Project: Capstone Project 2: Milestone Report 2
Predicting Cancer Mortality in United States Counties: 2015
Daniel Loew

Introduction and Problem Statement

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. Of course, the field of oncology research has a long and rich history, but using machine learning models to predict cancer outcomes is a fairly new field. The earliest paper found in a literature review on the subject used machine learning models in conjunction with electronic administrative records and a cancer registry to predict cancer survival (Gupta, Tran, Luo, Phung, Kennedy, Broad, Campbell, Kipp, Singh, Khasraw, Matheson, Ashley, Venkatesh, 2014). A paper just published in Scientific Reports in July 2020 used machine learning models to investigate the major effects of health-related quality of life on five-year survival prediction among lung cancer survivors (Sim, Kim, Kim, Lee, Kim, Shim, Zo, Yun, 2020).

This project aims to add to this body of literature by using a set of socioeconomic indicators, geographic locations, proximity to top oncology hospitals, major urban centers and cancerogenic locations, comorbidities and food environment features to attempt to predict cancer mortality at the U.S. county level for the year 2015. The goal for this project is to provide guidance to policy makers, medical professionals and the public at large on which risk factors have the greatest influence on cancer mortality by constructing a machine learning model and exploring the coefficients of its feature set that could assist with the creation of public health policy, medical practices, and behavior changes that can help mitigate these risk factors. These public bodies will be the “clients” for this project.

A slide deck summarizing the results can be found at:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Predicting%20Cancer%20Mortality%20in%20U.S.%20Counties_%202015.pdf

Data Introduction, Cleaning and Wrangling

The original data set of socioeconomic indicators and cancer mortality rates for 2015 of 3,047 out of 3,141 total U.S. counties and county equivalents (97%) came pre-assembled as a data science challenge from three sources: the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov. It was downloaded from https://data.world/exercises/linear-regression-exercise-1. The assemblage process can be examined at https://data.world/nrippner/cancer-trials. The data dictionaries are located at https://data.world/exercises/linear-regression-exercise-1/workspace/data-dictionary.

The data dictionary of the original dataset from data.world can be summarized by the following statistics:

- The target variable is the mean per capita (100,000) cancer mortalities by county, labeled as 'TARGET_deathRate'. It is a numerical feature.
- There are two categorical features. One is the county or county equivalent (labeled as 'geography'), which will serve as the row index. The other is median income per capita binned on the decile, which will be binarized into a series of binary features for machine learning purposes.
- There are 30 numerical features.
- The only special variable that involves location is the county feature named 'geography', but this won't be a special predictive feature as it will serve as the row index, as each row in the DataFrame is a county (or county equivalent). There are no special date, time, or image features. All of the data is for 2015, and there is no temporal granularity in the feature set beyond that year.

After removing a feature that contained the same information as the target variable in order to eliminate the chances for data leakage (detailed below), and after the binned median income string feature binned on deciles was binarized into separate binary features for each income category, an initial run of a basic Ordinary Least Squares (OLS) linear regression model returned a test accuracy with no hyperparameter tuning of 49.2%.

The full data dictionary used in this project is located in Appendix A.

The Jupyter notebook with the data cleaning and wrangling steps detailed below can be found at the following link:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_Data_Cleaning.ipynb

The following data cleaning and wrangling steps were performed on the "NYPD Complaint Data Historic" dataset:

1. The dataset was loaded in full into a Pandas DataFrame. 'Latin-1' encoding was necessary.
2. The presence of duplicate rows was checked for; there were none.
3. The presence of missing data in any of the columns was checked for; there were three columns with missing data:
    a. 'PctSomeCol18_24': Percent of county residents ages 18-24 highest education attained: some college
    b. 'PctEmployed16_Over': Percent of county residents ages 16 and over employed
    c. 'PctPrivateCoverageAlone': Percent of county residents with private health coverage alone (no public assistance)
4. New Boolean mask columns were added for each of these three columns designating which of their rows had missing values; a value of TRUE designated that the row had missing data and a value of FALSE designated that the row had data. These three Boolean mask columns were named 'PctSomeCol18_24_isnull', 'PctEmployed16_Over_isnull', and 'PctPrivateCoverageAlone_isnull'.
5. The median value of each of the three columns was calculated and their missing values were filled with their respective median value.

6. There were 30 counties whose ‘MedianAge’ value storing the median age of those counties was well over 100, in the 200-600 year old range. These values were obviously erroneous, and were replaced by the median value of the 'MedianAge' column.
7. It was also noticed that there were 61 counties where the average household size for occupied housing units (variable name 'AvgHouseholdSize') is less than one, ranging from 0.0221 to 0.0322. This is again erroneous, as the Census data that this is taken from explicitly states that the 'AvgHouseholdSize' variable stores the average household size for *occupied* housing units. Therefore, these counties had their ‘AvgHouseholdSize’ value filled by that variable’s median value.
8. In order to avoid data leakage in the machine learning models, the column called 'avgdeathsperyear' was dropped because it contains the mean number of reported fatalities due to cancer, while the target variable 'TARGET_deathRate' contains the mean per capita (100,000) cancer mortalities. The 'avgdeathsperyear' variable stores the same information as the target variable in a non-normalized form, and including a feature that stores the same information as the target variable in a machine learning model can result in a false accuracy score. Therefore, the ‘avgdeathsperyear’ feature was dropped.
9. The one native feature from the original DataFrame that was binarized for machine learning purposes was 'binnedInc', the median income per capita binned by decile. This variable was a string variable with the names of the median income deciles, so it was turned into a series of binary features.
10. The latitude/longitude centroids for the 3,047 counties were downloaded from https://www2.census.gov/geo/docs/maps-data/data/gazetteer/2019_Gazetteer/2019_Gaz_counties_national.zip and concatenated to the original DataFrame. 38 counties in the original DataFrame did not have county latitude/longitude centroid data, so they were looked up via Google search. These county latitude/longitude centroids were added to the core feature set. The square mileage of both landmass and water bodies was included in the Census data, so they were added to the feature set as well, out of interest. The ‘State’ feature was also added to the feature set.
11. The top 10 oncology hospitals were then found via https://health.usnews.com/best-hospitals/rankings/cancer, and their latitude/longitude locations were looked up via Google Search. L1 and L2 latitude/longitude distances to these hospitals were then calculated from each county centroid and added to the feature set, as were the L1 and L2 distances to the closest top 10 oncology hospital for each county.
12. Because large urban centers often have more healthcare resources, the latitude/longitude locations for eight major regional urban centers were looked up via Google search. These cities were Chicago, New York City, Atlanta, Dallas, Denver, Los Angeles, Seattle, and San Francisco. L1 and L2 latitude/longitude distances to these cities were then calculated from each county centroid and added to the core feature set, as were the L1 and L2 distances to the closest of these regional urban centers for each county.
13. As Environmental Protection Agency (EPA) designated Superfund Cleanup sites contain toxic and potentially carcinogenic materials, the latitude/longitude locations of these sites could serve as predictive features and were downloaded from the EPA's website at https://semspub.epa.gov/work/HQ/201261.pdf. The L1 and L2 latitude/longitude distances to the closest of these sites for each county were then calculated and added to the core feature set.

14. 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 farmer's markets in a county, were downloaded from the U.S. Department of Agriculture's Food Environment Atlas at https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/#August%202015%20Version. A set of 75 features of interest from this Atlas were added to the core feature set, including rates for serious health conditions other than cancer. These food environment variables are detailed in the data dictionary in Appendix A.
15. The presence of missing data in any of the features of the Food Environment Atlas was checked for; all but 14 of its 75 utilized features had at least one row with a missing value. These 14 features with no missing values are:
    a. 'PCT_DIABETES_ADULTS09': Adult diabetes rate, 2009
    b. 'PCT_DIABETES_ADULTS10': Adult diabetes rate, 2010
    c. 'PCT_OBESE_ADULTS09': Adult obesity rate (county), 2009
    d. 'PCT_OBESE_ADULTS10': Adult obesity rate (county), 2010
    e. 'PCT_OBESE_ADULTS13': Adult obesity rate, 2013
    f. 'RECFAC07': Recreation & fitness facilities, 2007
    g. 'RECFAC12': Recreation & fitness facilities, 2012
    h. 'RECFACPTH07': Recreation & fitness facilities/1,000 pop, 2007
    i. 'RECFACPTH12': Recreation & fitness facilities/1,000 pop, 2012
    j. 'PERPOV10': Persistent-poverty counties, 2010
    k. 'CHILDPOVRATE10': Child poverty rate, 2010
    l. 'PERCHLDPOV10': Persistent-child-poverty counties, 2010
    m. 'METRO13': Metro/nonmetro counties, 2010
    n. 'POPLOSS00': Population-loss counties, 2000
16. New Boolean mask columns were added for each of the 61 features with missing values designating which of their rows had missing values; a value of TRUE designated that the row had missing data and a value of FALSE designated that the row had data. These Boolean mask features were named as their respective corresponding feature with the suffix "_isnull" added.
17. The median value of each of the 61 features was calculated and their missing values were filled with their respective median value.

After adding all of these external features to the original dataset from data.world, a second run of a basic OLS linear regression model returned a test accuracy with no hyperparameter tuning of 59.4%, for an accuracy increase of 10.2% from 49.2%.

In order to capture nonlinear relationships between the target variable and the predictive feature set within the context of OLS linear regression, logarithmic and exponential transformations of a majority of the features were computed, plotted, and tested for their contribution to the OLS linear regression model's accuracy in a series of separate notebooks (named "2nd_Capstone_Cancer_LogExp_Expansion_1-4"). These four notebooks are found at the following GitHub locations:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_LogExp_Expansion_1.ipynb

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_LogExp_Expansion_2.ipynb

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_LogExp_Expansion_3.ipynb

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_LogExp_Expansion_4.ipynb

These notebooks first took each feature in isolation and ran a basic linear regression algorithm to determine that feature's accuracy on its own; plots of the actual values of the feature (Xx) and the target variable (y) were created with a line overlaid on this plot depicting the predicted value of the target y for each value of Xx. A second plot was also created showing each value of feature Xx and its residual. The logarithmic version of each feature Xx was then computed, and the two plots were shown again for the logarithmic version. The exponential (i.e., squared) version of each feature Xx was then computed, and the two plots were shown again for the exponential version. For many of the features, intriguing curves were created in place of the typical straight linear regression line, thereby capturing a nonlinear relationship between the feature Xx and the target variable y.

A logarithmic version of each feature Xx was then added to the overall feature set X, a train-test split was taken with a test size of 0.2 and a random state set, the linear regression algorithm was fitted to the training sets, and the accuracy was computed with the test sets. If there was an increase from the test accuracy of 55.3% noted above, the logarithmic version was kept in the feature set X and it was added to the cleaned 'cancer' DataFrame. If there was a decrease in test accuracy, it was dropped from the feature set X and *not* added to the cleaned 'cancer' DataFrame. This process was repeated for the exponential version of each feature Xx.

After a thorough process of experimentation, the addition of a subset of logarithmic and exponential features increased the OLS linear regression model's accuracy from 59.4% to 64.1%, for an additional accuracy increase of 4.7%. The list of which logarithmic and exponential features were added is included in the data dictionary in Appendix A. The spreadsheet that kept track of the effect on the OLS linear regression algorithm's accuracy that adding these logarithmic and exponential versions of the feature set had is stored on this project's GitHub repository at: https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_Log_Exp_Accuracy_Log.xls .

Visual Exploratory Data Analysis (EDA) & Data Story: Descriptive Statistics, Data Visualizations, Covariance Matrices, Correlations, and Plots of Linear and Nonlinear Relationships Between the Feature Set and the Target Variable

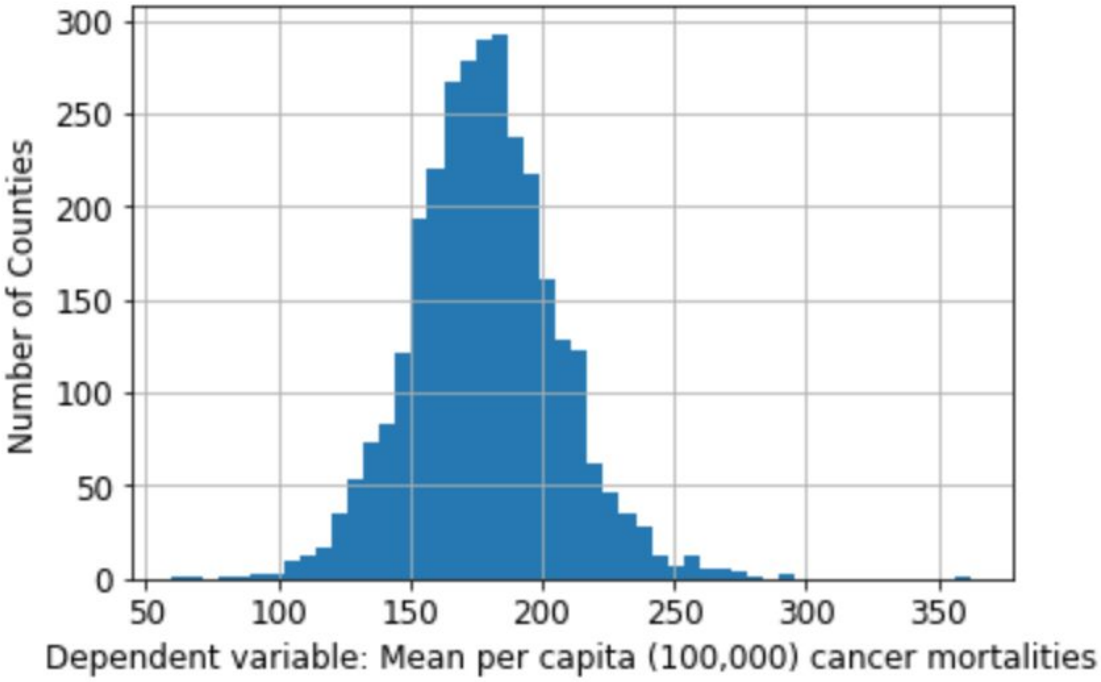
A comprehensive exploratory data analysis (EDA) of the cleaned and expanded cancer mortality DataFrame was carried out utilizing:

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 (details on this set are near the top of the Visual EDA notebook linked to below),
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.

This EDA is detailed in the following Jupyter notebook:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_Visual_EDA.ipynb

As a first step, the distribution of the target variable 'TARGET_deathRate' is visualized below. The distribution shows a roughly normal distribution with a high peak and relatively low variance.



The mean number of cancer mortalities in the United States in 2015 per 100,000 people was 178.664 and its standard deviation was 27.75. Of course, the population of U.S. counties included in this study range from 827 to 10.2 million people, so the actual count of cancer mortalities differed between counties based on county population size. Cancer mortality rates across all 3,047 counties in the DataFrame ranged from 59.7 in Pitkin County, Colorado (home of Aspen) to 362.8 in Union County, Florida. Five of the 10 counties with the lowest mortality rates were in Colorado, while five of the 10 counties with the highest mortality rates were in Kentucky.

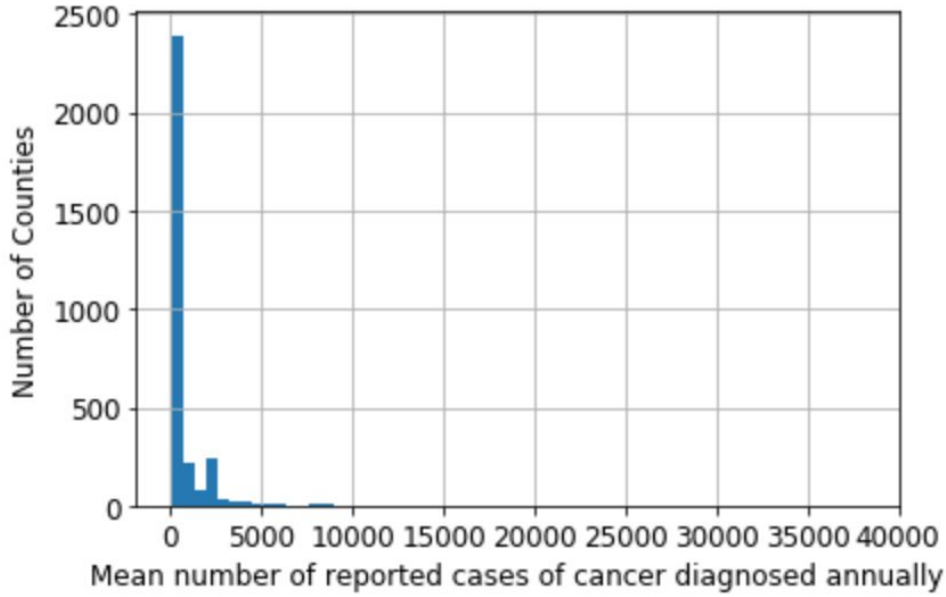
As mentioned above, only a subset of the feature set used for machine learning was explored using visual EDA. This was done for a few reasons:

1. The feature set consisted of 328 separate features, which is a large set to fully explore.
2. The feature set contains several dozen "_isnull" features which are simple Boolean features containing whether a referred column had a missing value or not in the original datasets.
3. Several features contain just slightly different information than those included in the visual EDA feature set.
4. Some features were only of interest from a machine learning perspective.

An even smaller subset of features contained interesting information pertinent to the project, whether that is a notable detail about the distribution of a feature's values, a sizable negative or positive correlation, or a noticeable increase in the accuracy of the OLS linear regression model due to a logarithmic and/or exponential transformation of a feature's values. For this reason, features are detailed in this report only if they meet at least one of three criteria: the feature had an unusual detail in its distribution of values, the feature had a negative or positive correlation of at least 0.1 with the target variable of per capita cancer mortality, or the feature's logarithmic and/or exponential version contributed an increase in the OLS linear regression model's accuracy of at least 0.001.

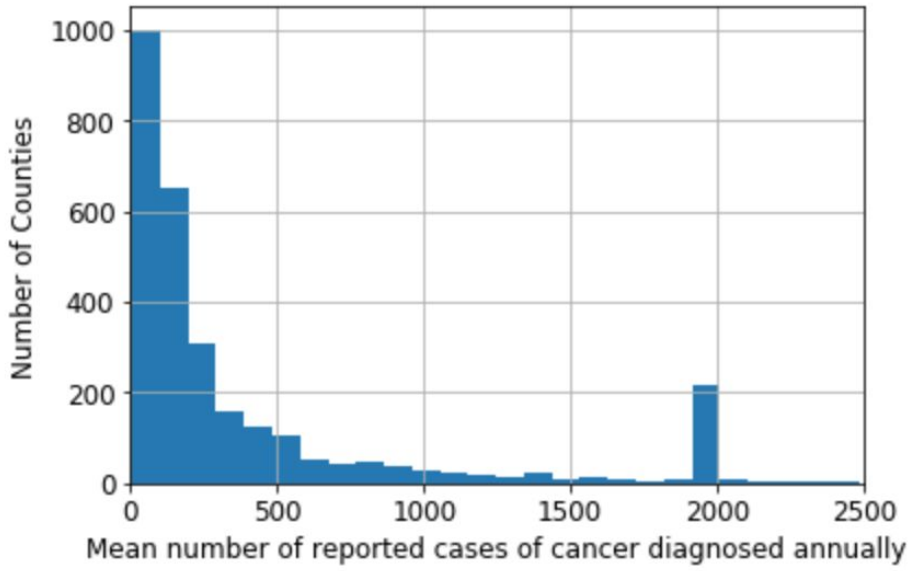
'avgAnnCount': Mean Number of Reported Cases of Cancer Diagnosed Annually

The first feature of interest is 'avgAnnCount', the mean number of reported cases of cancer diagnosed annually. The values of the 'avgAnnCount' feature range from 6 to 38,150, and have an average value of 606.3 with a standard deviation of 1,416. The full range with no X or Y limits imposed is visualized below. 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.



Focusing in on the main body of the distribution below, the majority of counties had between 6 and 600 reported cases of cancer diagnosed annually. There is an interesting peak at around 1,900

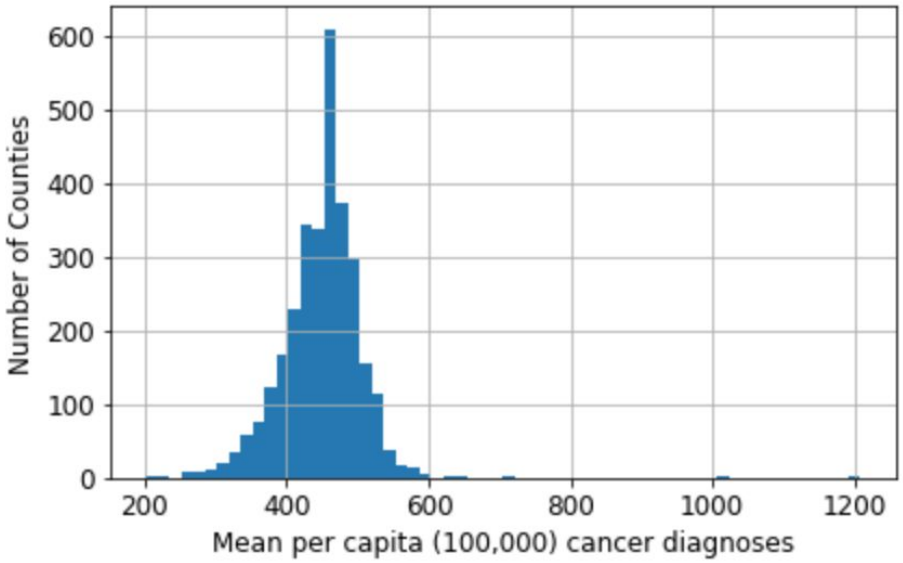
cases, where just over 200 counties have this number of cases diagnosed. Curiously, all of the cases in this peak have the exact same value for 'avgAnnCount', namely 1962.667684. This must be an imputational artifact from how the original dataset from data.world was constructed.



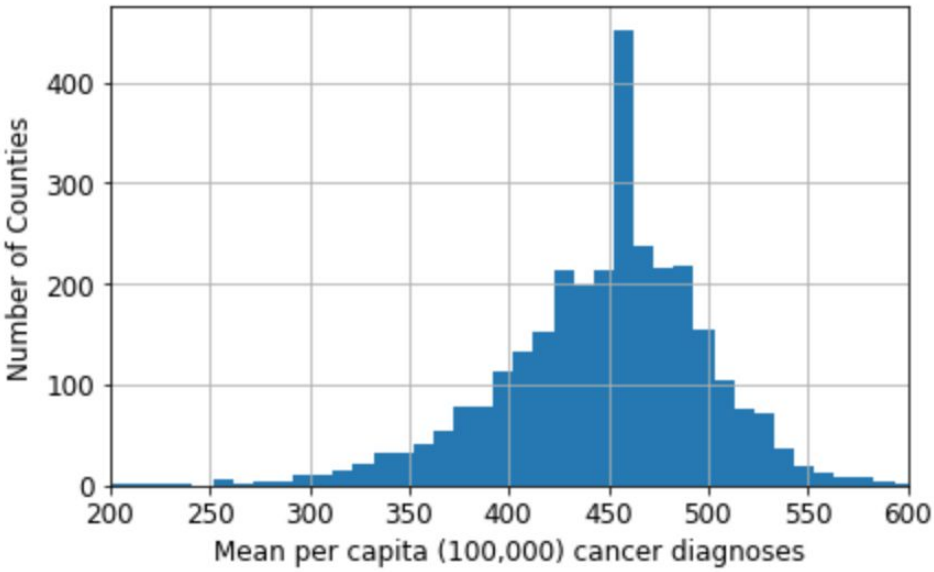
There is a slightly negative correlation of -0.14 between the number of diagnosed cancer cases and the number of cancer mortalities. This is a confusing correlation, as one would think that an increase of actual diagnosed cases would result in an increase in cancer mortality. This result could be due 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. Adding a logarithmic or exponential transformation of 'popEst2015' did not increase the model's overall accuracy.

'incidenceRate': Mean per capita (100,000) cancer diagnoses

The 'incidenceRate' feature is a per capita version of 'avgAnnCount', giving the mean per capita (100,000) cancer diagnoses per county. The feature's distribution ranges from 201 to 1,206 with a mean of 448.3 and a standard deviation of 54.6. The bulk of it lies between 200 and 600 diagnosed cases per 100,000 county residents. There are several outlier counties however, which are Powell and Bracken Counties in Kentucky, Petersburg City, Charlottesville City, and Williamsburg City in Virginia, and Union County, Florida.



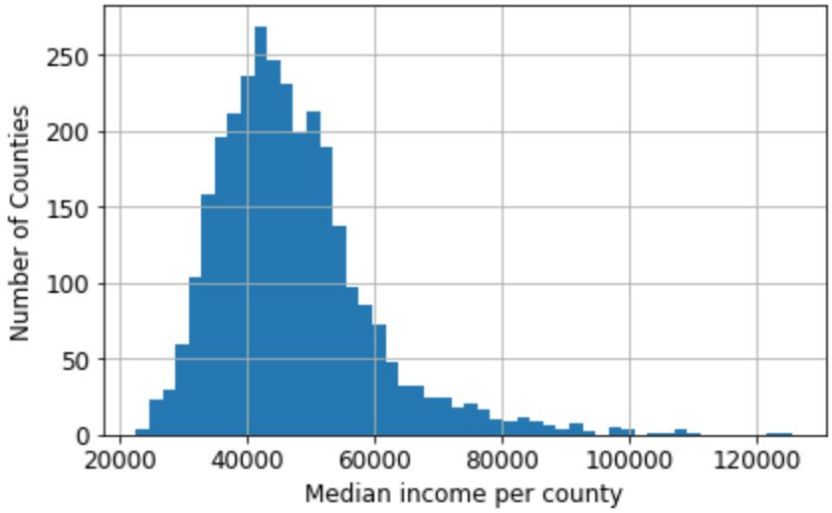
Looking at just the bulk of the distribution, one can see an unusual peak at just over 450 diagnosed cases per 100,000 people that stands high above the rest of the distribution. Calling the value counts of this range shows that 206 cases have the exact same incidence rate, 453.549422. This must be an artifact of the assemblage process of the original data.world dataset.



There was a moderately strong correlation of 0.45 between the cancer incidence rate and cancer mortality rate. This makes sense and may partially exist because, unlike the 'avgAnnCount' variable, the 'incidenceRate' and 'TARGET_deathRate' features are on the same scale, namely per capita (100,000) cases. Logarithmic and exponential transformations of the 'incidenceRate' did not result in an increase in model accuracy.

'medIncome': Median Income per County

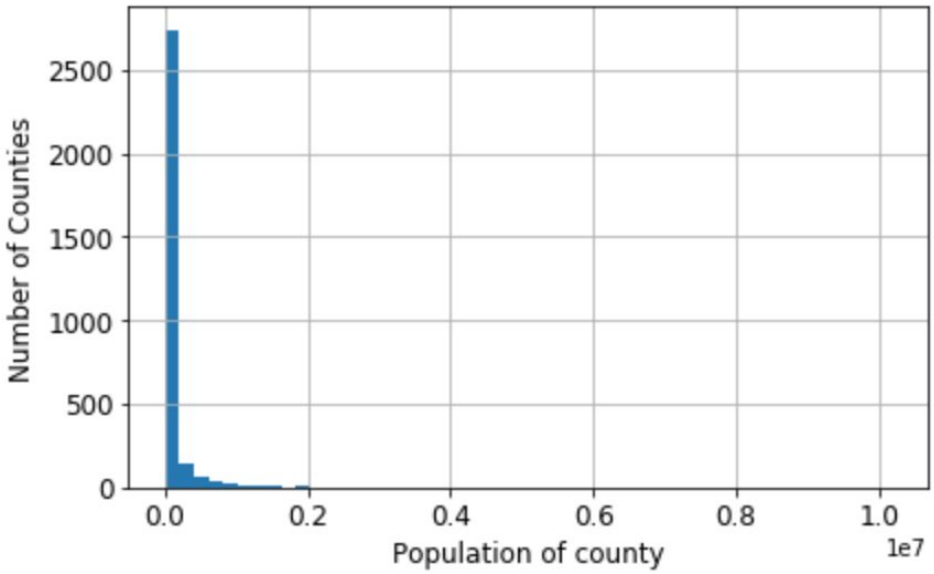
The median income feature had a distribution ranging from just over $20,000 a year to just over $120,000 a year, with a mean of $47,063 and a standard deviation of $12,040. The bulk of the distribution is between $20,000 and $80,000. The counties with the lowest median income are in Mississippi, Kentucky, Alabama, and West Virginia. The counties with the highest median income are in Colorado, New Mexico, and Virginia.

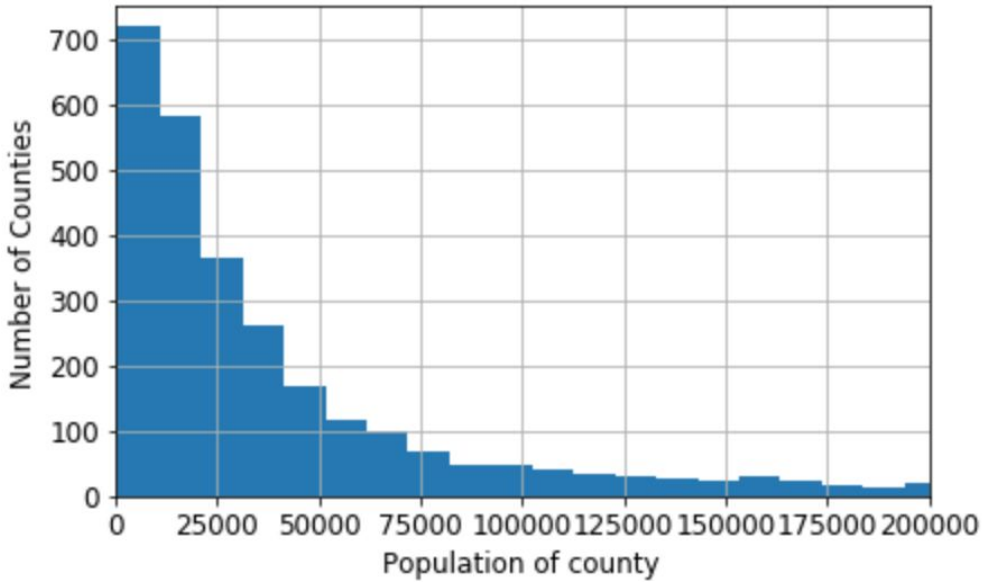


Median income and cancer mortality had a moderately negative correlation of -0.43, where cancer mortality in a county decreased as median income increased. Adding a logarithmic or exponential transformation of 'popEst2015' did not increase the model's overall accuracy.

'popEst2015': Population of County

The population of U.S. counties in this dataset ranges from 827 to 10.2 million people with an average value of 102,637 and a standard deviation of 329,059, although the bulk of the distribution ranges from 827 to roughly 200,000. The full distribution is visualized below, as is the bulk. As can be seen, the majority of counties have a population less than 25,000 people.

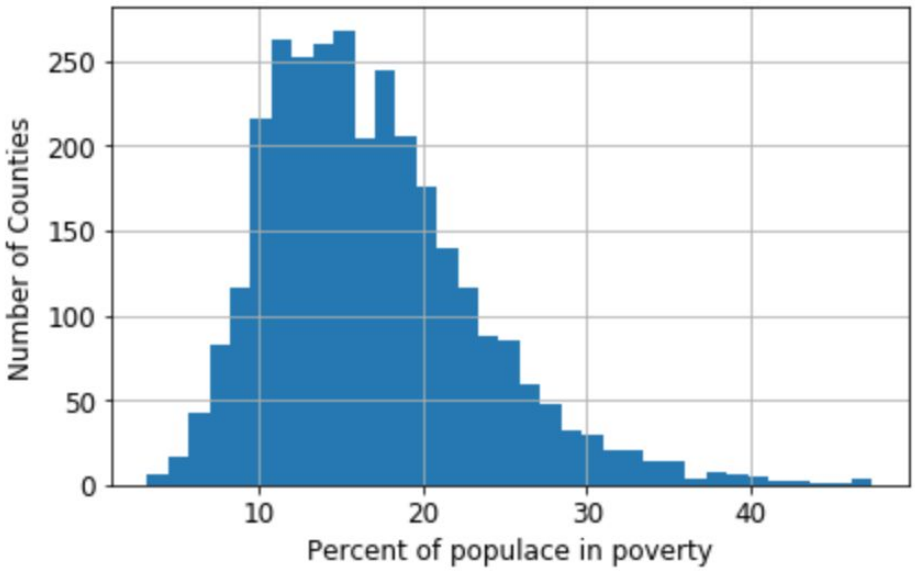




The 2015 county population estimate feature has a weak negative correlation with cancer mortality of -0.12, showing that as the county population increases cancer mortality slightly decreases. Adding a logarithmic or exponential transformation of 'popEst2015' did not increase the model's overall accuracy.

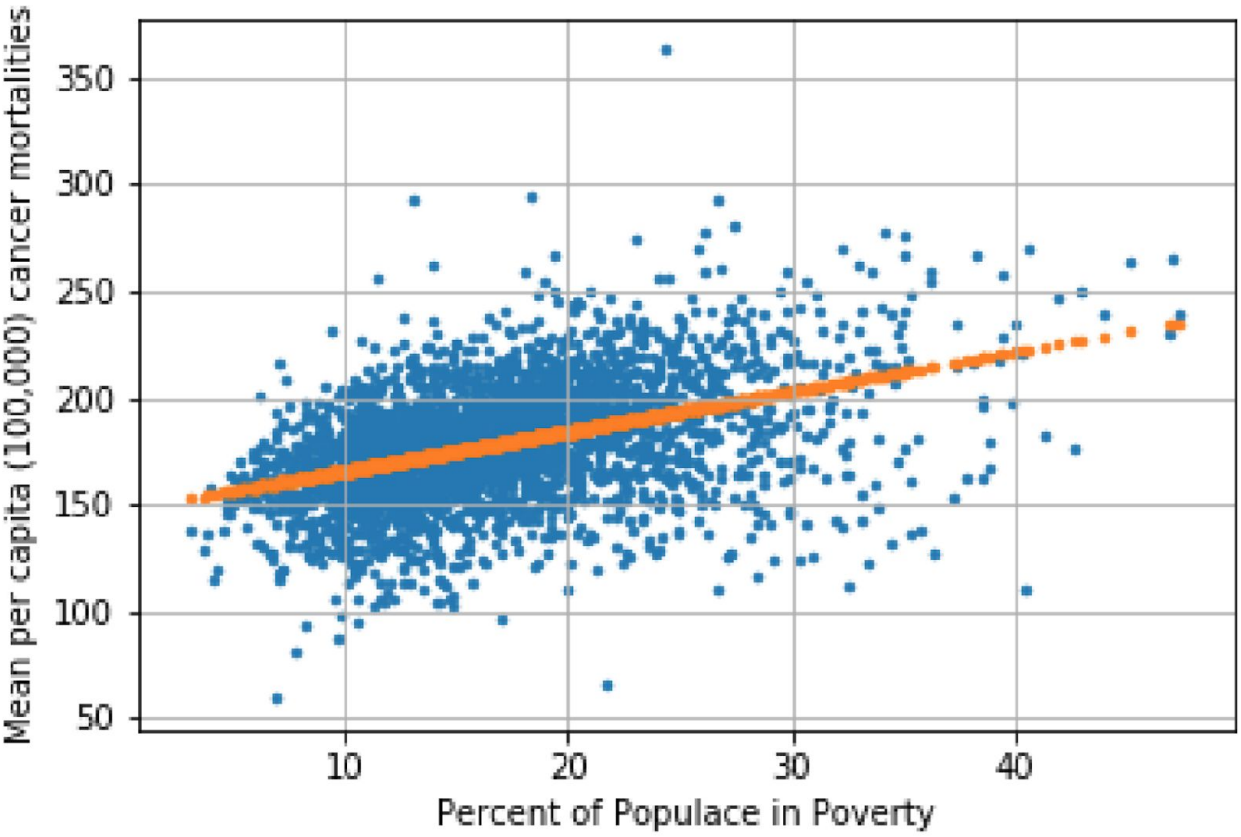
'povertyPercent': Percent of populace in poverty

The percent of the populace in U.S. counties who live at or under the poverty line ranges from 3.2% to 47.4%, and has an average value of 16.9% with a standard deviation of 6.4%. The five counties with the lowest poverty are in Virginia, Colorado, and New Mexico. The five counties with the highest poverty are in Mississippi, Kentucky, Alabama, and South Dakota.



There is a moderately strong correlation of 0.43 between poverty and cancer mortality. This shows that people diagnosed with cancer who have less financial means stand a greater chance of dying from cancer.

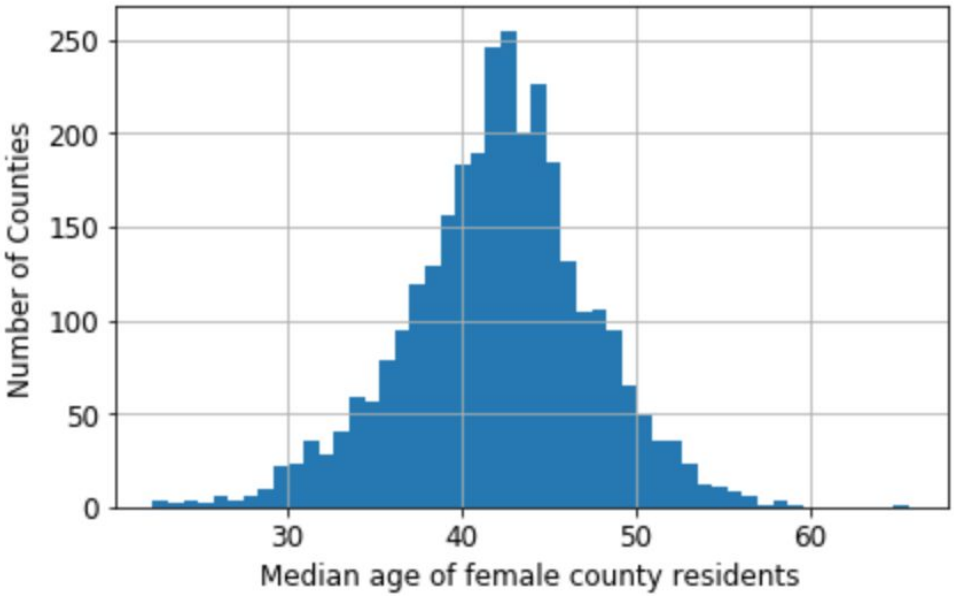
One can clearly see the moderate correlation between poverty and cancer mortality in the prediction line of the plot below.



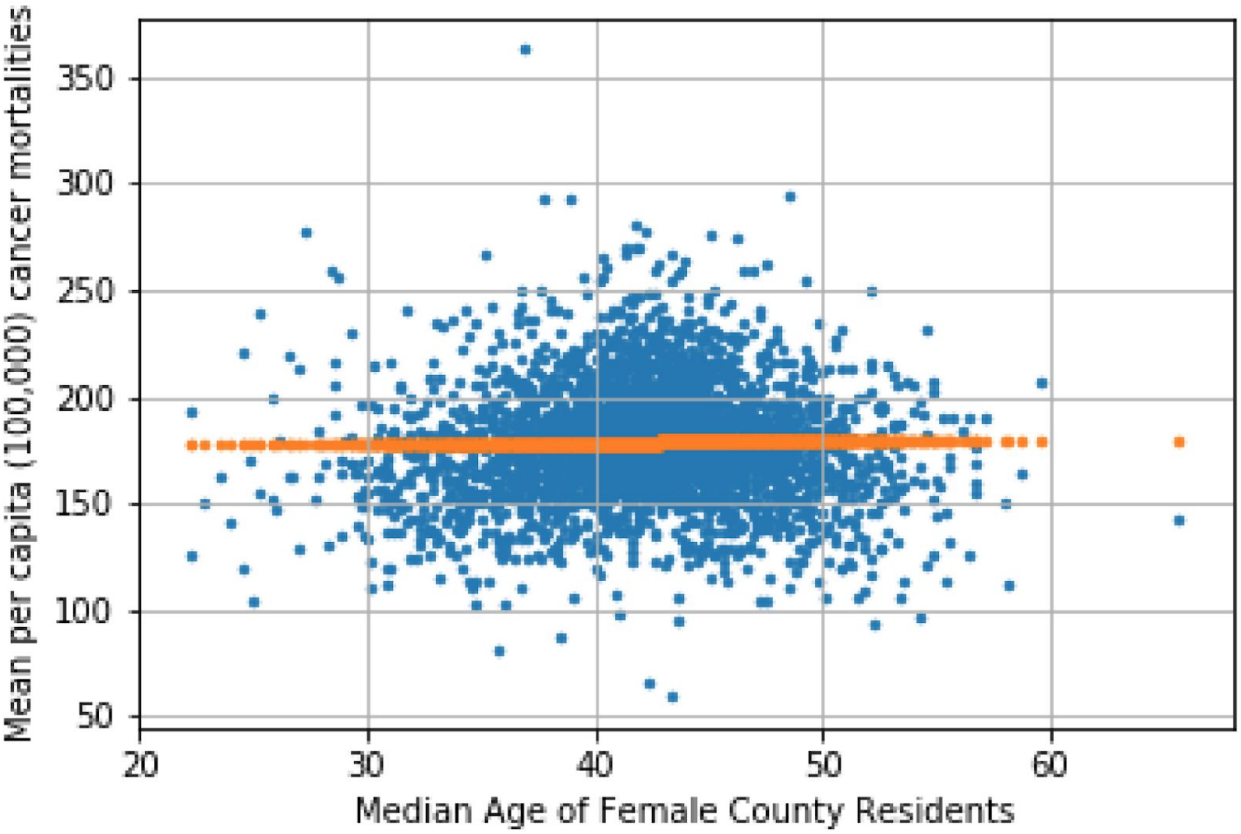
One can also see that adding the logarithmic and exponential transformations of 'povertyPercent' predicts several of the data points more accurately, and adding them increased the linear regression model's accuracy by 0.0008.

'MedianAgeFemale': Median age of female county residents

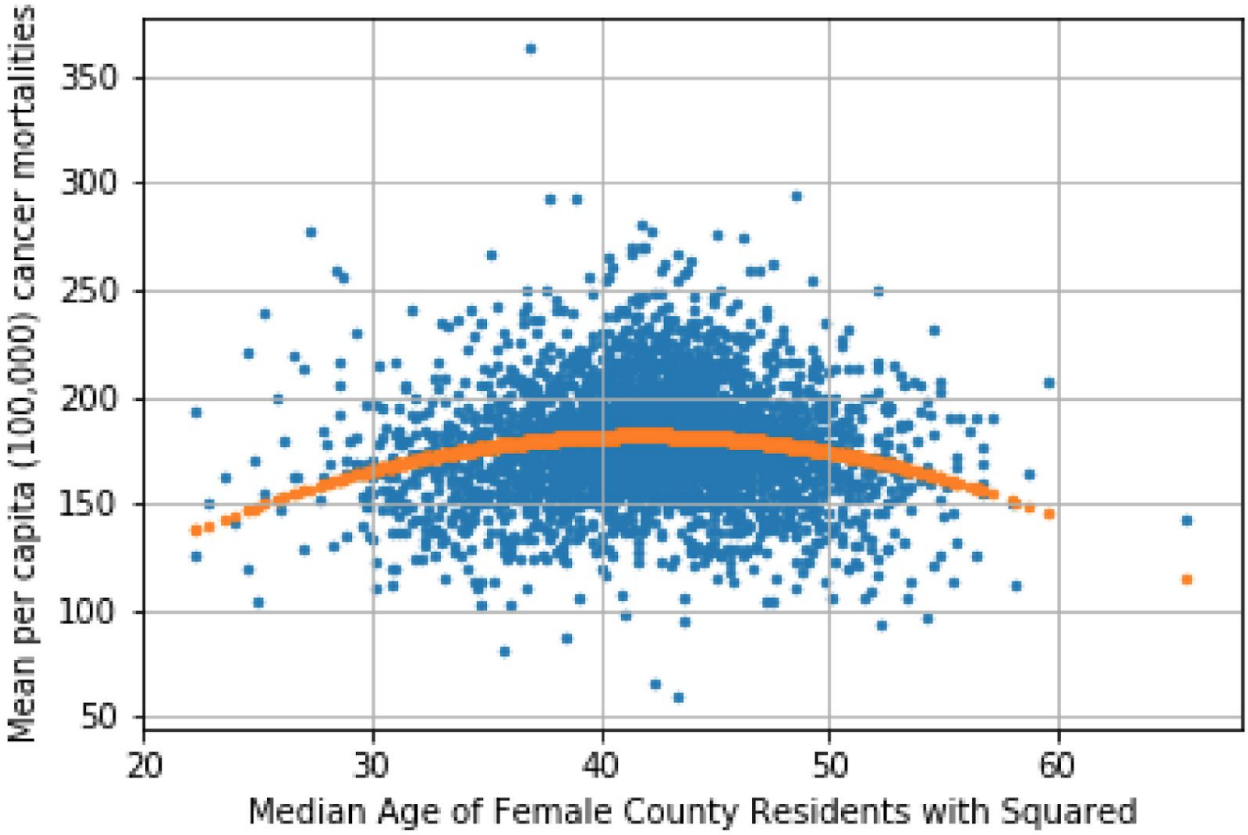
The female median age ranges from 22 to 66, and has an average value of 42 with a standard deviation of 5.3.



There is nearly no correlation (0.01) between female median age and cancer mortality. One can see the nearly flat correlation line between female median age and cancer mortality in the prediction line below.

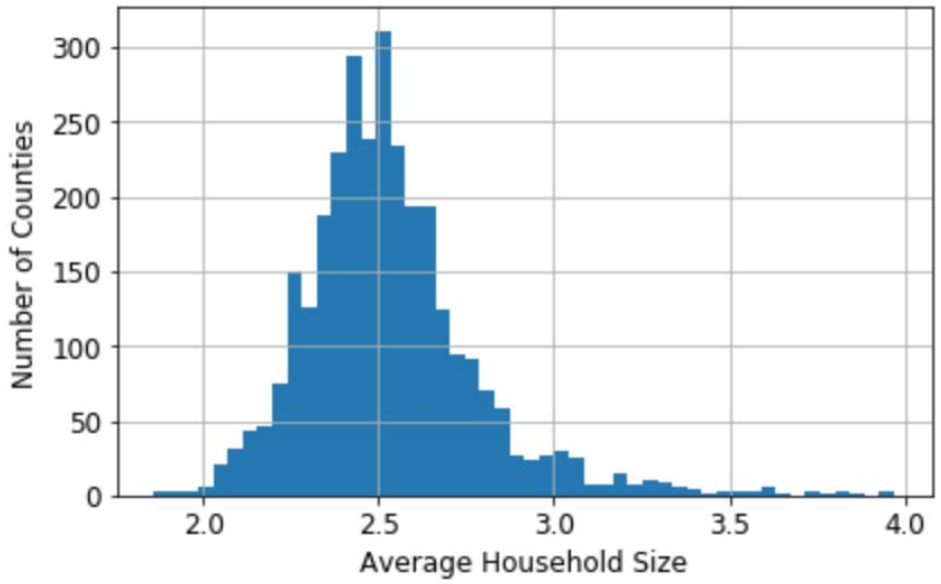


Adding the squared transformation of female median age reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.00005.



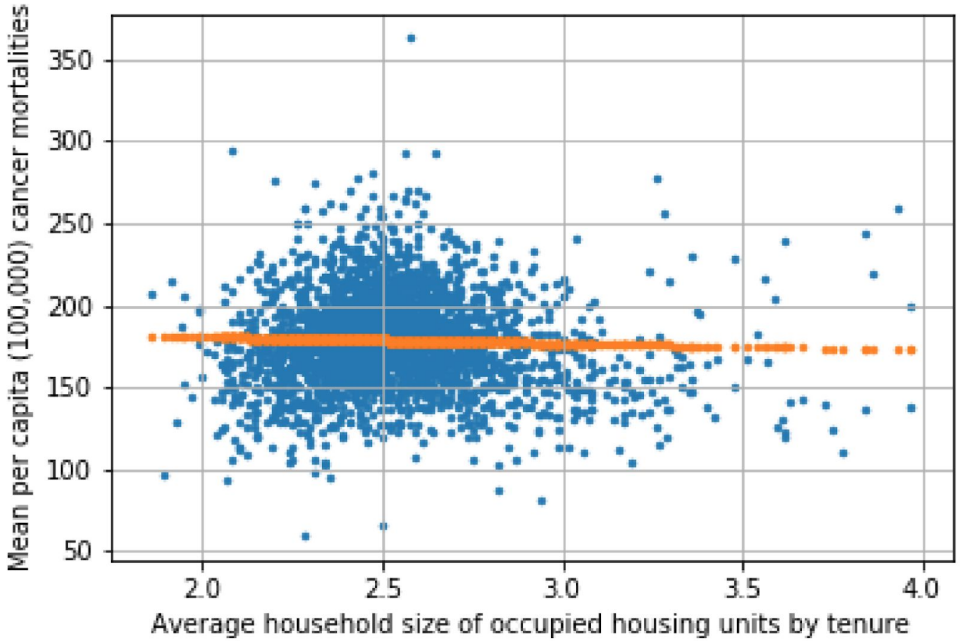
'AvgHouseholdSize': Average Household Size (occupied buildings)

The average household size ranges from 1.86 to 3.97, and has an average value of 2.5 with a standard deviation of 0.25.

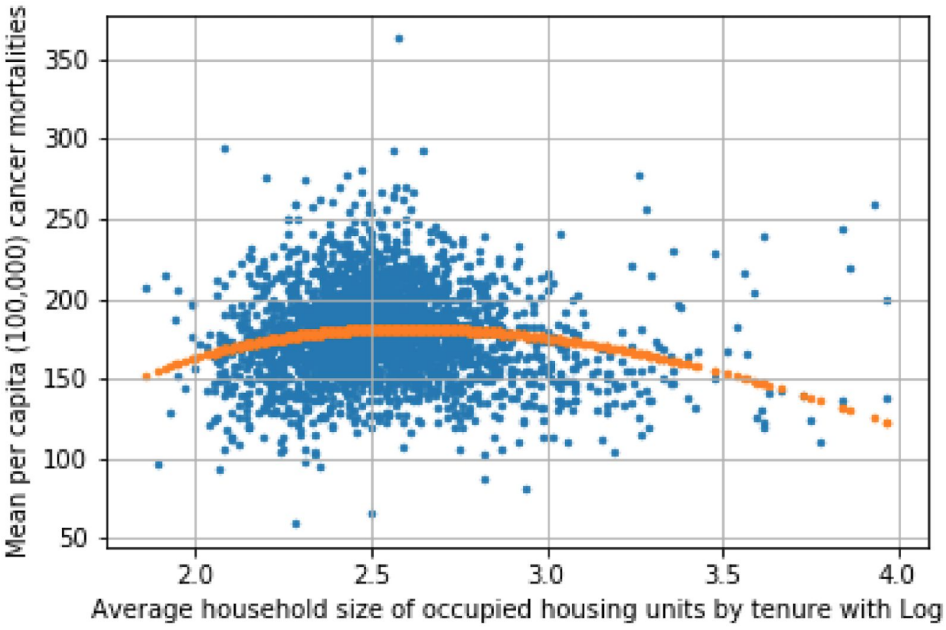


There is nearly no correlation between average household size and cancer mortality (-0.04).

One can see the nearly flat correlation between the average household size and cancer mortality in the prediction line of the first of the plots below.

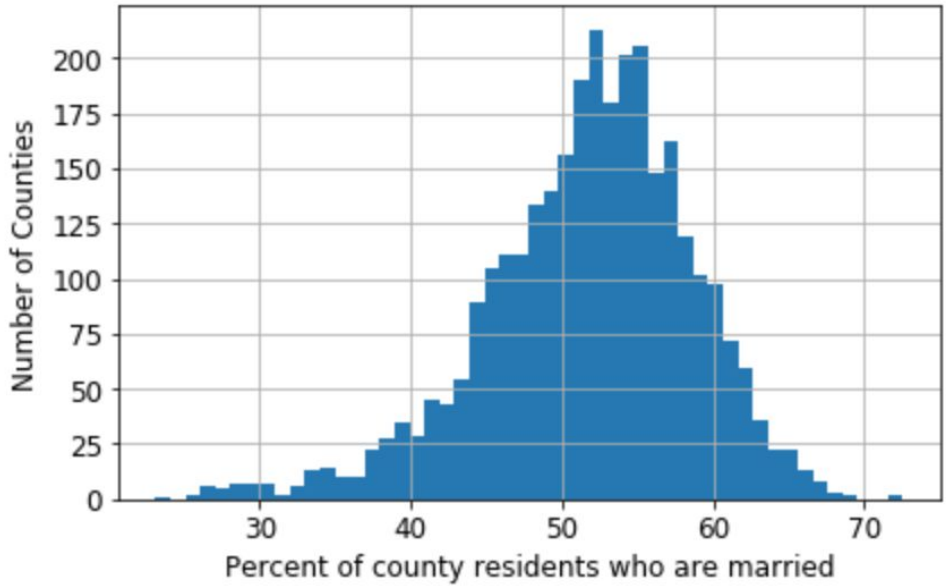


The error between the actual and predicted data points of cancer mortality is reduced when the logarithmic transformations of 'AvgHouseholdSize' is added, and doing so increased the model's accuracy by 0.0007.



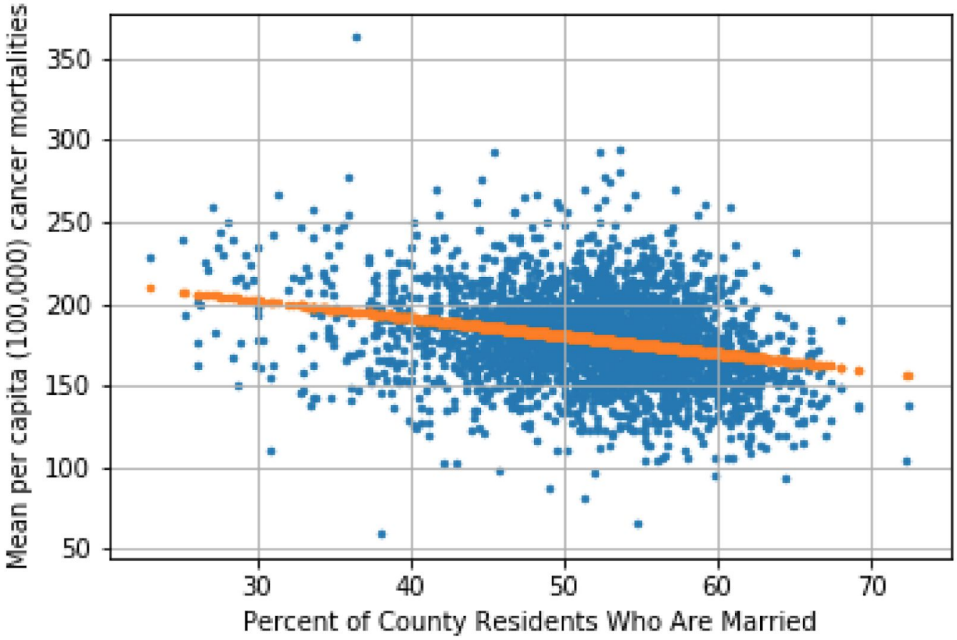
'PercentMarried': Percent of county residents who are married

The percentage of county residents who were married ranges from 23% to 73%, and has an average value of 52% with a standard deviation of 6.9%.

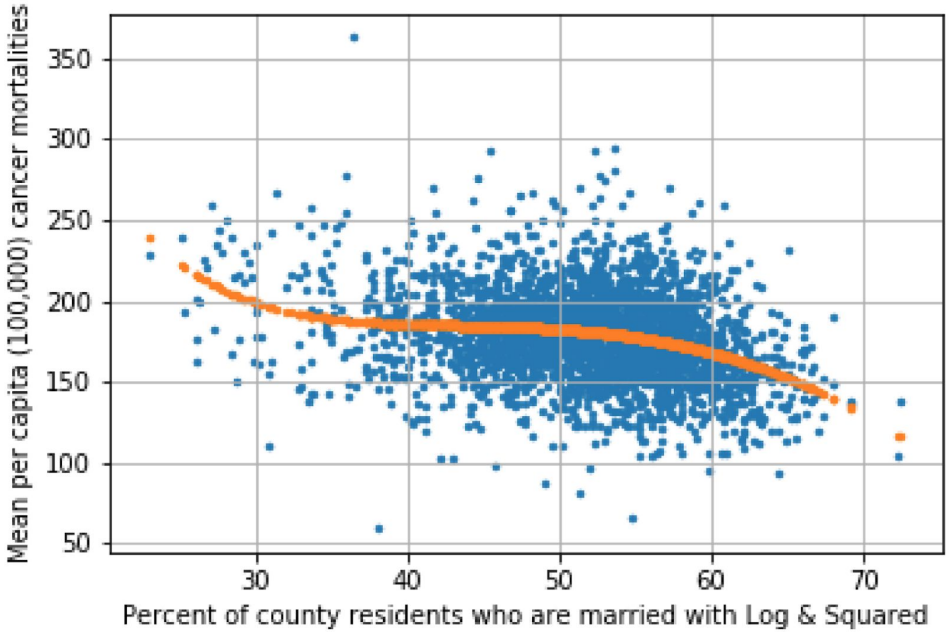


There is a somewhat weak negative correlation of -0.27 between the percentage of county residents who are married and cancer mortality rate. This suggests a weak relationship between being married and being less likely to die from cancer, but of course no causative relationship can be posited here.

One can see the weakly negative correlation between the percentage of county residents married and cancer mortality in the prediction line of the first of the plots below.

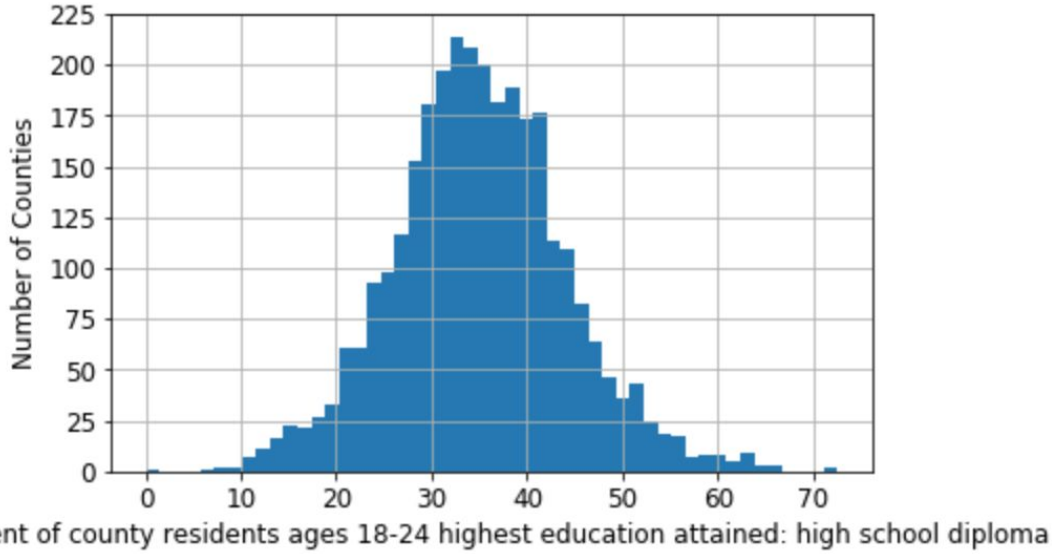


The error between the actual and predicted data points of cancer mortality is reduced when the logarithmic transformation of 'PercentMarried' is added, and doing so increased the model's accuracy by 0.003.



'PctHS18_24': Percent of county residents ages 18-24 highest education attained: high school diploma

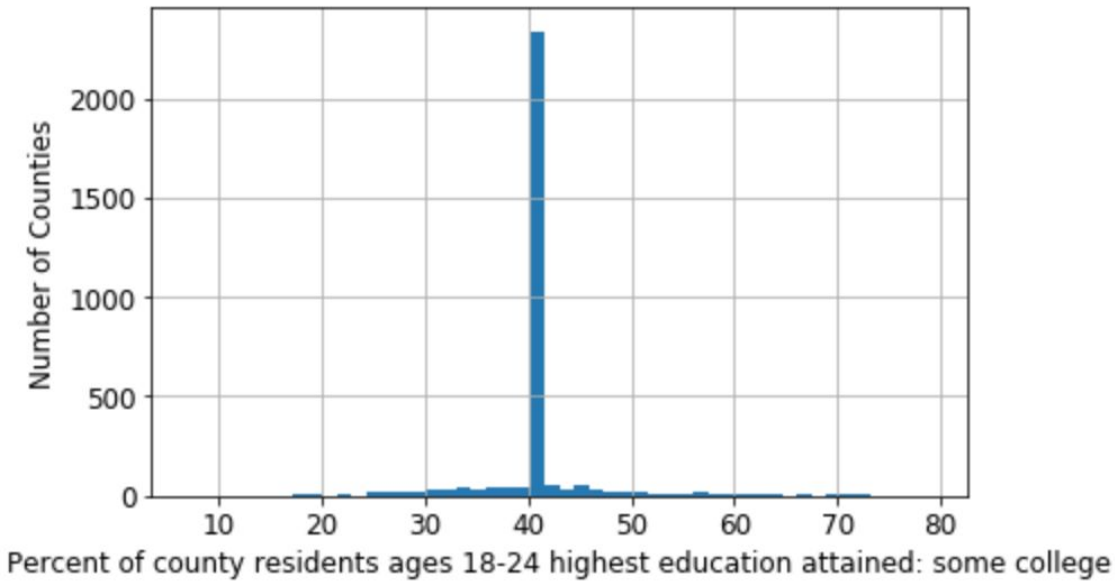
The percentage of county residents that are 18-24 years old whose highest education is a high school diploma ranges from zero to 72.5%, and has an average value of 35% with a standard deviation of 9.1%.



There is a 0.26 correlation between the percentage of 18-24 year old county residents whose highest education is a high school diploma and cancer mortality, showing some relationship between the prevalence of lower education and increased cancer mortality. Adding a logarithmic and exponential transformation of the 'PctHS18_24' feature did not increase the linear regression model's accuracy.

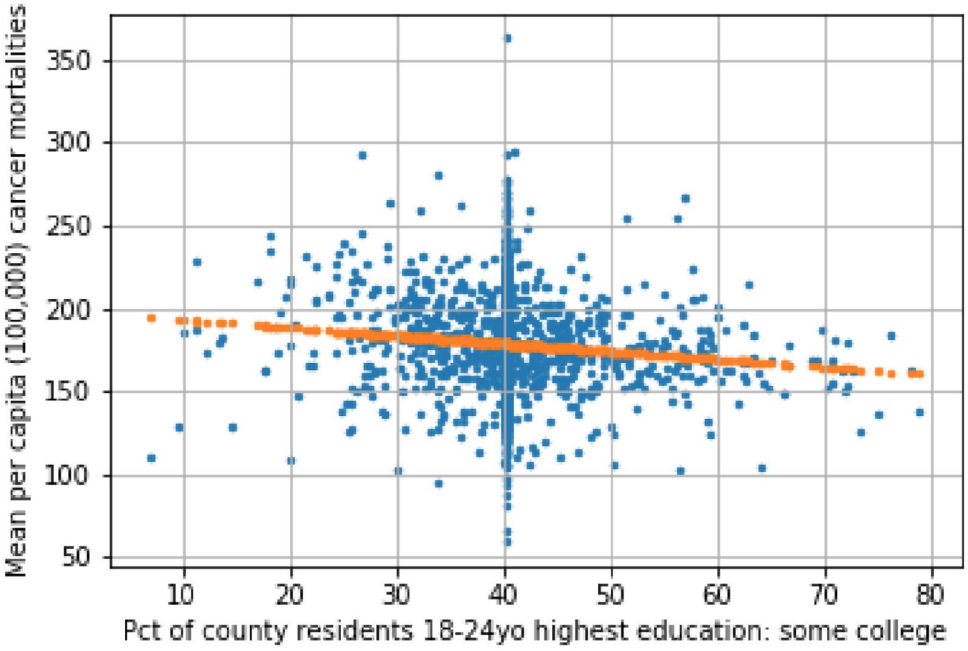
'PctSomeCol18_24': Percent of county residents ages 18-24 highest education attained: some college

The percentage of county residents that are 18-24 years old whose highest education is some college education ranges from 7.1% to 79%, and has a mean value of 40.5% with a standard deviation of 5.6%. The histogram below has such a strong peak at the median because nearly 75% of the counties initially had missing data in the original data set and the values for this variable were filled by the median (as described in the data cleaning section above).

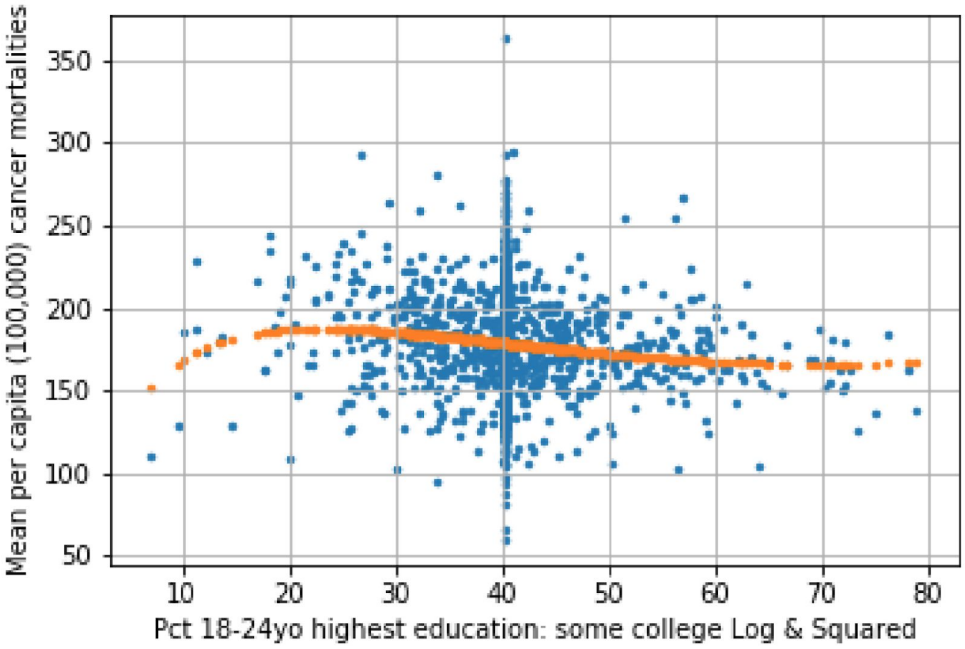


There is a weak negative correlation of -0.10 between 'PctSomeCol18_24' and cancer mortality, showing that as the percentage of 18-24 year olds in a county whose highest education is some college (but not a degree) increases, cancer mortality goes slightly down.

One can see the weakly negative correlation between the percentage of 18-24 year olds in a county whose highest education is some college and cancer mortality in the prediction line of the first of the plots below.

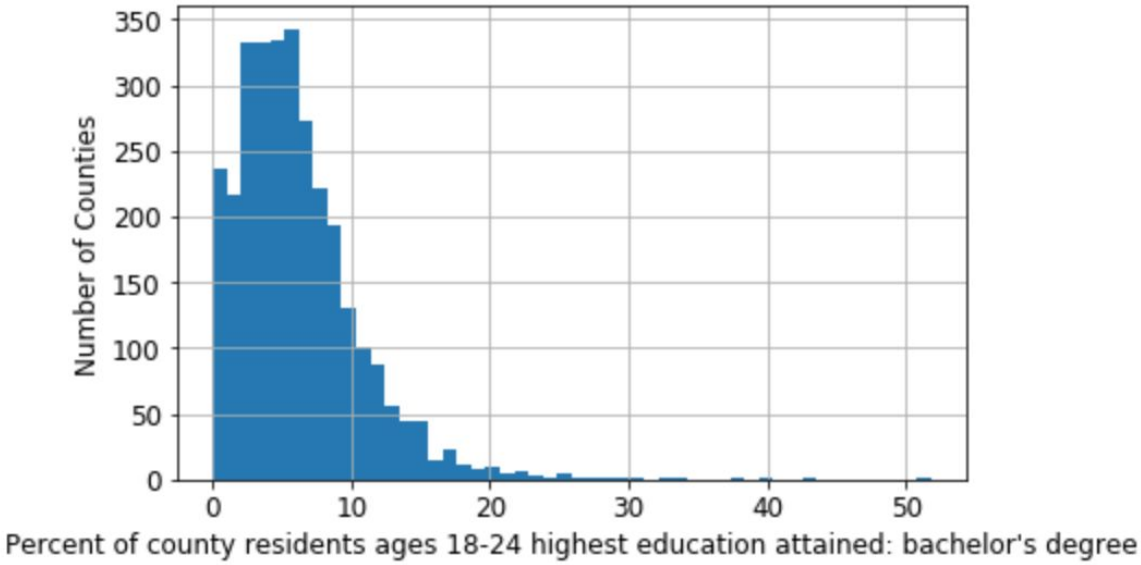


The error between the actual and predicted data points of cancer mortality is reduced when the logarithmic and exponential transformations of 'PctSomeCol18_24' are added, and doing so increased the model's accuracy by 0.0006.



'PctBachDeg18_24': Percent of county residents ages 18-24 highest education attained: bachelor's degree

The percentage of county residents that are 18-24 years old with their highest education being a Bachelor's degree ranges from zero to 51.8%, and has a mean value of 6.2% with a standard deviation of 4.5%.

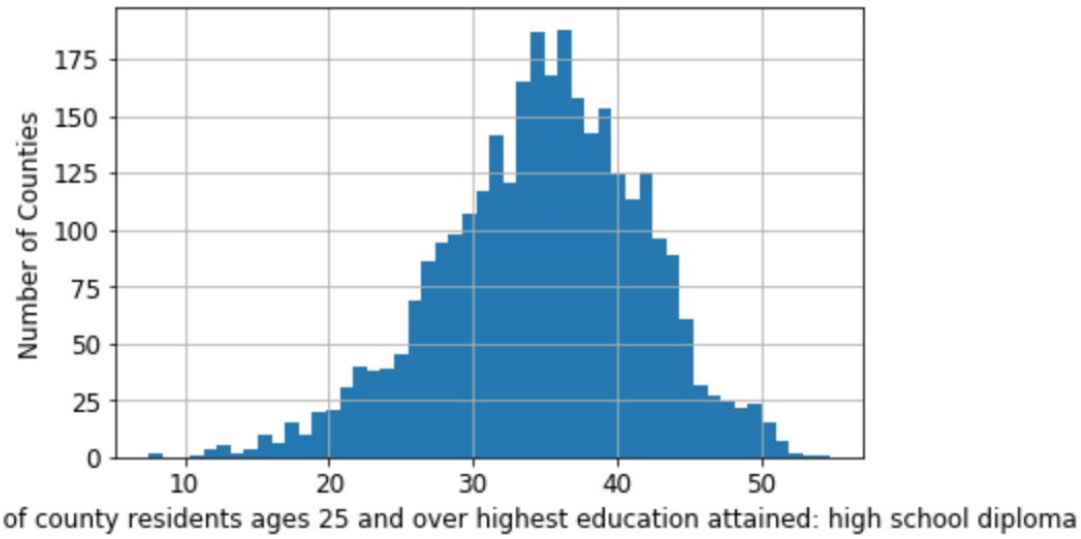


There is a moderate negative correlation of -0.29 between the percentage of 18-24 year old county residents with their highest education being a Bachelor's degree and cancer mortality. This again shows that there is a relationship between the level of education in a county and reduced cancer mortality, although a causative link can't be made between the two as part of this analysis.

Adding a logarithmic or exponential expansion of 'PctBachDeg18_24' did not add to the model's overall accuracy.

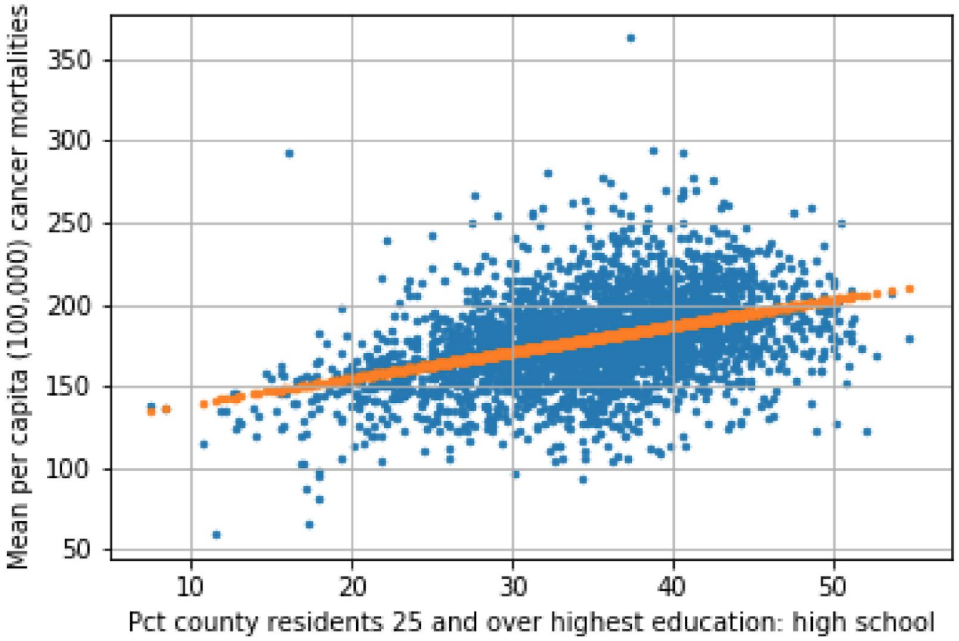
'PctHS25_Over': Percent of county residents ages 25 and over highest education attained: high school diploma

The percentage of county residents that are 25 years and over with their highest education being a high school degree ranges from 7% to 54.8%, and has a mean value of 34.8% with a standard deviation of 7%.

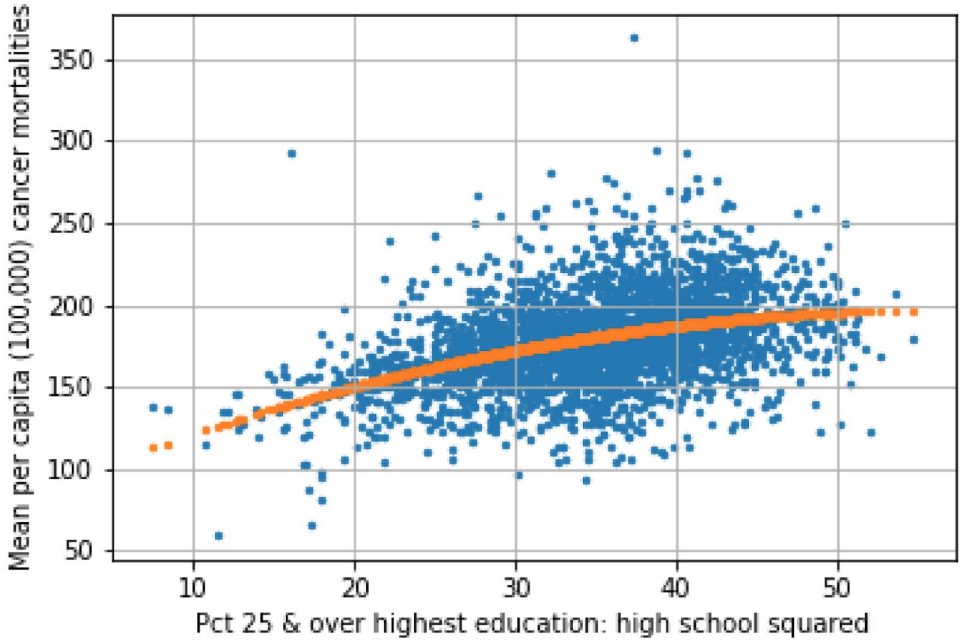


There is a moderately strong correlation of 0.41 between the percentage of county residents 25 years and over with their highest education being a high school degree and cancer mortality. This again shows the relationship between lower education and an increased rate of cancer mortality, because it is shown with this correlation that cancer mortality increases as the percentage of adults in a county with only a high school diploma increases.

One can see the moderately positive correlation between the percentage of county residents 25 years and over with their highest education being a high school degree and cancer mortality in the prediction line of the first of the plots below.

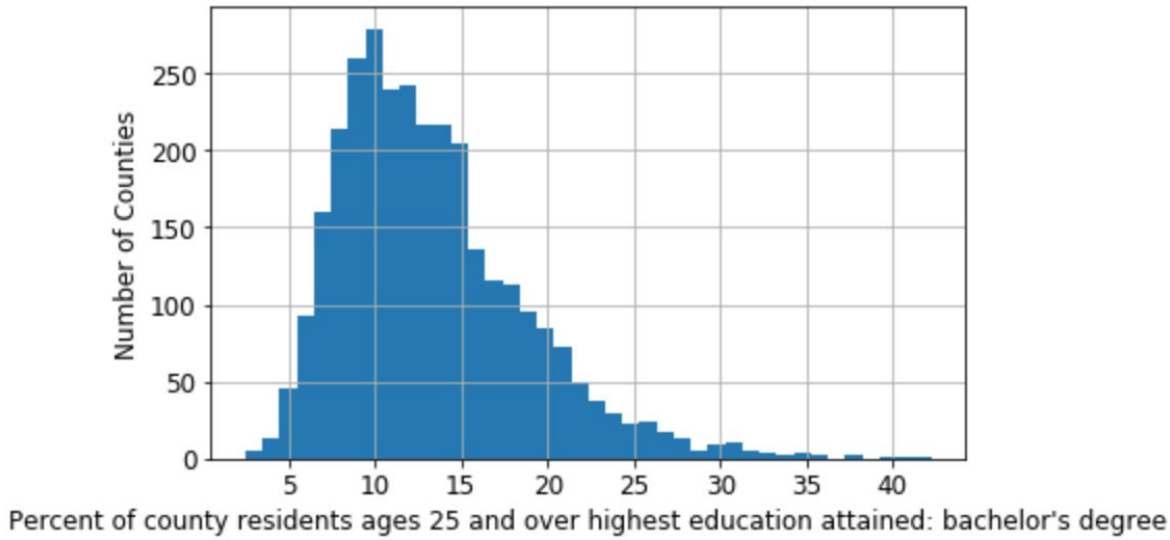


The error between the actual and predicted data points of cancer mortality is reduced when the exponential transformation of 'PctHS25_Over' is added, and doing so increased the model's accuracy by 0.0002.



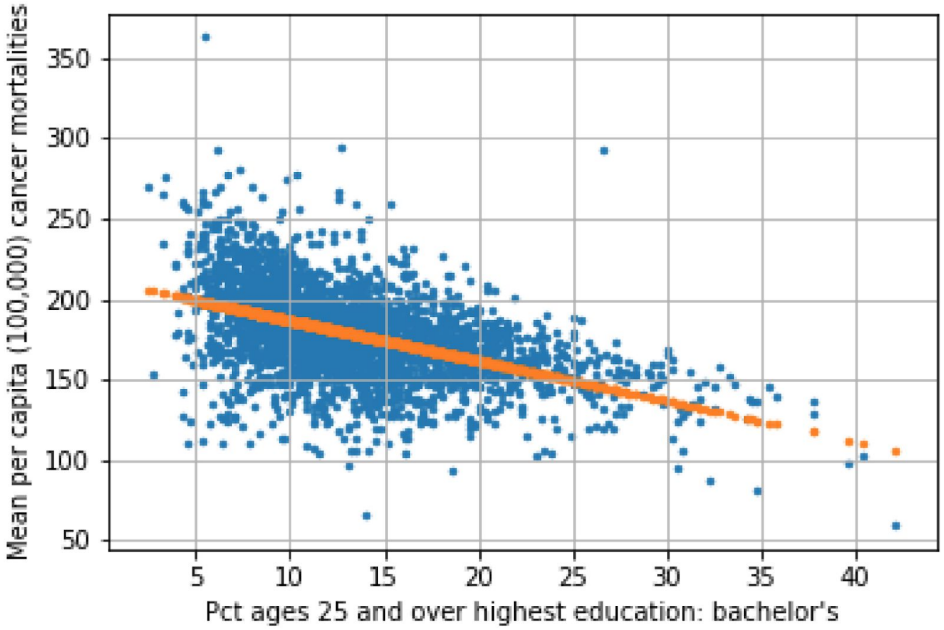
'PctBachDeg25_Over': Percent of county residents ages 25 and over highest education attained: bachelor's degree

The percentage of county residents that are 25 years and older with their highest education being a bachelor's degree ranges from 2.5% to 42.2%, and has a mean value of 13.3% with a standard deviation of 5.4%.

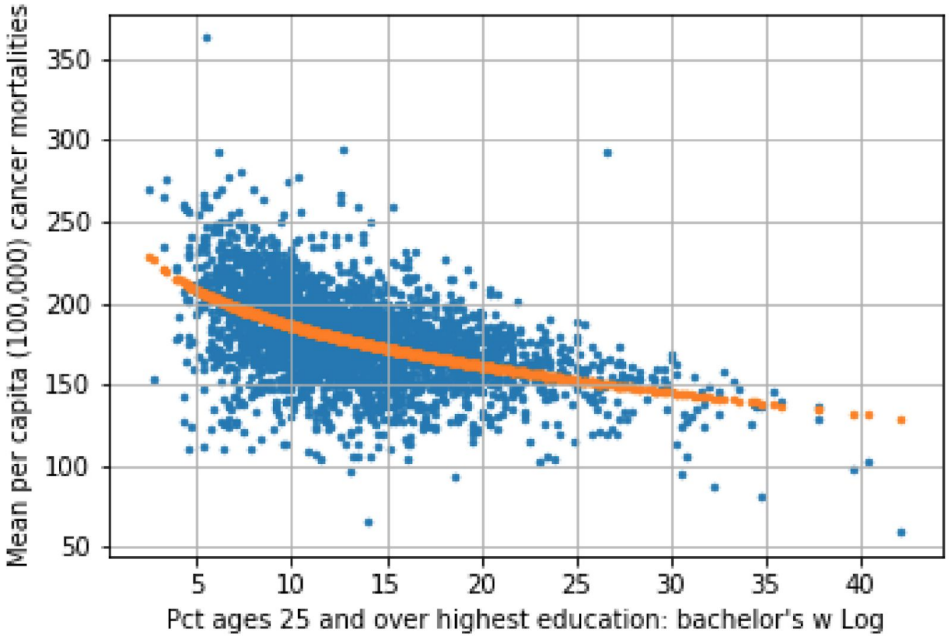


There is a moderately negative correlation of -0.49 between the percentage of county residents 25 years and older with their highest education being a Bachelor's degree and cancer mortality. This again shows the relationship between education and cancer mortality, because as that educated percentage goes up, cancer mortality goes down.

One can see the moderately positive correlation between the percentage of county residents 25 years and older with their highest education being a Bachelor's degree and cancer mortality in the prediction line of the first of the plots below.

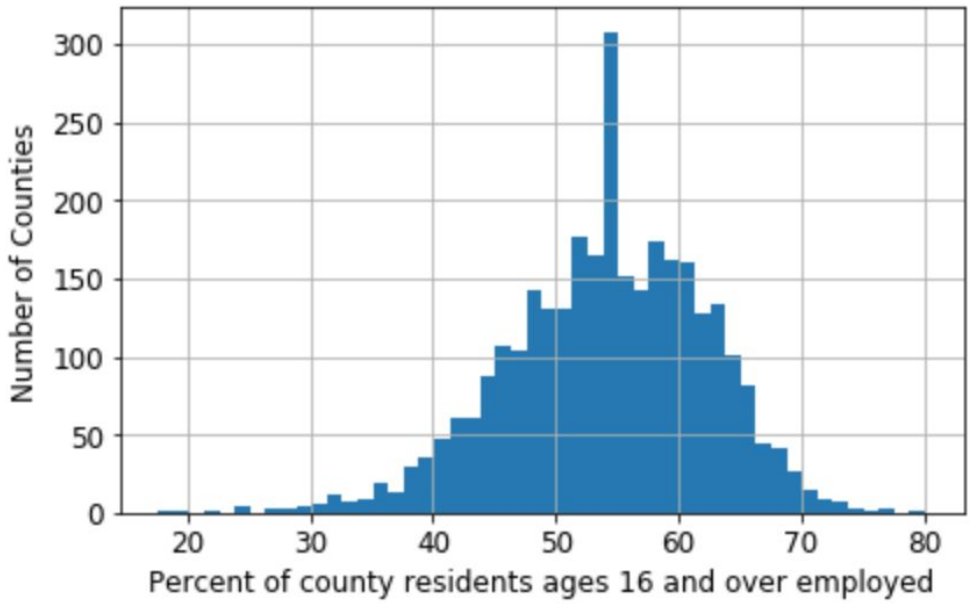


The error between the actual and predicted data points of cancer mortality is reduced when the logarithmic transformation of 'PercentMarried' is added, and doing so increased the model's accuracy by 0.0004.



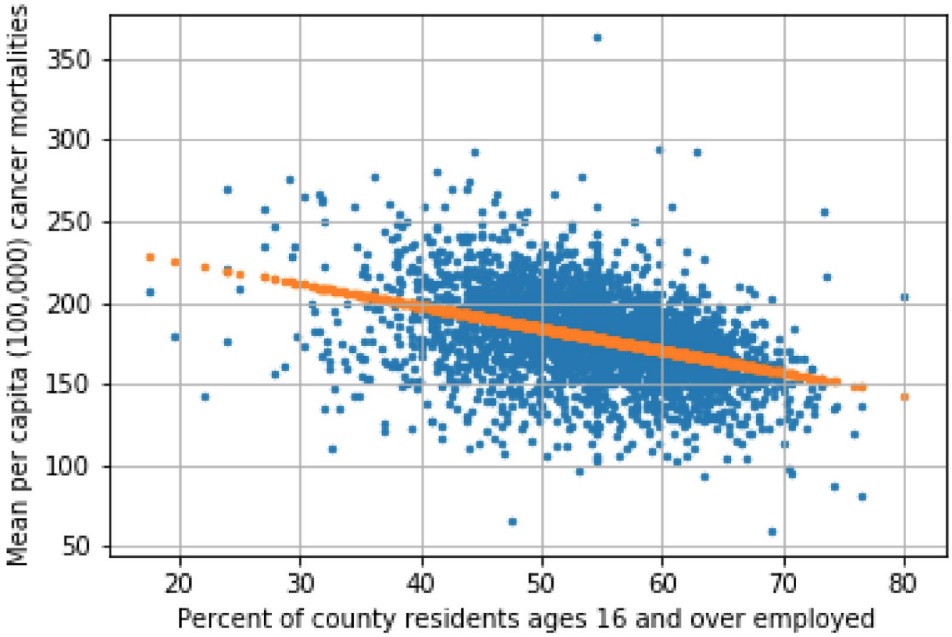
'PctEmployed16_Over': Percent of county residents ages 16 and over employed

The percentage of county residents that are 16 years and older and are employed ranges from 17.6% to 80.1%, and has a mean value of 54.2% with a standard deviation of 8.1%. One can see a high central measurement peak in the histogram below because 5% of the values for 'PctEmployed16_Over' were missing and these counties' values were filled in with the national median.

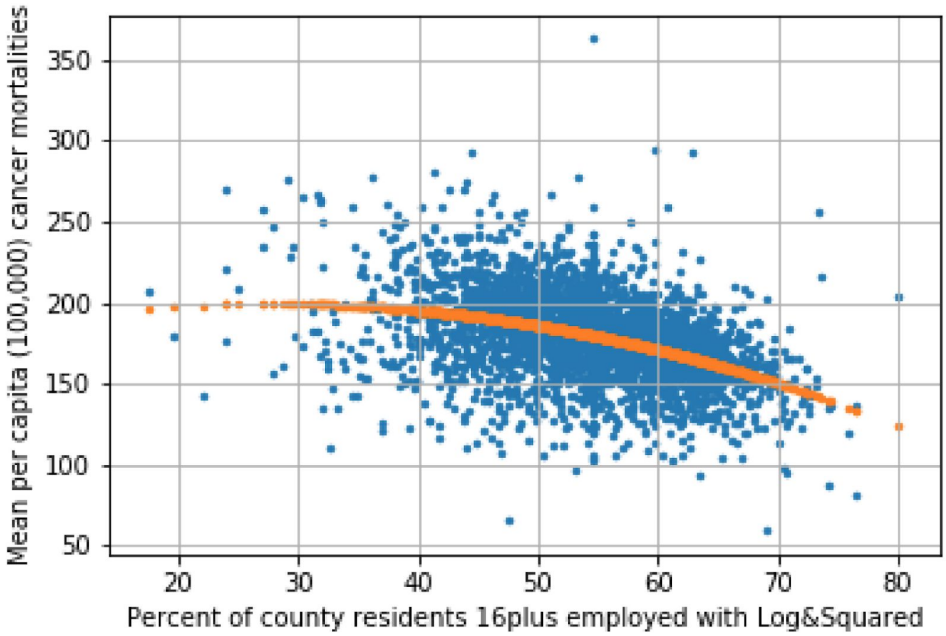


There is a moderately negative relationship of -0.40 between the percentage of county residents 16 years and older and cancer mortality. This shows that as the percentage of employed people in a county increases, cancer mortality decreases.

One can see the negative correlation line between the percentage of county residents 16 and older who are employed and cancer mortality in the prediction line below.

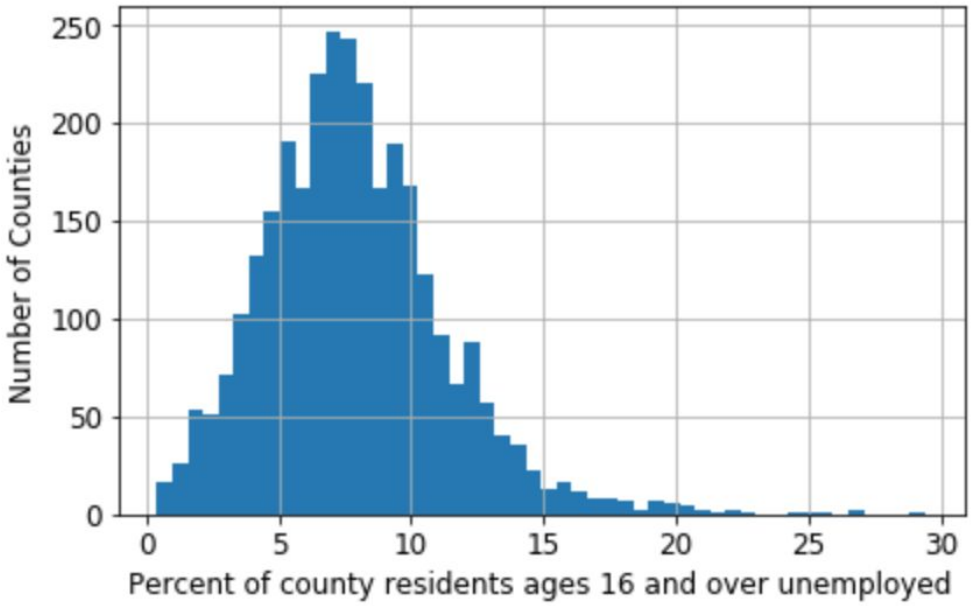


Adding the logarithmic and exponential transformations of this feature reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.006.



'PctUnemployed16_Over': Percent of county residents ages 16 and over unemployed

The percentage of county residents that are 16 years and over who are unemployed ranges from 0.4% to 29.4%, with a mean value of 7.9% and a standard deviation of 3.5%.

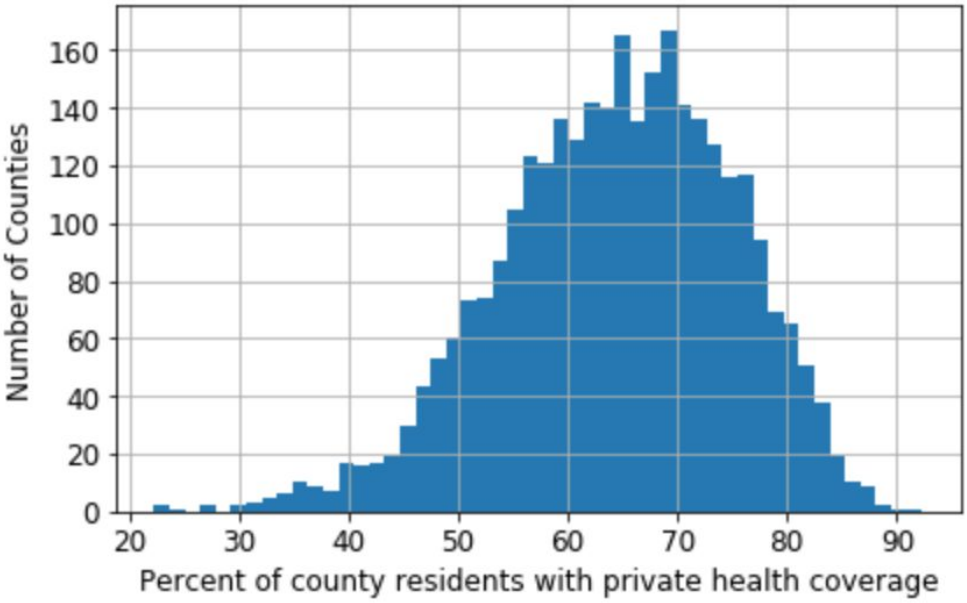


There is a moderately strong correlation of 0.38 between the percentage of county residents who were 16 years and older and unemployed and cancer mortality. This again shows the relationship between employment and cancer mortality, because as the percentage of unemployed individuals increases in a county, cancer mortality also increases.

Adding a logarithmic or exponential transformation of 'PctUnemployed16_Over' did not add to the accuracy of the overall model.

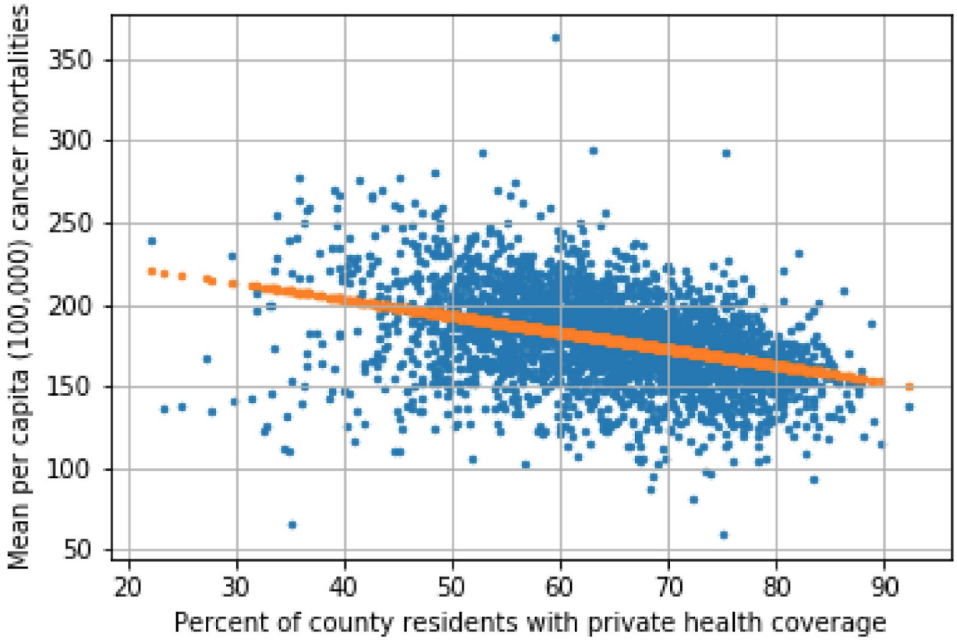
'PctPrivateCoverage': Percent of county residents with private health coverage

The percentage of county residents with private health coverage ranges from 22.3% to 92.3%, and has a mean value of 64.4% with a standard deviation of 10.6%.

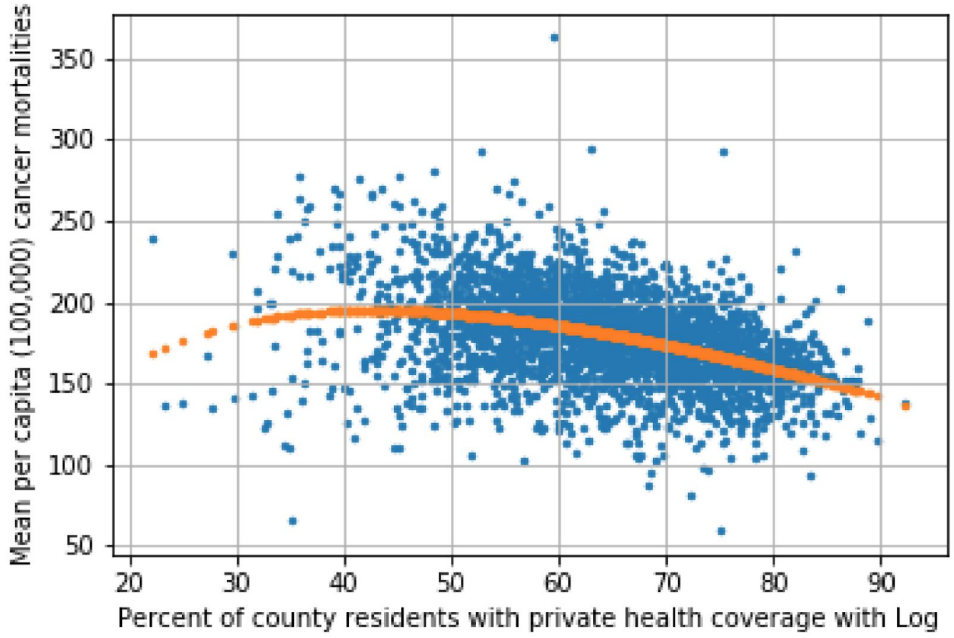


There is a moderate negative correlation of -0.39 between the percentage of county residents with private health coverage and cancer mortality. Although a causative link can't be made between the two, a county's cancer mortality rate decreases as the percentage of residents in a county with private health coverage increases.

One can see the negative correlation between the percentage of county residents with private health coverage and cancer mortality in the prediction line below.



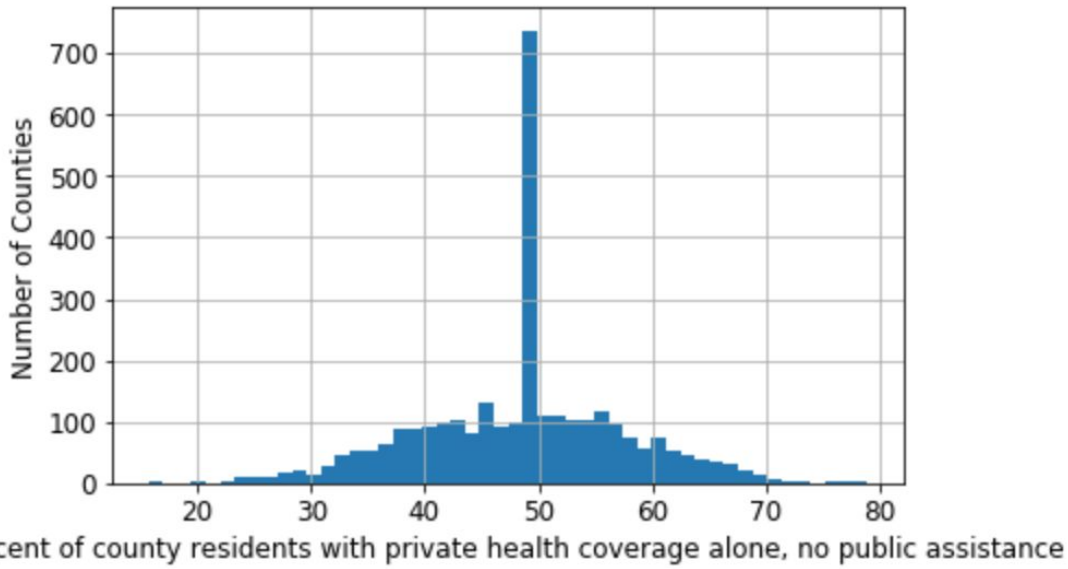
Adding the logarithmic transformation of 'PctPrivateCoverage' reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.008.



'PctPrivateCoverageAlone': Percent of county residents with private health coverage alone (no public assistance)

The percentage of county residents with private health coverage alone ranges from 15.7% to 78.9%, and has a mean value of 48.5% with a standard deviation of 9%. The high peak at the center of the

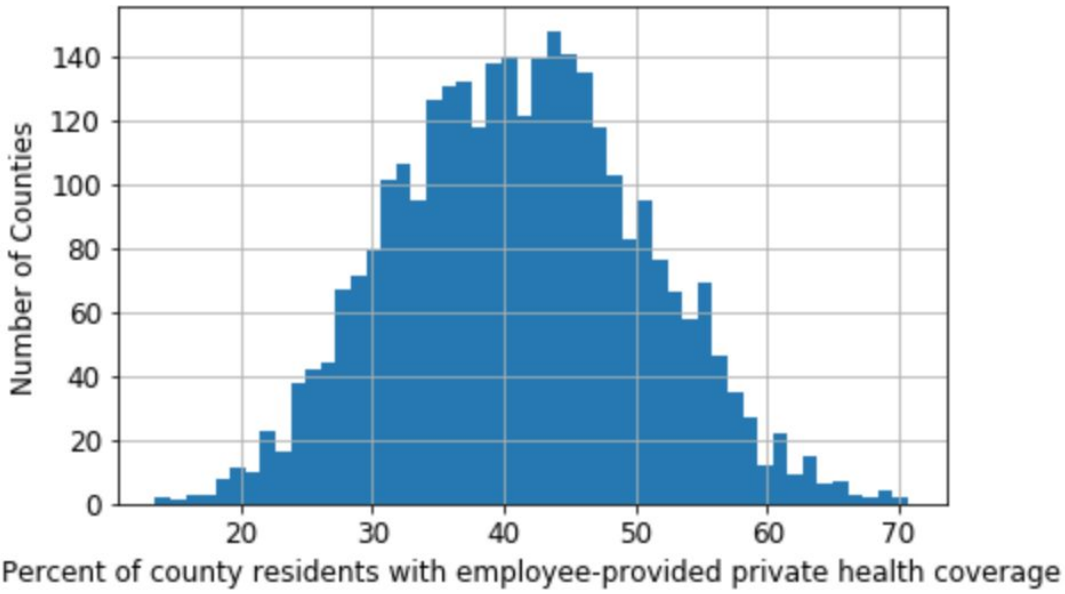
distribution comes from the fact that 20% of the counties had missing values for 'PctPrivateCoverageAlone' and were filled with the nationwide median value.



There is a moderate negative correlation of -0.33 between the percentage of county residents with only private health coverage and cancer mortality. Although a causative link can't be made between the two, a county's cancer mortality level decreases the more the percentage of residents in a county who have private health coverage alone increases. Adding a logarithmic or exponential expansion of 'PctPrivateCoverageAlone' did not add to the model's overall accuracy.

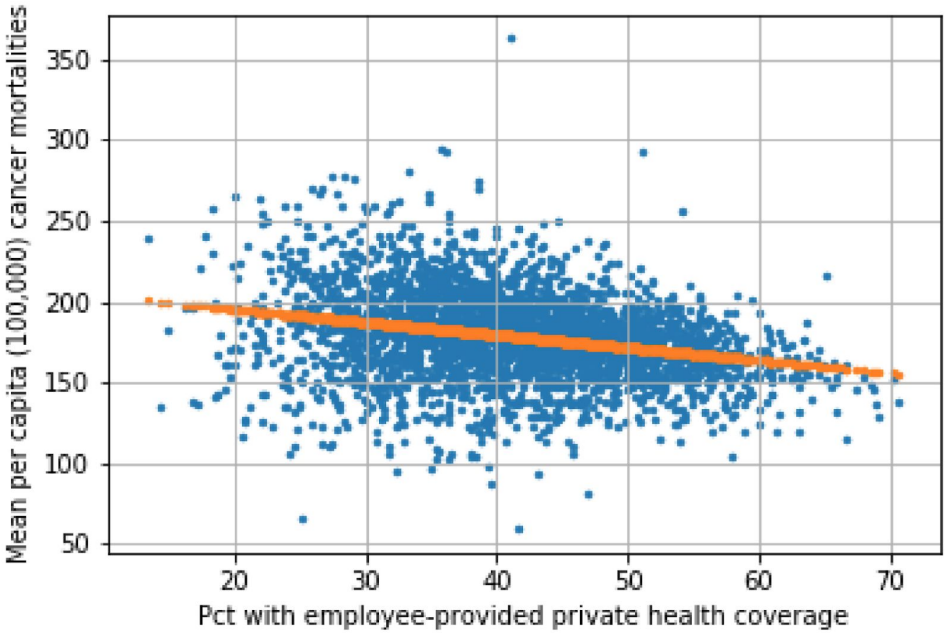
'PctEmpPrivCoverage': Percent of county residents with employee-provided private health coverage

The percentage of county residents with employee-provided private health coverage ranges from 13.5% to 70.7%, and has a mean value of 41.2% with a standard deviation of 9.5%.

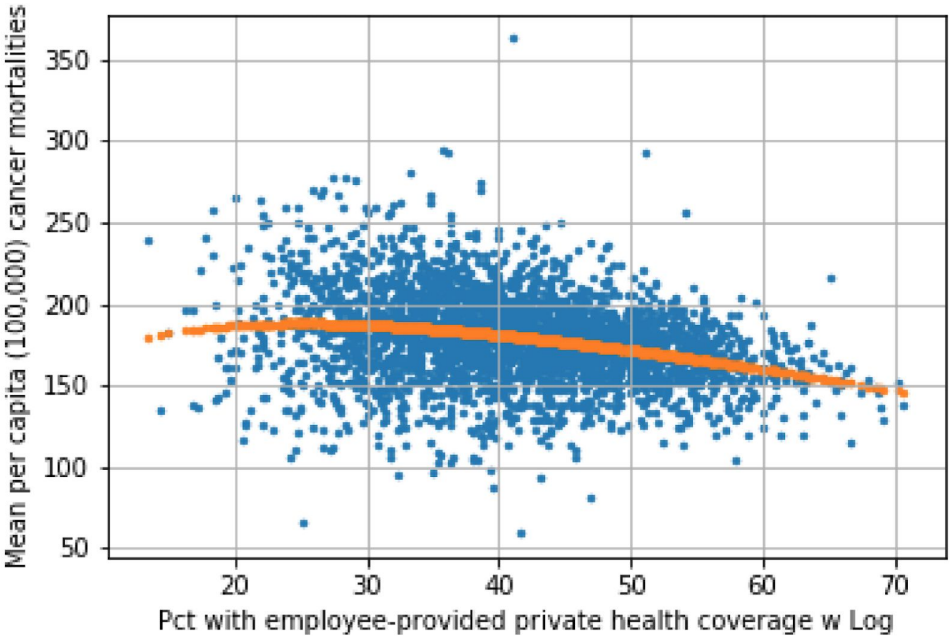


There is a moderate negative correlation of -0.27 between the percentage of county residents with employee-provided private health coverage and cancer mortality. Although a causative link can't be made between the two, a county's cancer mortality level decreases the more the percentage of residents in a county who have employee-provided private health coverage increases. Intriguingly, this negative correlation is weaker than what is seen with private health coverage and private health coverage alone, which possibly shows that employee-provided health insurance is not as effective as other forms of private health insurance. This is definitely an area for future research.

One can see the weakly negative correlation between the percentage of employee-provided private health coverage and cancer mortality in the prediction line below.

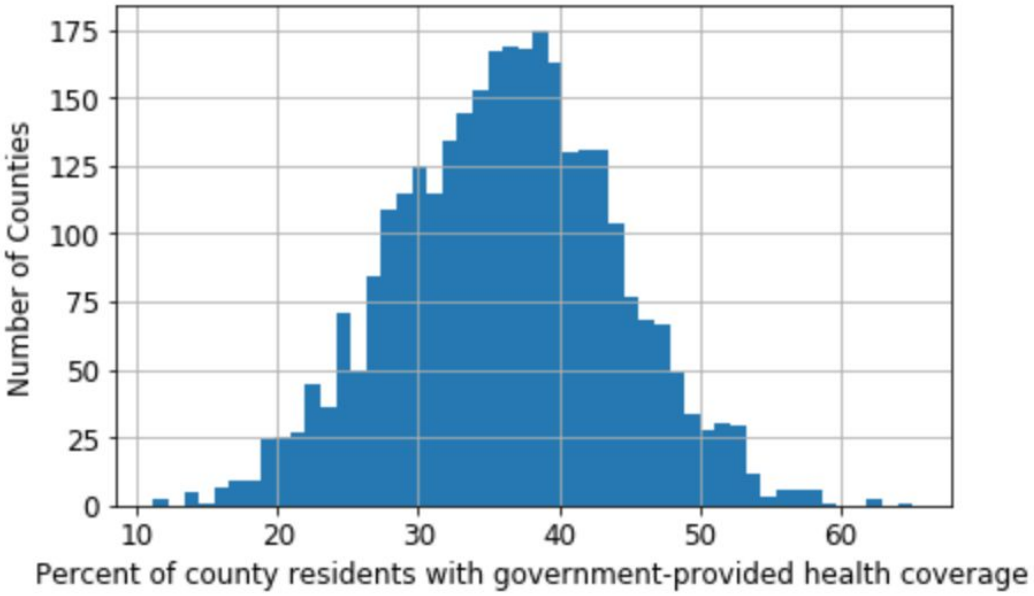


Adding the logarithmic transformation of 'PctEmpPrivCoverage' reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.0008.



'PctPublicCoverage': Percent of county residents with government-provided health coverage

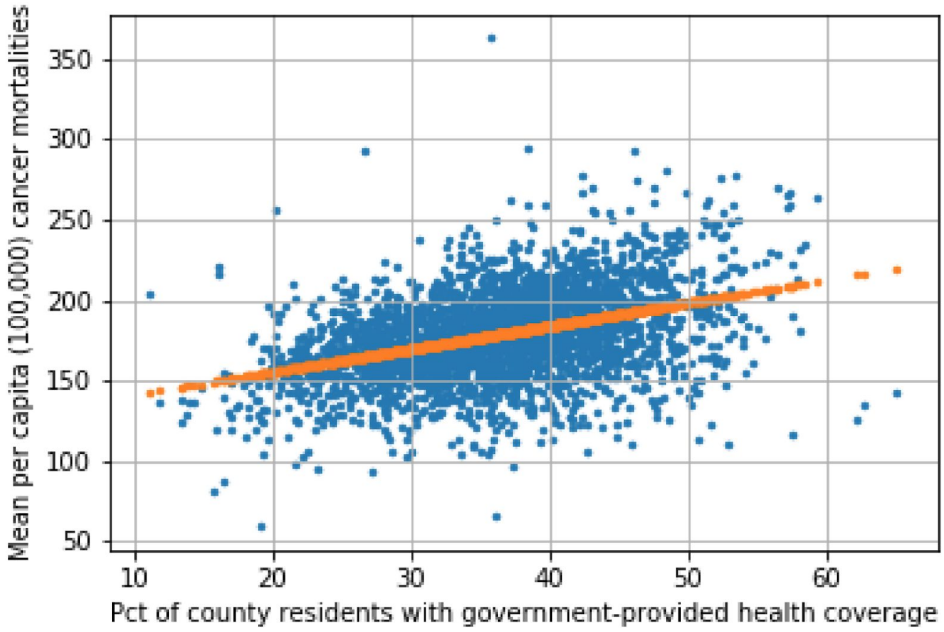
The percentage of county residents with government-provided health coverage ranges from 11.2% to 65.1%, and has a mean value of 36.3% with a standard deviation of 7.8%.



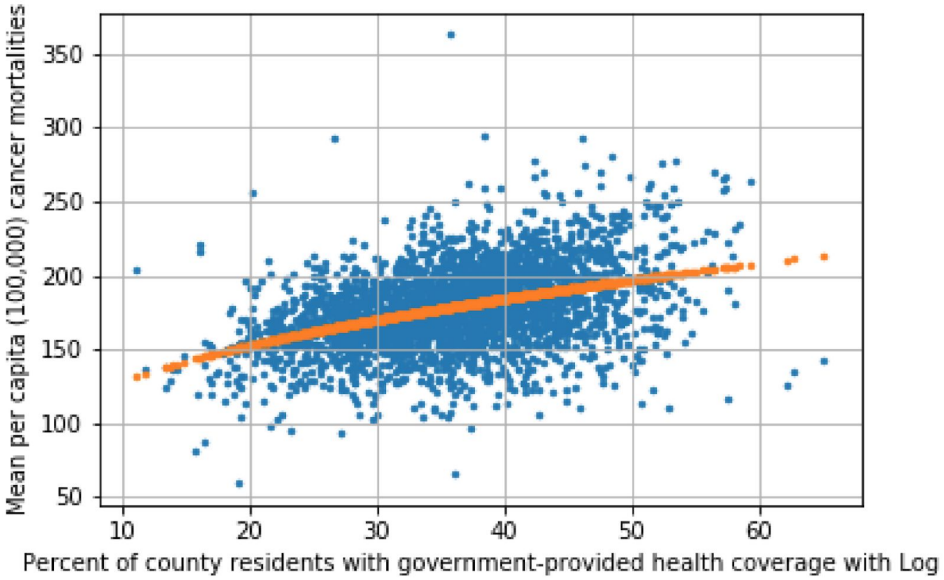
There is a moderately strong positive correlation of 0.41 between the percentage of county residents with public health coverage and cancer mortality. Therefore, as the percentage of county residents with public health coverage increases, cancer mortality increases. This is in marked opposition to the

negative correlation between the percentage of county residents with private health coverage and cancer mortality. This is an important finding that bolsters the need for further research into this connection and for better public health insurance generally.

One can see the moderately positive correlation between the percentage of county residents with public health coverage and cancer mortality in the prediction line of the first of the plots below.

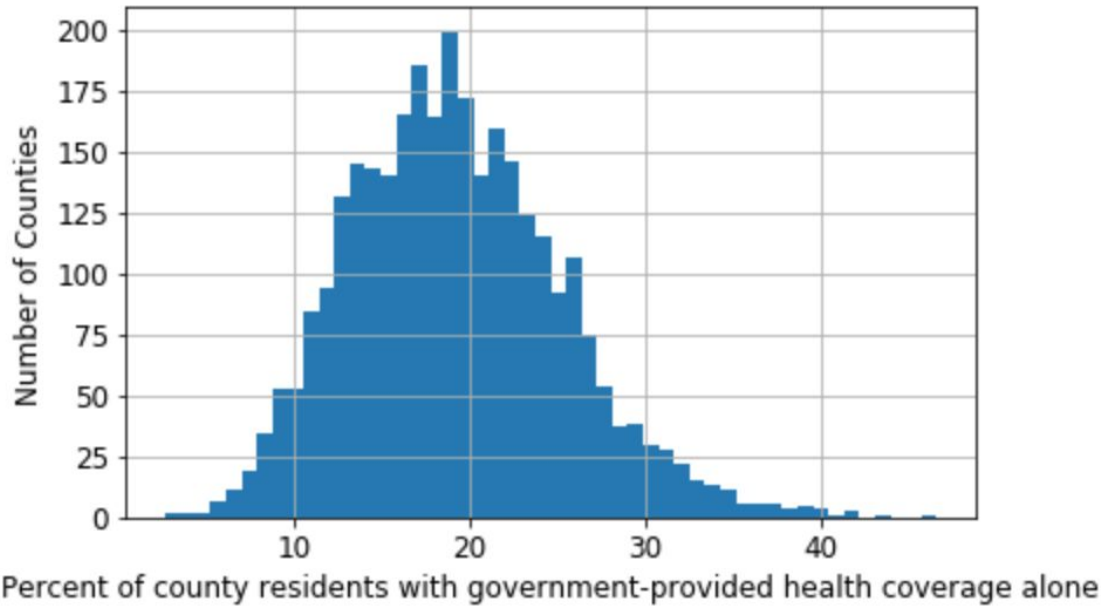


The error between the actual and predicted data points of cancer mortality is reduced when the logarithmic transformation of 'PctPublicCoverage' is added, and doing so increased the model's accuracy by 0.0002.



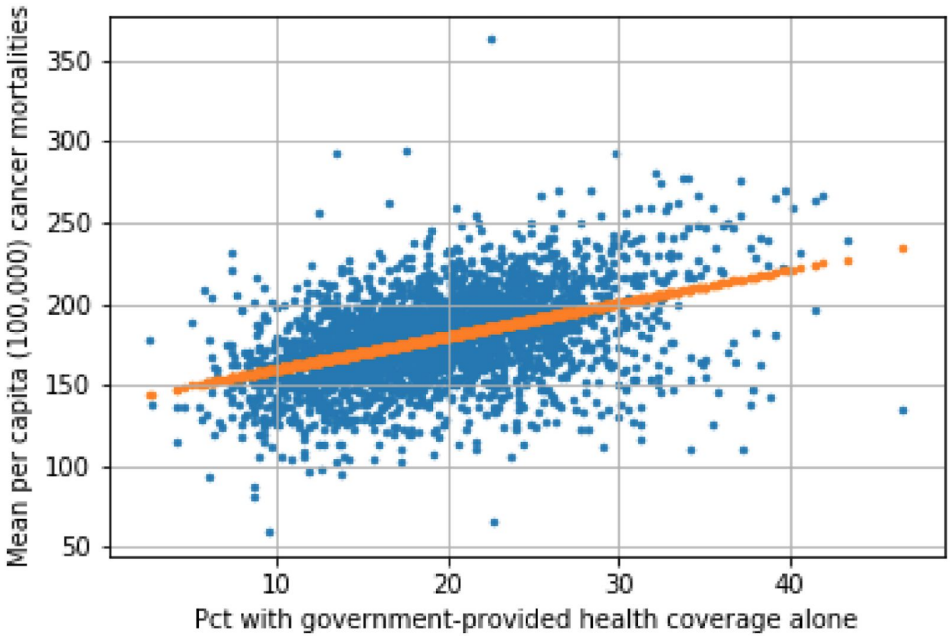
'PctPublicCoverageAlone': Percent of county residents with government-provided health coverage alone

The percentage of county residents with government-provided health coverage alone ranges from 2.6% to 46.6%, and has a mean value of 19.2% with a standard deviation of 6.1%.

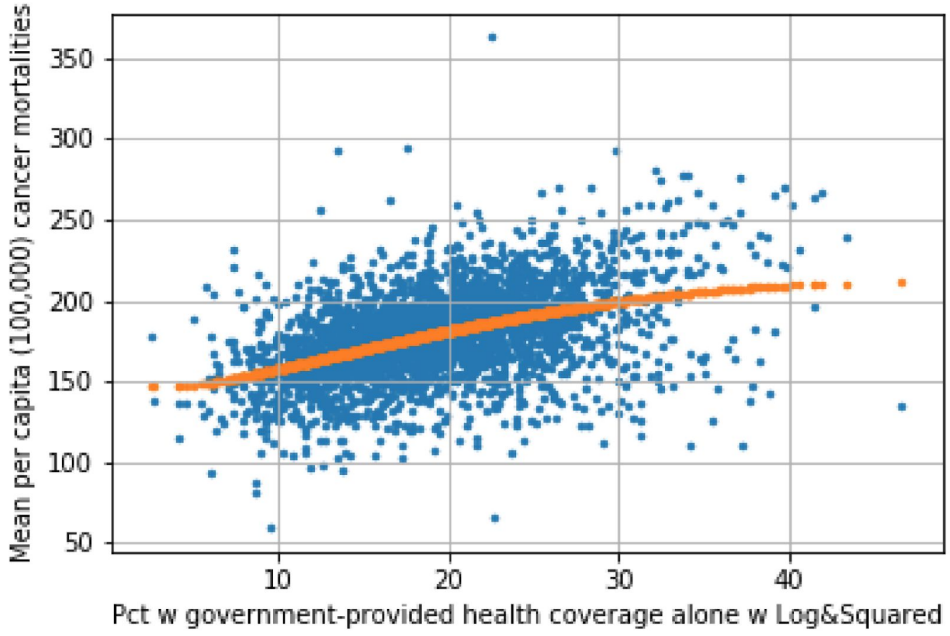


The correlation between the percentage of county residents with public health coverage ALONE and cancer mortality is slightly higher than the correlation between the percentage of county residents with public health coverage and cancer mortality (0.45 vs. 0.41). This again shows that more research and action needs to be enacted in this area to improve public health insurance.

One can see the positive correlation between the percentage of county residents with public health coverage alone and cancer mortality in the prediction line below.

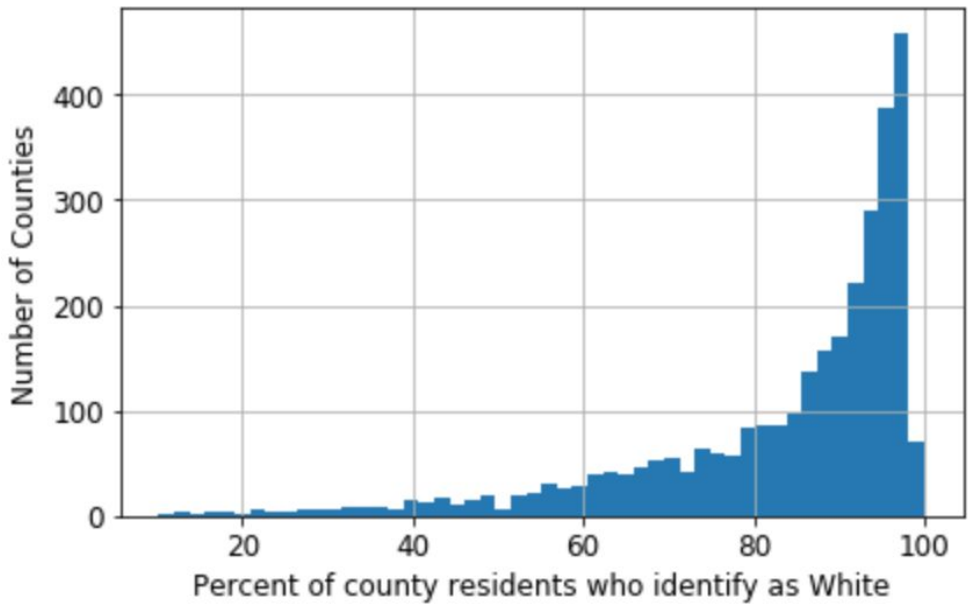


Adding the logarithmic and exponential transformations of the percentage of county residents with public health coverage alone reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.003.



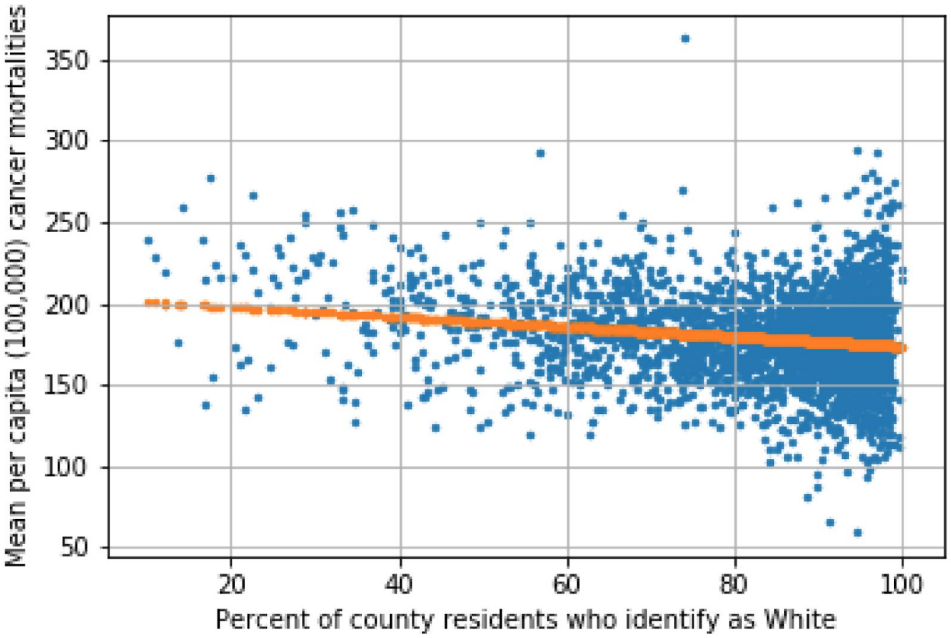
'PctWhite': Percent of county residents who identify as White

The percentage of county residents who identify as White ranges from 10.2% to 100%, and has a mean value of 83.7% with a standard deviation of 16.4%.

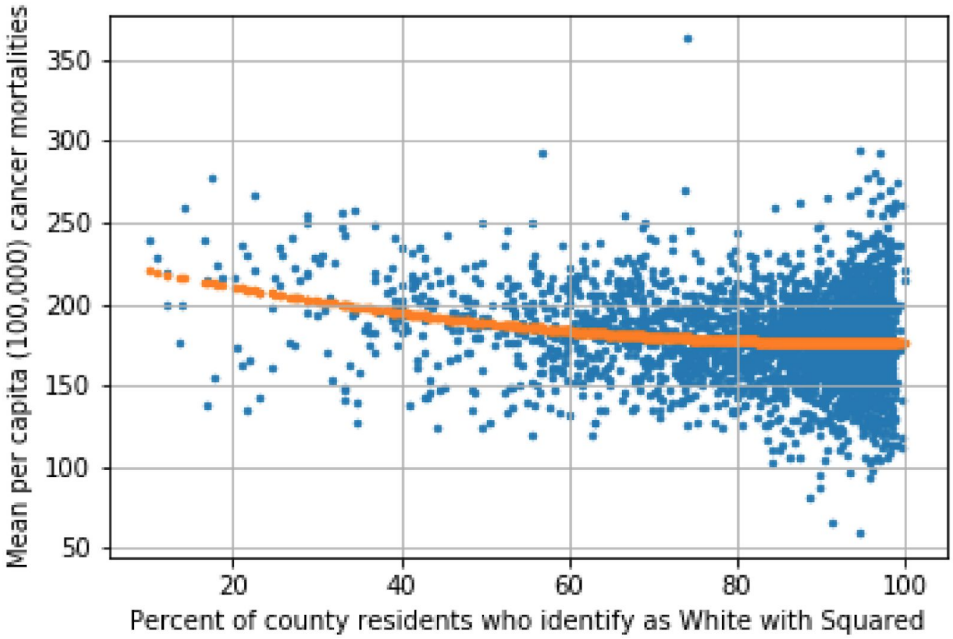


There is a weak negative correlation of -0.18 between the percentage of county residents who identify as white and cancer mortality. This shows that as the percentage of county residents who identify as "White" increases, cancer mortality slightly decreases.

One can see the weakly negative correlation between the percentage of county residents who identify as white and cancer mortality in the prediction line of the first of the plots below.

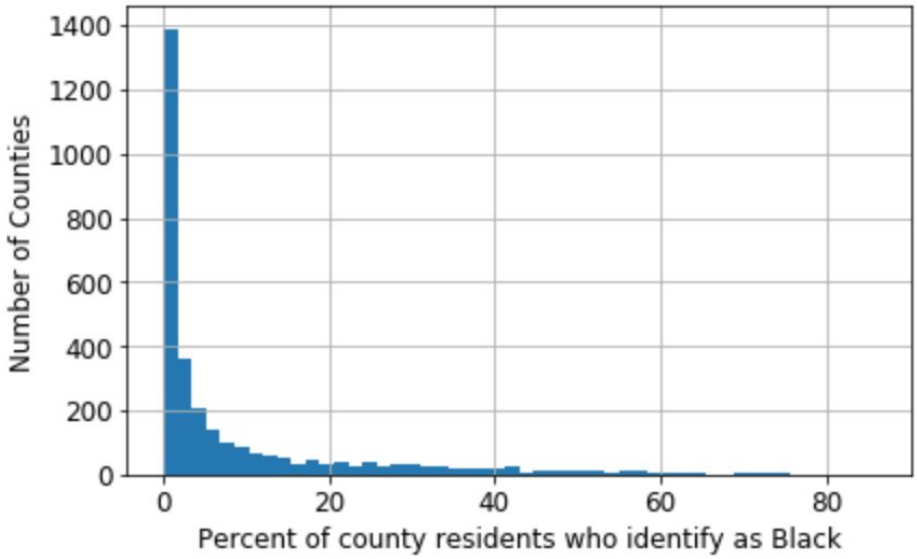


The error between the actual and predicted data points of cancer mortality is reduced when the exponential transformations of 'PctWhite' are added, and doing so increased the model's accuracy by 0.0003.



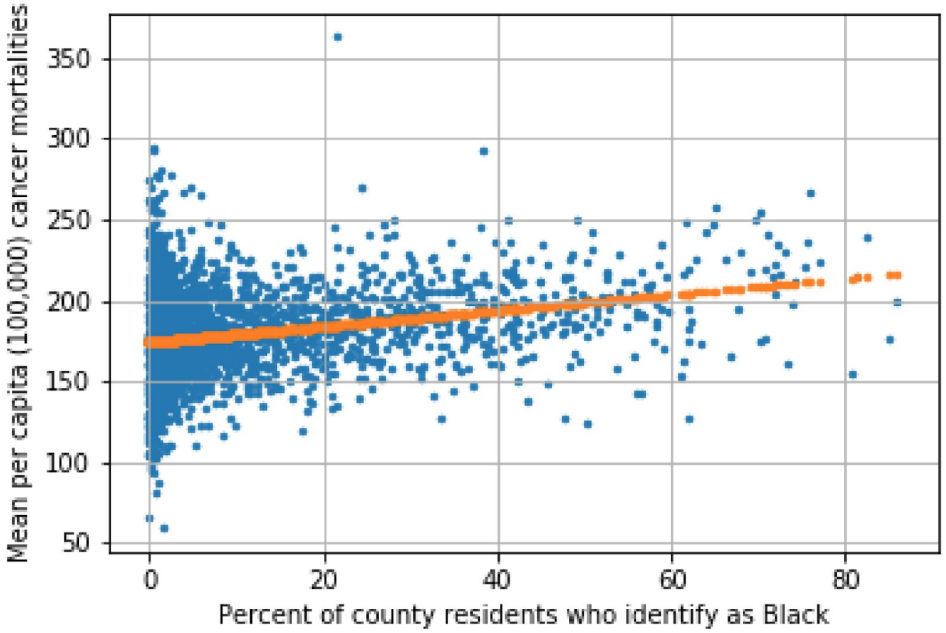
'PctBlack': Percent of county residents who identify as Black

The percent of county residents who identify as Black ranges from zero to 86%, and has a mean value of 9.1% with a standard deviation of 14.5%.

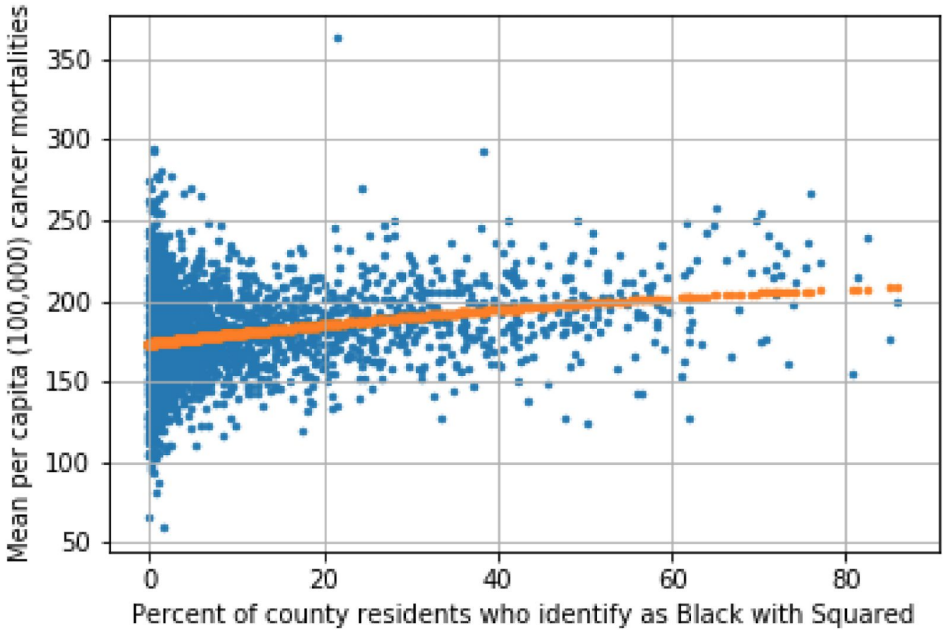


There is a moderate positive correlation of 0.26 between the percentage of county residents who identify as Black and cancer mortality. This shows that as the percentage of county residents who identify as "Black" increases, cancer mortality increases.

One can see the moderately positive correlation between the percentage of county residents who identify as Black and cancer mortality in the prediction line of the first of the plots below.

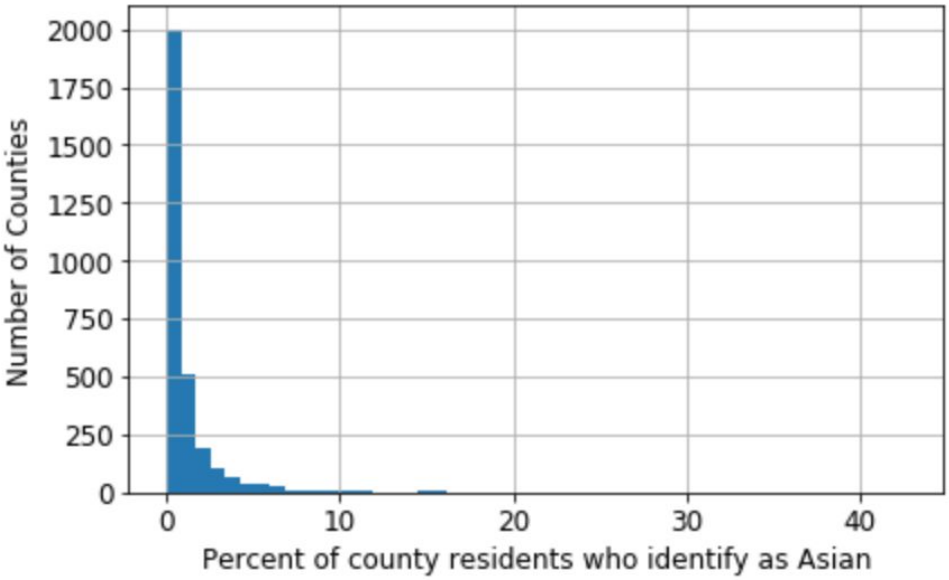


The error between the actual and predicted data points of cancer mortality is reduced when the exponential transformations of 'PctBlack' are added, and doing so increased the model's accuracy by 0.004.



'PctAsian': Percent of county residents who identify as Asian

The percent of county residents who identify as Asian ranges from zero to 43%, and has a mean value of 1.3% with a standard deviation of 2.6%.

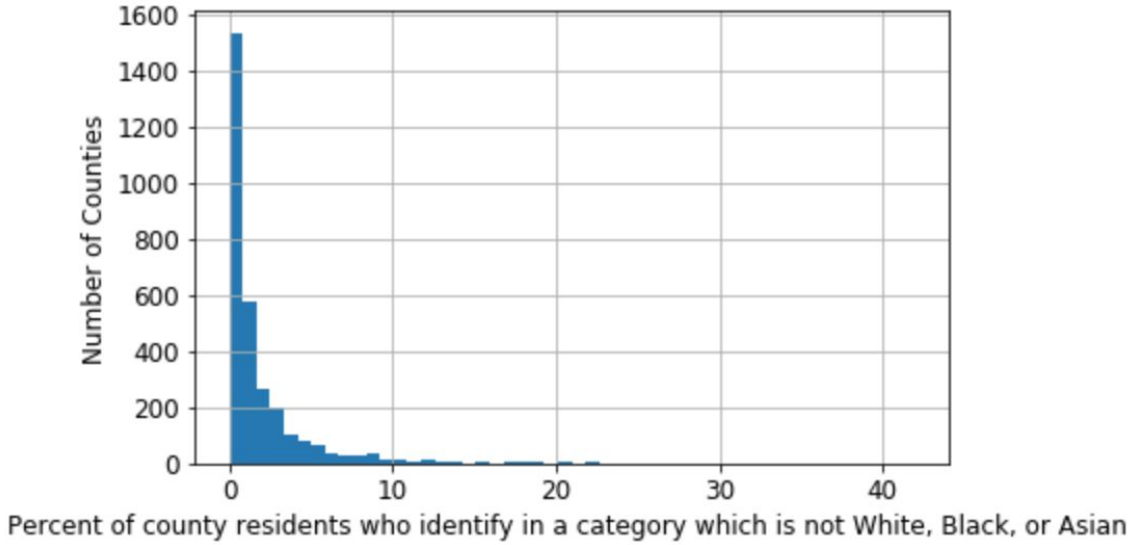


There is a moderate negative correlation of -0.19 between the percentage of county residents who identify as Asian and cancer mortality. This shows that as the percentage of county residents who identify as Asian increases, cancer mortality decreases.

Adding a logarithmic or exponential expansion of 'PctAsian' did not add to the model's overall accuracy.

'PctOtherRace': Percent of county residents who identify in a category which is not White, Black, or Asian

The percent of county residents who identify as a race other than white, black or Asian ranges from zero to 42%, and has a mean value of 2% with a standard deviation of 3.5%.



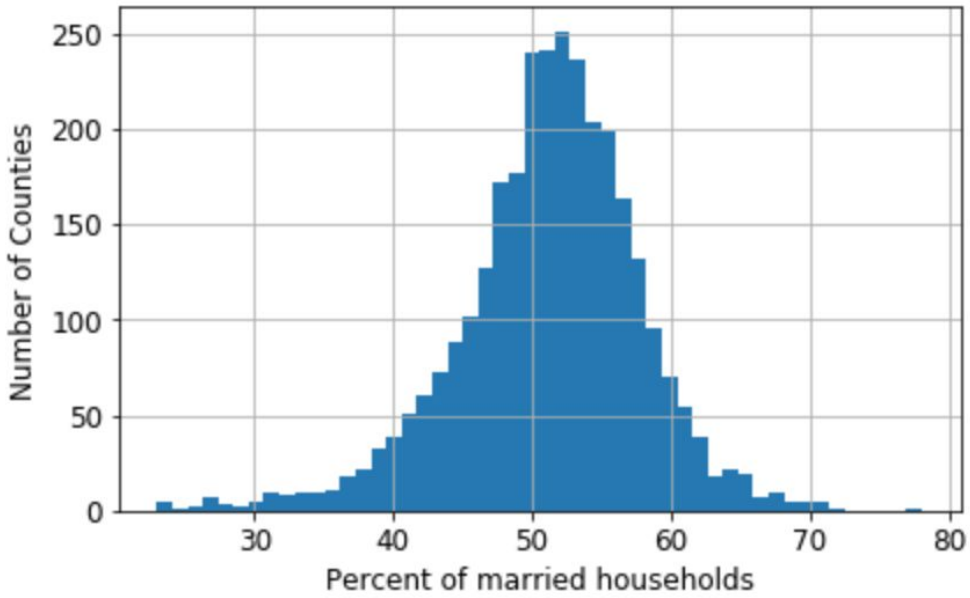
There is a moderate negative correlation of -0.19 between the percentage of county residents who identify as a race other than White, Black or Asian and cancer mortality. This shows that as the

percentage of county residents who identify as a race other than White, Black or Asian increases, cancer mortality decreases.

Adding a logarithmic or exponential expansion of 'PctOtherRace' did not add to the model's overall accuracy.

'PctMarriedHouseholds': Percent of married households

The percentage of households in a county that have married residents (versus simply the percentage of county residents who are married) ranges from 22.3% to 78.1%, and has a mean value of 51.2% with a standard deviation of 6.6%.

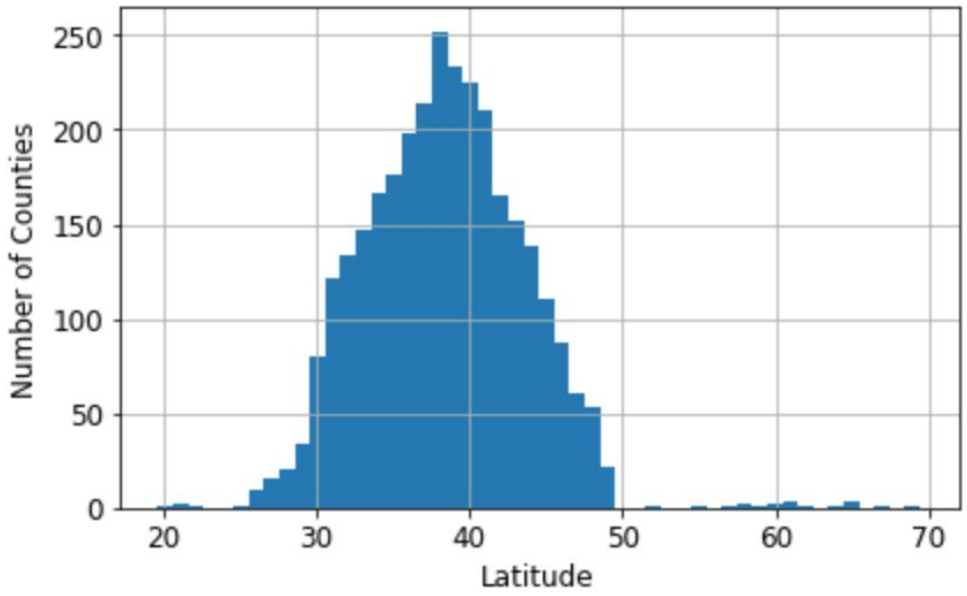


There is a moderate negative correlation of -0.29 between the percentage of married households and cancer mortality, showing that cancer mortality decreases as the percentage of married households in a county increases. Adding a logarithmic or exponential expansion of 'PctMarriedHouseholds' did not add to the model's overall accuracy.

'INTPTLAT': Latitude

The United States' latitude ranges from 19.6 to 70, from Hawaii to Alaska.

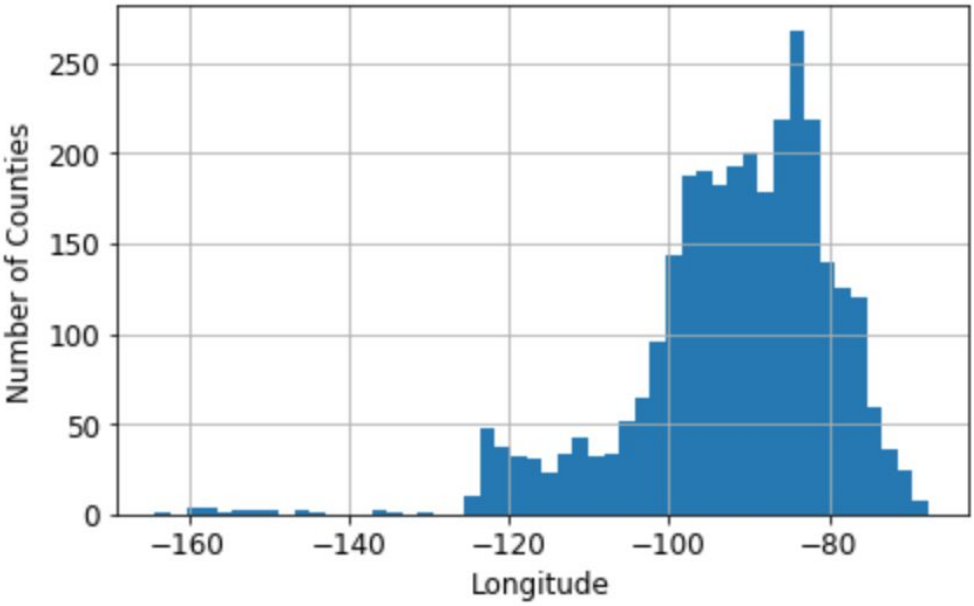
The distribution of counties across the range of latitudes is interesting as a measure of population. One can see that the measure of central tendency and greatest number of counties is found at a latitude of 38 (the north/south middle of the country), that the concentration of counties decreases the further away from the latitude of 38, and that Hawaii and Alaska occur as outliers on each tail of the distribution.



Latitude has a weakly negative correlation of -0.18 with cancer mortality, suggesting that generally the further north a county is the lower its cancer mortality rate is. Adding a logarithmic or exponential expansion of 'INTPTLAT' did not add to the model's overall accuracy.

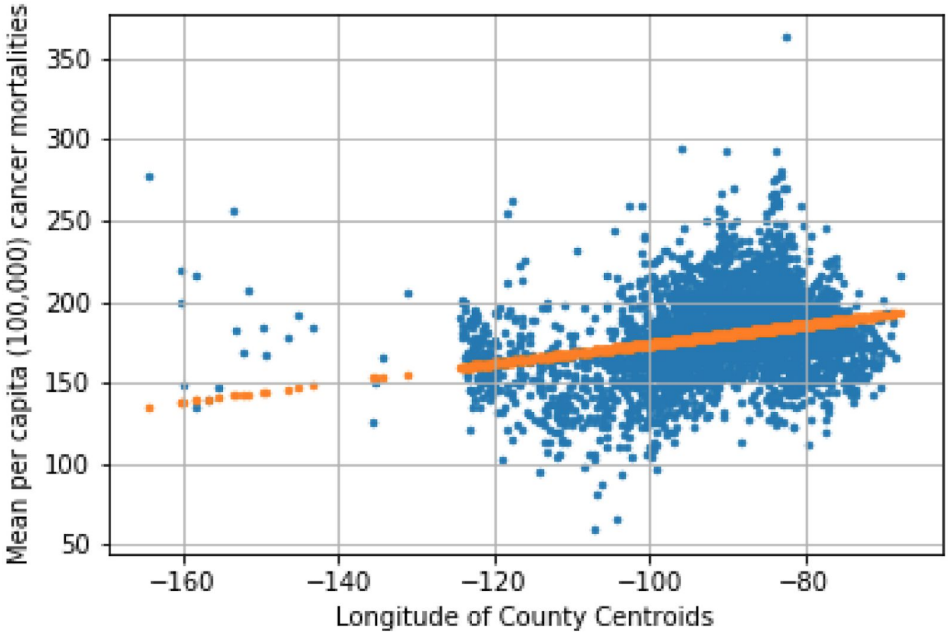
'INTPTLONG': Longitude

The United States' longitude ranges from -164.2 to -67.6. One can see that the greatest concentration of counties exists at the longitude of -91.9, in the Midwest.

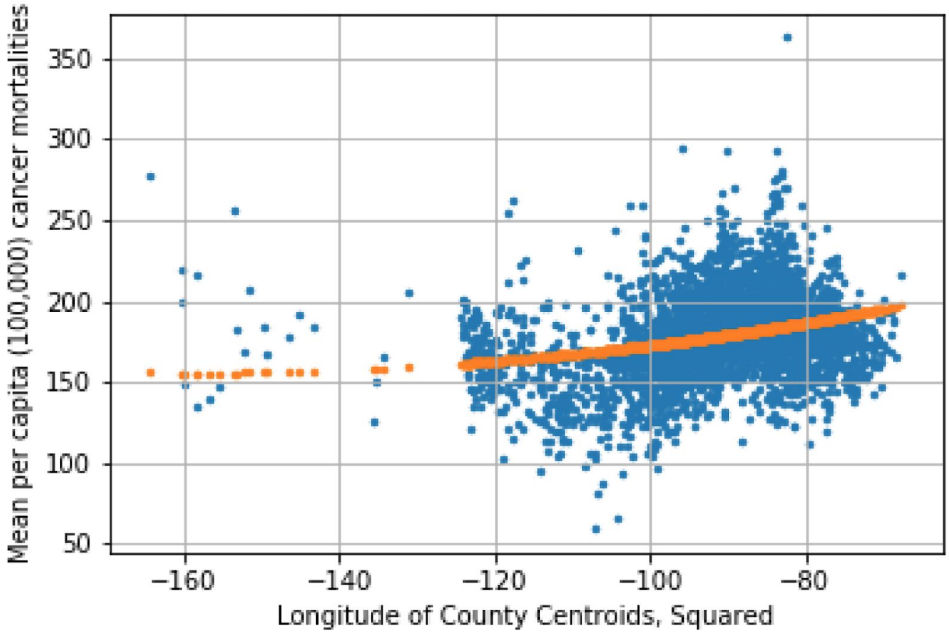


Longitude has a weakly positive correlation of 0.27 with cancer mortality, suggesting that generally the further east a county is the higher its cancer mortality rate is.

One can see the weakly positive correlation between longitude and cancer mortality in the prediction line below.



Adding the squared transformation of longitude reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.0005.



Individual L1 and L2 distance features and their contribution to the OLS linear regression model

There are a total of 36 individual features for the L1 and L2 distances to the top 10 oncology hospitals and eight urban centers. Running histograms of these features will simply repeat the

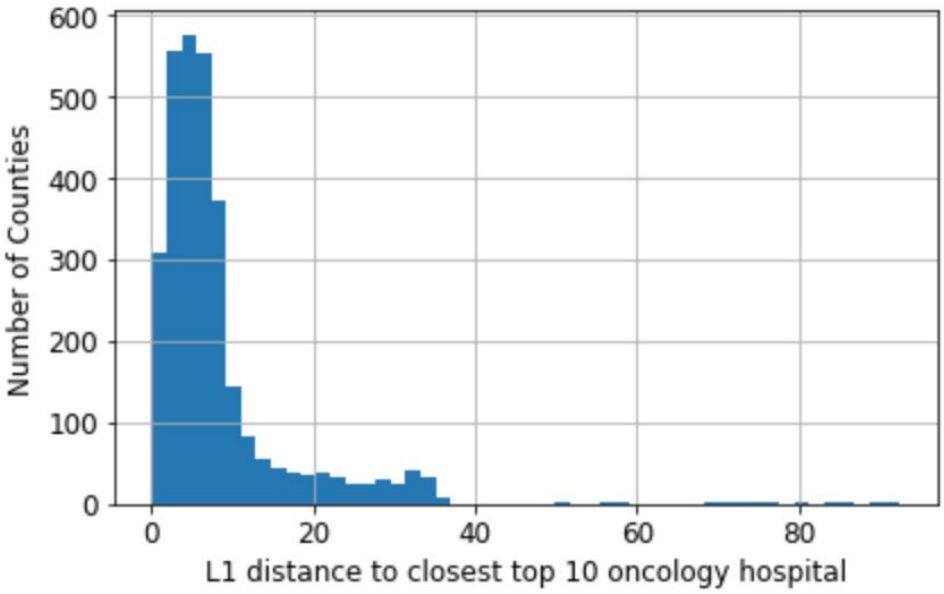
geographic layout of the country which is explored elsewhere, so they will not be created for the individual distance features. The scatterplots of the logarithmic and exponential transformations of these features is also not very interesting, so they will not be created here.

However, 16 of these 36 features did improve the OLS linear regression model's accuracy when their logarithmic and/or exponential transformations were added to the core feature set. These features' nonlinear contributions to the OLS linear regression model are detailed in Appendix B: Nonlinear Contributions of L1 & L2 Distance Features to Linear Regression Accuracy.

Correlations between each of the individual distance features and the target variable are run to see if any correlations stand out. Although some of these features have up to a 0.34 correlation with the target variable, most of them have between a .1 and .2 correlation. Because the hospitals and urban centers are spread throughout the country (as are all the country centroids), it is difficult to ascribe any meaning to these correlations, as a county may be far from one hospital or urban center but close to another. At any rate, none of these correlations truly stand out, and the L1 and L2 distances to the closest top 10 oncology hospital and urban center are covered later in the notebook. For reference, the correlations between these individual distance features and cancer mortality are detailed in Appendix C: Correlations Between L1 and L2 Distance Features and Cancer Mortality.

'onc_min_distsl1': L1 distance to closest top 10 oncology hospital

The L1 distance to the closest top 10 oncology hospital for each county ranges from 0.019 to 92.48 latitude/longitude units, and has a mean value of 8.28 with a standard deviation of 9.18. The counties with the shortest L1 distance to a top 10 oncology hospital were in New York, Maryland, Illinois, Pennsylvania and Massachusetts. The counties with the longest distance are all in Alaska.

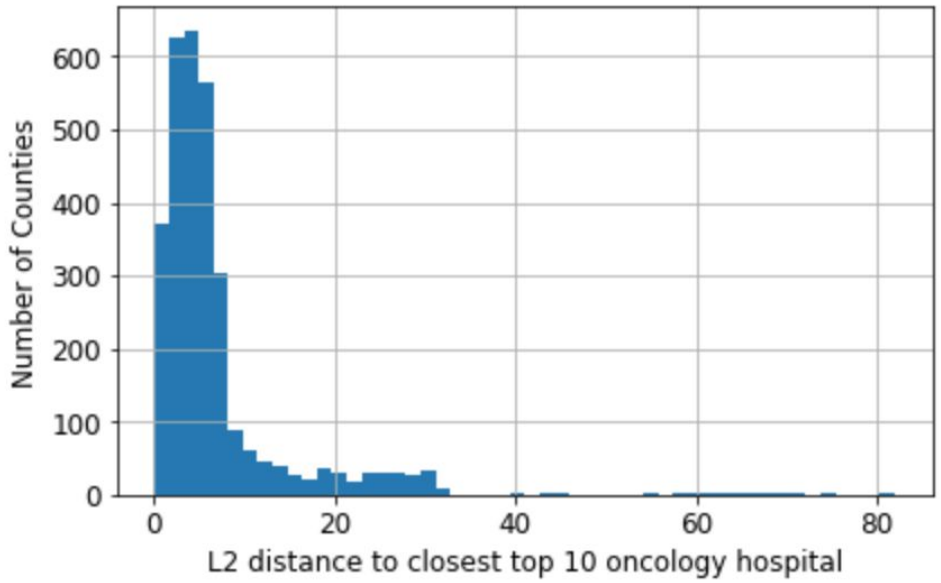


There is a weakly negative correlation of -0.16 between the L1 distance to the closest top 10 oncology hospital and cancer mortality, suggesting that there is a small decrease in cancer mortality associated with a county being closer to one of these hospitals.

Adding a logarithmic or exponential expansion of 'onc_min_distsl1' did not add to the model's overall accuracy.

'onc_min_distsl2': L2 distance to closest top 10 oncology hospital

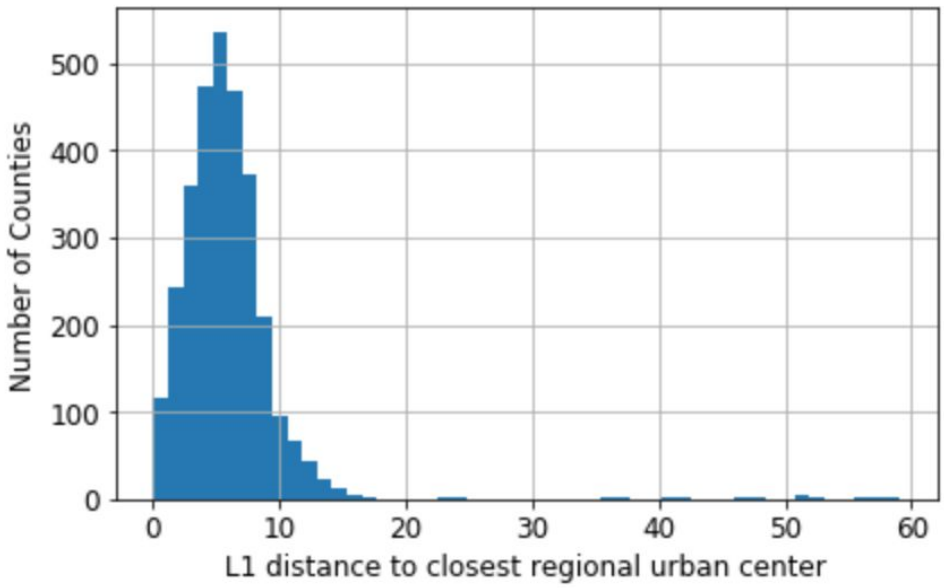
The L2 distance to the closest top 10 oncology hospital for each county ranges from 0.01 to 82 latitude/longitude units, and has a mean value of 6.78 with a standard deviation of 7.86.



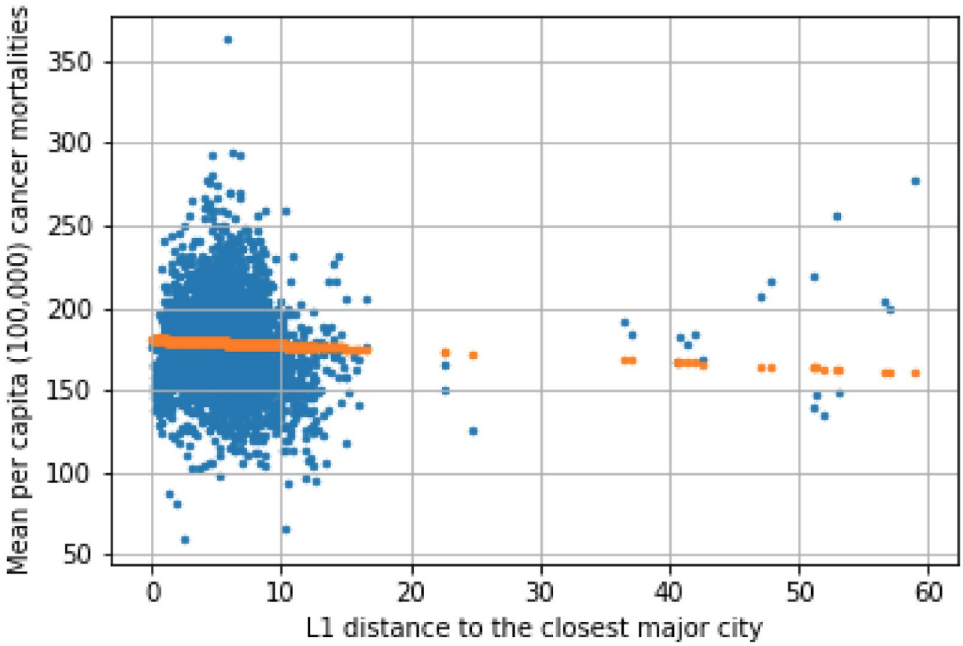
There is a weakly negative correlation of -0.17 between the L2 distance to the closest top 10 oncology hospital and cancer mortality, suggesting that there is a small decrease in cancer mortality associated with a county being closer to one of these hospitals. Adding a logarithmic or exponential expansion of 'onc_min_distsl2' did not add to the model's overall accuracy.

'city_min_distsl1': L1 distance to closest regional urban center

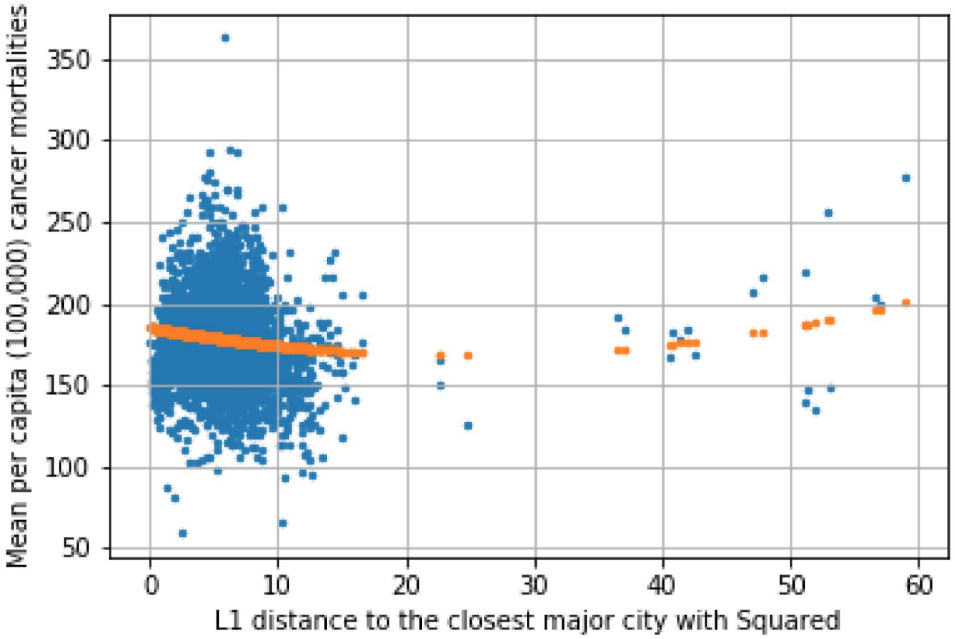
The shortest L1 distance to the closest of eight regional urban centers ranges from 0.03 to 59.03 latitude/longitude units, and has a mean value of 5.9 with a standard deviation of 4.3.



There is a very weakly negative correlation of -0.06 between the shortest L1 distance of a county to a regional urban center and cancer mortality. One can see the weakly negative correlation between the shortest L1 distance to a regional urban center and cancer mortality in the prediction line.

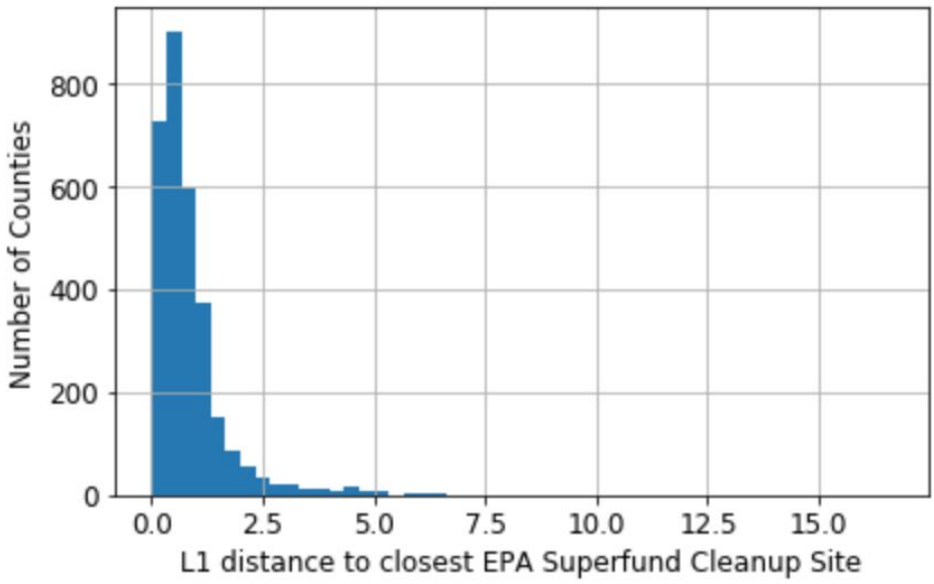


Adding the squared transformation of the shortest L1 distance to a regional urban center reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.001.



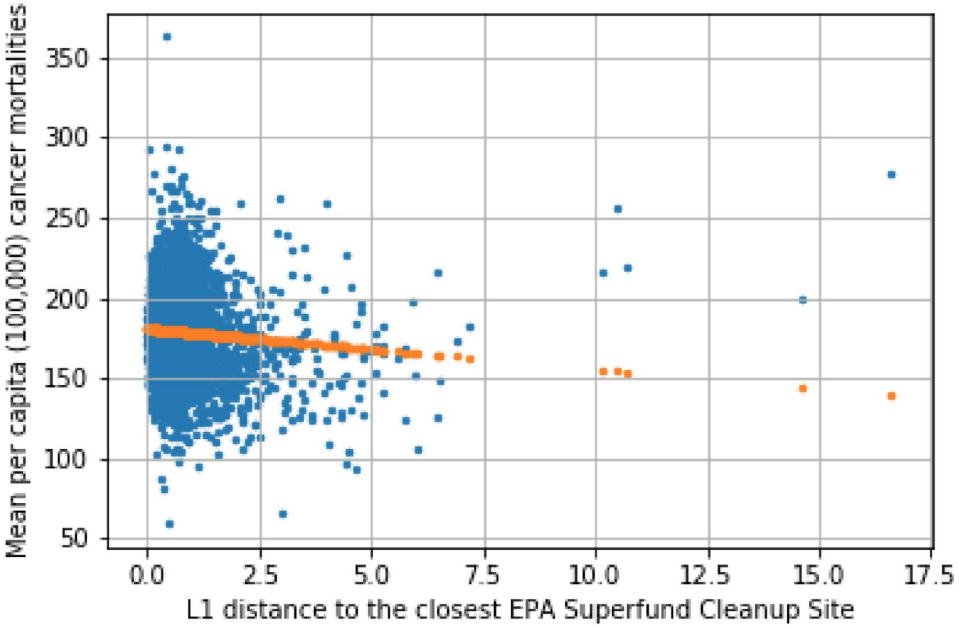
'sc_min_dists_l1': L1 distance to closest EPA Superfund Cleanup Site

The L1 distance to the closest EPA Superfund Cleanup site ranges from 0.004 to 16.62 latitude/longitude units, and has a mean value of 0.86 with a standard deviation of 0.96.

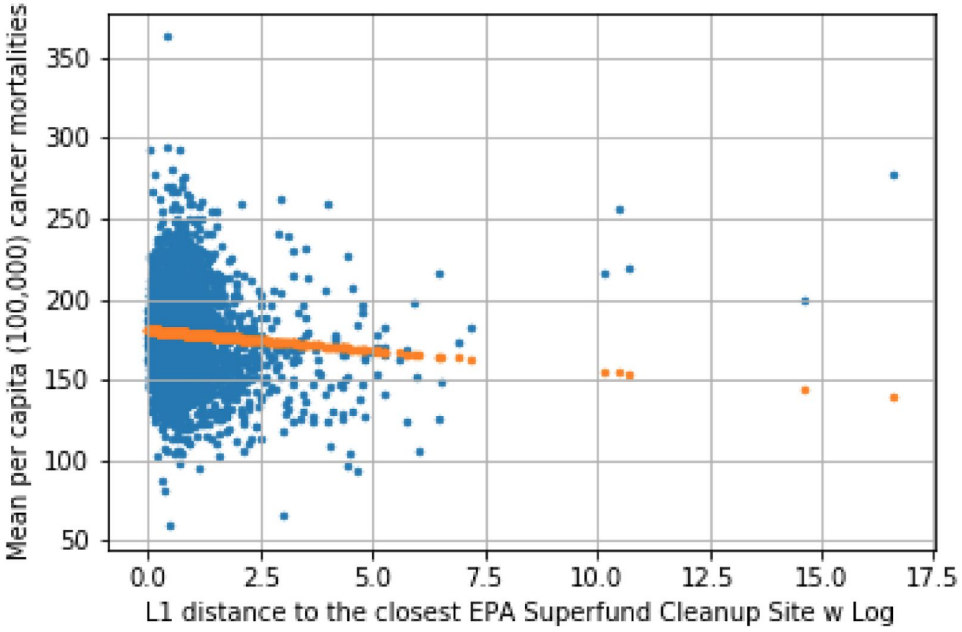


There is a very weakly negative correlation of -0.09 between the shortest L1 distance of a county to an EPA Superfund Cleanup site and cancer mortality.

One can see the very weak negative correlation between the shortest L1 distance to an EPA Superfund Cleanup Site and cancer mortality in the prediction line.

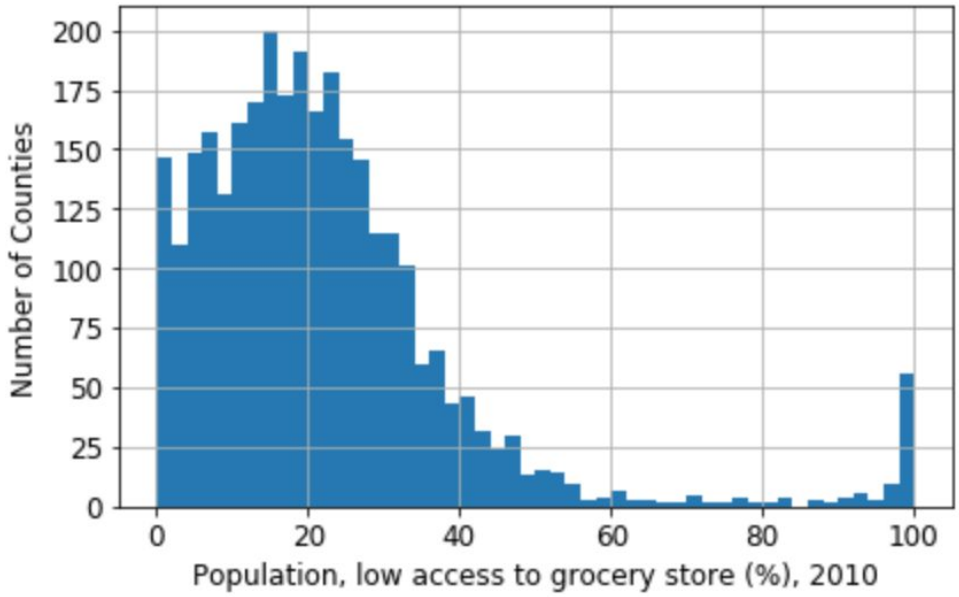


Adding the logarithmic transformation of the shortest L1 distance to an EPA Superfund Cleanup Site reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.0003.



'PCT_LACCESS_POP10': Population, low access to grocery store (%), 2010

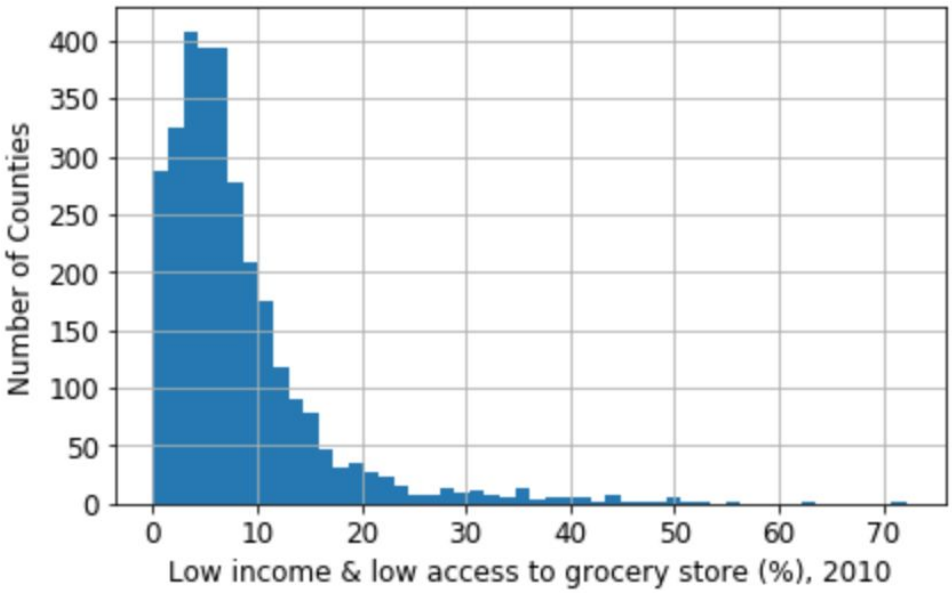
The county population percentage with low access to a grocery store (in 2010) ranges from zero to 100%, and has a mean value of 22.5% with a standard deviation of 18.22%.



There is a moderate negative correlation of -0.22 between the county population percentage with low access to a grocery store (in 2010) and cancer mortality. Therefore, as this percentage increases, cancer mortality decreases slightly. This is a surprising finding that warrants further research. Adding a logarithmic or exponential transformation of 'PCT_LACCESS_POP10' did not add to the overall accuracy of the OLS linear regression model.

'PCT_LACCESS_LOWI10': Low income & low access to grocery store (%), 2010

The county population percentage with low income and low access to a grocery store (in 2010) ranges from zero to 72.3%, and has a mean value of 8% with a standard deviation of 7.5%.



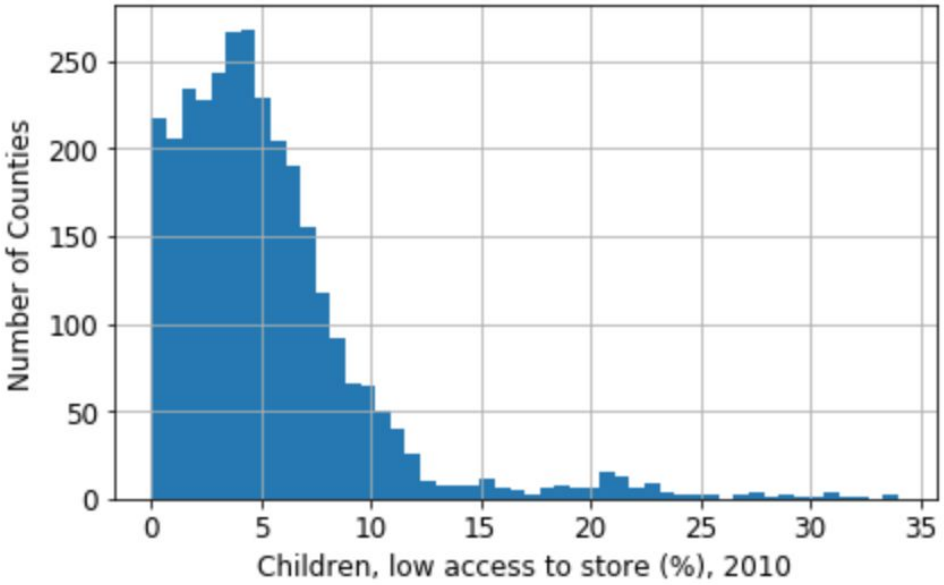
There is a weak negative correlation of -0.10 between the county population percentage with low income and low access to a grocery store (in 2010) and cancer mortality. Therefore, as this

percentage increases, cancer mortality very slightly decreases. This is a somewhat surprising finding, but less so than the 'PCT_LACCESS_10' correlation.

Adding a logarithmic or exponential expansion of 'PCT_LACCESS_LOWI10' did not add to the model's overall accuracy.

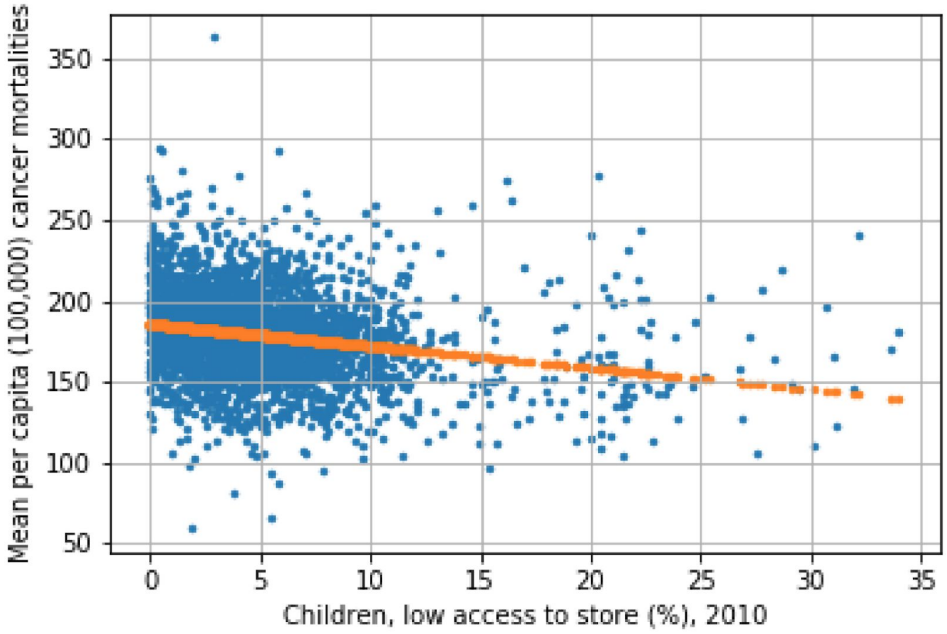
'PCT_LACCESS_CHILD10': Children, low access to store (%), 2010

The percentage of children with low access to a grocery store in 2010 ranges from zero to 34.02 latitude/longitude units, and has a mean value of 5.26 with a standard deviation of 4.44.

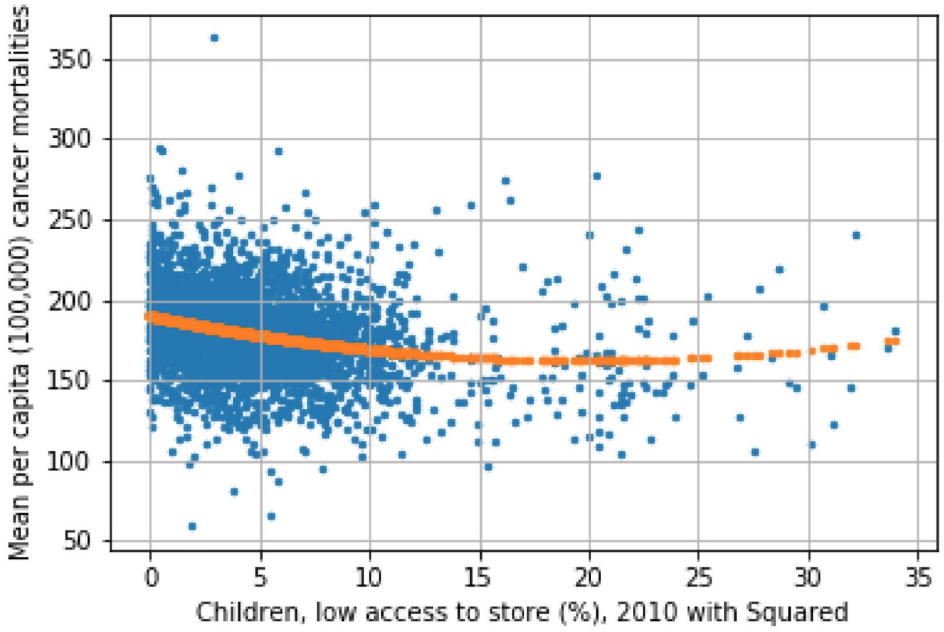


There is a weak negative correlation of -0.22 between the percentage of children with low access to a grocery store in 2010 and cancer mortality. This means that as this percentage increases in a county, cancer mortality slightly falls. This is a somewhat confusing correlation as other indicators of poverty or food deprivation are associated with a higher cancer mortality rate, so this correlation points towards the need for more research.

One can see the negative correlation between the percentage of county residents married and cancer mortality in the prediction line of the first of the plots below.

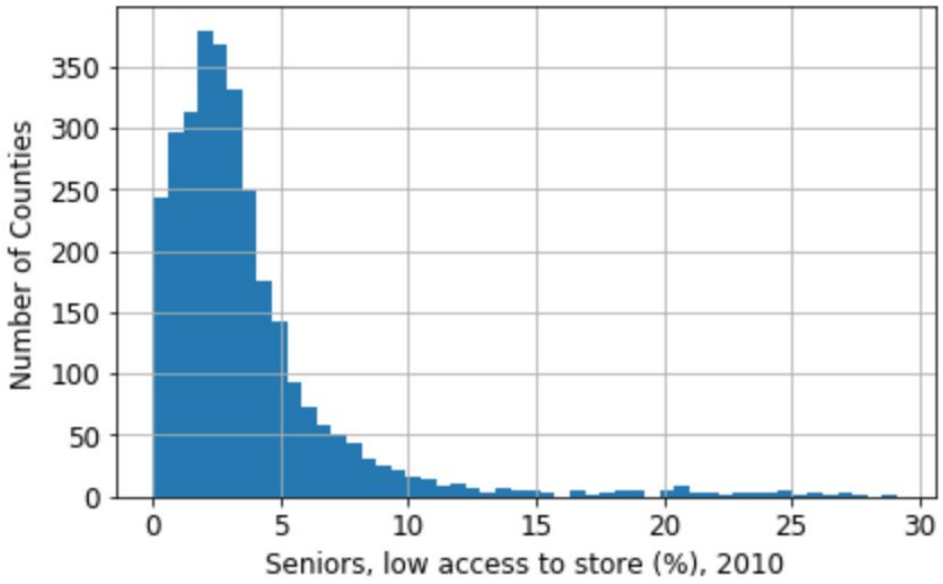


The error between the actual and predicted data points of cancer mortality is reduced when the exponential transformation of 'PCT_LACCESS_CHILD10' was added, and doing so increased the model's accuracy by 0.00009.



'PCT_LACCESS_SENIORS10': Seniors, low access to store (%), 2010

The county population percentage of seniors with low access to a grocery store (in 2010) ranges from zero to 29.2%, and has a mean value of 3.7% with a standard deviation of 3.8%.

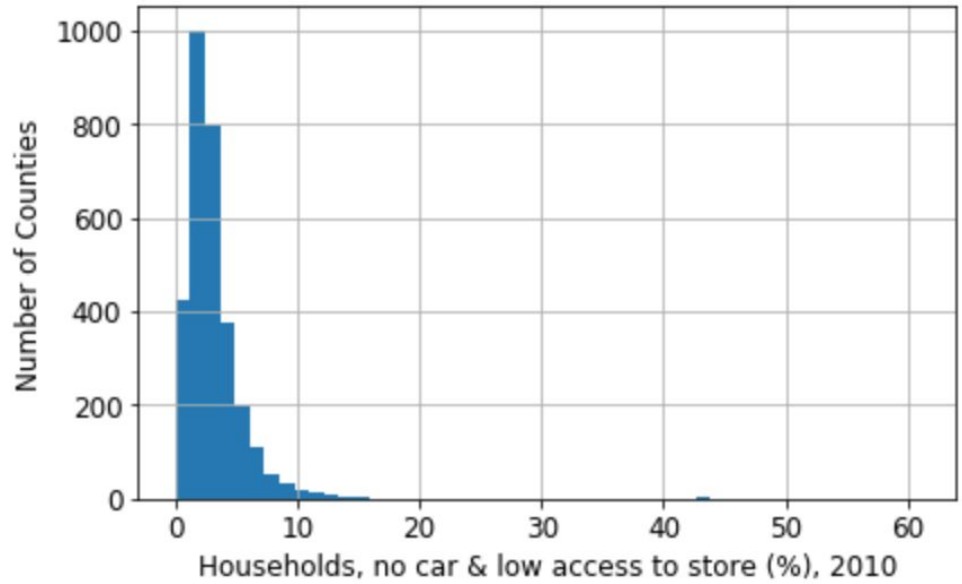


There is a weak negative correlation of -0.21 between the county population percentage of seniors with low access to a grocery store (in 2010) and cancer mortality. Therefore, as this percentage increases, cancer mortality very slightly decreases. Again, this is a somewhat surprising finding.

Adding a logarithmic or exponential expansion of 'PCT_LACCESS_SENIORS10' did not add to the model's overall accuracy.

'PCT_LACCESS_HHNV10': Households, no car & low access to store (%), 2010

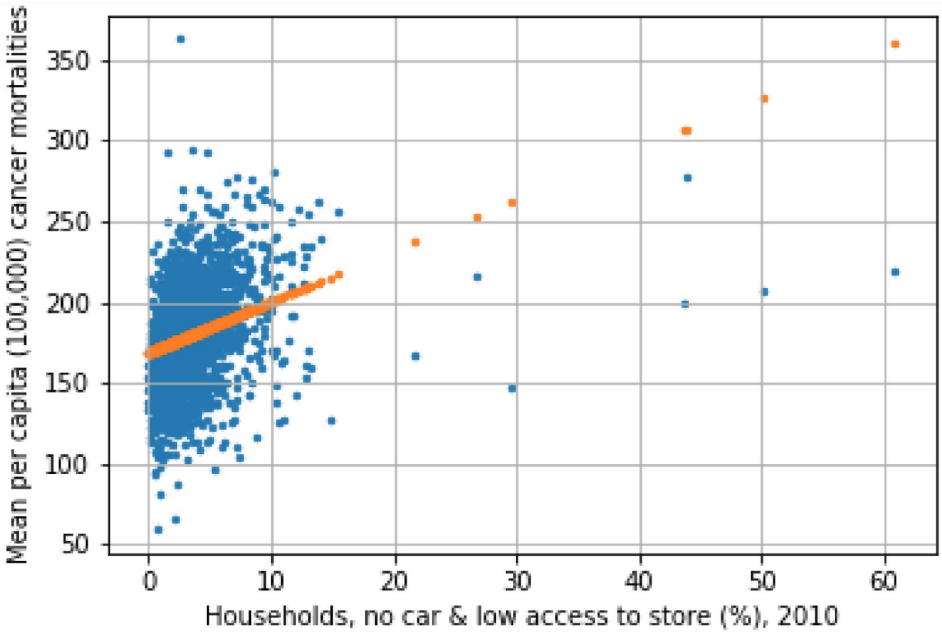
The county household percentage with no car and low access to a grocery store (in 2010) ranges from zero to 60.9%, and has a mean value of 3.1% with a standard deviation of 2.8%.



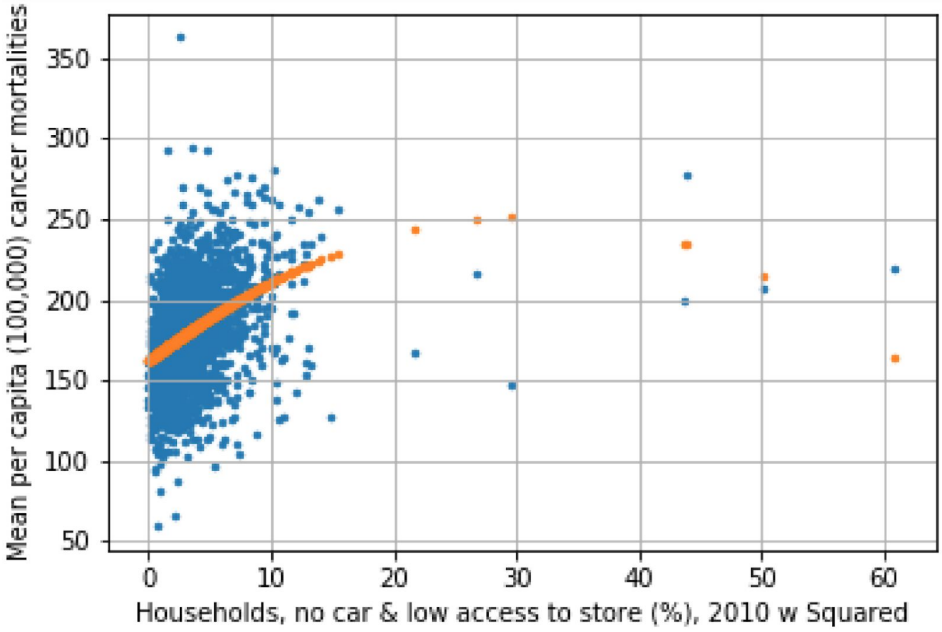
There is a moderately positive correlation of 0.31 between the county population percentage with no car and low access to a grocery store (in 2010) and cancer mortality. Therefore, as this percentage

increases, cancer mortality increases. This correlation makes more sense as one would assume that populations with low access to quality food would struggle more with cancer.

One can see the moderately positive correlation between the percentage of county residents with no car and low access to a grocery store and cancer mortality in the prediction line.

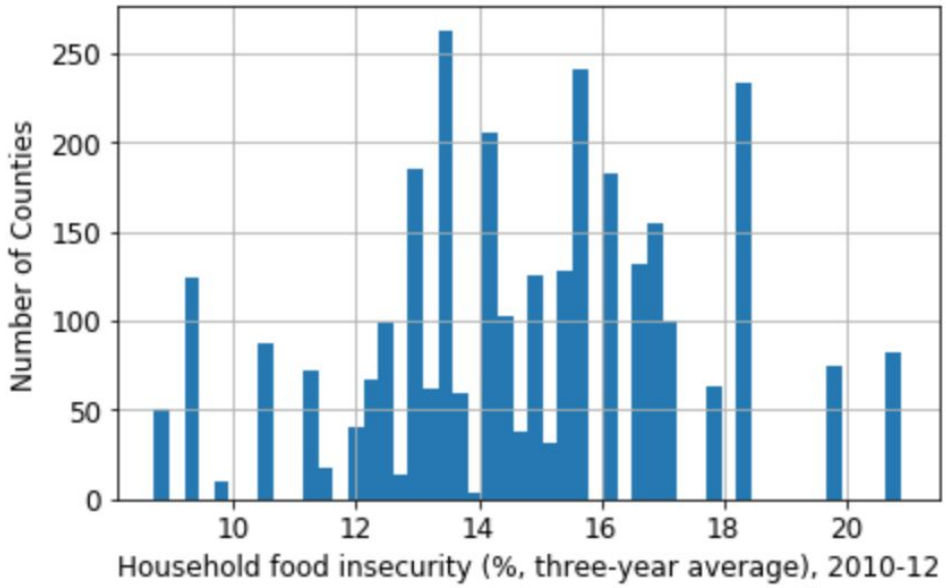


Adding the squared transformation of the percentage of county residents with no car and low access to a grocery store reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.00009.



'FOODINSEC_10_12': Household food insecurity (%, three-year average), 2010-12

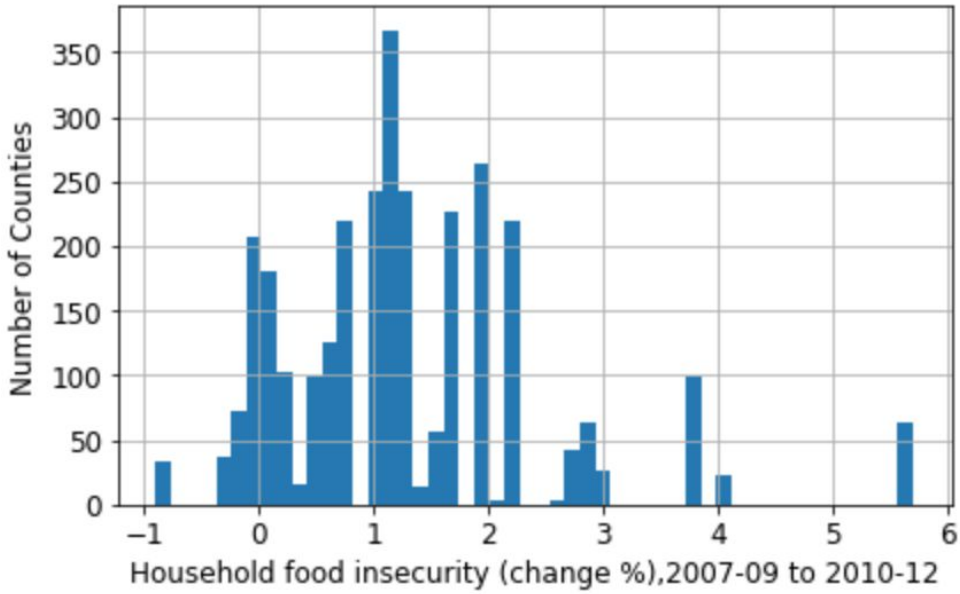
The three-year average county population percentage with household food insecurity (2010-2012) ranges from 8.7% to 20.9%, and has a mean value of 14.8% with a standard deviation of 2.7%.



There is a weakly positive correlation of 0.18 between the three-year average county population percentage with household food insecurity (2010-2012) and cancer mortality. Therefore, as this percentage increases, cancer mortality increases slightly. Adding a logarithmic or exponential expansion of 'FOODINSEC_10_12' did not add to the model's overall accuracy.

'CH_FOODINSEC_09_12': Household food insecurity (change %), 2007-09 to 2010-12

The percentage change of household food insecurity from 2007-2009 to 2010-2012 ranges from -0.9% to 5.7%, and has a mean value of 1.3% with a standard deviation of 1.2%.

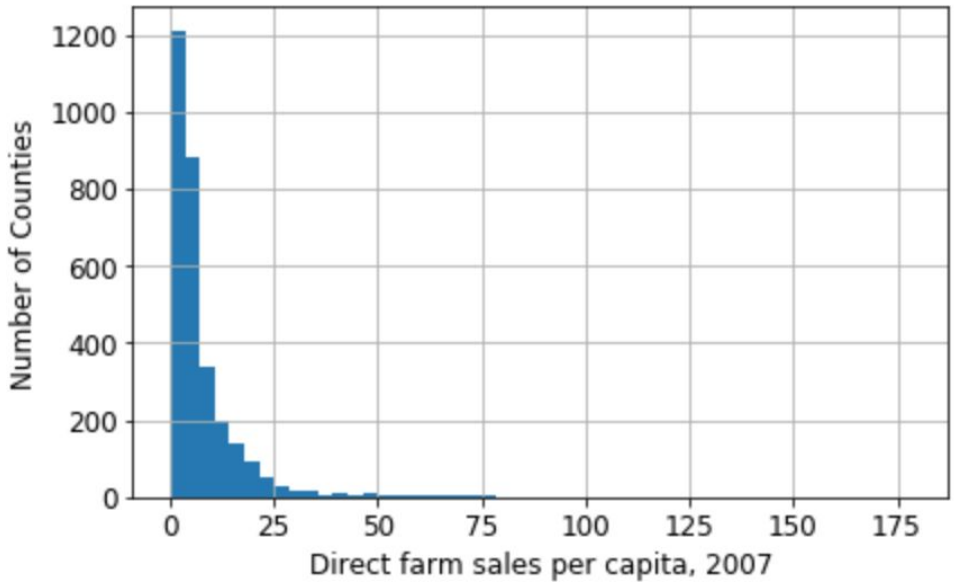


There is a weakly positive correlation of 0.12 between the percentage change of household food insecurity from 2007-2009 to 2010-2012 and cancer mortality. Therefore, as this percentage increases, cancer mortality slightly increases.

Adding a logarithmic or exponential expansion of 'CH_FOODINSEC_09_12' did not add to the model's overall accuracy.

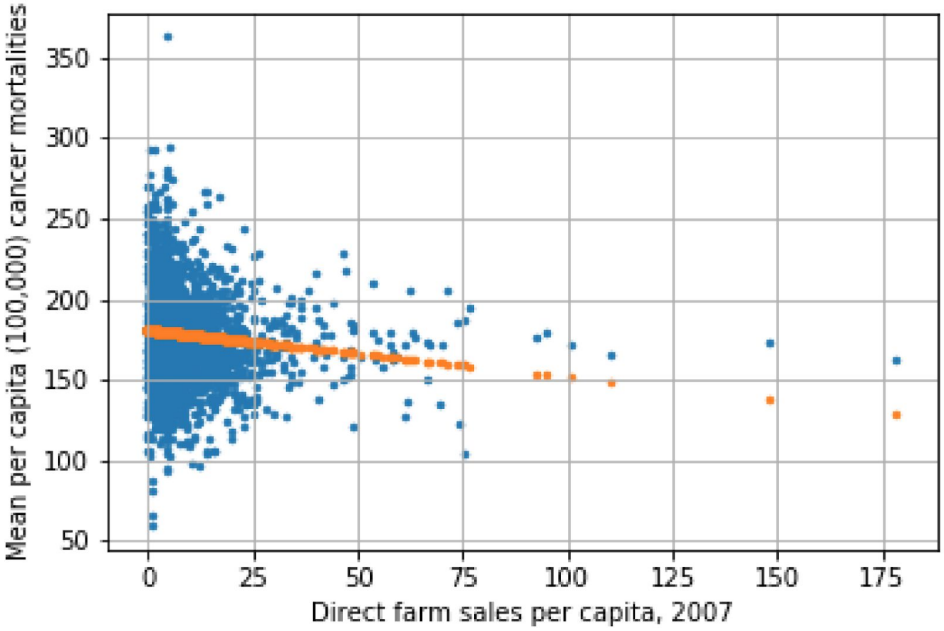
'PC_DIRSALES07': Direct farm sales per capita, 2007

The number of direct farm sales per capita (in 2007) ranges from zero to 178.3, with a mean value of 7.5 and a standard deviation of 10.6.

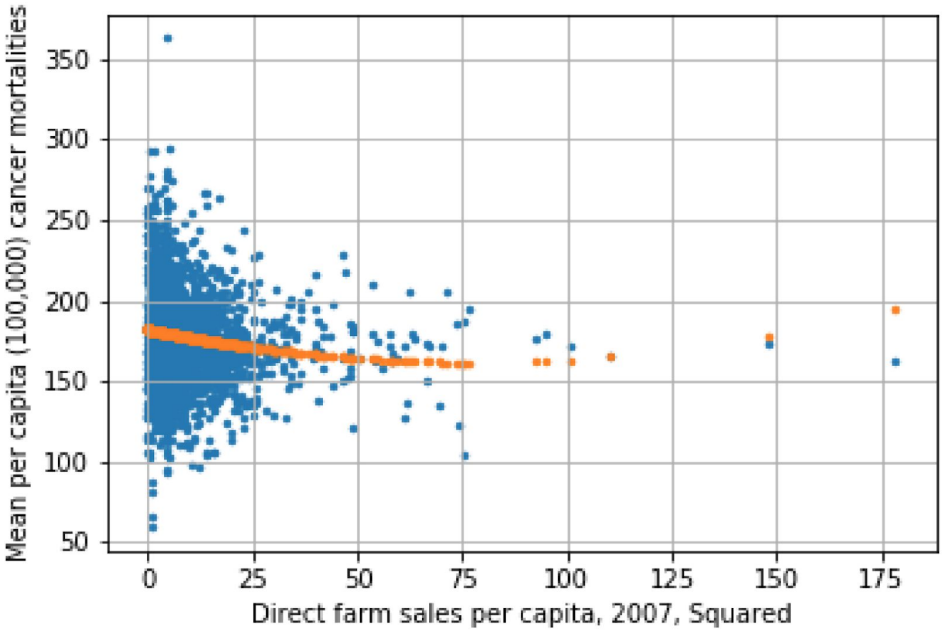


There is a weak negative correlation of -0.11 between the number of direct farm sales per capita (in 2007) and cancer mortality. Therefore, as this number increases, cancer mortality slightly decreases.

One can see the weakly negative correlation between the number of direct farm sales per capita (in 2007) and cancer mortality in the prediction line.

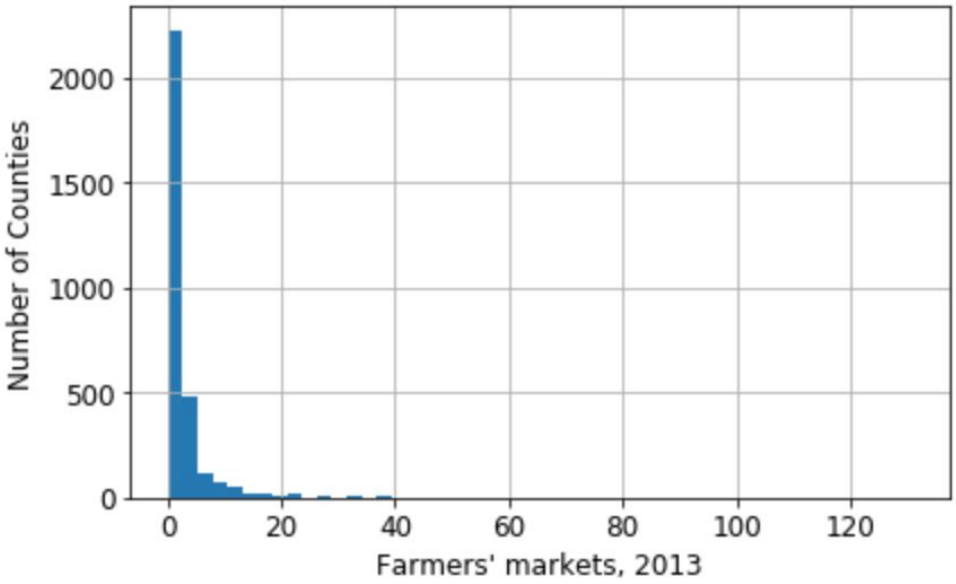


Adding the squared transformation reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.00009.



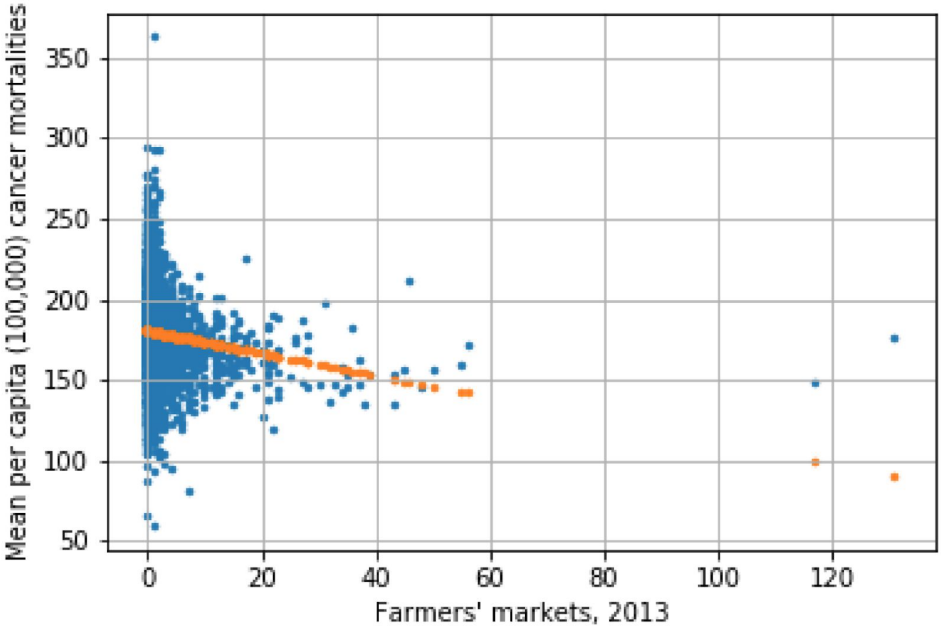
'FMRKT13': Farmers' markets, 2013

The number of farmers' markets per county (in 2013) ranges from zero to 131, with a mean value of 2.6 and a standard deviation of 5.7.

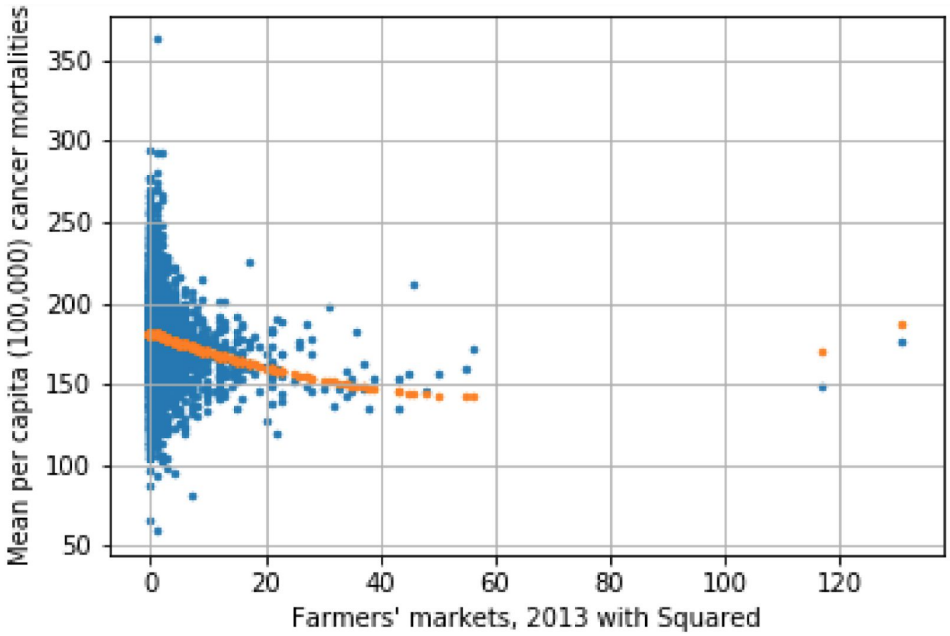


There is a weakly negative correlation of -0.14 between the number of farmers' markets per county (in 2013) and cancer mortality. Therefore, as this number increases, cancer mortality slightly decreases.

One can see the weakly negative correlation between the number of farmers' markets per county (in 2013) and cancer mortality in the prediction line of the first of the plots below.

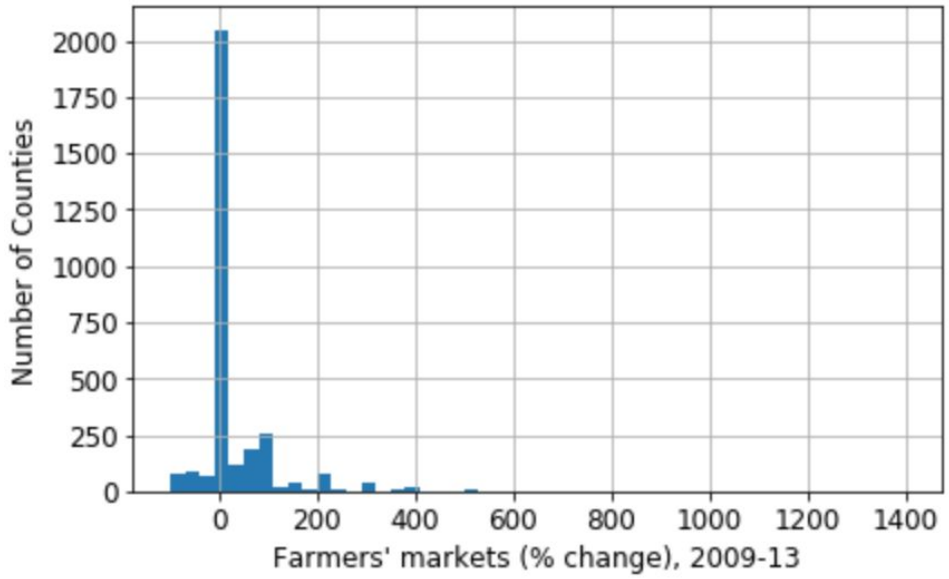


The error between the actual and predicted data points of cancer mortality is reduced when the exponential transformations of 'FMRKT13' are added, and doing so increased the model's accuracy by 0.0005.



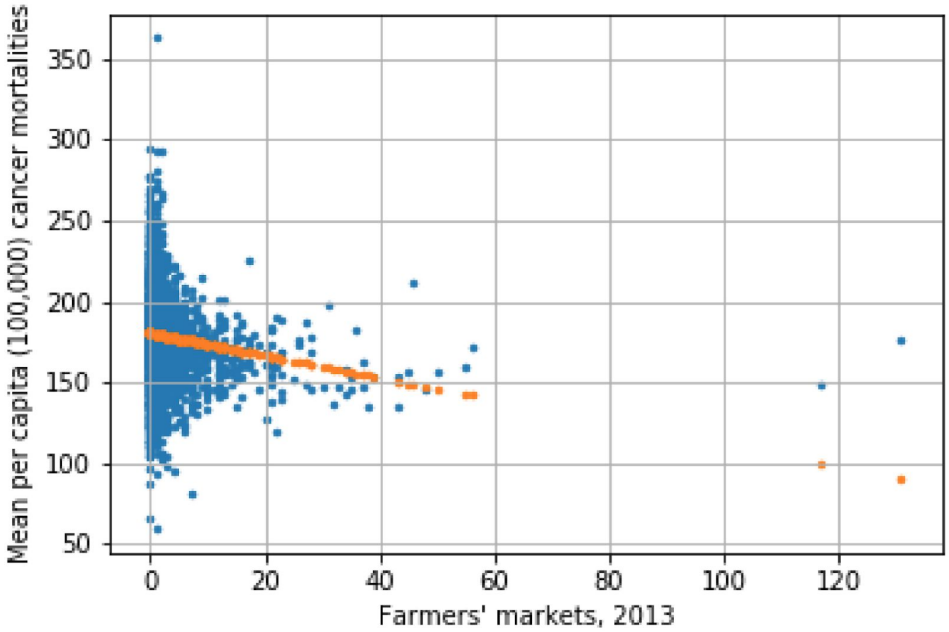
'PCH_FMRKT_09_13': Farmers' markets (% change), 2009-13

The percentage change in farmers' markets between 2009 and 2013 ranges from -100% to 1,400%, with a mean value of 26.4% and a standard deviation of 82.3%.

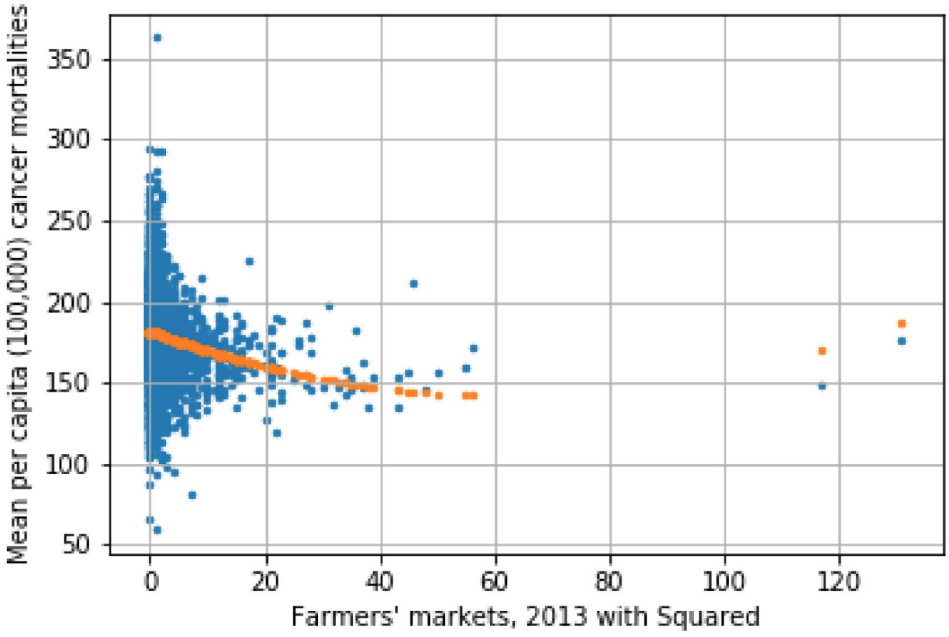


There is a very weak negative correlation of -0.08 between the percentage change in farmers' markets between 2009 and 2013 and cancer mortality. Therefore, as this percentage increases, cancer mortality very slightly decreases.

One can see the weakly negative correlation between percentage change in farmers' markets between 2009 and 2013 and cancer mortality in the prediction line of the first of the plots below.

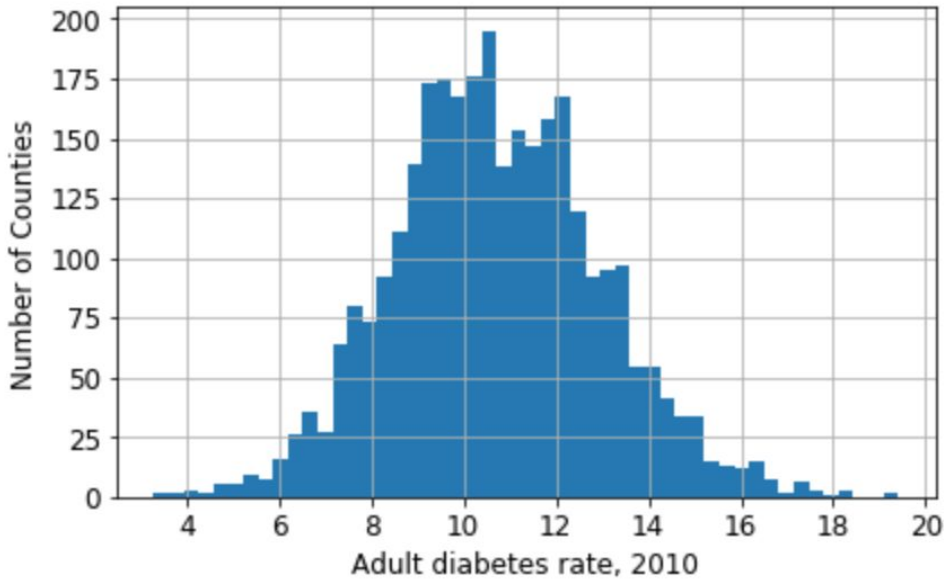


The error between the actual and predicted data points of cancer mortality is reduced when the exponential transformations of 'PCH_FMRKT_09_13' are added, and doing so increased the model's accuracy by 0.0000007.



'PCT_DIABETES_ADULTS10': Adult diabetes rate, 2010

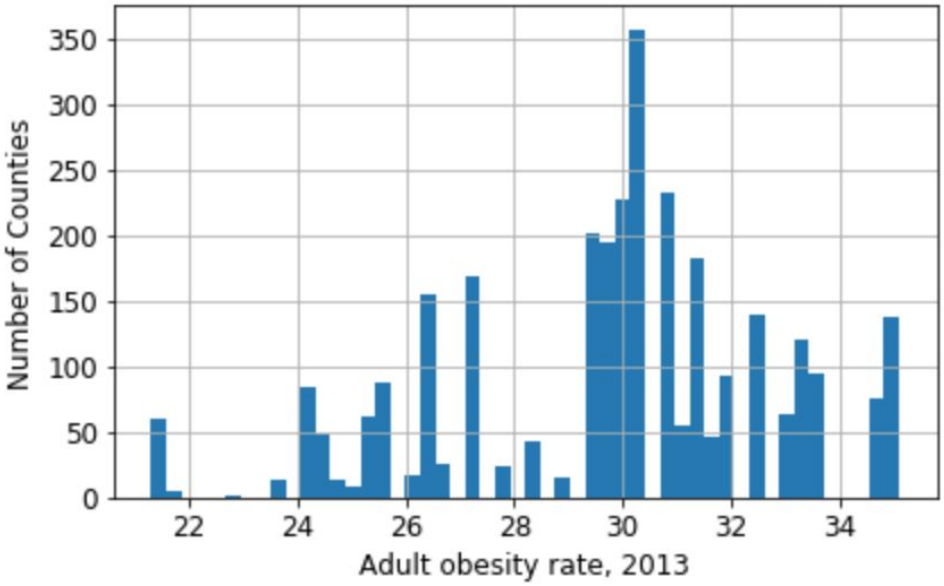
The percentage of adults in each county with diabetes (in 2010) ranges from 3.3% to 19.4%, with a mean value of 10.7% and a standard deviation of 2.3%.



There is a strongly positive correlation of 0.53 between the percentage of adults in each county with diabetes (in 2010) and cancer mortality. Therefore, as this percentage increases, cancer mortality dramatically increases. Adding a logarithmic or exponential expansion of 'PCT_DIABETES_ADULTS10' did not add to the model's overall accuracy.

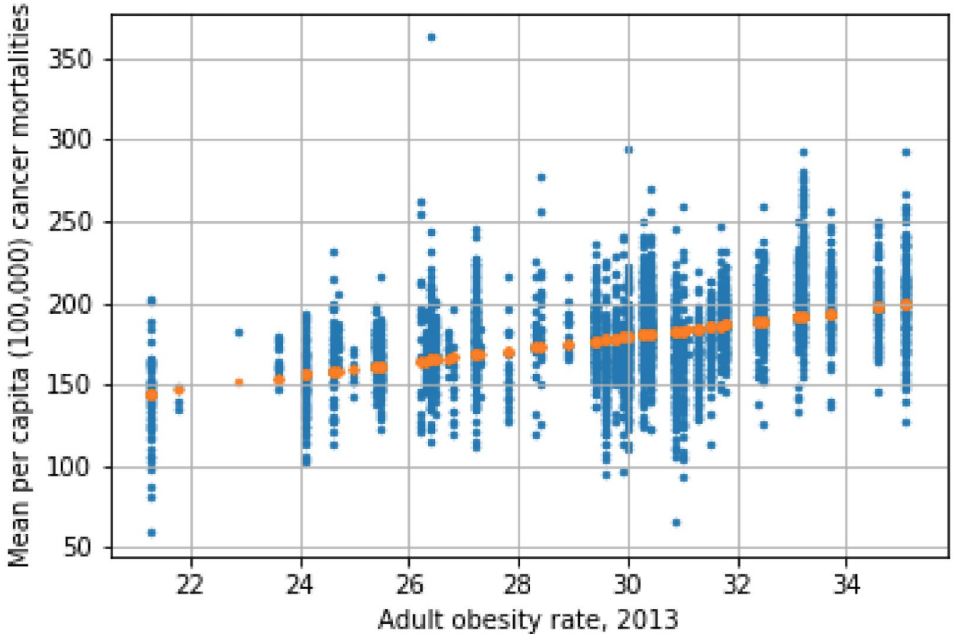
'PCT_OBESE_ADULTS13': Adult obesity rate, 2013

The percentage of adults in each county with obesity in 2013 ranges from 21.3% to 35.1%, with a mean value of 29.8% and a standard deviation of 3%.

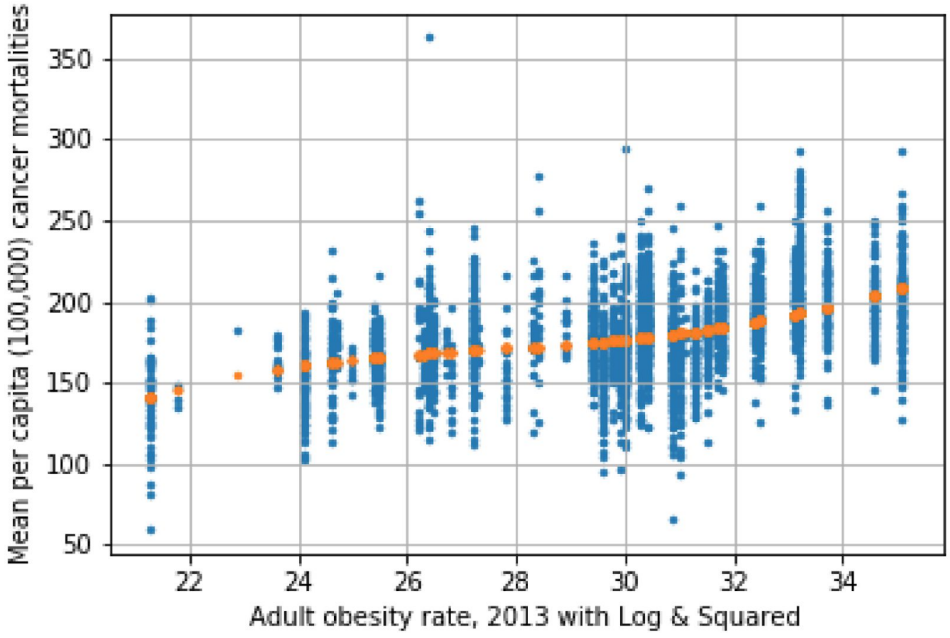


There is a moderate to strong positive correlation of 0.43 between the percentage of county adult residents with obesity (in 2013) and cancer mortality. Therefore, as this percentage increases, cancer mortality increases.

One can see the moderately positive correlation between the percentage of county adults with obesity (in 2013) and cancer mortality in the prediction line.

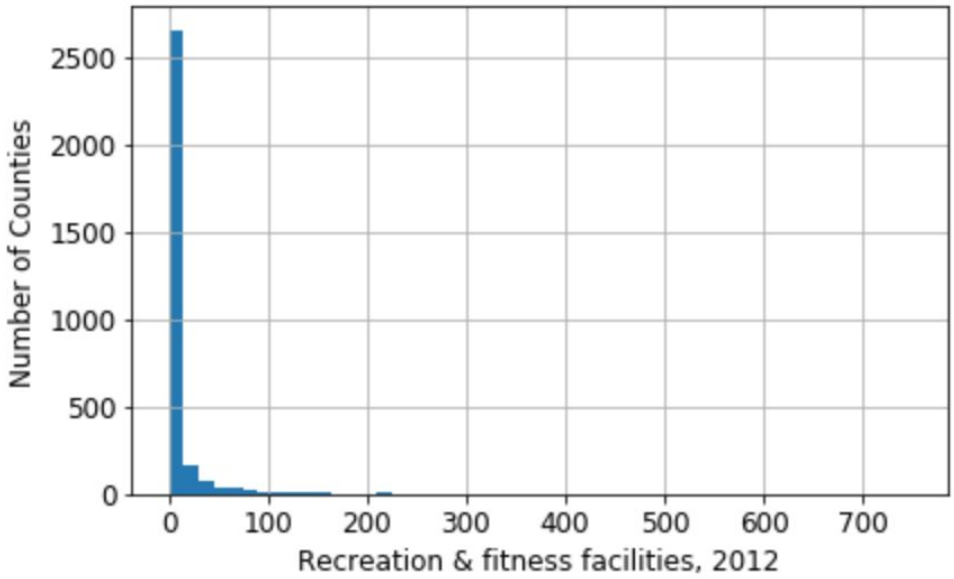


Adding the logarithmic and exponential transformation of the obesity percentage reduced the error between the actual and predicted values of cancer mortality, and increased the overall accuracy of the linear regression model by 0.000005.



'RECFAC12': Recreation & fitness facilities, 2012

The number of recreation and fitness facilities by county (in 2012) ranges from zero to 749, with a mean value of 9.4 and a standard deviation of 30.2.

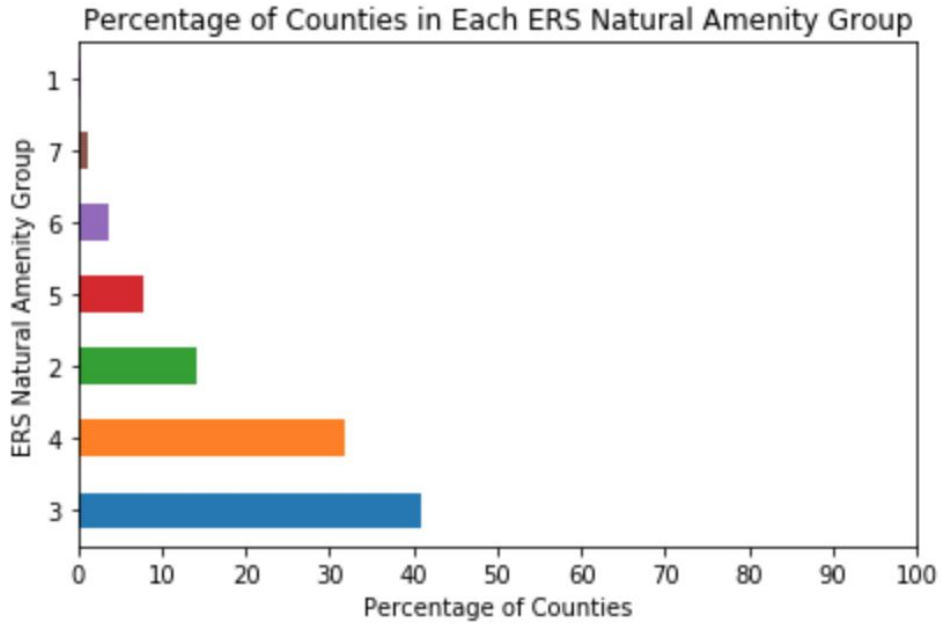


There is a slightly negative correlation of -0.14 between the percentage of recreation and fitness facilities by county (in 2012) and cancer mortality. Therefore, as this percentage increases, cancer mortality slightly decreases.

Adding a logarithmic or exponential expansion of 'REC_FAC_12' did not add to the model's overall accuracy.

'NATAMEN': ERS natural amenity index, 1999

According to the USDA, the ERS "natural amenities scale is a measure of the physical characteristics of a county area that enhance the location as a place to live. The scale was constructed by combining six measures of climate, topography, and water area that reflect environmental qualities most people prefer. These measures are warm winter, winter sun, temperate summer, low summer humidity, topographic variation, and water area. The data are available for counties in the lower 48 States." (https://www.ers.usda.gov/data-products/natural-amenities-scale/). The range of ERS natural amenity index values (in 1999) is from 1 to 7, with a mean value of 3.5 and a standard deviation of 1.04.

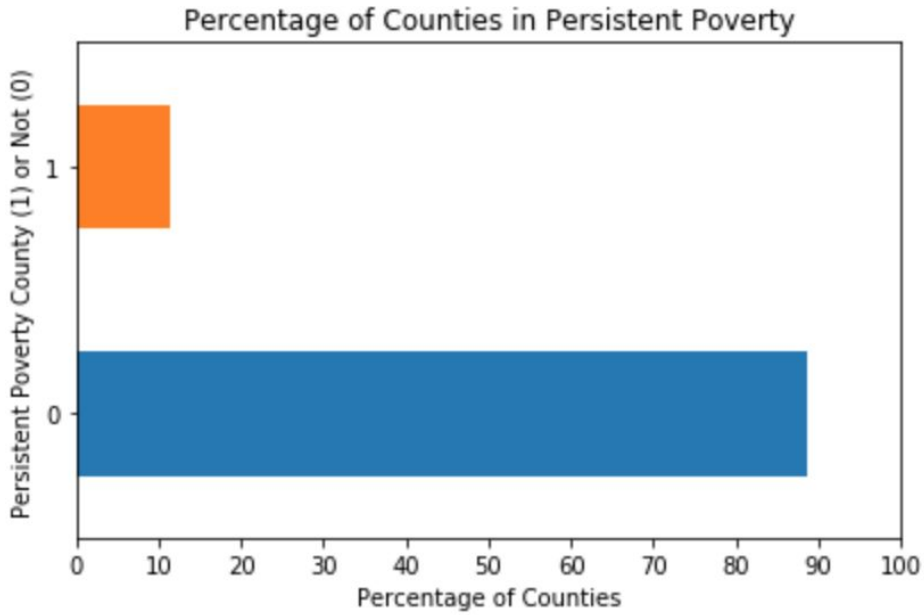


There is a slightly negative correlation of -0.17 between ERS Natural Amenity Index rankings (in 1999) and cancer mortality. Therefore, as this percentage increases, cancer mortality slightly decreases.

Adding a logarithmic or exponential expansion of 'NATAMEN' did not add to the model's overall accuracy.

'PERPOV10': Persistent-poverty counties, 2010

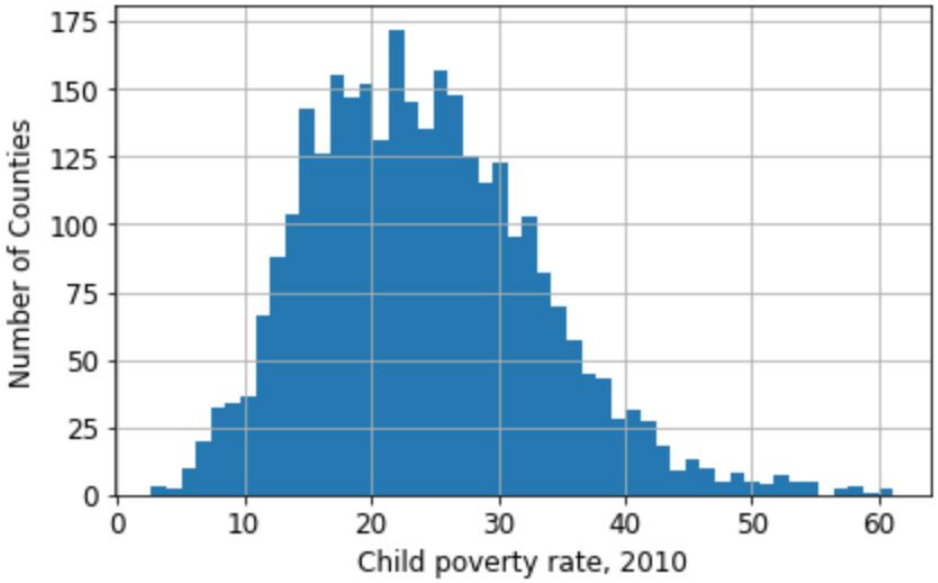
'PERPOV10', the binary feature describing whether a county has persistent poverty or not (in 2010) has a mean value of 0.11, translating to 11.4% of U.S. counties being classified as being in persistent poverty.



There is a moderately positive correlation of 0.27 between persistent poverty counties and cancer mortality. Therefore, as this percentage increases, cancer mortality increases. Adding a logarithmic or exponential expansion of 'PERPOV10' did not add to the model's overall accuracy.

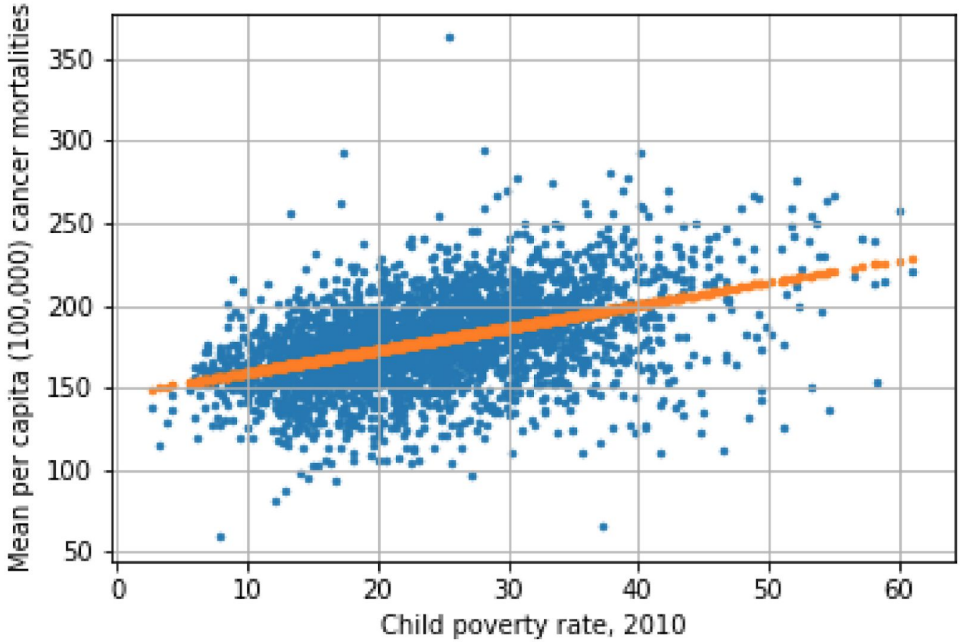
'CHILDPOVRATE10': Child poverty rate, 2010

The percentage of children in poverty per county (in 2010) ranges from 2.7% to 61.1%, with a mean value of 24.2% and a standard deviation of 9%.

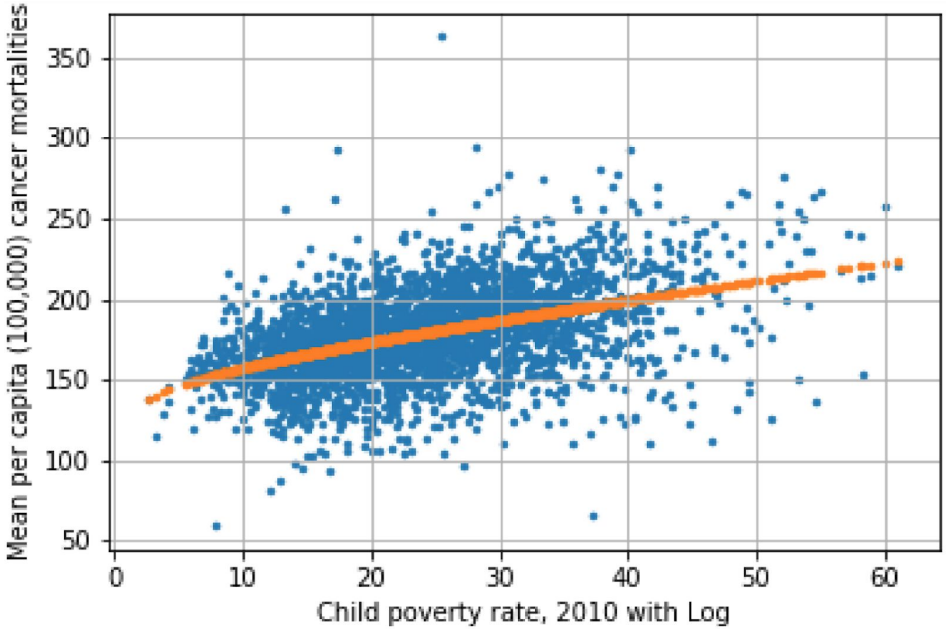


There is a moderately positive correlation of 0.45 between the percentage of children in poverty (in 2010) and cancer mortality. Therefore, as this percentage increases, cancer mortality increases.

One can see the strong positive correlation between the percentage of children in poverty (in 2010) and cancer mortality in the prediction line of the first of the plots below.

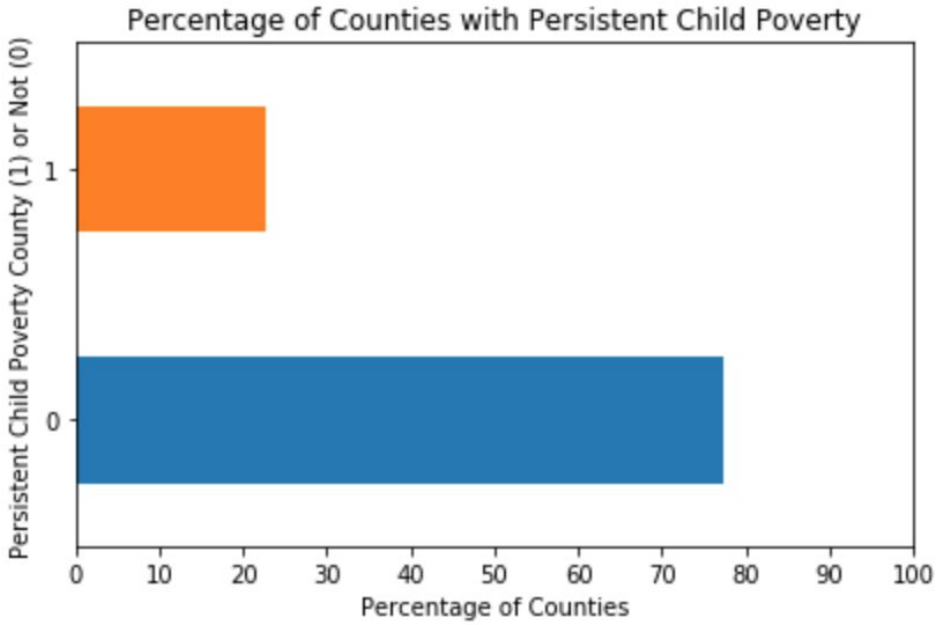


The error between the actual and predicted data points of cancer mortality is reduced when the logarithmic transformation of 'CHILDPOVRATE10' is added, and doing so increased the model's accuracy by a sizable 0.003.



'PERCHLDPOV10': Persistent-child-poverty counties, 2010

'PERCHLDPOV10', the binary feature describing whether a county has persistent child poverty or not (in 2010) has a mean value of 0.23, translating to 23% of U.S. counties being classified as having persistent child poverty.



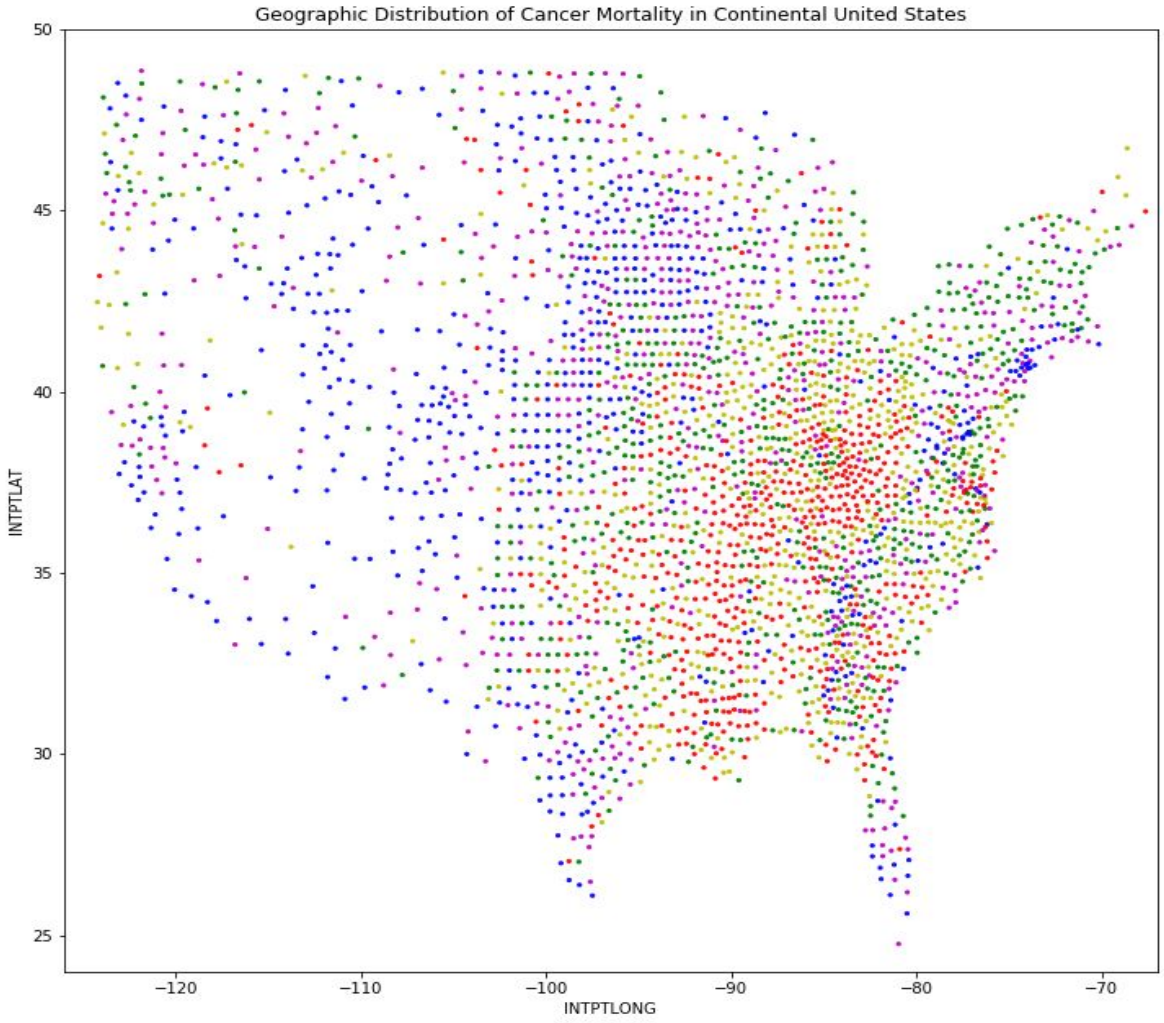
There is a moderately positive correlation of 0.30 between persistent child poverty and cancer mortality. Therefore, as this percentage increases, cancer mortality also increases. Adding a logarithmic or exponential expansion of 'PERPOV10' did not add to the model's overall accuracy.

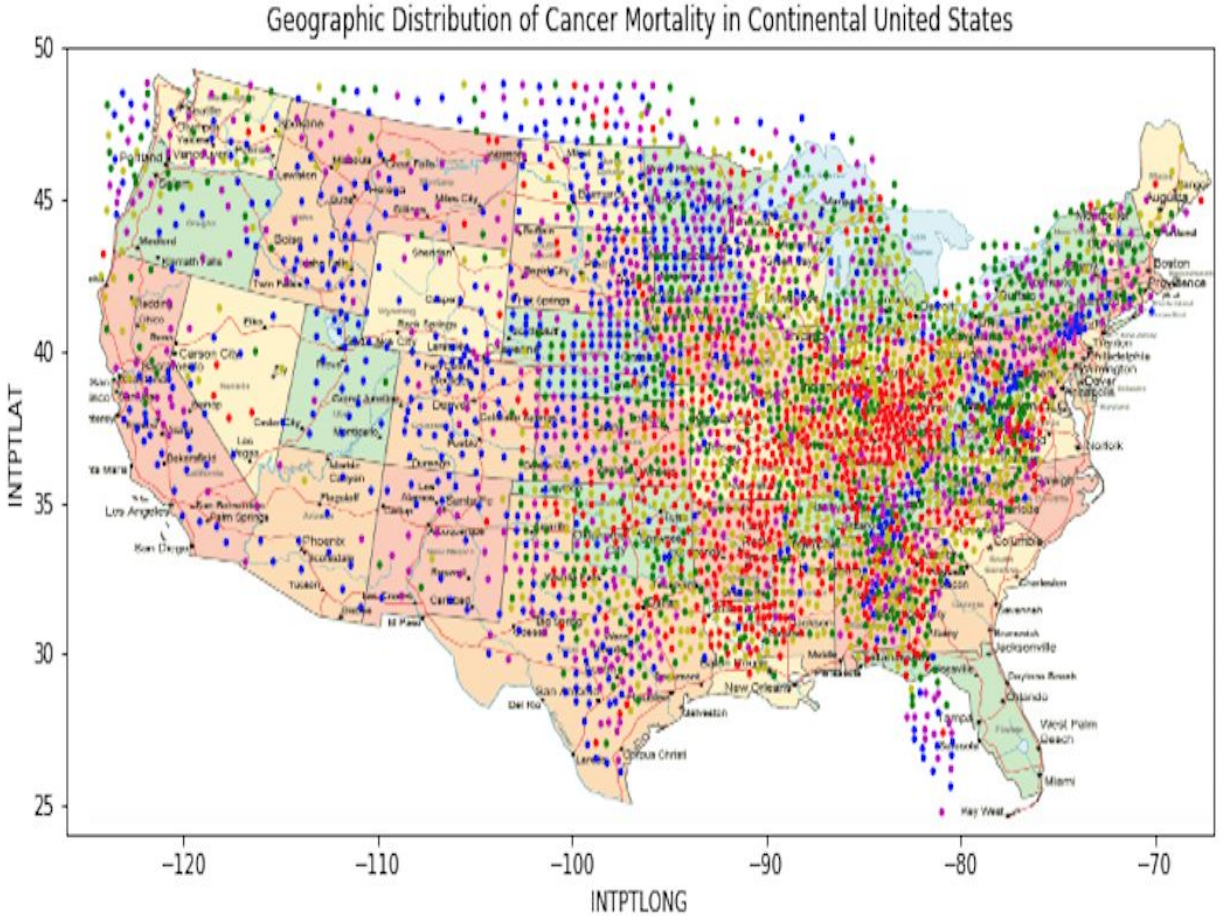
Geographic Distribution of Cancer Mortality in the Continental U.S.

The geographic distribution of cancer mortality rates is visualized in the two latitude/longitude scatterplot maps below. 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.

The maps have the following legend:

- Blue: Lowest Mortality (first quintile)
- Magenta: Low Mortality (second quintile)
- Green: Medium Mortality (third quintile)
- Yellow: High Mortality (fourth quintile)
- Red: Highest Mortality (fifth quintile)


Geographic Distribution of Cancer Mortality in Continental United States
INTPTLAT
INTPTLONG
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70



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.

Hypothesis Testing

The focus of this section is the running of hypothesis tests 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. The null hypotheses tested are detailed below.

The Jupyter notebook for these tests can be found here:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_DataStory_HT.ipynb

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 counties where over 50% of the population identified as White, the cancer mortality rate was also 179 per 100,000 people with a standard deviation of 27.2. For counties where over 50% of the population identified as Black, 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.

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

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 the unemployment feature: 'PctUnemployed16_Over'. 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.

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.

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 of each individual feature - the percentage of county residents ages 25 and over whose highest education attained was a high school diploma and the

percentage of county residents ages 25 and over whose highest education attained was a Bachelor's Degree. 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, population percentage with no car and low access to a grocery store, 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. By identifying the most salient predictors of cancer mortality, policy makers can use this study as a resource in which to guide public health policy as a component of the fight against cancer. Although these salient predictors cannot be identified as a cause of cancer mortality, identifying predictive features can help in the understanding of factors that contribute to cancer mortality. Random Forest was also used as a way to nonlinearly predict cancer mortality, but because the Random Forest method does not produce coefficients, it was not used to identify the most salient predictors of cancer mortality.

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. Generally speaking, 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), and the number of estimators (for Random Forest). 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 Jupyter notebooks for this machine learning section can be found at the following GitHub locations:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_ML_unscaled.ipynb

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_ML_scaled.ipynb

The first algorithm used for regression of cancer mortality was simple Ordinary Least Squares (OLS) linear regression. OLS linear regression estimates the unknown parameters of a linear function of a set of predictive features by the principle of least squares: minimizing the sum of the squares of the differences between the observed values of the target variable (e.g., cancer mortality) and its predicted values. An OLS linear regression model's root mean squared error (RMSE) is its evaluation metric. The Sci-Kit learn algorithm used in this project for OLS linear regression had no hyperparameters to tune, and was therefore simple. This algorithm performed very well with unscaled data.

The ridge regression algorithm was also experimented with. This algorithm is a version of OLS linear regression, but with L2 regularization added. L2 regularization performs the OLS loss function and adds the squared value of each coefficient multiplied by a constant value of the regularization hyperparameter alpha. In so doing, large positive and large negative coefficients are penalized. For this project, the ridge regression algorithm used the 'auto' solver, and alpha values of 0.001, 0.01, 0.1, 1, 10 and 100 were experimented with. Using unscaled data, the ridge regression algorithm with an alpha value of 0.001 was the only algorithm that performed slightly better than simple OLS linear regression.

The LASSO regression algorithm was attempted, but did not perform well. LASSO regression performs the OLS loss function and adds the absolute value of each coefficient multiplied by a constant value of the regularization hyperparameter alpha. As was the case with ridge regression, alpha values of 0.001, 0.01, 0.1, 1, 10 and 100 were experimented with. LASSO can be used to select important features of a dataset, because it shrinks the coefficients of less important features to zero. However, doing so with this project's feature set shrunk several features' coefficients to zero that were still contributing to the overall accuracy of the model. LASSO, ElasticNet, and the SGD Regressor are examples of algorithms that are involved in dimension reduction in an attempt to improve accuracy by removing unimportant features. However, it was obvious by the poor accuracy and high RMSE scores that this model's complexity requires dimension addition and not dimension reduction.

The ElasticNet regression algorithm was also utilized, but performed unsatisfactorily. ElasticNet is a hybrid algorithm combining Ridge and LASSO regression in a combination quantified by its 'L1_ratio', which was experimented with at the values of 0.25, 0.5 and 0.75. An 'L1_ratio' value of zero is pure ridge regression and a value of one is pure LASSO regression. For each 'L1_ratio' value experimented with, alpha values of 0.001, 0.01, 0.1, 1, 10 and 100 were attempted. However, all combinations of ElasticNet's hyperparameters resulted in lower accuracy and higher RMSE scores than simple OLS linear regression and ridge regression with an alpha value of 0.001. ElasticNet's

performance is another example of why the strategy of dimension reduction was not appropriate for this model.

Stochastic Gradient Descent, or SGD, Regression was tried as a way to use the learning rate of gradient descent to minimize loss, or error. With SGD Regression, “the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate)” (SciKit-Learn documentation). This algorithm uses either the L1, L2, or ElasticNet regularization penalty that is added to the loss function that reduces model parameters towards the zero vector. The ‘Huber’ loss function was used and the L1, L2 and ElasticNet penalty were all tried. Alpha values of 0.0001, 0.001, 0.01, 0.1, 1, 10 and 100 were experimented with. An ‘L1_ratio’ value of 0.5 was tried for the ElasticNet penalty. Not all combinations of these hyperparameters were tried, as it was obvious that negative accuracy scores were being returned for all iterations of this algorithm using unscaled data. This seemed to be another example of the ill fit of dimension reduction strategies with this model, but SGD Regression did perform fairly well with scaled data and a constant learning rate. Using the very low Alpha value of 0.00001, the L2 penalty and various values of epsilon up to 5, the accuracy score got up to 0.5 for the training set and 0.46 for the test set.

As was explained earlier, nonlinear relationships between the predictive feature set and the target variable of cancer mortality were explored through logarithmic and exponential transformations of several features. Because the “kernel trick” is used in Support Vector Machines to capture nonlinear relationships, the Kernel Regression algorithm was tried. This algorithm returned accuracy and RMSE scores comparable to OLS linear regression, but did not perform quite as well. The auto solver was used, and alpha values of 0.001, 0.01, 0.1, 1, 10 and 100 were experimented with.

The Random Forest algorithm was also tried as an attempt to further capture nonlinear relationships, but it did not perform as well as OLS linear regression with 10, 100 and 1,000 estimators. The training set accuracy scores were around 0.93, but the test accuracy scores were around 0.55.

The best performing regression algorithm was Ridge Regression using the ‘auto’ solver and an Alpha of 0.001, with a training accuracy score of 0.6465, a training RMSE of 16.59, a test accuracy score of 0.6408, and a test RMSE of 16.2.

The top 10 performing models are detailed in the following table. All of them used unscaled data and the ‘auto’ solver.

Model	**Alpha Regularization Parameter Value**	**Training Set Accuracy/ R-Squared**	**Training Set Root Mean Squared Error**	**Test Set Accuracy/ R-Squared**	**Test Set Root Mean Squared Error**
Ridge Regression	0.001	0.6465	16.59	0.6408	16.2
OLS Linear	N/A	0.6466	16.59	0.6406	16.2

Regression					
Ridge Regression	0.01	0.6461	16.6	0.6398	16.22
Ridge Regression	0.1	0.6446	16.64	0.6342	16.34
Ridge Regression	1.0	0.6386	16.77	0.6219	16.62
Kernel Ridge Regression	1.0	0.6387	16.77	0.6215	16.63
Kernel Ridge Regression	0.01	0.63	16.97	0.6181	16.7
Kernel Ridge Regression	10.0	0.6317	16.94	0.613	16.81
Ridge Regression	10.0	0.6317	16.94	0.613	16.81
Ridge Regression	100.0	0.621	17.18	0.6077	16.93

The complete hyperparameter tuning table with accuracy and RMSE scores of all models is located at:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/HP%20tuning%20table%20-%202nd%20Capstone_Loew.xlsx

Discussion of Best Performing Regressor's Coefficients

After the best performing regressor (unscaled Ridge Regression with an alpha of 0.001) was identified through the iterative process of hyperparameter tuning, its regression coefficients were examined in an effort to discover further significant contributors to cancer mortality that weren't identified by the predictive feature set's Pearson's correlation coefficients or by hypothesis testing.

After examining the feature set's Ridge Regression coefficients, it was clear that they often did not have similar values to the Pearson's correlation coefficients, frequently not even sharing the same polarities (positive or negative). In fact, the proportion of the predictive feature set whose correlation coefficients and Ridge Regression coefficients did share the same polarity (positive or negative) was 0.54. This was an interesting finding suggesting that statistical relationships between predictive features and the target feature within a whole model can be quite different than one-on-one statistical relationships between individual features and the target feature.

This proportion of the predictive feature set whose correlation coefficients and Ridge Regression coefficients shared the same polarity, or “same sign” proportion, was further explored through modification of the feature set and the Ridge Regression regressor algorithm. A list of these modifications follow:

- First, the logarithmic and exponential transformations of features which contributed to the model's accuracy were removed, as the nonlinear statistical relationships they uncovered could have been causing volatility in the higher dimensional feature set. The “same sign” proportion was 0.525.

- Second, the Ridge Regression algorithm was re-run with just the predictive features that had the strongest correlations with the target feature. To do this, a Boolean mask was created assigning a 'True' value to those features whose absolute value coefficient was greater than 0.3. The “same sign” proportion was 0.55.

- Third, the Ridge Regression algorithm was re-run with the full feature set using scaled data implemented by Sci-Kit Learn’s MinMax Scaler. The “same sign” proportion was 0.52.

- Fourth, the Ridge Regression algorithm was re-run with internal normalization and the full feature set. The “same sign” proportion was 0.53.

- Fifth, the Ridge Regression algorithm was re-run using the ‘svd’ solver and the full feature set. The “same sign” proportion was 0.54.

- Sixth, the Ridge Regression algorithm was re-run using the ‘cholesky’ solver and the full feature set. The “same sign” proportion was 0.54.

- Seventh, the only slightly worse performing Ordinary Least Squares (OLS) linear regression algorithm was rerun, to see if the Ridge Regression algorithm suppressed extreme coefficients to the degree where the “same sign” proportion was only around 0.5. The “same sign” proportion was 0.54.

However, the “same sign” proportion remained remarkably consistent through all of these modifications, ranging between 0.52 and 0.55. Therefore, the feature set’s Ridge Regression coefficients were accepted as representing one partial, but real, model of the relationship between socioeconomic, environmental, economic, and geographic features and the target variable of per-capita cancer mortality rates (per 100,000 people) in the United States at the county level in the year 2015. The Ridge Regression coefficients were first sorted in descending order, with the strongest positive coefficients at the top of the series and the strongest negative coefficients at the bottom of the series. Although there surely is a complex web of interconnections and relationships between the features, one can look at the coefficients of each feature individually to see what impact they had on cancer mortality in 2015. Doing so can identify the features with the strongest impact on cancer mortality, which in turn can help inform policy interventions that could help reduce cancer mortality.

The Jupyter notebook with the code for both the “same sign” proportion exploration and the discussion of the Ridge Regression coefficients can be found at: https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_Best_Model_Coefficients.ipynb

In order to understand how the Ridge Regression coefficients can identify the features with the strongest impact on cancer mortality, a brief technical explanation of regression coefficients is in order. Binary features' Ridge Regression coefficients can show that if a county's value is true for that binary feature, one can expect a change in the cancer mortality rate per 100,000 equal to the Ridge Regression coefficient of that feature. For example, the 'State_Nevada' feature can be examined. This feature stores data on whether a county is in Nevada (1) or not (0). If a county is in Nevada, one can expect an increase of 43 cancer deaths per 100,000 people.

For a continuous feature, an increase of one unit for that feature will result in a change equal to that feature's Ridge Regression coefficient. For example, the 'RECFACPTH12' feature stores the number of recreation and fitness facilities per 1,000 people in 2012. For every percentage point increase in the value of this feature for any given county, one can (counter-intuitively) expect an increase of 21 cancer mortalities in 2015.

Features in the feature set are in different scales - they are measured in percentages, per capita rates (per 100,000 people), binary flags, and in real numbers. Because of this, the feature set was necessarily normalized as a whole in order to return meaningful coefficients as to the relative effect of each feature on the target variable.

In order to identify the features with the strongest relationship to both increasing and decreasing cancer mortality, arbitrary cut-off coefficient values of 10 or higher for positive relationships and -10 or lower for negative relationships were selected.

It should be noted that the Ridge Regression coefficient for every individual feature in the predictive feature set can be found in the Jupyter notebook located at:

https://github.com/danloew/Springboard-2nd-Capstone/blob/master/Cancer_Best_Model_Coefficients.ipynb

Predictive Features with Positive Normalized Coefficients:

A table of the predictive features with the strongest positive relationships with the cancer mortality target variable follows:

Feature Name	**Feature Description**	**Ridge Regression Coefficient**
'PctPrivateCovera	Logarithmic transformation of the percentage of county	182

ge_log'	residents with private health coverage	
‘State_Nevada’	Whether a county is in Nevada or not	43
'AvgHouseholdSize_log'	Logarithmic transformation of the average household size	38
'State_Alaska'	Whether a county is in Alaska or not	35
'State_California'	Whether a county is in California or not	30
‘PctPublicCoverage_log’	Logarithmic transformation of the percentage of county residents with public health coverage	22
'RECFACPTH12'	Number of recreation and fitness facilities per 1,000 people in 2012	21
'State_Florida'	Whether a county is in Florida or not	20
'FMRKT09_isnull'	Whether a county has a missing value for the number of farmer's markets it had in 2009	19
'State_Arizona'	Whether a county is in Arizona or not	18
'State_Utah'	Whether a county is in Utah or not	17
'State_District of Columbia'	Whether a county is Washington, DC or not	14
‘povertyPercent_log’	Logarithmic transformation of the percentage of county residents who are in poverty	14
'NATAMEN_isnull'	Whether there are missing values in the 'NATAMEN' feature which stores the quality of life 'ERS Natural Amenity Index' (for 1999)	11
'State_Oklahoma'	Whether a county is in the state of Oklahoma or not	11

As can be seen above, the individual features with the strongest positive correlations with cancer mortality fall into the following categories:

- State that the county is in
- Recreation facility-related feature

Logarithmic transformations of features:

- Health insurance features (private and public)
- Poverty-related features
- Average household size

Missing value features:

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

It should be pointed out that at least one individual feature had a positive Ridge Regression coefficient that conflicted with its Pearson's correlation coefficient. The Ridge Regression coefficient for 'PctPrivateCoverage_log' denotes that for each whole number increase, there was an increase in cancer mortality of 182 deaths per capita (100,000 people) in 2015. This conflicts with the negative Pearson's correlation coefficient that the non-transformed version of this feature has, which suggests an area for future research. **UPDATE: DUE TO THE FACT THAT IT'S A LOGARITHMIC FUNCTION? EXPAND.**

Predictive Features with Negative Coefficients:

A table of predictive features with the strongest negative relationships with the cancer mortality target variable follows:

Feature Name	**Feature Description**	**Ridge Regression Coefficient**
'PctEmployed16_Over_log'	Logarithmic transformation of the feature which stores the percentage of county residents ages 16 and over who were employed	-122
'State_Hawaii'	Whether a county is in the state of Hawaii or not	-43
'PctEmpPrivCoverage_log'	Logarithmic transformation of the percentage of county residents with employee-provided private health coverage	-27
'RECFACPTH07'	Number of recreation and fitness facilities per 1,000 people in 2007	-25
'PercentMarried_log'	Logarithmic transformation of the feature which stores the percentage of county residents who are married	-25
'MedianAge_log'	Logarithmic transformation of the feature which stores the median age of county residents	-22
'State_Rhode Island'	Whether a county is in the state of Rhode Island or not	-19
'State_Connecticut'	Whether a county is in the state of Connecticut or not	-17

'State_Iowa'	Whether a county is in the state of Iowa or not	-15
'State_Illinois'	Whether a county is in the state of Illinois or not	-14
'State_North Carolina'	Whether a county is in the state of North Carolina or not	-14
'State_Ohio'	Whether a county is in the state of Ohio or not	-13
'State_Michigan'	Whether a county is in the state of Michigan or not	-11
'State_New Hampshire'	Whether a county is in the state of New Hampshire or not	-10
CHILDPOVRATE10_log	Logarithmic transformation of the rate of children in poverty in 2010	-10

As can be seen above, the features with the strongest negative correlations with cancer mortality fall into the following categories:

- State that the county is in
- Recreation facility-related feature

Logarithmic transformations of features:

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

It should be pointed out that the Ridge Regression coefficient for the logarithmic transformation of the feature that stores the percentage of children living in poverty in 2010 ('CHILDPOVRATE10_log') denotes that for each whole number increase in this feature's value, there would be a decrease in cancer mortality of 30 deaths per capita (100,000) in 2015. This contradicts other poverty-related features' positive Pearson's correlations, warranting further research.

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

Because there are so many 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" (e.g. distance from major urban centers, healthcare-related features, etc.). This grouping strategy involved taking the sum of all positive coefficients in the feature set, then summing the positive coefficients for each "feature family", and then dividing the sum of each positive "feature family" by the total positive coefficient sum to uncover the proportion of

the total predictive value that each "feature family" had on increasing cancer mortality. This strategy was then repeated for negative coefficients to uncover the proportion of the total predictive value that each "feature family" has on decreasing cancer mortality.

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):

Feature Family Name	**Feature Family Description**	**Proportion of Feature Families' Total Positive Influence of Increasing Cancer Mortality**
State	United States state each county is in	0.3642
Health Insurance	Types of health insurance for each county's populace	0.3026
Missing Value Feature	Missing values	0.1338
Average Household Size	Average household size of each county	0.0571
Distances to Top 10 Oncology Hospitals	L1 and L2 distances from county centroids to top 10 oncology hospitals	0.033
Recreation and Fitness	Recreation and fitness facilities in each county	0.0317
Poverty-related	Poverty	0.0231
Food Environment	Food environment of each county	0.0123
Income	Financial income of each county's populace	0.0121
Education	Education levels of each county's populace	0.0082
Distance to Major Urban Centers	L1 and L2 distances from county centroids to major cities	0.069
Population Loss	Counties' significant population loss as of the year 2000	0.006
Erroneous data indicator	Erroneous data indicator referencing average household size	0.0059
Comorbid Health Conditions	Health conditions comorbid with cancer	0.0016

Metropolitan indicator	Indicator of whether a county is in a metropolitan area or not	0.0012
Environmental Health	L2 distance to closest EPA Superfund Cleanup site	0.0002
Employment	Employment status of each county's populace	0.0002
Age	Age of each county's populace	0.0001
Marital feature	Marital status of county's populace	0.0001
Race	Race of each county's populace	0.000005
Cancer Diagnoses	Rate of cancer diagnoses in each county	0.0000003
Clinical Cancer Trials	Per capita number of cancer-related clinical trials per county	0.00000000003
Geography	Square mileage of land mass for each county	0.0000000000001

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):

Feature Family Name	**Feature Family Description**	**Proportion of Feature Families' Total Negative Influence of Decreasing Cancer Mortality**
State	United States state each county is in	0.4341
Employment	Employment status of each county's populace	0.2161
Missing Value Feature	Missing values	0.0687
Health Insurance	Types of health insurance for each county's populace	0.0591
Recreation and Fitness	Recreation and fitness facilities in each county	0.0441
Marital feature	Marital status of county's populace	0.0441
Age	Age of each county's populace	0.0393
Distances to Top 10 Oncology Hospitals	L1 and L2 distances from county centroids to top 10 oncology hospitals	0.0215

Food Environment	Food environment of each county	0.0182
Poverty-related	Poverty	0.0179
Distances to Major Urban Center	L1 and L2 distances from county centroids to major cities	0.0146
Income	Financial income of each county's populace	0.0144
Average Household Size	Average household size of each county	0.0055
Erroneous data indicator	Erroneous data indicator referencing median age	0.0009
Latitude Longitude	Latitude and Longitude	0.0006
Environmental Health	L1 distance to closest EPA Superfund Cleanup site	0.0004
Quality of Life	Quality of life index	0.0002
Comorbid Health Conditions	Health conditions comorbid with cancer	0.0001
Race	Race of each county's populace	0.0001
Birth Rate	Birth Rate	0.0001
Education	Education levels of each county's populace	0.00003
Cancer Diagnoses	Rate of cancer diagnoses in each county	0.0000000002
Geography	Square mileage of water for each county	0.0000000005
Population	Population of county	0.00000000000000005

Summary

This project has been an attempt at building a predictive model of the socioeconomic, educational, geographic, environmental and dietary features that increase or decrease cancer mortality in the United States at the county level. The data is from 2015, which is inherently a limiting factor in its generalizability to future cancer mortality rates. Also, the 64.1% test accuracy score shows that a stronger model could be built with the acquisition of other county-level data. However, the process of increasing the test accuracy score by 15 percentage points from its original form through the addition of features from external data sets, through feature engineering, and through the creation of logarithmic and exponential versions of features draws a solid blueprint for the building of future models of cancer mortality in the United States (and potentially other countries).

The model was built off of data from multiple sources. The original data set of socioeconomic indicators and cancer mortality rates for 2015 of 3,047 out of 3,141 total U.S. counties and county equivalents (97%) came pre-assembled as a data science challenge from the website data.world. Latitude/longitude centroids for the 3,047 counties were mostly downloaded from the Census' 2019 Gazetteer dataset, and a few were necessarily found via Google search. The top 10 rated oncology hospitals were looked up and their latitude/longitude locations were found via Google search, and the latitude/longitude location of eight major regional urban centers were found via Google search. The latitude/longitude location of all EPA-designated Superfund cleanup sites were downloaded from the EPA's website, and a large set of food environment and disease rate features were downloaded from the USDA's Food Environment Atlas. Logarithmic and exponential transformations of the feature set were tested for their contribution to the overall model's accuracy, and added to the feature set if they increased this accuracy.

After identifying a series of features that had particularly strong correlations (detailed further below), a series of hypothesis tests were developed about cancer mortality rates. All tests were t-tests of null hypotheses. The null hypotheses were:

- The differing cancer mortality rates seen in majority White counties and majority Black counties was due to chance. Majority White counties had an average cancer mortality rate of 179 deaths per 100,000 people, while majority Black counties had an average rate of 202 deaths per 100,000 people.
- The differing cancer mortality rates seen in counties where the majority of the populace has private health insurance and counties where the majority of the populace has public health insurance is due to chance. Majority private health insurance counties had an average cancer mortality rate of 177 deaths per 100,000 people, while majority public health insurance counties had an average cancer mortality rate of 199 deaths per 100,000 people.
- 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. Counties with a high unemployment rate had a cancer mortality rate of 188 out of 100,000 people, while counties with a low unemployment rate had a cancer mortality rate of 169 out of 100,000 people.
- 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, while the counties with a median income above the national median had a cancer mortality rate of 168 per 100,000 people.
- 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. The "high school counties" had a cancer mortality rate of 188 out of 100,000 people, while "college degree counties" had a cancer mortality rate of 168 out of 100,000 people.

All hypotheses were rejected at a p-value of less than 1%, warranting further research into the causes of these differences in cancer mortality rates.

The distribution of cancer mortality rates was split into quintiles and visualized. Although counties in the top quintile were in every region, they were highly concentrated in the Southern states in the Eastern half of the country, as well as in eastern Midwestern states.

The best performing regression model was discovered through an iterative exploration of scaling and hyperparameter tuning of a set of regression algorithms. After experimenting with unscaled and scaled data, as well as with Ordinary Least Squares (OLS) Linear Regression, Ridge Regression, LASSO Regression, ElasticNet Regression, Stochastic Gradient Descent Regression, Kernel Ridge Regression, and Random Forest, a hyperparameter tuning process uncovered that the best performing regression model used Ridge Regression with a regularization 'alpha' parameter of 0.001. This model had a training accuracy of 0.6465, a training Root Mean Squared Error (RMSE) of 16.59, a test accuracy of 0.6408, and a test RMSE of 16.2.

After discovering this best performing regression model, its coefficients were called to explore the features which had the strongest relationship with the cancer mortality target variable, in either a positive or negative direction. In order to do this, the feature set was normalized due to the fact that features in the feature set were in different scales - they were measured in percentages, per capita rates (per 100,000 people), binary flags, and in real numbers.

Intriguingly, the model often delivered Ridge Regression coefficients that differed from the Pearson's correlation coefficients, even in whether they had a relationship with increased or decreased cancer mortality. This dynamic was explored through a series of modifications made to the best performing regression model, but it seemed to be a fundamental characteristic of the model. Therefore, when taken in isolation, the best performing regression model shows that individual features can have very different relationships with cancer mortality when they are included in a complete systemic model than when they are analyzed in isolation.

Although the Visual EDA section above discusses the Pearson's correlation coefficients for a subset of the feature set, the table below details the features that have a correlation coefficient of greater than 0.3 (illustrating features that have a strong relationship with increased cancer mortality) or lower than -0.3 (illustrating features that have a strong relationship with decreased cancer mortality). Public health insurance, the incidence rate of cancer diagnoses, diabetes, public health insurance, poverty, obesity, low levels of education, unemployment, having no car and low access to grocery stores, and distance from Seattle and San Francisco all had a strong relationship with increased cancer rates. High levels of education, higher income, employment, private health insurance, distance from Atlanta, and distances from a few top 10 oncology hospitals all had a strong relationship with decreased cancer rates.

Feature Name	**Feature Description**	**Pearson's Correlation Coefficient**
PCT_DIABETES_ADULTS10	Percentage of adult county residents in 2010 who had diabetes	0.528
incidenceRate	Mean per capita (100,000) cancer diagnoses	0.449
PctPublicCoverageAlone	Percent of county residents with government-provided health coverage alone	0.449
CHILDPOVRATE10	Child poverty rate, 2010	0.445
povertyPercent	Percent of populace in poverty	0.429
PCT_OBESE_ADULTS13	Adult obesity rate, 2013	0.429
PctHS25_Over	Percent of county residents ages 25 and over highest education attained: high school diploma	0.405
PctPublicCoverage	Percent of county residents with government-provided health coverage	0.405
PctUnemployed16_Over	Percent of county residents ages 16 and over unemployed	0.378
seattle_l1	L1 distance to Seattle	0.342
PCT_LACCESS_HHNV10	Households with no car & low access to store (%), 2010	0.314
seattle_l2	L2 distance to Seattle	0.312
san_fran_l2	L2 distance to San Francisco	0.303
hlmcc_l2	L2 distance to H. Lee Moffitt Cancer Center and Research Institute	-0.326
PctPrivateCoverageAlone	Percent of county residents with private health coverage alone (no public assistance)	-0.326
hlmcc_l1	L1 distance to H. Lee Moffitt Cancer Center and Research Institute	-0.328
atlanta_l2	L2 distance to Atlanta	-0.342
atlanta_l1	L1 distance to Atlanta	-0.345

PctPrivateCoverage	Percent of county residents with private health coverage	-0.386
PctEmployed16_Over	Percent of county residents ages 16 and over employed	-0.397
medIncome	Median income per county	-0.429
PctBachDeg25_Over	Percent of county residents ages 25 and over highest education attained: bachelor's degree	-0.486

Several features did not meet a correlation coefficient of greater than 0.3 or less than -0.3, but were still of interest. The percentage of county residents who identify as Black or African-American had a relationship with increased cancer mortality (correlation coefficient of 0.257), while the percentage of county residents who identify as White, Asian, or another race all had a relationship with decreased cancer mortality (correlation coefficients of -0.177, -0.186, and -0.19, respectively). Longitude had a positive correlation of 0.263, showing that the further east a county is, the higher its cancer mortality rate was in 2015. The percentage of married households in a county also had a notable negative correlation of -0.293, showing that being married has a relationship with a lower risk of cancer mortality.

The Ridge Regression coefficients provided a differing set of insights about the relationships between the predictive feature set and cancer mortality (as described in detail in the “Discussion of Best Performing Regressor’s Coefficients” section). In looking at non-transformed features that had a relationship with increasing cancer rates and Ridge Regression coefficients greater than 10, one sees several features describing the state that each county is in, as well as a single feature describing the number of recreational facilities in a county and two features describing missing values for the number of farmer’s markets and for the ERS Natural Amenity Index. The states associated with an increased cancer mortality rate were Nevada, Alaska, California, Florida, Arizona, Utah, Oklahoma and also Washington, D.C. Logarithmic transformations of features who had Ridge Regression coefficients greater than 10 included private and public health insurance features, a poverty-related feature, and a feature describing average household size. Non-transformed features with a negative coefficient of less than -10 included again the state the county was in, as well as the per capita number of recreational facilities in each county. The states associated with a decreased cancer mortality rate were Hawaii, Rhode Island, Connecticut, Iowa, Illinois, North Carolina, Ohio, Michigan and New Hampshire. Logarithmic transformations of features who had Ridge Regression coefficients less than -10 included a feature describing private health insurance rates, the median age of the populace, the percentage of the populace who are married, the percentage of children in poverty, and the percentage of people 16 years and older who are employed.

As described in the “Relationship of ‘Feature Families’ to the Target Variable of Per-Capita Cancer Mortality” section above, features that had to do with the same topic were grouped into “feature families” to look at the predictive value of features at a more abstract level of analysis. All positive coefficients were summed, and then the coefficient sum of each “feature family” was divided by the overall positive coefficient sum to calculate the proportion that each “feature family” had of all

relationships involved with increasing cancer mortality. This process was repeated for all negative coefficients. Therefore, some “feature families” had members in both the positive coefficient pool as well as the negative coefficient pool.

In descending order of proportion of the total positive coefficient sum, the “feature families” associated with an increased cancer mortality rate were the U.S. state the county is in, health insurance, missing value features, average household size, distances to top 10 oncology hospitals, recreation and fitness, poverty, food environment, income, education, distances to major urban centers, population loss, an erroneous data indicator, comorbid health conditions, metropolitan indicator, environmental health, employment, age, marriage, race, cancer diagnoses, clinical cancer trials and total land mass of each county. The vast majority of this proportion came from the U.S. state and the type of health insurance most prevalent in each county. This truly underscores the importance of proper health insurance in reducing cancer mortality.

In descending order of proportion of the total negative coefficient sum, the “feature families” associated with an decreased cancer mortality rate were the U.S. state the county is in, employment, missing value features, health insurance, recreation and fitness, marriage, age, distances to top 10 oncology hospitals, food environment, poverty, distances to major urban centers, income, average household size, an erroneous data indicator, latitude and longitude, environmental health, quality of life, comorbid health conditions, race, birth rate, education, cancer diagnoses, water area, and population. The vast majority of this proportion came from the U.S. state that each county is in and employment levels.

This project, although imperfect, built a moderately strong model of cancer mortality at the county level in the year 2015. The feature engineering undertaken with this model illustrates the importance of merging multiple data sources and creating logarithmic and exponential transformations of one’s feature set in order to capture nonlinear relationships inside of linear regressors. Although a more comprehensive feature set and complete data sources from the current year would be helpful, the model built in this project serves as a powerful blueprint of how to help predict cancer mortality through socioeconomic and geographic features of our world. In doing so, it has helped highlight the roles that health insurance, comorbid health conditions, fitness, low access to grocery stores, poverty, race, income, education and geographical location play in increasing or decreasing cancer mortality. The findings of this project can support two goals: helping to build stronger models of cancer mortality while providing hard data that can provide guidance for policy makers to design public health interventions to reduce cancer mortality. This project is another example of how machine learning can allow us to break down the totality of what we know about a phenomenon and outline the differential influences that its features have on important issues and problems central to the challenges our society faces.

References:

Sim, J., Kim, Y., Kim, J., Lee, J., Kim, M., Shim, Y., Zo, J., Yun, Y. (2020, July). The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Scientific Reports 2020; 10: 10693.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7329866/

Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., Matheson, L., Ashley, D., Venkatesh, S. (2014, March). Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open. 2014; 4(3): e004007. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3963101/

SciKit-Learn documentation on SGD Regressor: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor

Appendix A: Full Data Dictionary

Features that are included in the EDA part of the analysis are marked with an asterisk.

Variable	**Definition**	**Data Type**	**Had Nulls?**
TARGET_deathRate	Dependent variable. Mean per capita (100,000) cancer mortalities	Continuous	No
avgAnnCount*	Mean number of reported cases of cancer diagnosed annually	Continuous	No
incidenceRate*	Mean per capita (100,000) cancer diagnoses	Continuous	No
medIncome*	Median income per county	Continuous	No
popEst2015*	Population of county	Continuous	No
povertyPercent*	Percent of populace in poverty	Continuous	No
studyPerCap*	Per capita number of cancer-related clinical trials per county	Continuous	No
MedianAge*	Median age of county residents	Continuous	No
MedianAgeMale*	Median age of male county residents	Continuous	No
MedianAgeFemale*	Median age of female county residents	Continuous	No
AvgHouseholdSize*	Average Household Size (occupied buildings)	Continuous	No
PercentMarried*	Percent of county residents who are married	Continuous	No
PctNoHS18_24*	Percent of county residents ages	Continuous	No

	18-24 highest education attained: less than high school		
PctHS18_24*	Percent of county residents ages 18-24 highest education attained: high school diploma	Continuous	No
PctSomeCol18_24*	Percent of county residents ages 18-24 highest education attained: some college	Continuous	Yes
PctBachDeg18_24*	Percent of county residents ages 18-24 highest education attained: bachelor's degree	Continuous	No
PctHS25_Over*	Percent of county residents ages 25 and over highest education attained: high school diploma	Continuous	No
PctBachDeg25_Over*	Percent of county residents ages 25 and over highest education attained: bachelor's degree	Continuous	No
PctEmployed16_Over*	Percent of county residents ages 16 and over employed	Continuous	Yes
PctUnemployed16_Over*	Percent of county residents ages 16 and over unemployed	Continuous	No
PctPrivateCoverage*	Percent of county residents with private health coverage	Continuous	No
PctPrivateCoverageAlone*	Percent of county residents with private health coverage alone (no public assistance)	Continuous	Yes
PctEmpPrivCoverage*	Percent of county residents with employee-provided private health coverage	Continuous	No
PctPublicCoverage*	Percent of county residents with government-provided health coverage	Continuous	No
PctPublicCoverageAlone*	Percent of county residents with government-provided health coverage alone	Continuous	No
PctWhite*	Percent of county residents who	Continuous	No

	identify as White		
PctBlack*	Percent of county residents who identify as Black	Continuous	No
PctAsian*	Percent of county residents who identify as Asian	Continuous	No
PctOtherRace*	Percent of county residents who identify in a category which is not White, Black, or Asian	Continuous	No
PctMarriedHouseholds*	Percent of married households	Continuous	No
BirthRate*	Number of live births relative to number of women in county	Continuous	No
PctSomeCol18_24_isnull	Missing value indicator for 'PctSomeCol18_24' feature	Binary	No
PctEmployed16_Over_isnull	Missing value indicator for 'PctEmployed16_Over' feature	Binary	No
PctPrivateCoverageAlone_isnull	Missing value indicator for 'PctPrivateCoverageAlone' feature	Binary	No
age_gt_100	Indicator that erroneous data in 'MedianAge' was corrected by replacing it with its mother feature's median	Binary	No
household_lt_1	Indicator that erroneous data in 'AvgHouseholdSize' was corrected by replacing it with its mother feature's median	Binary	No
binnedInc	Median Income per capita binned by decile, one binarized feature for each decile	String, binarized for Machine Learning	No
States	Individual binarized features for each of the 50 U.S. states	Binary	No
ALAND_SQMI	Land Area Square Mileage	Continuous	No
AWATER_SQMI	Water Area Square Mileage	Continuous	No
INTPTLAT*	Latitude	Continuous	No

INTPTLONG*	Longitude	Continuous	No
utmda_l1	L1 distance to University of Texas MD Anderson Cancer Center	Continuous	No
mskcc_l1	L1 distance to Memorial Sloan Kettering Cancer Center	Continuous	No
mayo_l1	L1 distance to Mayo Clinic	Continuous	No
hopkins_l1	L1 distance to Johns Hopkins Hospital	Continuous	No
dfb_l1	L1 distance to Dana Farber/Brigham and Women's Cancer Center	Continuous	No
cleveland_l1	L1 distance to Cleveland Clinic	Continuous	No
upmcps_l1	L1 distance to UPMC Presbyterian Shadyside	Continuous	No
hlmcc_l1	L1 distance to H. Lee Moffitt Cancer Center and Research Institute	Continuous	No
mgs_l1	L1 distance to Massachusetts General Hospital	Continuous	No
nw_mem_l1	L1 distance to Northwestern Memorial Hospital	Continuous	No
chi_l1	L1 distance to Chicago	Continuous	No
nyc_l1	L1 distance to New York City	Continuous	No
atlanta_l1	L1 distance to Atlanta	Continuous	No
dallas_l1	L1 distance to Dallas	Continuous	No
denver_l1	L1 distance to Denver	Continuous	No
los_ang_l1	L1 distance to Los Angeles	Continuous	No
seattle_l1	L1 distance to Seattle	Continuous	No
san_fran_l1	L1 distance to San Francisco	Continuous	No
utmda_l2	L2 distance to University of Texas MD Anderson Cancer Center	Continuous	No
mskcc_l2	L2 distance to Memorial Sloan	Continuous	No

	Kettering Cancer Center		
mayo_l2	L2 distance to Mayo Clinic	Continuous	No
hopkins_l2	L2 distance to Johns Hopkins Hospital	Continuous	No
dfb_l2	L2 distance to Dana Farber/Brigham and Women's Cancer Center	Continuous	No
cleveland_l2	L2 distance to Cleveland Clinic	Continuous	No
upmcps_l2	L2 distance to UPMC Presbyterian Shadyside	Continuous	No
hlmcc_l2	L2 distance to H. Lee Moffitt Cancer Center and Research Institute	Continuous	No
mgs_l2	L2 distance to Massachusetts General Hospital	Continuous	No
nw_mem_l2	L2 distance to Northwestern Memorial Hospital	Continuous	No
chi_l2	L2 distance to Chicago	Continuous	No
nyc_l2	L2 distance to New York City	Continuous	No
atlanta_l2	L2 distance to Atlanta	Continuous	No
dallas_l2	L2 distance to Dallas	Continuous	No
denver_l2	L2 distance to Denver	Continuous	No
los_ang_l2	L2 distance to Los Angeles	Continuous	No
seattle_l2	L2 distance to Seattle	Continuous	No
san_fran_l2	L2 distance to San Francisco	Continuous	No
onc_min_distsl1*	L1 distance to closest top 10 oncology hospital	Continuous	No
onc_min_distsl2*	L2 distance to closest top 10 oncology hospital	Continuous	No
city_min_distsl1*	L1 distance to closest regional urban center	Continuous	No
city_min_distsl2*	L2 distance to closest regional urban	Continuous	No

	center		
sc_min_dists_l1*	L1 distance to closest EPA Superfund Cleanup Site	Continuous	No
sc_min_dists_l2*	L2 distance to closest EPA Superfund Cleanup Site	Continuous	No
PCT_LACCESS_POP10*	Population, low access to grocery store (%), 2010	Continuous	Yes
PCT_LACCESS_LOWI10*	Low income & low access to grocery store (%), 2010	Continuous	Yes
PCT_LACCESS_CHILD10	Children, low access to store (%), 2010	Continuous	Yes
PCT_LACCESS_SENIORS10	Seniors, low access to store (%), 2010	Continuous	Yes
PCT_LACCESS_HHNV10*	Households, no car & low access to store (%), 2010	Continuous	Yes
PCT_LACCESS_POP10_isnull	Missing value indicator for 'PCT_LACCESS_POP10' feature	Binary	No
PCT_LACCESS_LOWI10_isnull	Missing value indicator for 'PCT_LACCESS_LOWI10' feature	Binary	No
PCT_LACCESS_CHILD10_isnull	Missing value indicator for 'PCT_LACCESS_CHILD10' feature	Binary	No
PCT_LACCESS_SENIORS10_isnull	Missing value indicator for 'PCT_LACCESS_SENIORS10' feature	Binary	No
PCT_LACCESS_HHNV10_isnull	Missing value indicator for 'PCT_LACCESS_HHNV10' feature	Binary	No
FOODINSEC_00_02	Household food insecurity (%, three-year average), 2000-02	Continuous	Yes
FOODINSEC_07_09	Household food insecurity (%, three-year average), 2007-09	Continuous	Yes
FOODINSEC_10_12*	Household food insecurity (%, three-year average), 2010-12	Continuous	Yes
CH_FOODINSEC_02_12	Household food insecurity (change	Continuous	Yes

	%),2000-02 to 2010-12		
CH_FOODINSEC_09_12*	Household food insecurity (change %),2007-09 to 2010-12	Continuous	Yes
VLFOODSEC_00_02	Household very low food security (%, three-year average), 2000-02	Continuous	Yes
VLFOODSEC_07_09	Household very low food security (%, three-year average), 2007-09	Continuous	Yes
VLFOODSEC_10_12	Household very low food security (%, three-year average), 2010-12	Continuous	Yes
CH_VLFOODSEC_02_12	Household very low food security (change %),2000-02 to 2010-12	Continuous	Yes
CH_VLFOODSEC_09_12	Household very low food security (change %),2007-09 to 2010-12	Continuous	Yes
FOODINSEC_CHILD_01_07	Child food insecurity (% households, multiple-year average), 2001-07	Continuous	Yes
FOODINSEC_CHILD_03_11	Child food insecurity (% households, multiple-year average), 2003-11	Continuous	Yes
FOODINSEC_00_02_isnull	Missing value indicator for 'FOODINSEC_00_02' feature	Binary	No
FOODINSEC_07_09_isnull	Missing value indicator for 'FOODINSEC_07_09' feature	Binary	No
FOODINSEC_10_12_isnull	Missing value indicator for 'FOODINSEC_10_12' feature	Binary	No
CH_FOODINSEC_02_12_isnull	Missing value indicator for 'CH_FOODINSEC_02_12' feature	Binary	No
CH_FOODINSEC_09_12_isnull	Missing value indicator for 'CH_FOODINSEC_09_12' feature	Binary	No
VLFOODSEC_00_02_isnull	Missing value indicator for 'VLFOODSEC_00_02' feature	Binary	No
VLFOODSEC_07_09_isnull	Missing value indicator for 'VLFOODSEC_07_09' feature	Binary	No
VLFOODSEC_10_12_isnull	Missing value indicator for 'VLFOODSEC_10_12' feature	Binary	No

CH_VLFOODSEC_02_12_isnull	Missing value indicator for 'CH_VLFOODSEC_02_12' feature	Binary	No
CH_VLFOODSEC_09_12_isnull	Missing value indicator for 'CH_VLFOODSEC_09_12' feature	Binary	No
FOODINSEC_CHILD_01_07_isnull	Missing value indicator for 'FOODINSEC_CHILD_01_07' feature	Binary	No
FOODINSEC_CHILD_03_11_isnull	Missing value indicator for 'FOODINSEC_CHILD_03_11' feature	Binary	No
PCT_LOCLFARM07*	Farms with direct sales (%), 2007	Continuous	Yes
PCT_LOCLSALE07	Direct farm sales (%), 2007	Continuous	Yes
PC_DIRSALES07	Direct farm sales per capita, 2007	Continuous	Yes
FMRKT09	Farmers' markets, 2009	Continuous	Yes
FMRKT13*	Farmers' markets, 2013	Continuous	Yes
PCH_FMRKT_09_13*	Farmers' markets (% change), 2009-13	Continuous	Yes
FMRKTPTH09	Farmers' markets/1,000 pop, 2009	Continuous	Yes
FMRKTPTH13	Farmers' markets/1,000 pop, 2013	Continuous	Yes
PCH_FMRKTPTH_09_13	Farmers' markets/1,000 pop (% change), 2009-13	Continuous	Yes
PCT_FMRKT_SNAP13	Farmers' markets that report accepting SNAP (%), 2013	Continuous	Yes
PCT_FMRKT_WIC13	Farmers' markets that report accepting WIC (%), 2013	Continuous	Yes
PCT_FMRKT_WICCASH13	Farmers' markets that report accepting WIC Cash (%), 2013	Continuous	Yes
PCT_FMRKT_SFMNP13	Farmers' markets that report accepting SFMNP (%), 2013	Continuous	Yes
PCT_FRMKT_FRVEG13	Farmers' markets that report selling fruit & vegetables (%), 2013	Continuous	Yes
PCT_FRMKT_ANMLPROD13	Farmers' markets that report selling	Continuous	Yes

	animal products (%), 2013		
PCT_FMRKT_OTHER13	Farmers' markets that report selling other products (%), 2013	Continuous	Yes
VEG_FARMS07	Vegetable farms, 2007	Continuous	Yes
VEG_ACRES07	Vegetable acres harvested, 2007	Continuous	Yes
VEG_ACRESPTH07	Vegetable acres harvested/1,000 pop, 2007	Continuous	Yes
FRESHVEG_FARMS07*	Farms with vegetables harvested for fresh market, 2007	Continuous	Yes
FRESHVEG_ACRES07	Vegetable acres harvested for fresh market, 2007	Continuous	Yes
FRESHVEG_ACRESPTH07	Vegetable acres harvested for fresh market/1,000 pop, 2007	Continuous	Yes
ORCHARD_FARMS07*	Orchard farms, 2007	Continuous	Yes
ORCHARD_ACRES07	Orchard acres, 2007	Continuous	Yes
ORCHARD_ACRESPTH07	Orchard acres/1,000 pop, 2007	Continuous	Yes
BERRY_FARMS07	Berry farms, 2007	Continuous	Yes
BERRY_ACRES07	Berry acres, 2007	Continuous	Yes
BERRY_ACRESPTH07	Berry acres/1,000 pop, 2007	Continuous	Yes
SLHOUSE07	Small slaughterhouse facilities, 2007	Continuous	Yes
GHVEG_FARMS07*	Greenhouse vegetable and fresh herb farms, 2007	Continuous	Yes
GHVEG_SQFT07	Greenhouse veg and fresh herb sq feet, 2007	Continuous	Yes
GHVEG_SQFTPTH07	Greenhouse veg and fresh herb sq feet/1,000 pop, 2007	Continuous	Yes
FOODHUB12	Food hubs, 2012	Continuous	Yes
CSA07	CSA farms, 2007	Continuous	Yes
AGRITRSM_OPS07	Agritourism operations	Continuous	Yes

AGRITRSM_RCT07	Agritourism receipts	Continuous	Yes
FARM_TO_SCHOOL	Farm to school program, 2009	Continuous	Yes
PCT_LOCLFARM07_isnull	Missing value indicator for 'PCT_LOCLFARM07' feature	Binary	No
PCT_LOCLSALE07_isnull	Missing value indicator for 'PCT_LOCLSALE07' feature	Binary	No
PC_DIRSALES07_isnull	Missing value indicator for 'PC_DIRSALES07' feature	Binary	No
FMRKT09_isnull	Missing value indicator for 'FMRKT09' feature	Binary	No
FMRKT13_isnull	Missing value indicator for 'FMRKT13' feature	Binary	No
PCH_FMRKT_09_13_isnull	Missing value indicator for 'PCH_FMRKT_09_13' feature	Binary	No
FMRKTPTH09_isnull	Missing value indicator for 'FMRKTPTH09' feature	Binary	No
FMRKTPTH13_isnull	Missing value indicator for 'FMRKTPTH13' feature	Binary	No
PCH_FMRKTPTH_09_13_isnull	Missing value indicator for 'PCH_FMRKTPTH_09_13' feature	Binary	No
PCT_FMRKT_SNAP13_isnull	Missing value indicator for 'PCT_FMRKT_SNAP13' feature	Binary	No
PCT_FMRKT_WIC13_isnull	Missing value indicator for 'PCT_FMRKT_WIC13' feature	Binary	No
PCT_FMRKT_WICCASH13_isnull	Missing value indicator for 'PCT_FMRKT_WICCASH13' feature	Binary	No
PCT_FMRKT_SFMNP13_isnull	Missing value indicator for 'PCT_FMRKT_SFMNP13' feature	Binary	No
PCT_FRMKT_FRVEG13_isnull	Missing value indicator for 'PCT_FRMKT_FRVEG13' feature	Binary	No
PCT_FRMKT_ANMLPROD13_isnull	Missing value indicator for 'PCT_FRMKT_ANMLPROD13' feature	Binary	No

PCT_FMRKT_OTHER13_isnull	Missing value indicator for 'PCT_FMRKT_OTHER13' feature	Binary	No
VEG_FARMS07_isnull	Missing value indicator for 'VEG_FARMS07' feature	Binary	No
VEG_ACRES07_isnull	Missing value indicator for 'VEG_ACRES07' feature	Binary	No
VEG_ACRESPTH07_isnull	Missing value indicator for 'VEG_ACRESPTH07' feature	Binary	No
FRESHVEG_FARMS07_isnull	Missing value indicator for 'FRESHVEG_FARMS07' feature	Binary	No
FRESHVEG_ACRES07_isnull	Missing value indicator for 'FRESHVEG_ACRES07' feature	Binary	No
FRESHVEG_ACRESPTH07_isnull	Missing value indicator for 'FRESHVEG_ACRESPTH07' feature	Binary	No
ORCHARD_FARMS07_isnull	Missing value indicator for 'ORCHARD_FARMS07' feature	Binary	No
ORCHARD_ACRES07_isnull	Missing value indicator for 'ORCHARD_ACRES07' feature	Binary	No
ORCHARD_ACRESPTH07_isnull	Missing value indicator for 'ORCHARD_ACRESPTH07' feature	Binary	No
BERRY_FARMS07_isnull	Missing value indicator for 'BERRY_FARMS07' feature	Binary	No
BERRY_ACRES07_isnull	Missing value indicator for 'BERRY_ACRES07' feature	Binary	No
BERRY_ACRESPTH07_isnull	Missing value indicator for 'BERRY_ACRESPTH07' feature	Binary	No
SLHOUSE07_isnull	Missing value indicator for 'SLHOUSE07' feature	Binary	No
GHVEG_FARMS07_isnull	Missing value indicator for 'GHVEG_FARMS07' feature	Binary	No
GHVEG_SQFT07_isnull	Missing value indicator for 'GHVEG_SQFT07' feature	Binary	No
GHVEG_SQFTPTH07_isnull	Missing value indicator for	Binary	No

	'GHVEG_SQFTPTH07' feature		
FOODHUB12_isnull	Missing value indicator for 'FOODHUB12' feature	Binary	No
CSA07_isnull	Missing value indicator for 'CSA07' feature	Binary	No
AGRITRSM_OPS07_isnull	Missing value indicator for 'AGRITRSM_OPS07' feature	Binary	No
AGRITRSM_RCT07_isnull	Missing value indicator for 'AGRITRSM_RCT07' feature	Binary	No
FARM_TO_SCHOOL_isnull	Missing value indicator for 'FARM_TO_SCHOOL' feature	Binary	No
PCT_DIABETES_ADULTS09	Adult diabetes rate, 2009	Continuous	No
PCT_DIABETES_ADULTS10*	Adult diabetes rate, 2010	Continuous	No
PCT_OBESE_ADULTS09	Adult obesity rate (county), 2009	Continuous	No
PCT_OBESE_ADULTS10	Adult obesity rate (county), 2010	Continuous	No
PCT_OBESE_ADULTS13*	Adult obesity rate, 2013	Continuous	No
PCT_OBESE_CHILD08	Low-income preschool obesity rate, 2006-08	Continuous	Yes
PCT_OBESE_CHILD11	Low-income preschool obesity rate, 2009-11	Continuous	Yes
PCH_OBESE_CHILD_08_11	Low-income preschool obesity rate (% change), 2006-08 to 2009-11	Continuous	Yes
PCT_HSPA09	High schoolers physically active (%), 2009	Continuous	Yes
RECFAC07*	Recreation & fitness facilities, 2007	Continuous	No
RECFAC12*	Recreation & fitness facilities, 2012	Continuous	No
PCH_RECFAC_07_12*	Recreation & fitness facilities (% change), 2007-12	Continuous	Yes
RECFACPTH07	Recreation & fitness facilities/1,000 pop, 2007	Continuous	No
RECFACPTH12	Recreation & fitness facilities/1,000	Continuous	No

	pop, 2012		
PCH_RECFACPTH_07_12	Recreation & fitness facilities/1,000 pop (% change), 2007-12	Continuous	Yes
NATAMEN*	ERS natural amenity index, 1999	Continuous	Yes
PCT_OBESE_CHILD08_isnull	Missing value indicator for 'PCT_OBESE_CHILD08' feature	Binary	No
PCT_OBESE_CHILD11_isnull	Missing value indicator for 'PCT_OBESE_CHILD11' feature	Binary	No
PCH_OBESE_CHILD_08_11_isnull	Missing value indicator for 'PCH_OBESE_CHILD_08_11' feature	Binary	No
PCT_HSPA09_isnull	Missing value indicator for 'PCT_HSPA09' feature	Binary	No
PCH_RECFAC_07_12_isnull	Missing value indicator for 'PCH_RECFAC_07_12' feature	Binary	No
PCH_RECFACPTH_07_12_isnull	Missing value indicator for 'PCH_RECFACPTH_07_12' feature	Binary	No
NATAMEN_isnull	Missing value indicator for 'NATAMEN' feature	Binary	No
PERPOV10*	Persistent-poverty counties, 2010	Binary	No
CHILDPOVRATE10*	Child poverty rate, 2010	Continuous	No
PERCHLDPOV10*	Persistent-child-poverty counties, 2010	Binary	No
METRO13*	Metro/nonmetro counties, 2010	Binary	No
POPLOSS00*	Population-loss counties, 2000	Binary	No
povertyPercent_log*	Percent of populace in poverty LOG	Continuous	No
povertyPercent_sqrd*	Percent of populace in poverty SQUARED	Continuous	No
MedianAge_log*	Median age of county residents LOG	Continuous	No
MedianAgeFemale_sqrd*	Median age of female county residents SQUARED	Continuous	No

AvgHouseholdSize_log*	Average Household Size (occupied buildings) LOG	Continuous	No
PercentMarried_log*	Percent of county residents who are married LOG	Continuous	No
PercentMarried_sqrd*	Percent of county residents who are married SQUARED	Continuous	No
PctSomeCol18_24_log*	Percent of county residents ages 18-24 highest education attained: some college LOG	Continuous	No
PctSomeCol18_24_sqrd*	Percent of county residents ages 18-24 highest education attained: some college SQRD	Continuous	No
PctHS25_Over_sqrd*	Percent of county residents ages 25 and over highest education attained: high school diploma SQUARED	Continuous	No
PctBachDeg25_Over_log*	Percent of county residents ages 25 and over highest education attained: bachelor's degree	Continuous	No
PctEmployed16_Over_log*	Percent of county residents ages 16 and over employed LOG	Continuous	No
PctEmployed16_Over_sqrd*	Percent of county residents ages 16 and over employed SQUARED	Continuous	No
PctPrivateCoverage_log	Percent of county residents with private health coverage LOG	Continuous	No
PctEmpPrivCoverage_log	Percent of county residents with employee-provided private health coverage LOG	Continuous	No
PctPublicCoverage_log*	Percent of county residents with government-provided health coverage LOG	Continuous	No
PctPublicCoverageAlone_log*	Percent of county residents with government-provided health coverage alone LOG	Continuous	No
PctPublicCoverageAlone_sqrd*	Percent of county residents with government-provided health coverage alone SQUARED	Continuous	No

PctWhite_sqrd*	Percent of county residents who identify as White SQUARED	Continuous	No
PctBlack_sqrd*	Percent of county residents who identify as Black SQUARED	Continuous	No
INTPTLONG_sqrd*	Longitude SQUARED	Continuous	No
mskcc_l1_log*	L1 distance to Memorial Sloan Kettering Cancer Center LOG	Continuous	No
mayo_l1_log*	L1 distance to Mayo Clinic LOG	Continuous	No
mayo_l1_sqrd*	L1 distance to Mayo Clinic SQUARED	Continuous	No
dfb_l1_log*	L1 distance to Dana Farber/Brigham and Women's Cancer Center LOG	Continuous	No
dfb_l1_sqrd*	L1 distance to Dana Farber/Brigham and Women's Cancer Center SQUARED	Continuous	No
cleveland_l1_log*	L1 distance to Cleveland Clinic LOG	Continuous	No
cleveland_l1_sqrd*	L1 distance to Cleveland Clinic SQUARED	Continuous	No
upmcps_l1_log*	L1 distance to UPMC Presbyterian Shadyside LOG	Continuous	No
mgs_l1_log*	L1 distance to Massachusetts General Hospital LOG	Continuous	No
atlanta_l1_log*	L1 distance to Atlanta LOG	Continuous	No
denver_l1_sqrd*	L1 distance to Denver SQUARED	Continuous	No
los_ang_l1_sqrd*	L1 distance to Los Angeles SQUARED	Continuous	No
seattle_l1_log*	L1 distance to Seattle LOG	Continuous	No
hopkins_l2_log*	L2 distance to Johns Hopkins Hospital LOG	Continuous	No
dfb_l2_log*	L2 distance to Dana Farber/Brigham and Women's Cancer Center LOG	Continuous	No
cleveland_l2_log*	L2 distance to Cleveland Clinic LOG	Continuous	No

upmcps_l2_log*	L2 distance to UPMC Presbyterian Shadyside LOG	Continuous	No
mgs_l2_log*	L2 distance to Massachusetts General Hospital LOG	Continuous	No
atlanta_l2_log*	L2 distance to Atlanta LOG	Continuous	No
city_min_distsl1_sqrd*	L1 distance to closest regional urban center SQUARED	Continuous	No
sc_min_dists_l1_log*	L1 distance to closest EPA Superfund Cleanup Site LOG	Continuous	No
PCT_LACCESS_CHILD10_sqrd*	Children, low access to store (%), 2010 SQUARED	Continuous	No
PCT_LACCESS_HHNV10_sqrd*	PCT_LACCESS_HHNV10 SQUARED	Continuous	No
PC_DIRSALES07_sqrd*	Direct farm sales per capita, 2007 SQUARED	Continuous	No
FMRKT09_sqrd*	Farmers' markets, 2009	Continuous	No
PCT_FRMKT_FRVEG13_sqrd*	Farmers' markets that report selling fruit & vegetables (%), 2013	Continuous	No
PCT_OBESE_ADULTS13_log*	PCT_OBESE_ADULTS13 LOG	Continuous	No
PCT_OBESE_ADULTS13_sqrd*	PCT_OBESE_ADULTS13 SQUARED	Continuous	No
CHILDPOVRATE10_log*	Child poverty rate, 2010 LOG	Continuous	No

Appendix B: Nonlinear Contributions of L1 & L2 Distance Features to Linear Regression Accuracy

Feature Name	Feature Description	Improvement to Accuracy	Type of Improvement
mskcc_l1	L1 distance to Memorial Sloan Kettering Cancer Center	0.002	Logarithmic
mayo_l1	L1 distance to Mayo Clinic	0.00005	Both Logarithmic & Exponential
dfb_l1	L1 distance to Dana Farber/Brigham and Women's Cancer Center	0.0003	Both Logarithmic & Exponential
cleveland_l1	L1 distance to Cleveland Clinic	0.004	Both Logarithmic & Exponential
upmcps_l1	L1 distance to UPMC Presbyterian Shadyside	0.00005	Logarithmic
mgs_l1	L1 distance to Massachusetts General Hospital	0.0001	Logarithmic
atlanta_l1	L1 distance to Atlanta	0.002	Logarithmic

denver_l1	L1 distance to Denver	0.0009	Exponential
los_ang_l1	L1 distance to Los Angeles	0.0006	Exponential
seattle_l1	L1 distance to Seattle	0.0004	Logarithmic
hopkins_l2	L2 distance to Johns Hopkins Hospital	0.0003	Logarithmic
dfb_l2	L2 distance to Dana Farber/Brigham and Women's Cancer Center	0.0004	Logarithmic
cleveland_l2	L2 distance to Cleveland Clinic	0.00007	Logarithmic
upmcps_l2	L2 distance to UPMC Presbyterian Shadyside	0.0004	Logarithmic
mgs_l2	L2 distance to Massachusetts General Hospital	0.0004	Logarithmic
atlanta_l2	L2 distance to Atlanta	0.00005	Logarithmic

Appendix C: Correlations Between L1 and L2 Distance Features and Cancer Mortality

Feature Name	Feature Description	Correlation with Cancer Mortality
utmda_l1	L1 distance to University of Texas MD	-0.14

	Anderson Cancer Center	
mskcc_l1	L1 distance to Memorial Sloan Kettering Cancer Center	-0.22
mayo_l1	L1 distance to Mayo Clinic	-0.07
hopkins_l1	L1 distance to Johns Hopkins Hospital	-0.26
dfb_l1	L1 distance to Dana Farber/Brigham and Women's Cancer Center	-0.2
cleveland_l1	L1 distance to Cleveland Clinic	-0.25
upmcps_l1	L1 distance to UPMC Presbyterian Shadyside	-0.26
hlmcc_l1	L1 distance to H. Lee Moffitt Cancer Center and Research Institute	-0.33
mgs_l1	L1 distance to Massachusetts General Hospital	-0.2
nw_mem_l1	L1 distance to Northwestern Memorial Hospital	-0.23
chi_l1	L1 distance to Chicago	-0.23
nyc_l1	L1 distance to New York City	-0.22
atlanta_l1	L1 distance to Atlanta	-0.35
dallas_l1	L1 distance to Dallas	-0.1
denver_l1	L1 distance to Denver	0.24
los_ang_l1	L1 distance to Los Angeles	0.22
seattle_l1	L1 distance to Seattle	0.34
san_fran_l1	L1 distance to San Francisco	0.26
utmda_l2	L2 distance to University of Texas MD Anderson Cancer Center	-0.16
mskcc_l2	L2 distance to Memorial Sloan Kettering Cancer Center	-0.26
mayo_l2	L2 distance to Mayo Clinic	-0.09
hopkins_l2	L2 distance to Johns Hopkins Hospital	-0.28

dfb_l2	L2 distance to Dana Farber/Brigham and Women's Cancer Center	-0.25
cleveland_l2	L2 distance to Cleveland Clinic	-0.29
upmcps_l2	L2 distance to UPMC Presbyterian Shadyside	-0.29
hlmcc_l2	L2 distance to H. Lee Moffitt Cancer Center and Research Institute	-0.33
mgs_l2	L2 distance to Massachusetts General Hospital	-0.25
nw_mem_l2	L2 distance to Northwestern Memorial Hospital	-0.25
chi_l2	L2 distance to Chicago	-0.25
nyc_l2	L2 distance to New York City	-0.26
atlanta_l2	L2 distance to Atlanta	-0.34
dallas_l2	L2 distance to Dallas	-0.10
denver_l2	L2 distance to Denver	0.24
los_ang_l2	L2 distance to Los Angeles	0.29
seattle_l2	L2 distance to Seattle	0.31
san_fran_l2	L2 distance to San Francisco	0.3