0.001, with a training accuracy score of 0.647, a training RMSE of 16.59, a test accuracy score of 0.6408, and a test RMSE of 16.2.

Individual Features with Strongest Positive Correlations with Cancer Mortality (Part I)

The individual features with the strongest positive correlations with cancer mortality fall into the following categories:

- L1 and L2 distances of counties from top 10 oncology hospitals, real number and logarithmic versions
- State that the county is in
- Recreation facility-related feature

Individual Features with Strongest Positive Correlations with Cancer Mortality (Part II)

Logarithmic transformations of features:

- Health insurance features (private)
- Education-related: percentage some college 18-24
- Poverty-related features
- Average household size

Missing value features:

- Farmer’s market related features
- ERS Natural Amenity Index

Individual Features with Strongest Negative Correlations with Cancer Mortality (Part I)

The features with the strongest negative correlations with cancer mortality fall into the following categories:

- L1 and L2 distances of counties from top 10 oncology hospitals, real number and logarithmic versions
- State that the county is in
- Recreation facility-related feature

Individual Features with Strongest Negative Correlations with Cancer Mortality (Part II)

Logarithmic transformations of features:

- Health insurance features (public and private)
- Median Age
- Percentage of populace who are married
- Percentage of children in poverty
- Percentage of populace 16 years and older who are employed

Relationship of "Feature Families" to the Target Variable of Per-Capita Cancer Mortality, Part I

Because there are over 300 features in the feature set, the interpretability of these features' Ridge Regression coefficients with the target variable per capita cancer mortality is supported by grouping the features into "feature families" and calculating the proportion of the positive and negative influence that each of these “feature families” had on the target variable of per-capita (100,000) cancer mortality in the United States on the county level in 2015.

This calculation was done separately for features with positive coefficients and negative coefficients.

Relationship of Positive "Feature Families" to the Target Variable of Per-Capita Cancer Mortality, Part I

The proportions that each "feature family" had of the total **positive** influence of increasing cancer mortality in 97% of the counties in the United States during 2015 are as follows (in descending order):

- L1 and L2 distances from county centroids to top 10 oncology hospitals: 0.471
- Types of health insurance for each county's populace: 0.188
- United States state each county is in: 0.158
- Missing values: 0.058
- L1 and L2 distances from county centroids to major cities: 0.031
- Poverty: 0.025
- Average household size of each county: 0.0188
- Recreation and fitness facilities in each county: 0.0114

Relationship of Positive "Feature Families" to the Target Variable of Per-Capita Cancer Mortality, Part II

- Education levels of each county's populace: 0.0097
- Food environment of each county: 0.0072
- Financial income of each county's populace: 0.0052
- Employment status of each county's populace: 0.0043
- Health conditions comorbid with cancer: 0.003
- Counties' significant population loss as of the year 2000: 0.0026
- Erroneous data indicator referencing average household size: 0.0025
- Age of each county's populace: 0.0015
- Marital status of county's populace: 0.0014
- L2 distance to closest EPA Superfund Cleanup site: 0.0011
- Indicator of whether a county is in a metropolitan area or not: 0.0005
- Race of each county's populace: 0.0002
- Rate of cancer diagnoses in each county: 0.0001
- Per capita number of cancer-related clinical trials per county: 0.0000001
- Square mileage of land mass for each county: 0.000000003

Relationship of Negative "Feature Families" to the Target Variable of Per-Capita Cancer Mortality, Part I

The proportions that each "feature family" had of the total **negative** influence of decreasing cancer mortality in 97% of the counties in the United States during 2015 are as follows (in descending order):

- L1 and L2 distances from county centroids to major cities: 0.377
- United States state each county is in: 0.1702
- L1 and L2 distances from county centroids to top 10 oncology hospitals: 0.1341
- Employment status of each county's populace: 0.1245
- Types of health insurance for each county's populace: 0.0458
- Missing values: 0.0269
- Poverty: 0.0226
- Latitude and Longitude: 0.0213
- Marital status of county's populace: 0.0203

Relationship of Negative "Feature Families" to the Target Variable of Per-Capita Cancer Mortality, Part II

- Age of each county's populace: 0.017
- Recreation and fitness facilities in each county: 0.0133
- Food environment of each county: 0.0099
- Financial income of each county's populace: 0.0056
- Average household size of each county: 0.0045
- L1 distance to closest EPA Superfund Cleanup site: 0.0022
- Health conditions comorbid with cancer: 0.0014
- Race of each county's populace: 0.0014
- Education levels of each county's populace: 0.0006
- Quality of life index: 0.0006
- Birth Rate: 0.0004
- Erroneous data indicator referencing median age: 0.0004
- Rate of cancer diagnoses in each county: 0.000002
- Square mileage of water for each county: 0.000002
- Population of county: 0.0000000002