Predicting Heart Disease Mortality

Andrew Dugan, July 2018

Overview

This report provides an analysis of heart disease mortality data across United States counties. It is based on observations of 3,198 counties, and the observations include area, economic, health, and demographic data for each county. The observations in this data were taken over the course of two years, and the year each observation was made is listed in the data. This data set is labeled with the rate of heart disease mortality per 100,000 residents in each county, and it was used for analysis and training a predictive model. A second set of values (unlabeled) from a different 3,198 counties was then run through the predictive model in order to predict what heart disease mortality rate each of those counties had per 100,000 residents.

First, this report describes and visualizes detailed factors that are correlated to heart disease mortality within the following categories for each county: area, economics, health, and demographics. Finally, the report describes the predictive model that was created to predict the rate of heart disease mortality per 100,000 residents in each U.S. county. The most effective model was determined to be a multivariate linear regression model, and the root-mean-square-error metric (RMSE), which was used to judge the model's accuracy, predicted with an RMSE of 32.2696. The goal of an RMSE metric is to achieve an accuracy score as close to zero as possible, and the degree of accuracy that the 32.2696 number represents will be discussed later in the "Regression" section of this report.

Through the analysis, the author found that some of the factors most significantly correlated with heart disease mortality are:

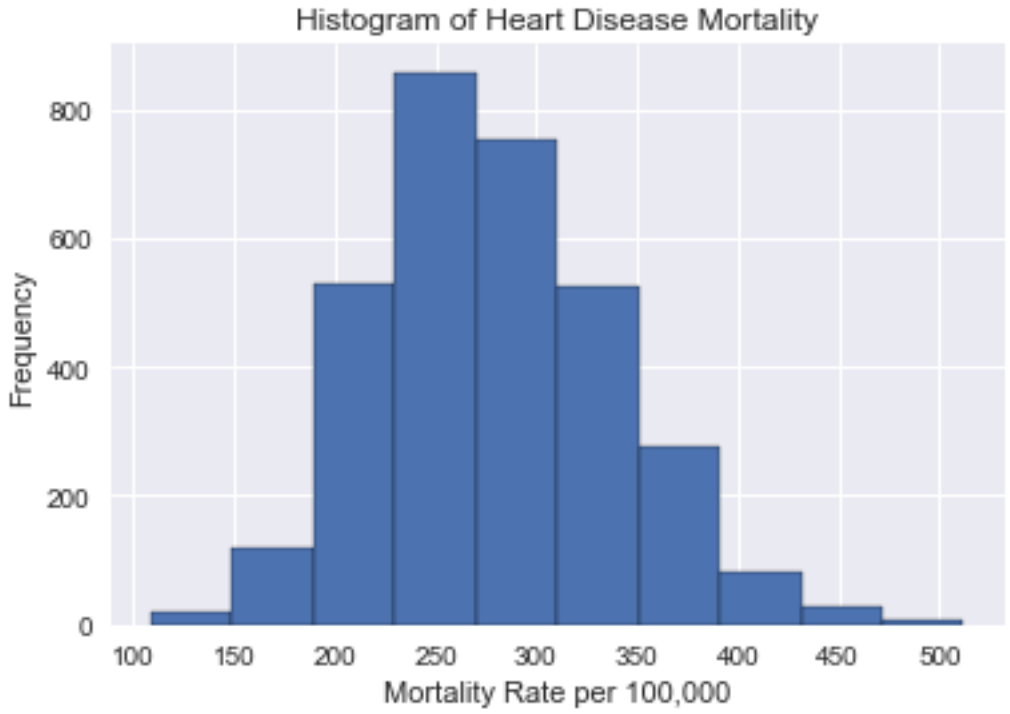
- Health factors - Adult obesity, smoking, diabetes, and physical inactivity all appear to have a positive correlation with heart disease mortality, while alcohol consumption appears to have little effect.
- Economic factors - Higher percentages of uninsured adults seem to have a positive correlation with heart disease mortality, while recreational forms of employment seem to have a negative correlation.
- Demographic factors – High school completion rates appear to have a positive correlation, while bachelor degree completion has a negative correlation. This could be contributed to other socio-economic factors that impact health or access to healthcare.

Data Exploration and Visualization

First, summary statistics, including mean, standard deviation, minimum, and maximum values were calculated for heart disease mortality rates per 100,000 residents. On average, approximately 279 people for every 100,000 residents die from heart disease annually. The highest rate for counties measured was 512 per 100,000.:

	heart_disease_mortality_per_100k
count	3198.000000
mean	279.369293
std	58.953338
min	109.000000
25%	237.000000
50%	275.000000
75%	317.000000
max	512.000000

The histogram below shows the frequency that the observed counties have mortality rates within the different ranges or “bins”. The heart disease mortality rate data is slightly skewed right.



The data was primarily made up of numeric features, such as percentage of the population descriptions; however, there were some categorical features such as:

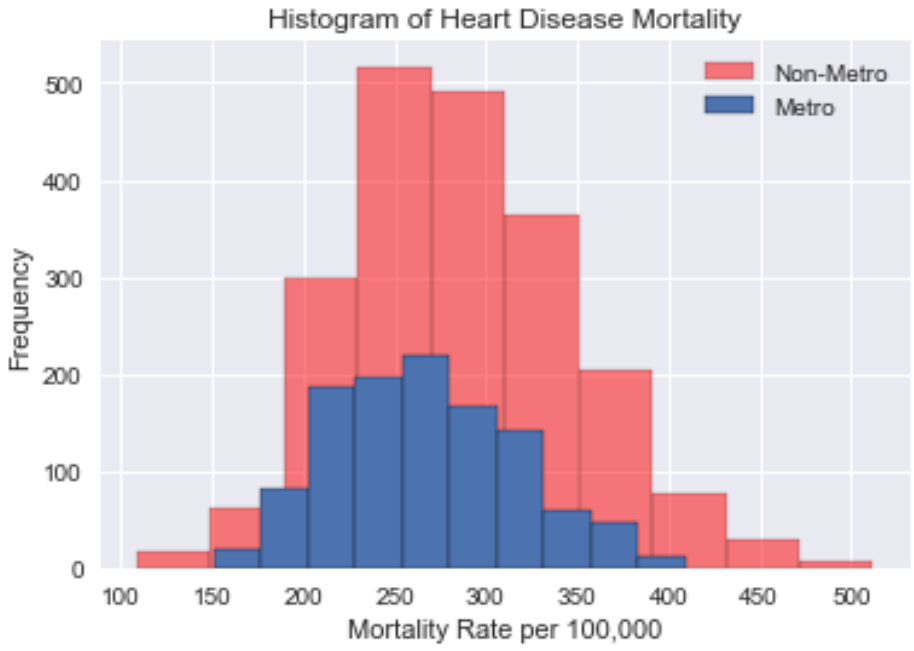
- Rural-Urban Continuum Codes – One of 9 codes used by the Office of Management and Budget to describe the degree of urbanization and adjacency to a metro area.
- Urban influence Codes – These classify metropolitan counties based on metro area population size and nonmetropolitan counties by the size of the largest city or proximity to metro/micropolitan areas.
- Economic Typology – Economic classifications that describe the counties' economic dependence, such as farming, mining, manufacturing, recreation, etc.

Area

In studying area, non-metro areas showed a slightly higher rate of heart disease mortality, according to their mean and median differences. The chart below shows on average, non-metro counties see 20 more mortalities per 100,000 residents.

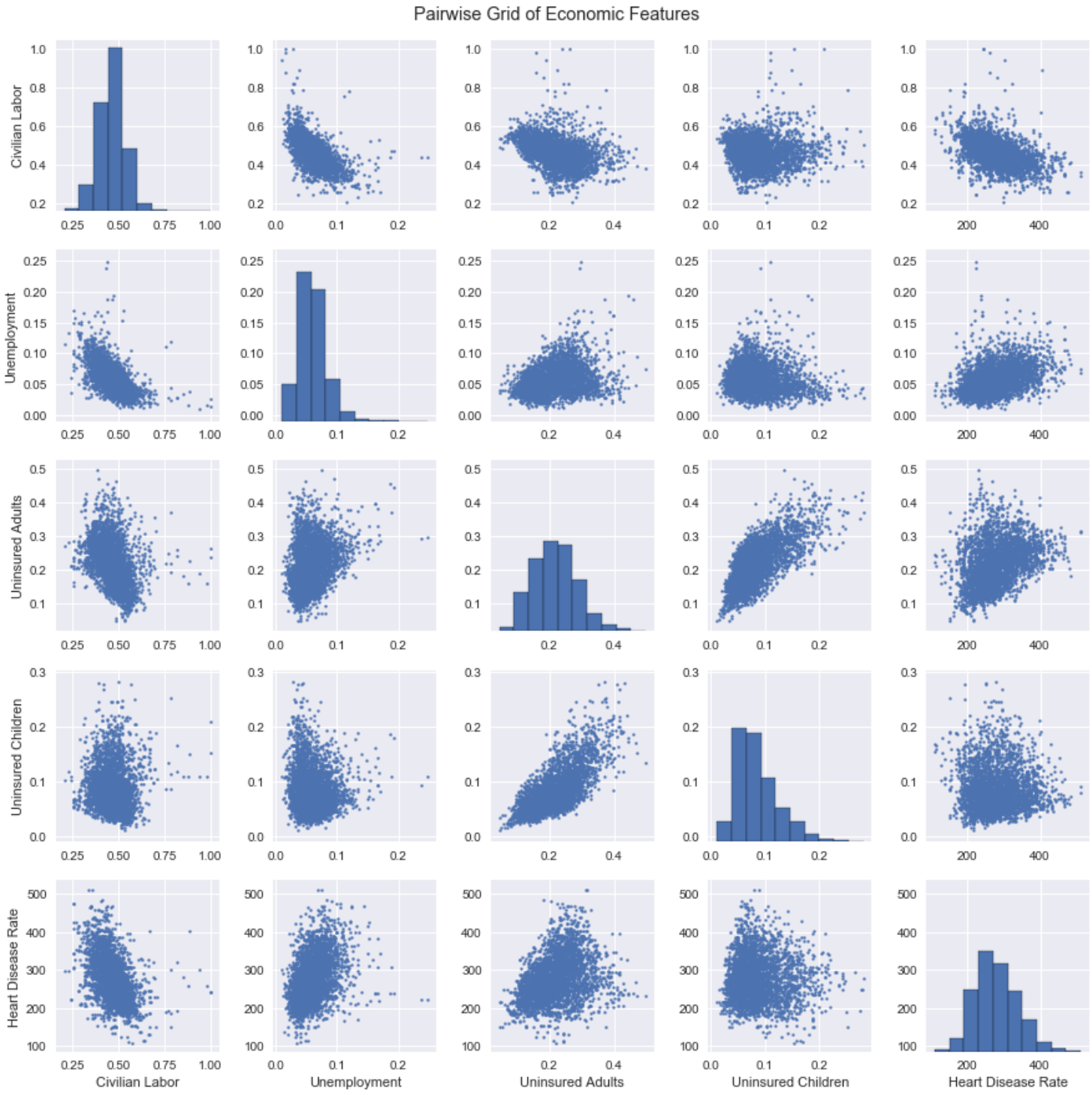
	Heart Disease Mortality per 100k Mean	Median
Non-Metro	286.550242	282.0
Metro	266.191489	261.0

While the histogram below shows frequency of each non-metro rate was higher, this was due to the greater total number of non-metro counties in the U.S. The distributions were similar, although the non-metro data seems slightly more skewed to the right.



Economics

Some of the economic factors that seemed most significantly correlated with the heart disease mortality rate were the percentages of the population that were part of the civilian labor force, unemployed, and uninsured. the correlation coefficients between heart disease and these three factors were -0.48 (% Civilian Labor), 0.37 (% Unemployed), 0.33 (% Uninsured). According to these coefficients, the higher the % of the population that is part of the civilian work force, the lower the rate of heart disease mortality, but the higher the rate of unemployed and uninsured, the higher the rate of heart disease mortality. The plots on the bottom row show the different features plotted with the rate of heart disease mortality.

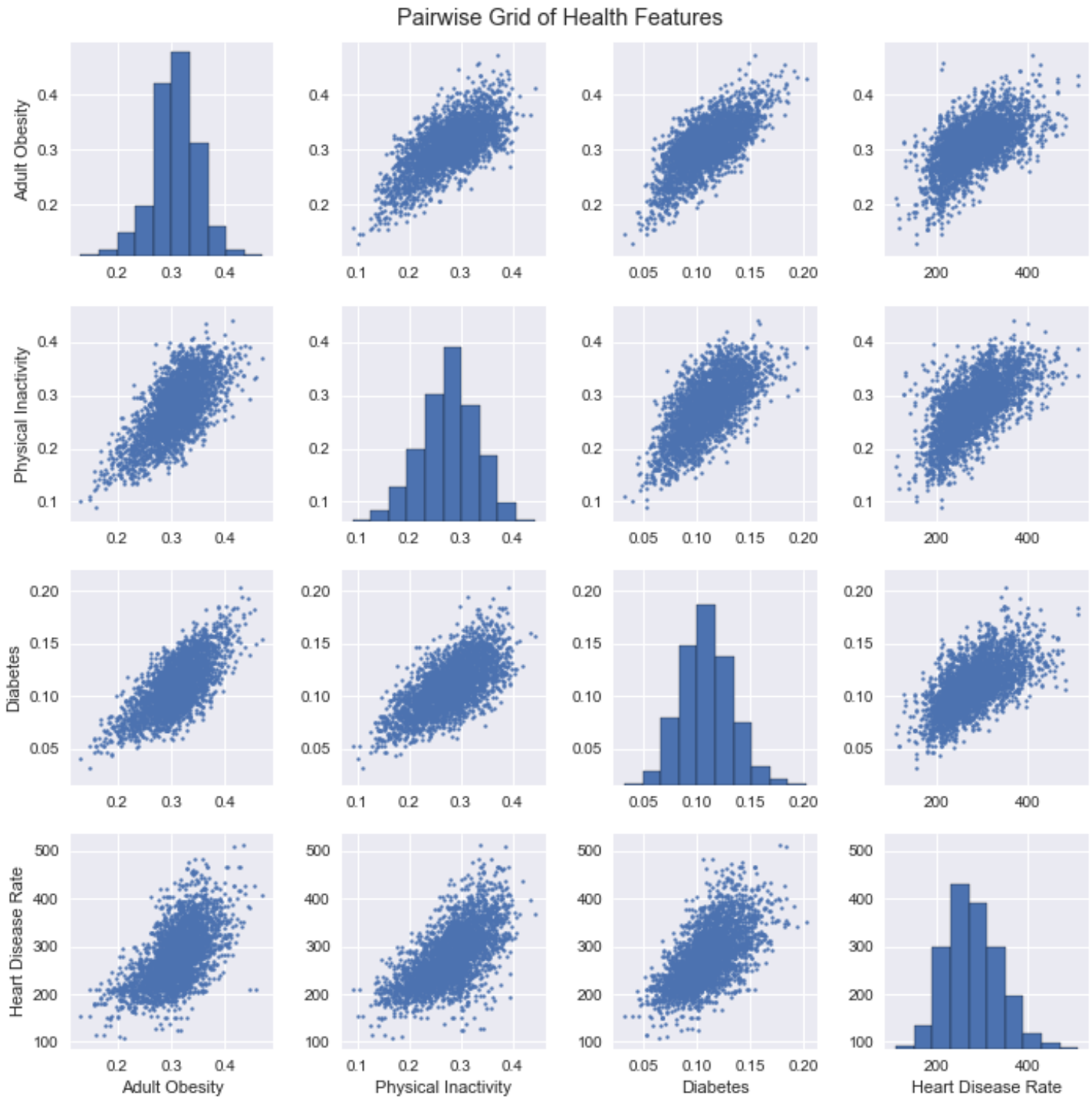


Among the plots above, it appears that there is some multicollinearity between the independent variables, such as uninsured adults and uninsured children. When the percentage of uninsured adults increases, so does the number of uninsured children, which is to be expected. However, with linear regression models, correlation between these independent variables can often affect the variables' effectiveness at predicting the dependent variable (rate of heart disease mortality).

In the regression model that was created and is described later in this report, the author attempted to select variables that were not correlated with other independent variables but that were correlated with heart disease mortality (the dependent variable). However, the best results came from a multivariate linear regression analysis on all variables as opposed to smaller selected sets of variables. Therefore, the multivariate linear regression analysis was performed on *all* of the variables to achieve the best RMSE.

Health

There were a large number of health factors to take into account in this data. However, the most correlated with heart disease mortality were the rates of obesity, diabetes, and inactivity. These features were highly correlated with heart disease mortality – each of these three had which had correlation coefficients with mortality of 0.59 (obesity), 0.63 (diabetes), and 0.64 (inactivity). Smoking also seemed to have some correlation, but other factors, such as air pollution, drinking, and population per primary care physician had less of a correlation that expected. Notice on the bottom row of the following plots that as the rates of diabetes, physical inactivity, and adult obesity increase, so does the heart disease mortality rate.

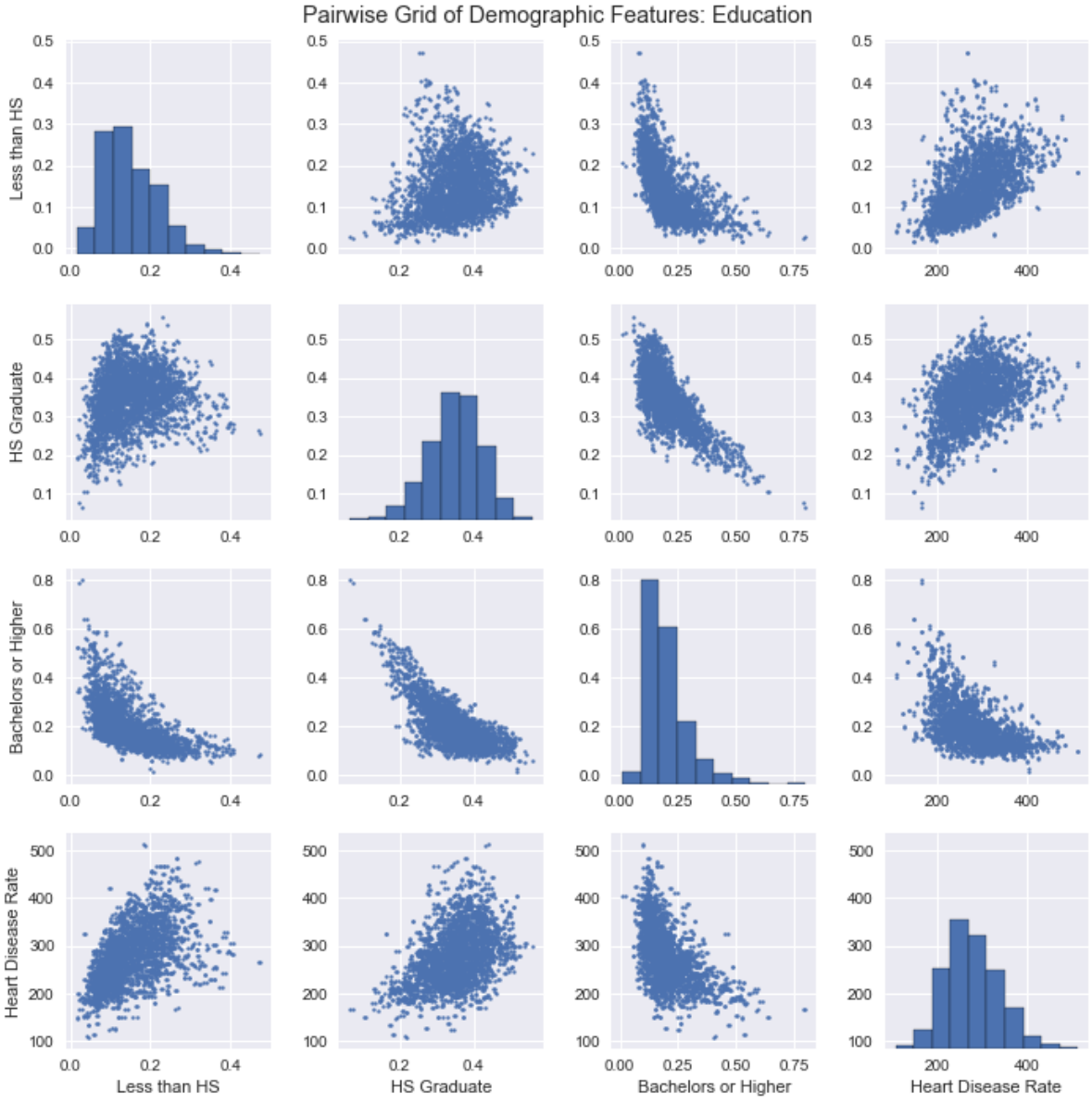


As with the economic features' plots, there seems to be some multicollinearity among all of the variables in the above plot. This is especially obvious with the correlation coefficients for each feature in the chart below. Each correlation coefficient between all pairs of features for these health factors is above 0.59. As physical inactivity and diabetes increase, so does the rate of adult obesity. With correlation among these independent variables, it is more difficult to understand specifically which variables are explaining the heart disease mortality rate.

	Adult Obesity	Physical Inactivity	Diabetes	Heart Disease Rate
Adult Obesity	1.000000	0.683851	0.701160	0.593324
Physical Inactivity	0.683851	1.000000	0.674470	0.649810
Diabetes	0.701160	0.674470	1.000000	0.631285
Heart Disease Rate	0.593324	0.649810	0.631285	1.000000

Demographics

Some of the most significant and surprising correlations between demographic factors and heart disease mortality were related to education. Counties with a higher percentage of university graduates seemed to have lower rates of heart disease mortality than counties with higher percentages of high school graduates or residents who didn't finish high school.



The reason for this surprising correlation could be the access to better healthcare or hire quality of living that college graduates may be afforded. These socioeconomic factors could provide college graduates with earlier detection of heart disease or better treatment.

Regression

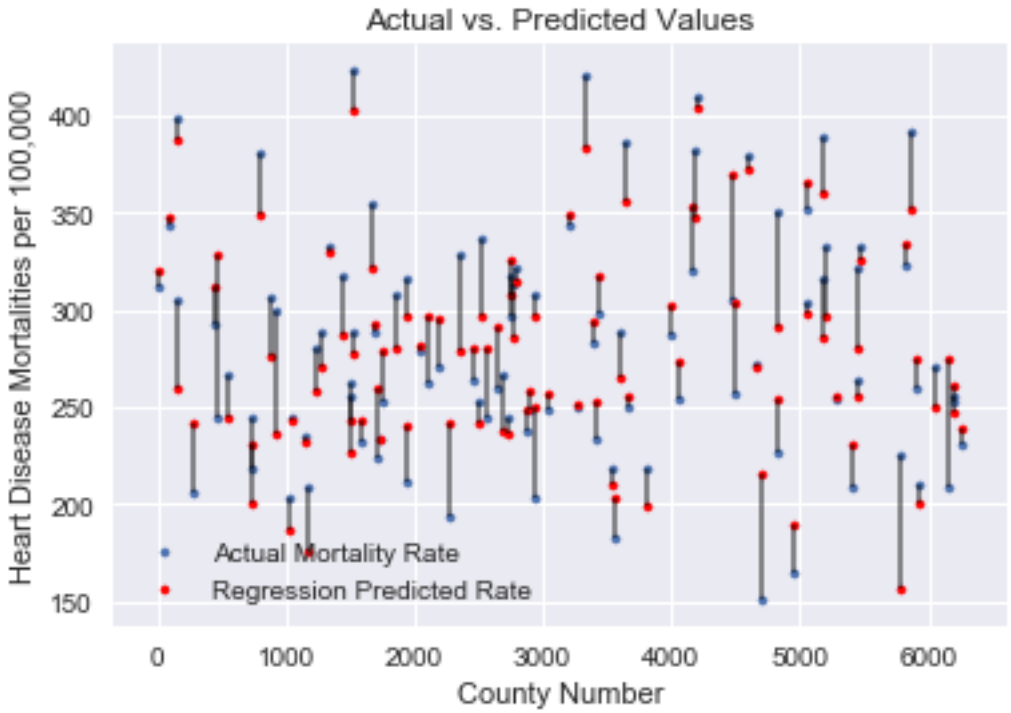
A regression model was created to predict the heart disease rate based on the area, economic, health, and demographic features. The model was trained using 80% of the provided data, and 20% of the data was used to test the model. Its accuracy was evaluated with the root-mean-square error metric, in which the goal is to minimize the RMSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

After assigning dummy variables to non-numerical area and economic data points, filling the empty data values with their features' mean values, and finally dropping the "year" feature from the data set, the multivariate linear regression model was trained and run on the test data set. The resulting RMSE was 32.2696.

Work was done to improve the RMSE by selecting different feature sets based on the individual features' correlation to the heart disease mortality rate, based on their correlation to other features, and based on the features' ranking using recursive feature elimination (RFE), but the RMSE did not improve. Furthermore, various regression models, such as ridge regression, lasso regression, linear and non-linear support vector regression, and random forest repressor models were trained and tested, but they did not produce an RMSE lower than 33.

After training the model on 80% of the labeled data, the model was tested on the other 20% of the labeled data. The plot below shows the output values of the predictive model for that 20% and compares those output values to the actual labeled values for the 20%. This plot shows the first 100 counties out of the 20%. The lines connect the output values of the predictive model and the actual values for each county. The lines are meant to illustrate the differences (errors) between the predicted value and the actual value for each county and therefore visualize the accuracy of the model.



According to this plot, the model seems to have been fairly close on a large number of the counties.

Conclusion

This analysis and report have shown that a multivariate linear regression model can be successful at predicting heart disease mortalities per 100,000 residents in counties across the U.S. While health and demographic factors appear to play a major role, economic and area factors also have some influence.

For further understanding heart disease mortality rates, it is recommended that more data is collected for new features such as: average proximity to a hospital, geographic location within the U.S., and prevalence of non-lethal heart disease. This would allow researchers to understand the role that nearby health care facilities play in preventing heart disease mortalities and the role that geographic factors such as climate may have on heart disease.

Furthermore, many of the features that were included in this data set were highly correlated to each other, whereas the most effective linear regression models use independent variables that are not highly correlated to other independent variables and are instead highly correlated to the dependent variable (heart disease mortality). A greater number of features may allow for more successfully selecting a smaller set features that are not correlated with each other but are correlated with heart disease mortality.