PREDICTING HEART DISEASE MORTALITY RATE

By David Edelman
DAT102x Capstone Project

EXECUTIVE SUMMARY

Heart disease is one of the leading causes of death in the United States. This document presents an analysis of county-by-county heart disease mortality rates (per 100,000 individuals) and uses various demographic statistics to build a model that predicts the heart disease mortality rate for that county.

The given data set, containing known heart disease mortality rates, comprises 3198 observations, representing 1599 counties and data collected over two separate years. Both categorical and numeric data are present in the feature set, and after some initial data exploration, several other features were calculated. A predictive regression model was then created using select features in order to predict the heart disease mortality rate of 3080 county/year pairs, representing 1540 distinct counties.

While building and tuning the predictive model, the following features were deemed significant to the predictive power of the model

AREA INFORMATION

- Area RUCC (Rural-Urban Continuum Codes) – classifies each county into one of 9 mutually exclusive categories that identifies (a) the population size, (b) the degree of urbanization and (c) proximity to a metropolitan area.
    - While the RUCC itself did now show any strong correlation to heart disease mortality rate, by combining one or more RUCC values into one of 5 groups based solely on population size, the data showed 3 population groups that skewed below the mean heart disease mortality rate and 1 that skewed above the mean.
    - Additionally, classifying the RUCC values into another calculated feature as “Metro” or “Non-Metro” showed the data being skewed below the mean for Metro and above the men for Non-Metro, adding an enhancement to the predictive model (data source: USDA Economic Research Service)

ECONOMIC FACTORS

- Economic Typology – classifies each county into one of 6 mutually exclusive categories of economic dependence; two categories show a distribution skewed below the mean heart disease mortality rate and two categories are skewed above (source: USDA Economic Research Service).
- Percent of Civilian Labor – the annualized percent of the county’s population classified as civilian; analysis shows a moderate *negative* correlation to heart disease mortality rate (source: Bureau of Labor Statistics)

DEMOGRAPHICS

- Percent of non-Hispanic African Americans – analysis shows the strongest positive correlation among the 5 race/ethnicity groups (source: US Census Population Estimates)

- Percent of adults with less than a high-school diploma; percent of adults with a Bachelor's degree (or higher) – the former shows a strong positive correlation, while the latter shows an equally strong negative correlation (source: US Census Population Estimates)

HEALTH FACTORS

- Air Pollution Particulate Matter – measured in $\mu g/m^3$; the average concentration of fine particulate matter as measured over the year. Lower concentrations skewed below the population mean while higher concentrations skewed above (source: CDC WONDER).
- Percent of physical inactivity – percent of adult population that self-identifies as physically inactive (source: National Center for Chronic Disease Prevention and Health Promotion); physical inactivity showed a moderately strong correlation to heart disease mortality rate
- Percent of adult obesity; percent of diabetes; percent of adult smoking – each of the three factors individually showed a strong correlation with heart disease mortality rate, but with a thought experiment and some data transformation into combined factors, an even stronger correlation was found (sources NCCDPHP; NCCDPHP, Division of Diabetes Translation; Behavioral Risk Factor Surveillance System)
- Homicide, Motor Vehicle Death rates per 100,000 – two separate death rates (per 100,000 individuals); by scaling each rate to a logarithmic scale (base 10), a strong correlation to heart disease mortality rate was found (source: National Center for Health Statistics)

DATA EXPLORATION AND INITIAL ANALYSIS

HEART DISEASE MORTALITY RATE

DESCRIPTIVE STATISTICS

Heart Disease Mortality Rate (herein shown with a unit of "deaths per 100,000 population") in the given data set showed a mean of 279.4, median of 275.0, and standard deviation of 59.0 with a range of 109.0 – 512.0. A histogram with 40 equal bins (calculated based on the minimum and maximum values) and a box plot both show that heart disease mortality rate is approximately normal (skewed very slightly positive), with only approximately 10-15 outliers (values outside of the Inner Quartile Range of 237 – 317).

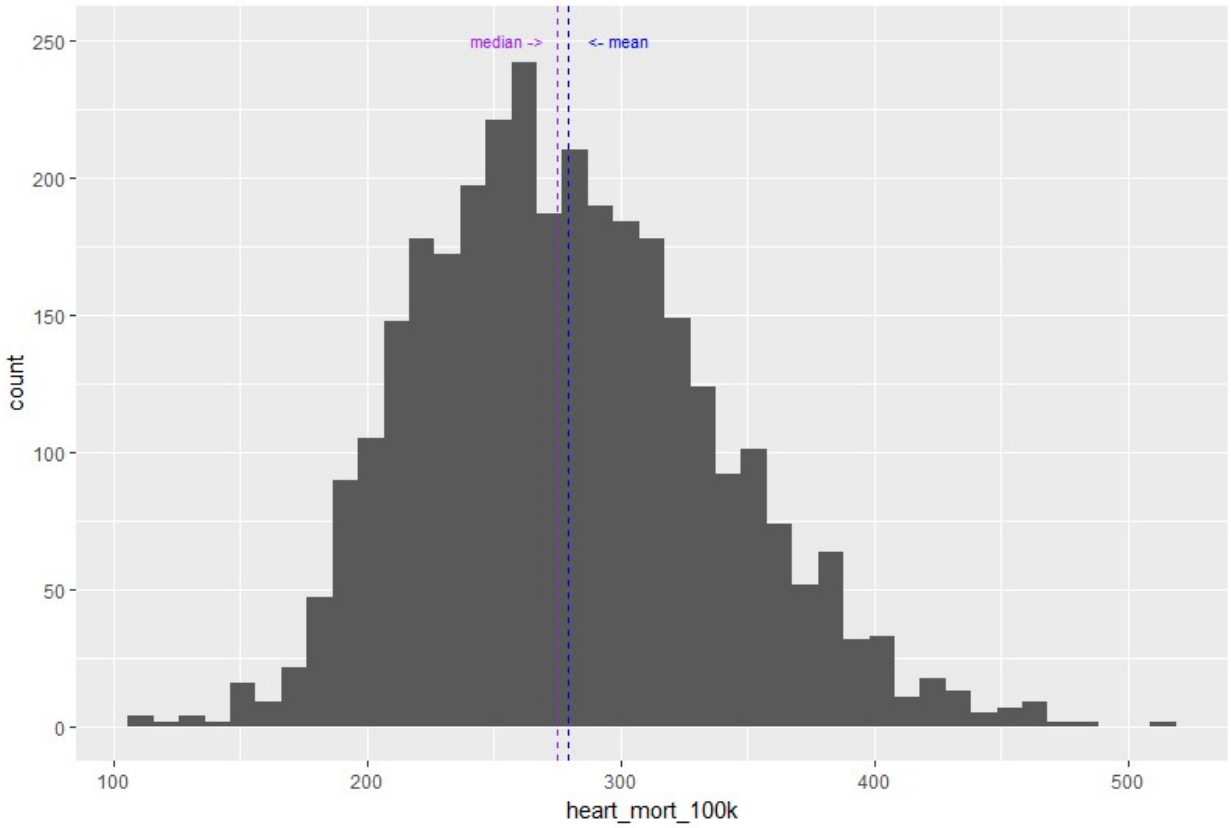

Figure 1 - Heart Mortality Rate Histogram

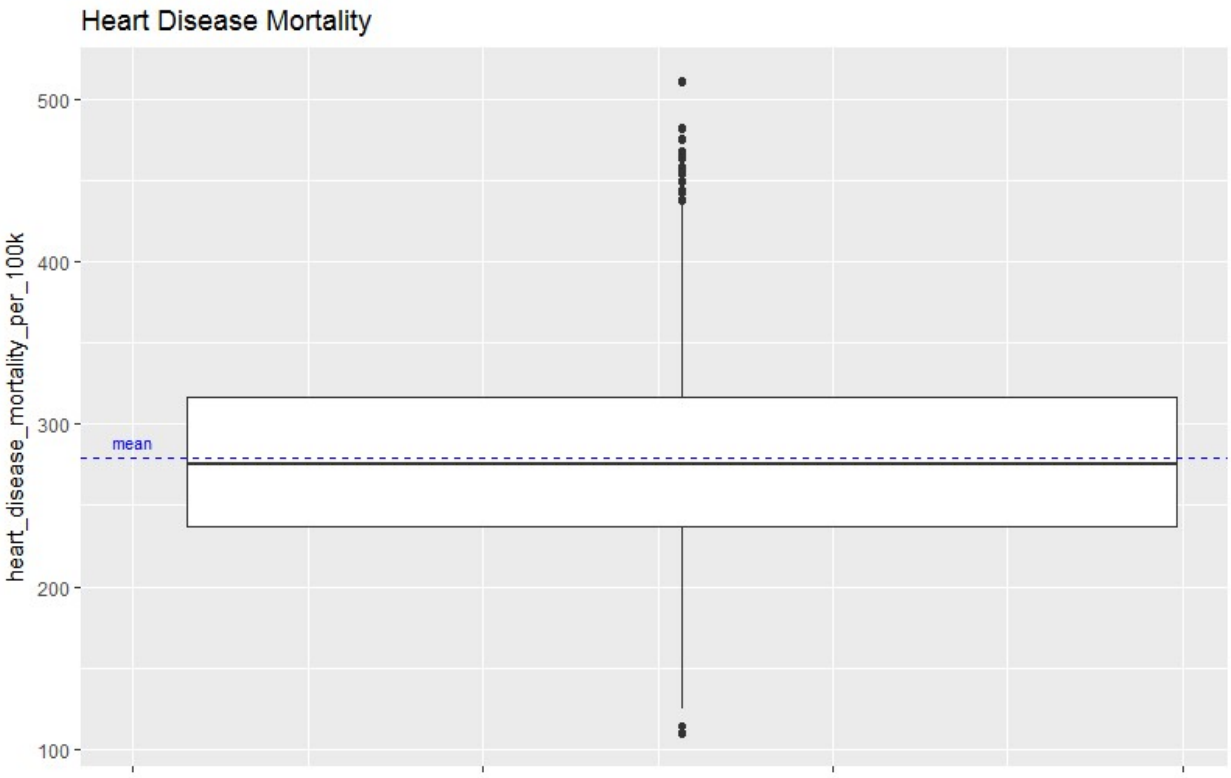

Figure 2 - Heart Mortality Rate Boxplot

DISTRIBUTIONS ACROSS CATEGORICAL FEATURES

ECONOMIC TYPOLOGY

The 6 economic typologies used by the USDA Economic Research Service to assign to each county are

- Farm-dependent
- Federal/State government-dependent
- Manufacturing-dependent
- Mining-dependent
- Non-specialized
- Recreation

The typologies speak to the main type of industries of a particular county. So while we don't have the specific industries in each county (knowing, intellectually, that certain industries have been linked in the past to specific conditions, including heart disease), we can use the typology as a good representative proxy for specific industries. To identify any potential relationship between typology and heart disease mortality, we use histograms and box plots to examine each typology for an abnormal or skewed distribution.

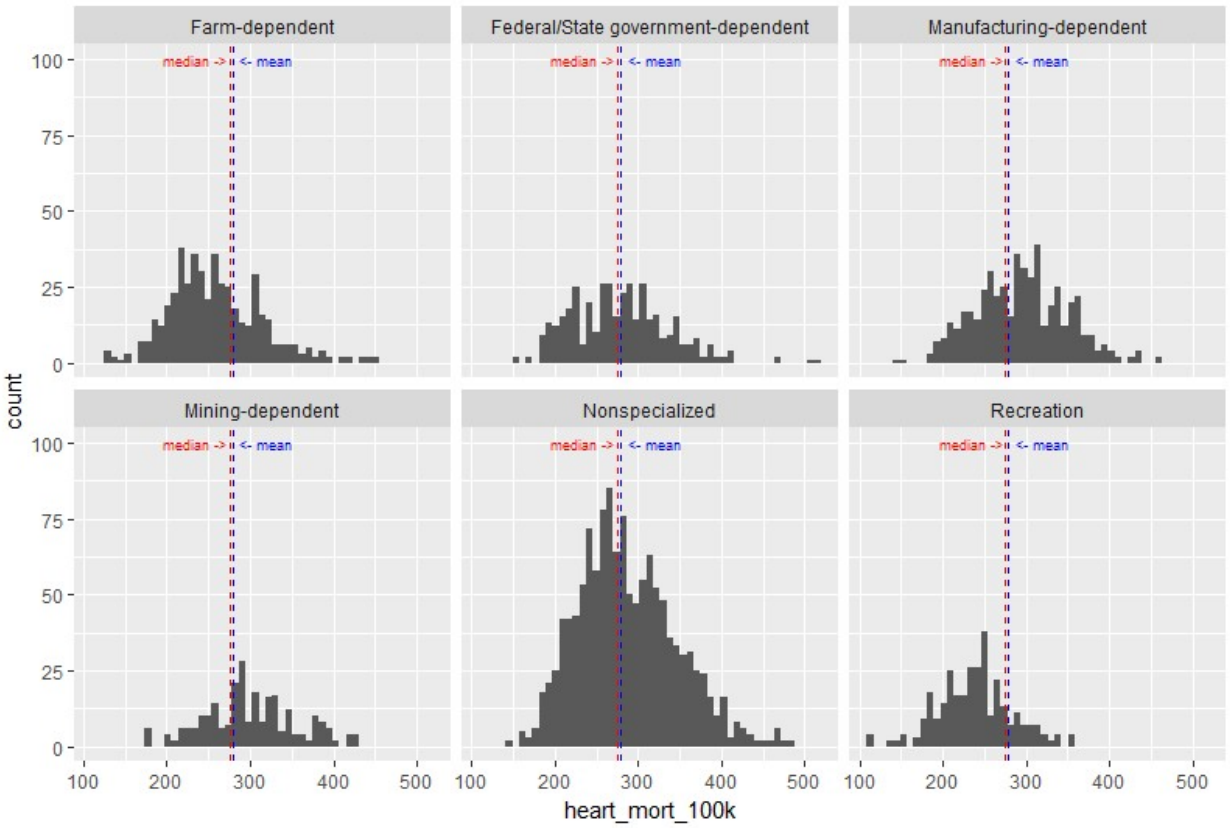

Figure 3 - Histograms by Economic Typology

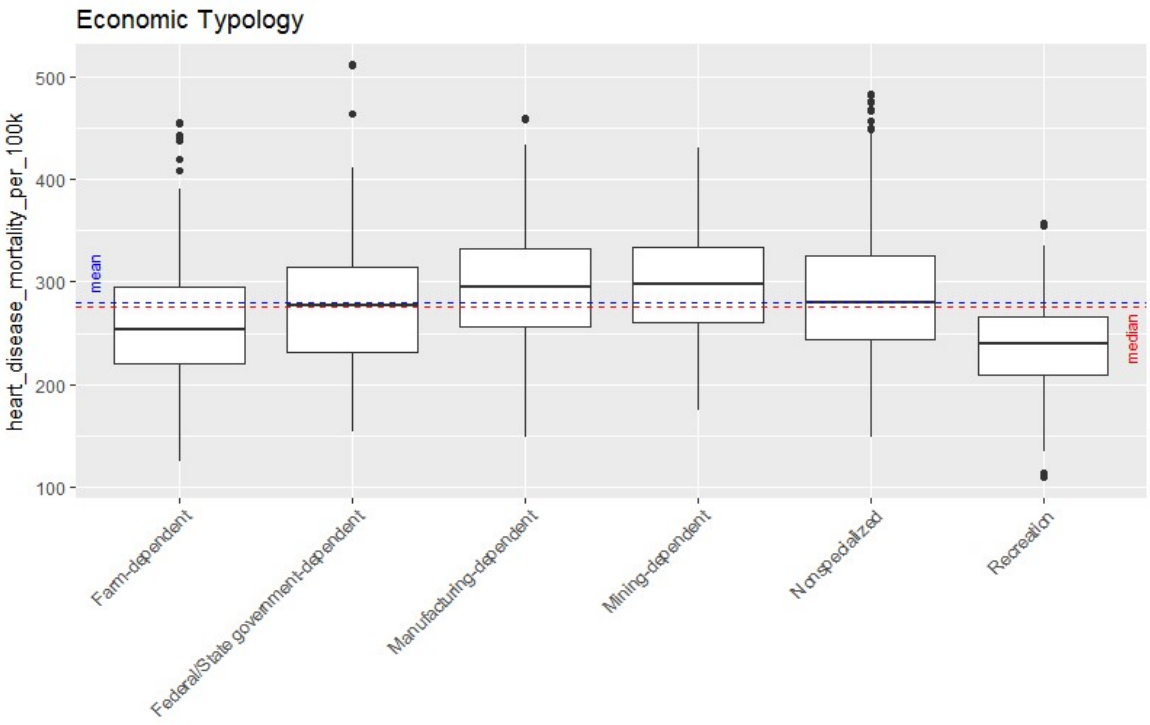

Figure 4 - Box Plots by Economic Typology

The histograms shows that all of the typologies appear to be approximately normal. They also show that the mean from Recreation and from Farm-dependent (to a lesser degree) to be lower than the population mean. None of the typologies appear to have significantly higher means than the population average. However, the box plots give a better picture of the distributions, showing that indeed Recreation and Farm-Dependent have their median value below the population average, while Manufacturing-dependent and Mining-dependent have medians well above the populate average (Non-specialized shows a very slight variation from the population mean and median). The same results are shown below:

Pop Mean = 279.7, Pop Median = 275.0			
Typology	**Mean**	**Median**	
Farm-dependent	258	253	Below Pop > 1%
Federal/State government-dependent	278	277	
Manufacturing-dependent	295	295	
Mining-dependent	301	297.5	Above Pop > 1%
Nonspecialized	286	280	
Recreation	238	239	

Table 1 - Summary Statistics per Economic Typology

Due to heart disease mortality rates being distributed differently across the economic typologies, this feature is a good candidate for use in a predictive model.

POPULATION SPREAD

The individual values for “Area_RUCC” have populations ranging from < 100 to > 600, and histograms show that many distributions do not appear to be approximately normal.

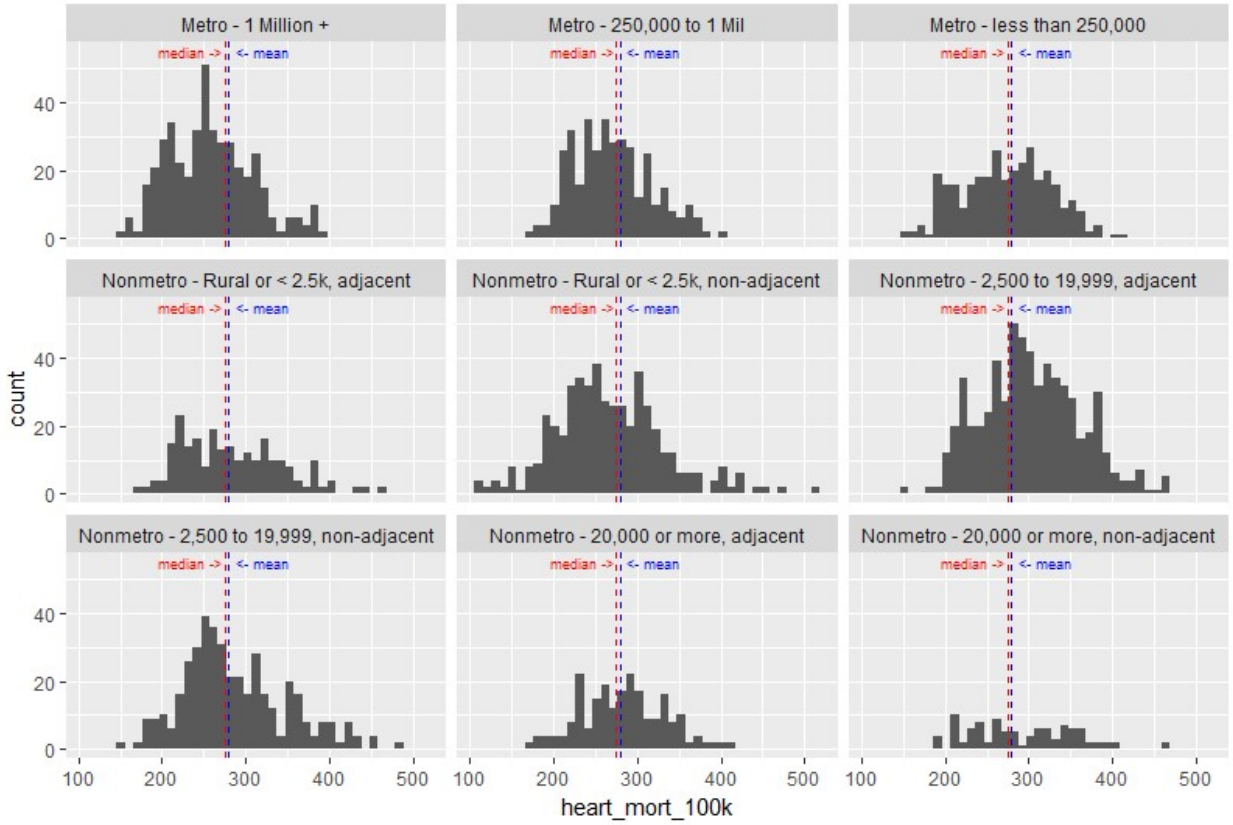

Figure 5 - Histograms for Area_RUCC[1]

However, we can transform them into population groups as a proxy for Area_RUCC. The transformation is

Area RUCC	Population
Metro - 1 Million +	1M+
Metro - 250,000 to 1 Mil	250k – 1M
Metro - less than 250,000	20 – 250K
Nonmetro - Rural or < 2.5k, adjacent	under 2,500
Nonmetro - Rural or < 2.5k, non-adjacent	under 2,500
Nonmetro - 2,500 to 19,999, adjacent	2.5 – 20K
Nonmetro - 2,500 to 19,999, non-adjacent	2.5 – 20K

[1] Adjacent/non-adjacent refers to being adjacent to a metro area

Nonmetro - 20,000 or more, adjacent	20 – 250K
Nonmetro - 20,000 or more, non-adjacent	20 – 250K

Table 2 - Transformations Area_RUCC --> Population

Looking at the histograms we now see a more even spread of data and more normal distributions

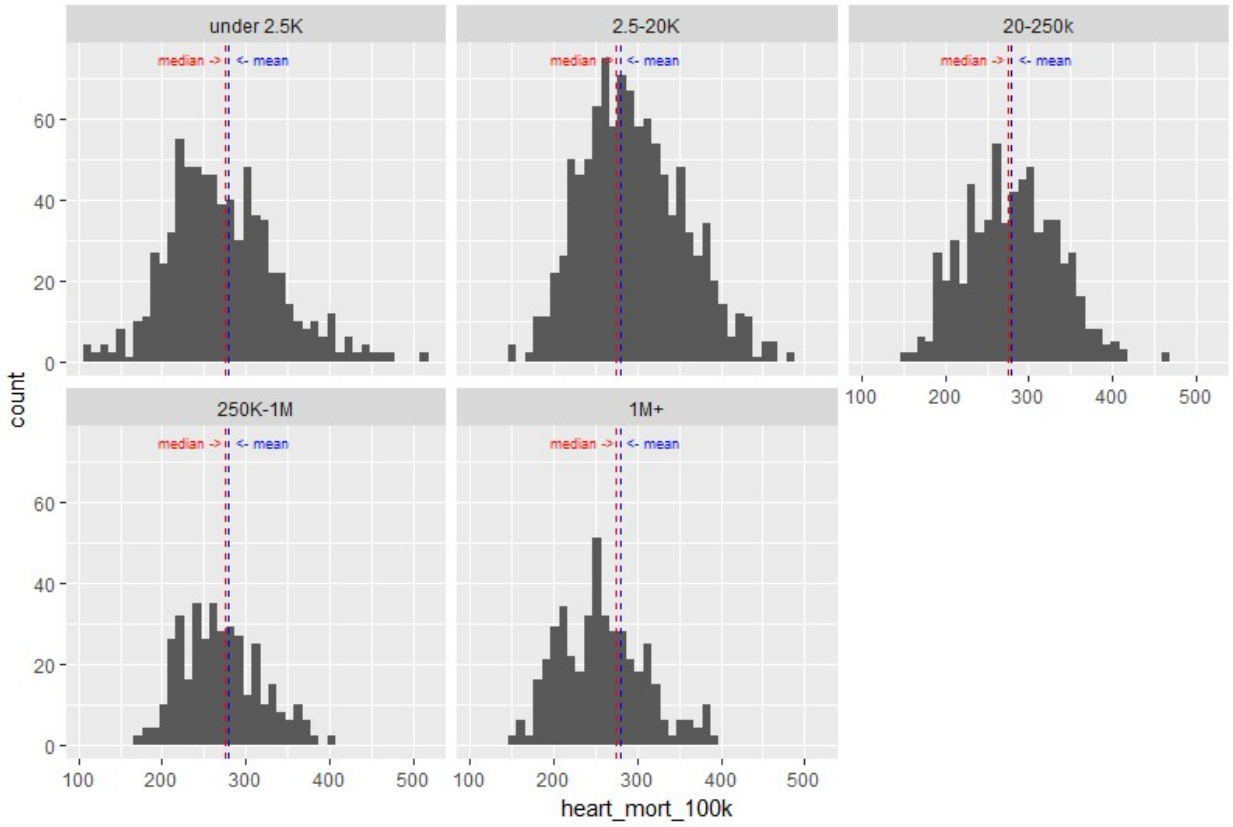

Figure 6 - Histograms by Population Groups

The “2.5-20K” group appears to be skewed somewhat positive, while the “1M+” group is skewed negative, even though the peak is below the population mean and median. Box plots show the distributions more cleanly.

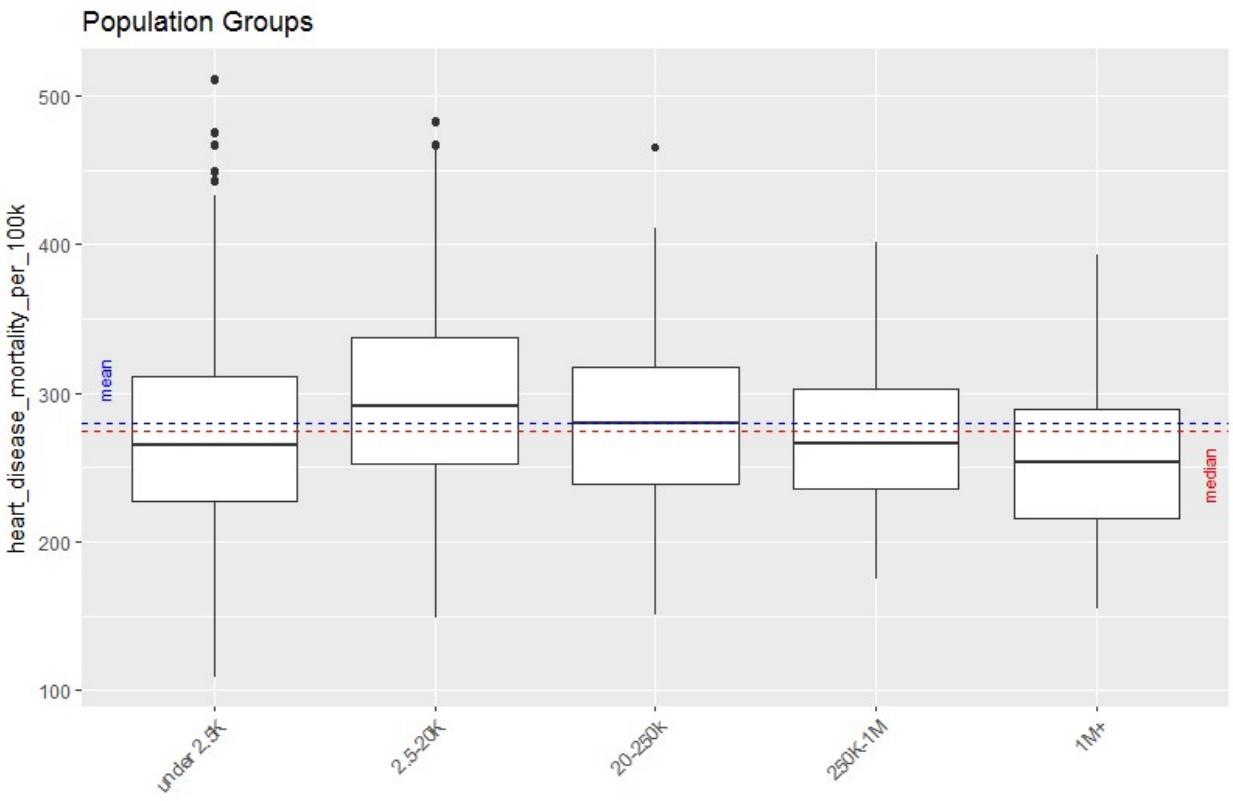

Figure 7 - Box plots for Population Groups

Here we can more clearly see that the median for “under 2.5K”, “250K-1M” and “1M+” are all below the population mean and median. The group “2.5-20K” has a median well above the population mean and median, while the “20-250K” group median is only slightly higher than the population.

Pop Mean = 279.7, Pop Median = 275.0			
Population Group	Mean	Median	
under 2.5K	272	265	Below Pop > 1%
2.5-20K	296	291	
20-250k	279	280	
250K-1M	269	266	Above Pop > 1%
1M+	257	253	

Table 3 - Summary Statistics for Population Groups

With population groups showing a different distribution for heart mortality rates, this calculated feature is another good candidate for use in a predictive model.

The analysis went one step further and grouped the RUCC values based on whether or not the county was considered to be a metropolitan area or not. Based solely on the first word in the RUCC category title (see Table 2), each county was classified as either “Metro” or “Non-Metro”.

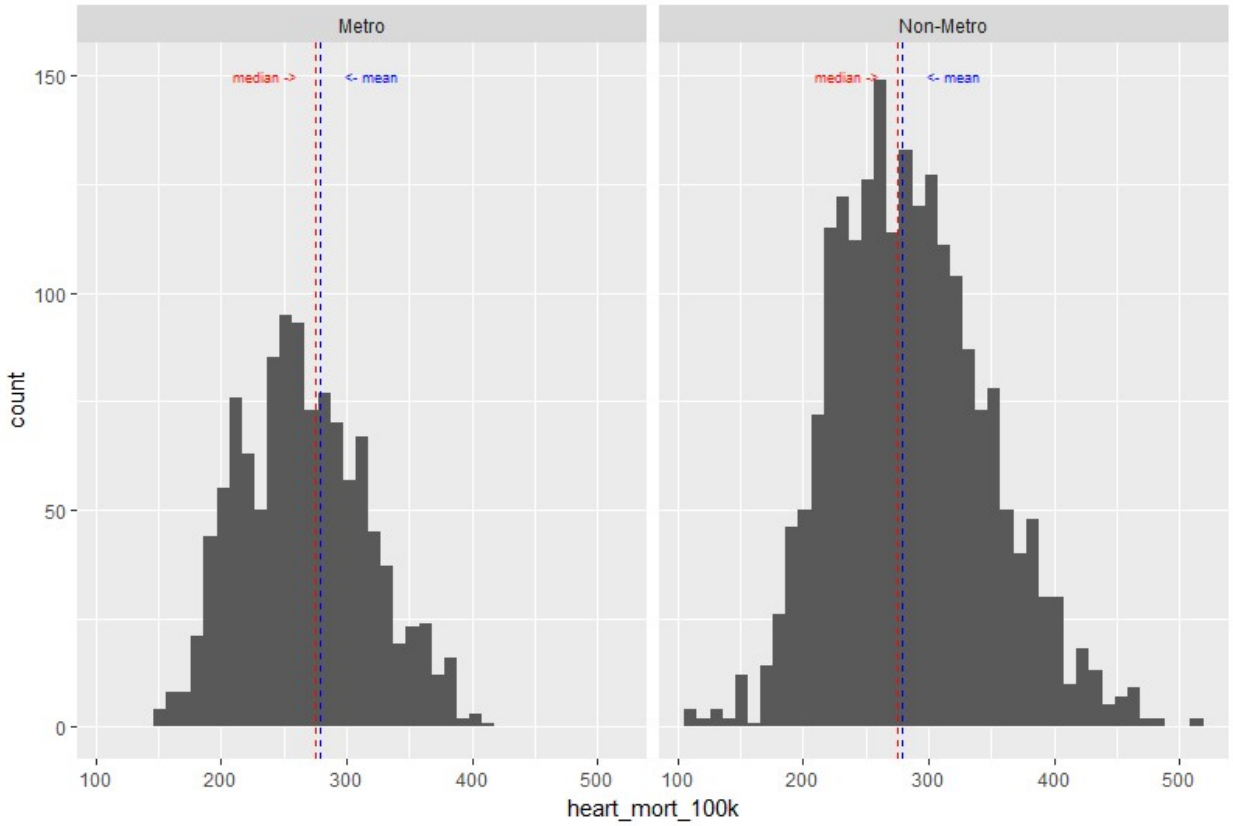

Figure 8 - Histogram for Metropolitan Type

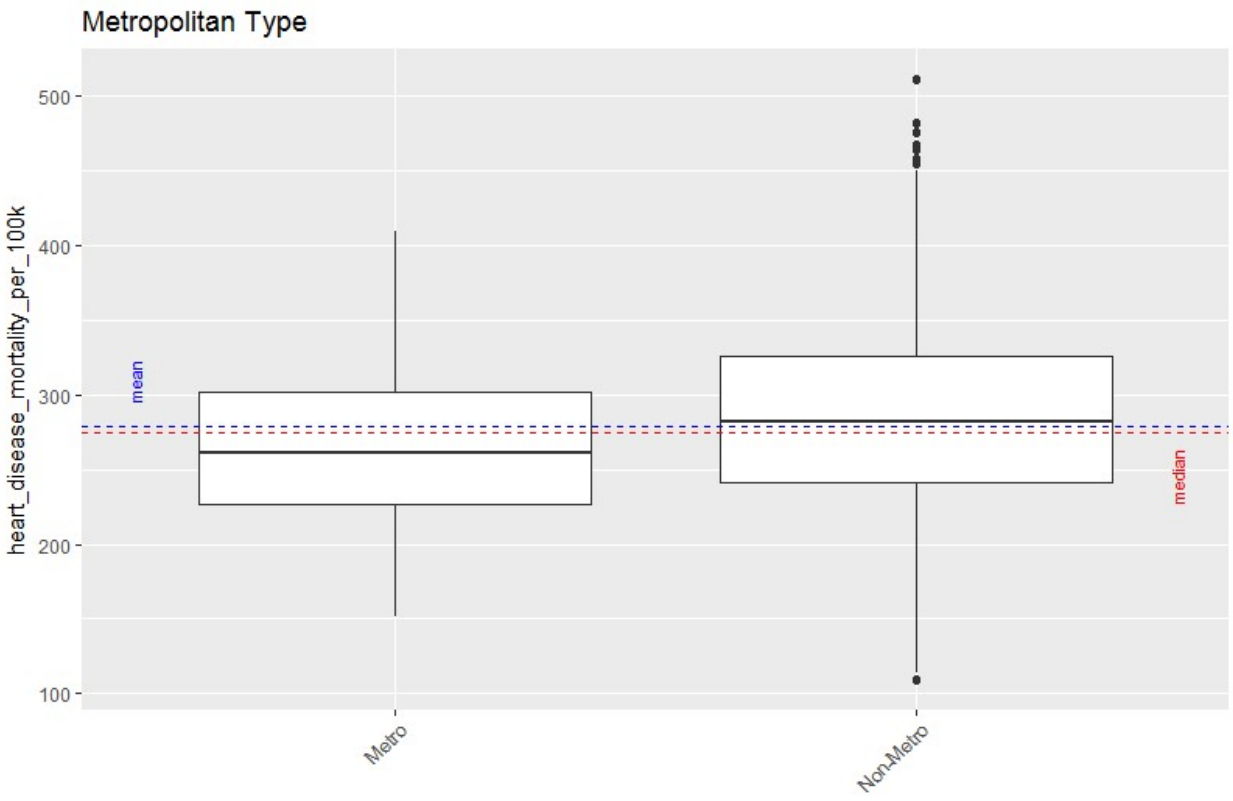

Figure 9 - Box Plots for Metropolitan Type

The box plots clearly show that metropolitan areas have a lower median than the population, while non-metropolitan areas skew higher than the population as a whole.

Pop Mean = 279.7, Pop Median = 275.0			
Metropolitan Type	Mean	Median	
Metro	266	261	Below Pop > 1%
Non-Metro	286	282	Above Pop > 1%

Table 4- Summary Statistics for Metropolitan Type

A two-category classification that shows one subset being clearly below the mean and one clearly above the mean leads to the conclusion that Metropolitan Type is a strong candidate for the predictive model.

AIR POLLUTION

Ambient air pollution for the county is presented in the data set as a measure of the concentration of fine particulate matter, as measured in µg/m3. These figures were presented as positive integers, so as a feature of the data set it is more akin to a categorical variable. The initial spread of values showed a large discrepancy in population size, especially in the min and max regions (Figure 10A), so values were grouped into a smaller number of populations that were closer in size (Figure 10B)

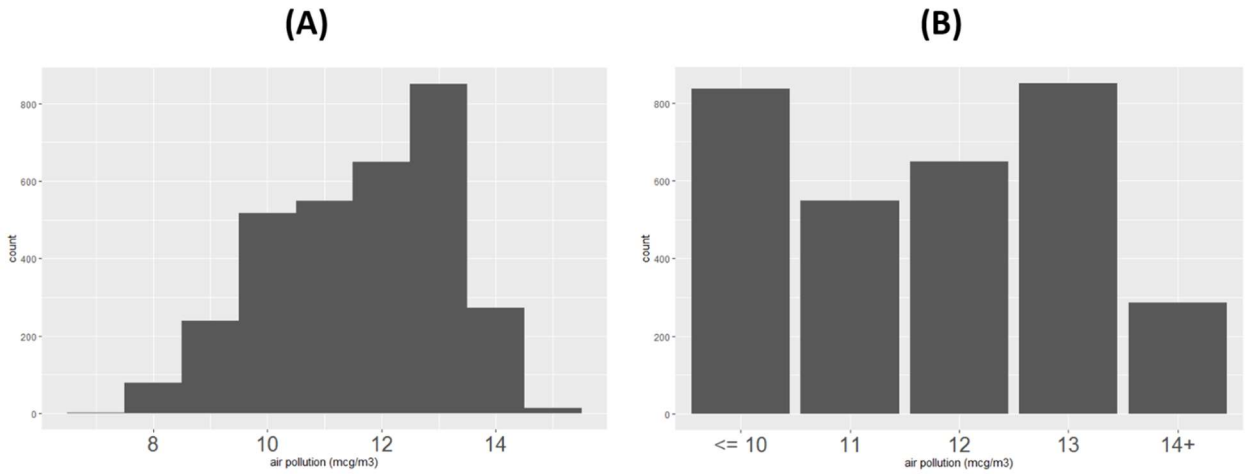

Figure 10 - Air Pollution. (A) Original population spread; (B) population spread after grouping

Using the new groups, histograms and boxplots show the spreads to be approximately normal, with some groups skewing below the population mean and some above. The group with 12 µg/m^3 shows a mean just at the population mean, with a median slightly lower.

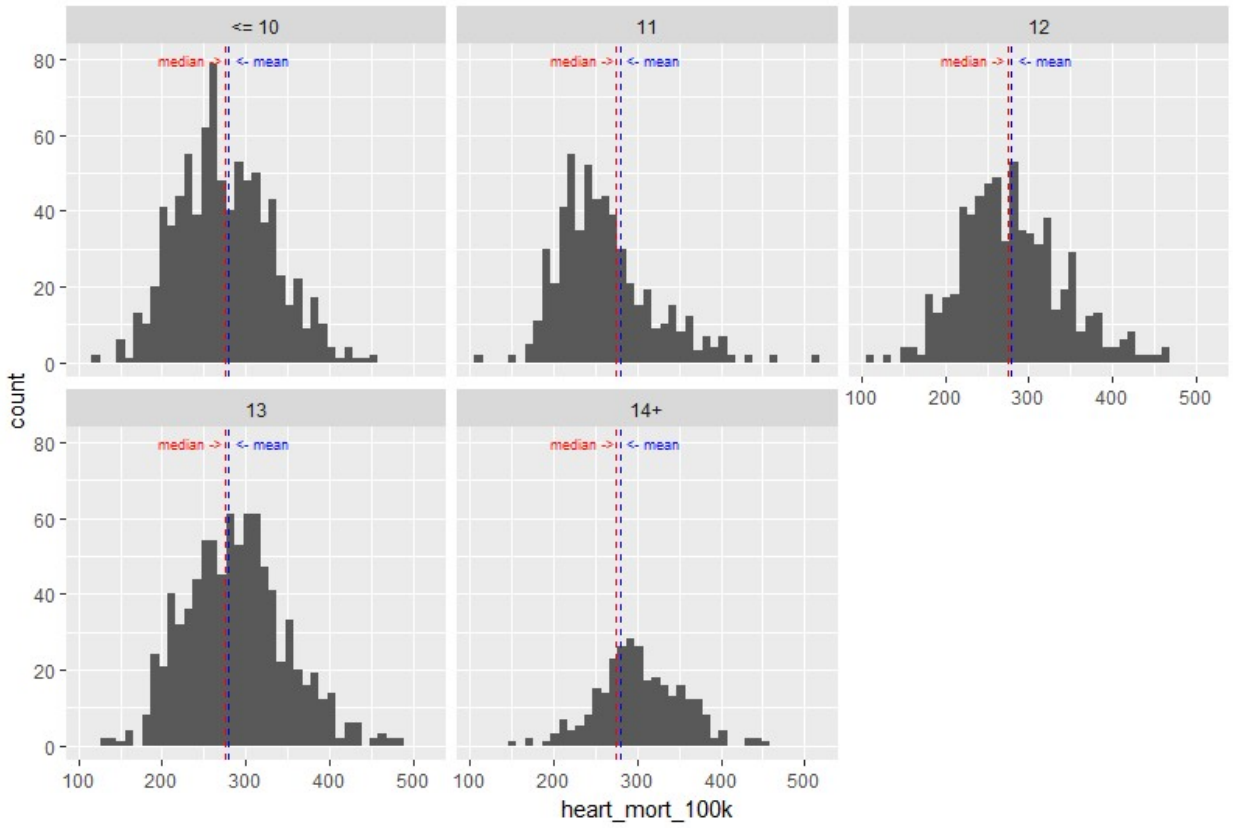

Figure 11 - Histograms for Air Pollution

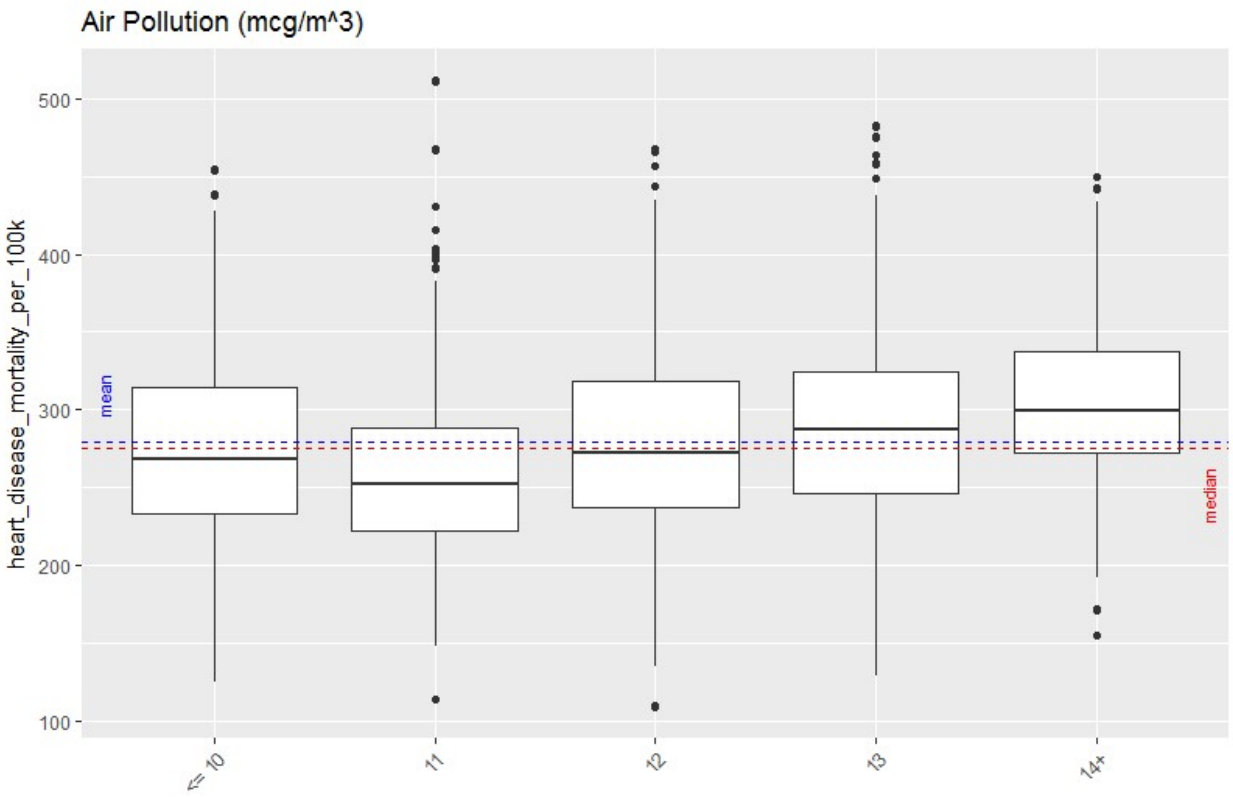

Figure 12 - Box plots for Air Pollution

A statistical summary of the air pollution groups confirms the interpretation of the histogram and box plots, and makes this feature a suitable candidate for a predictive model.

Pop Mean = 279.7, Pop Median = 275.0			
Air Pollution (µg/m³)	Mean	Median	
<= 10	275	268.5	Below Pop > 1%
11	261	252	
12	280	272	
13	288	287	Above Pop > 1%
14+	303	299	

CORRELATION[2] BETWEEN NUMERIC FEATURES AND HEART DISEASE MORTALITY

ECONOMIC FACTORS

[2] All correlation calculations in this analysis are performed after data records with missing values for the feature in questions are removed

The remaining economic factors in the dataset (after already looking at typology) are numeric, so a correlation analysis is done to determine if any of these are candidates for use in a predictive model.

- % Civilian Labor
- % Unemployment
- % Uninsured Adults
- % Uninsured Children

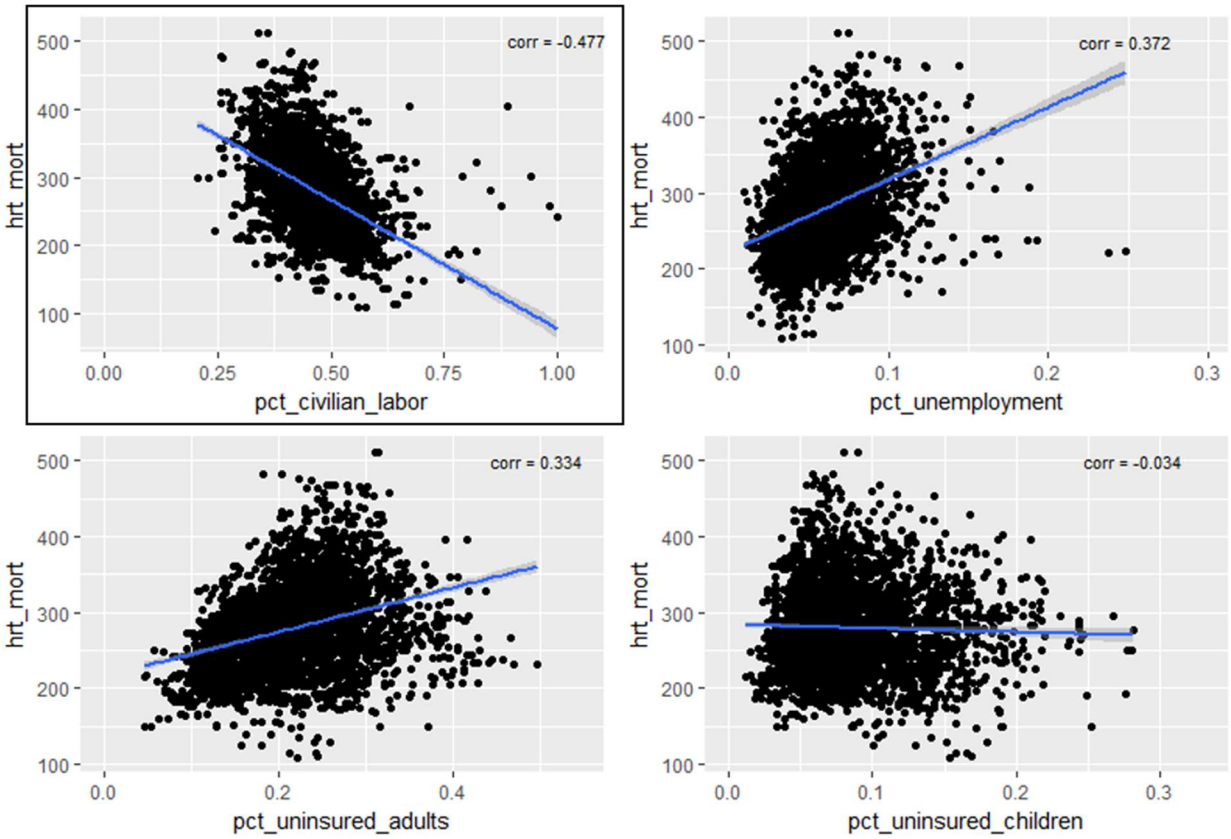

Figure 13 - Correlation between Economic Factors and Heart Disease Mortality

As seen and highlighted in Figure 8, "Percent of Civilian Labor" shows a moderate negative correlation with heart disease mortality (despite the number of outliers appearing to be greater than the other factors) and is a good candidate for use in a predictive model. The remaining factors do not show a strong enough correlation (>= 0.45) to be considered good candidates.

DEMOGRAPHICS

The dataset contains 12 demographic features that are reported as percentages of the population. A correlation analysis of those features in a similar manner to the economic factors. (NOTE: there are 2 numeric features that are represented as rates per 1000 people, and act more like categorical features as they are distinct whole numbers; a cursory exploration of these two factors was performed, but nothing of interest resulted from the exploration)

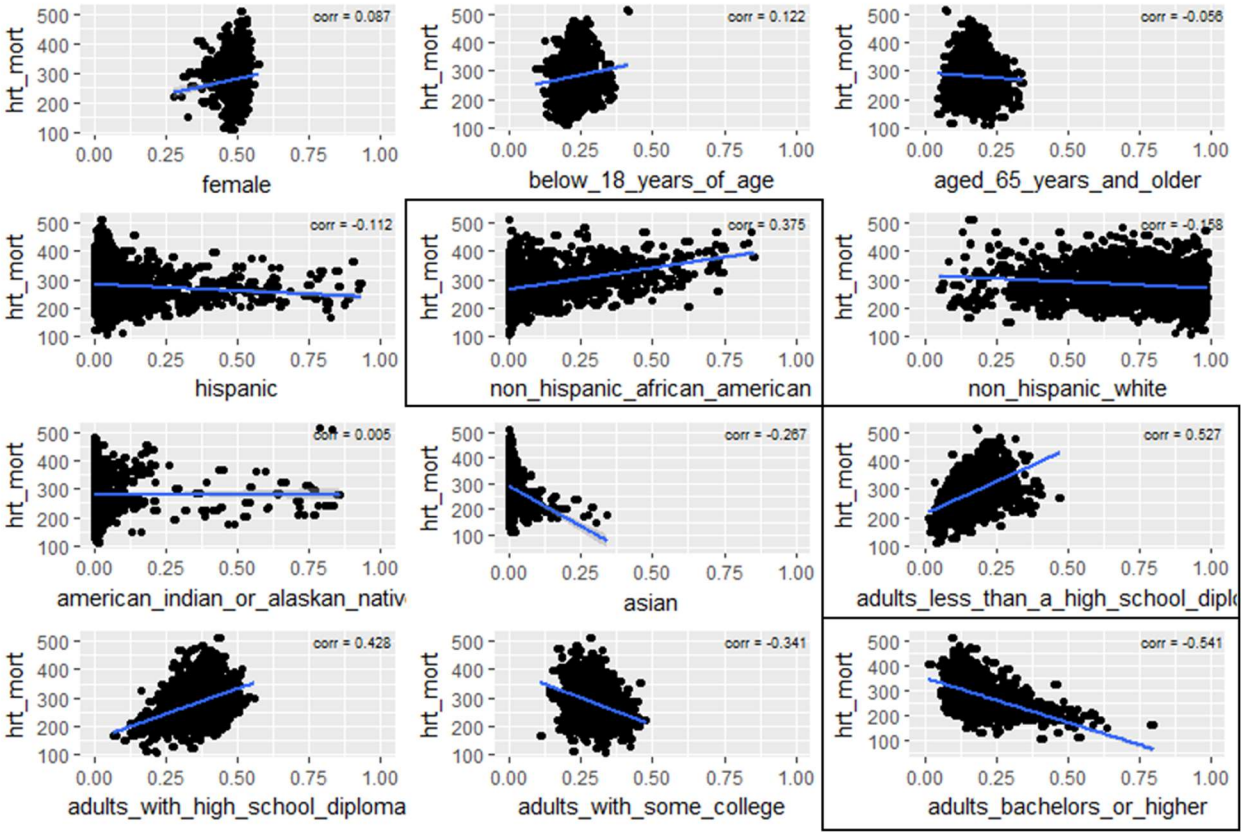

Figure 14 - Correlation between Demographic %s and Heart Disease Mortality

Analysis of the correlation scatter plots showed that there were several candidates for model features. “Percent of non-Hispanic African Americans” was chosen as it had the strongest correlation in the subset of race/ethnicity demographics, even though its actual correlation was only moderate (< .45). “Adults with less than a High School diploma” and “Adults with a Bachelor’s Degree or higher” were chosen for their strong positive and negative correlations, respectively.

The other two demographic categories related to education level (“High School Diploma”, “Some College”) were considered as features, but in practice they added no predictive power to the model. Two combination features, one of which added together “High School” and “Less than High School” and the other “Some College” and “Bachelor’s or higher” were considered as well. While the combination features had higher correlations with heart disease mortality than the individual features, in practice the model seemed to suffer from overfitting when the combination features were used in place of the individual ones.

Because of their extremely weak correlations (abs value < .15), demographic features related to gender and age were not considered for the model.

HEALTH STATISTICS

There are 10 more health-related factors present in the dataset (after air pollution). All are numerical, 6 are presented as percentage of the population, two are rates per 100,000 people, and two are rates related to the number of people per medical professional (primary care doctor, dentist). The 2 features related to medical

professional did not show a strong correlation (> 0.45) to heart disease mortality rate, so they were not considered for the predictive model.

HEALTH PERCENTAGES

A scatter plot and correlation analysis was done for the 6 numerical features presented as population percentages:

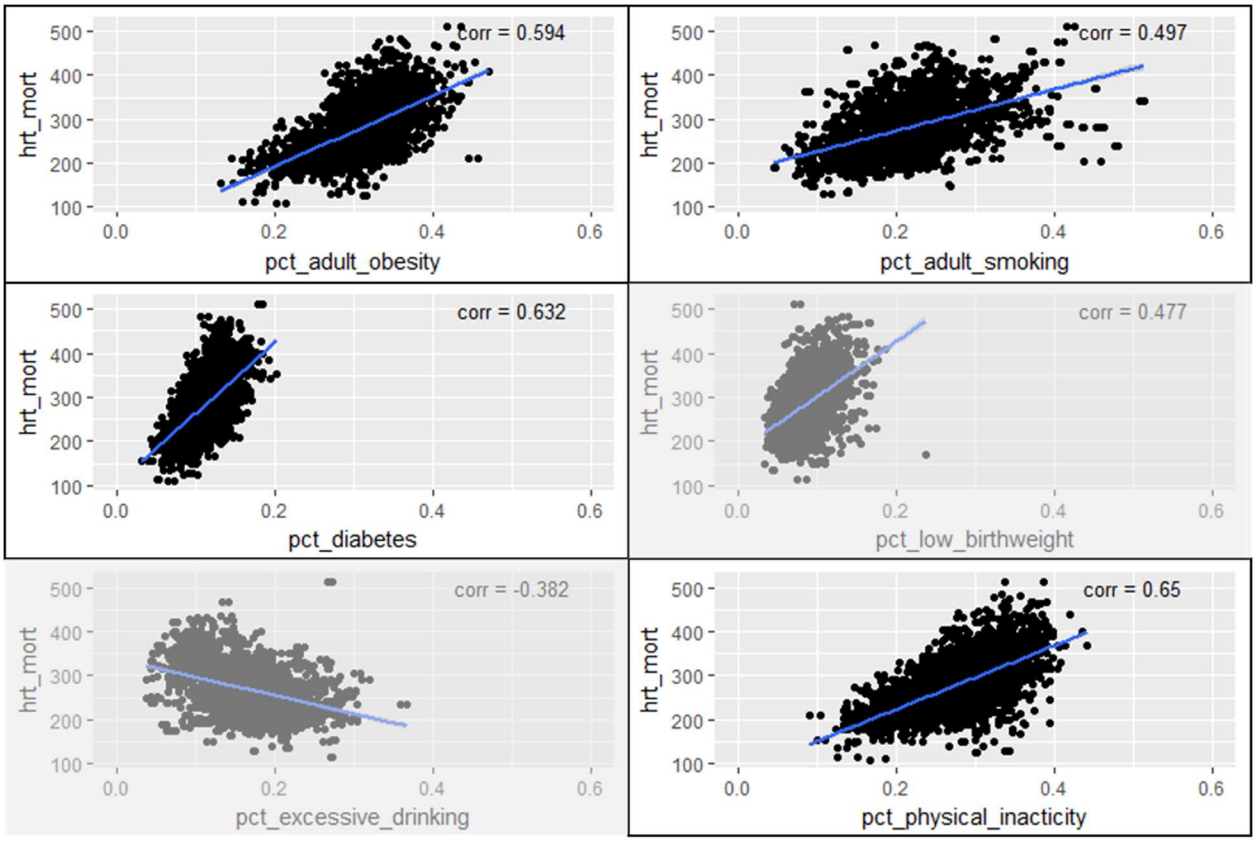

Figure 15 - Correlation between Health Percentages and Heart Disease Mortality

Percent of Physical Inactivity (shown as "pct_physical_inacticity [sic]" shows the strongest correlation with heart disease mortality among the health percentages, indicating a strong candidate for the predictive model. The other 3 percentages highlighted in Figure 15 also showed strong correlation with heart disease mortality, but some further thought and analysis produced features with even stronger correlations.

COMBINING HEALTH PERCENTAGES

Thinking about the 3 health percentages:

- Percent of adult obesity
- Percent of adult smoking
- Percent of diabetes

we can say that each of the percentages is the same as the *probability* of a randomly chosen individual in that county having the specific condition. Let us further assume that each of the events (being obese, smoking, having diabetes) are independent of each other. According to probability theory, the probability of two independent events occurring is simply the product of the two individual probabilities. Therefore

- P(diabetes &) = P(diabetes) * P(obese)
- P(diabetes & smoking) = P(diabetes) * P(smoking)
- P(obese & smoking) = P (obese) * P(smoking)
- P(obese & smoking & diabetes) = P(obese) * P(smoking) * P(diabetes)

By calculating these combined probabilities, we find features that have even stronger correlation that each individual feature

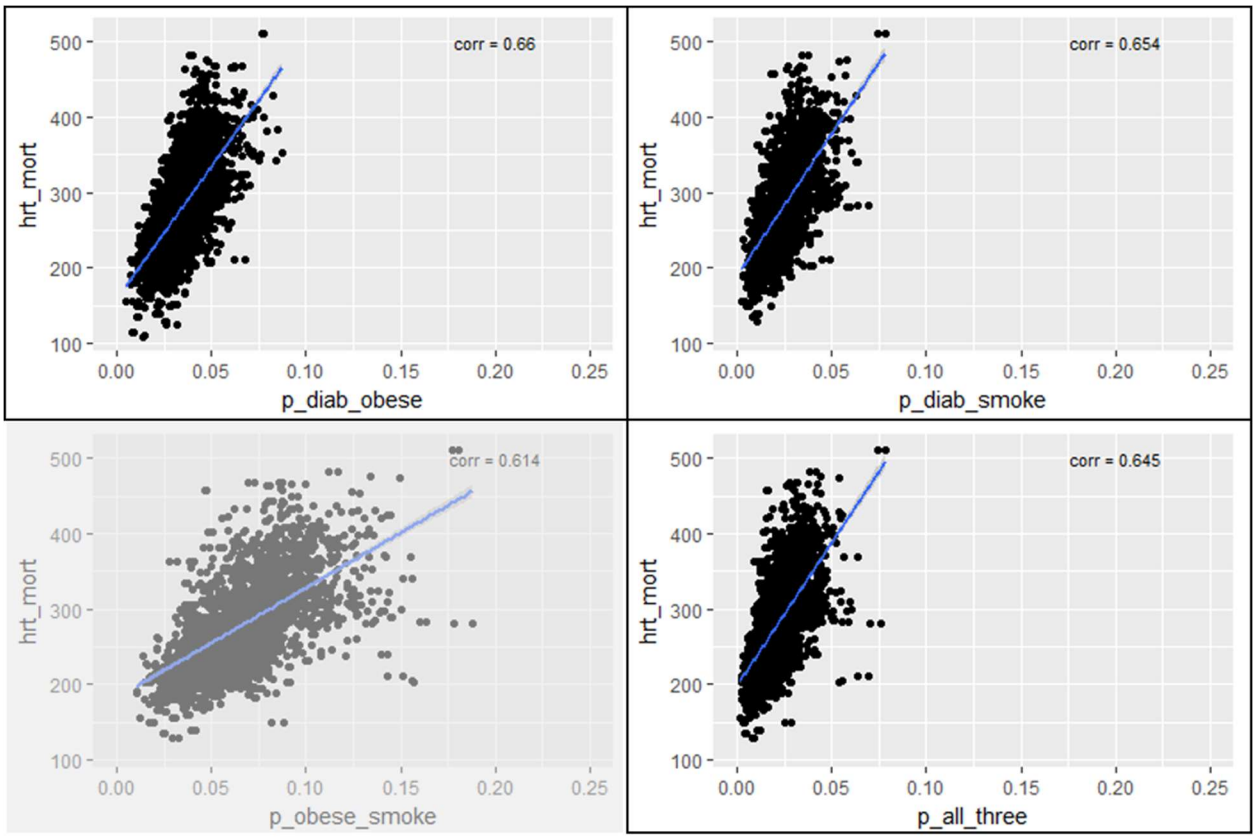

Figure 16 - Correlations between Combined Health Percentages and Heart Disease Mortality

The 3 combinations highlighted in Figure 16 have the strongest correlations (> 0.64) and are good candidates for the predictive model. In practice, the 4th combination did not add any predictive power to the model, so it was omitted.

DEATH RATES

Homicide rate and Motor Vehicle Death rate (both per 100,000 individuals) each showed moderate correlation with heart disease mortality (0.441 and 0.460, respectively). When each factor was converted to a $\log_{10}$ scale, the correlation got stronger (.501, .504, respectively).

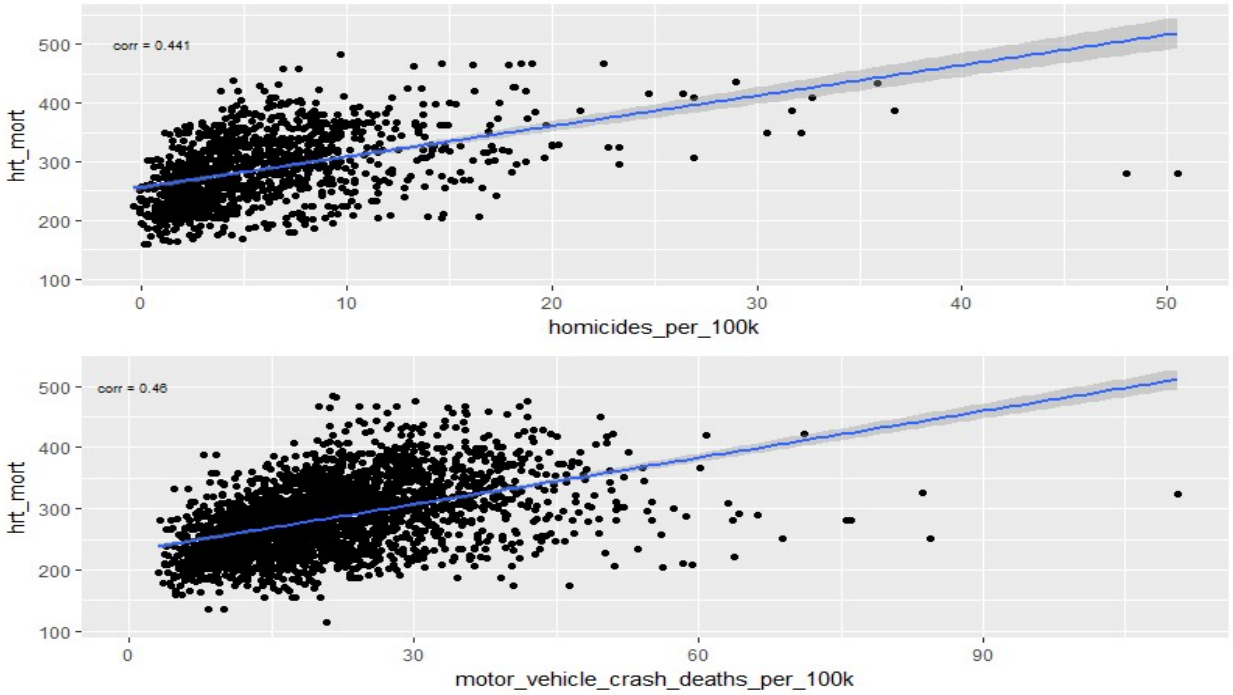

Figure 17 - Correlation between Death Rates and Heart Disease Mortality

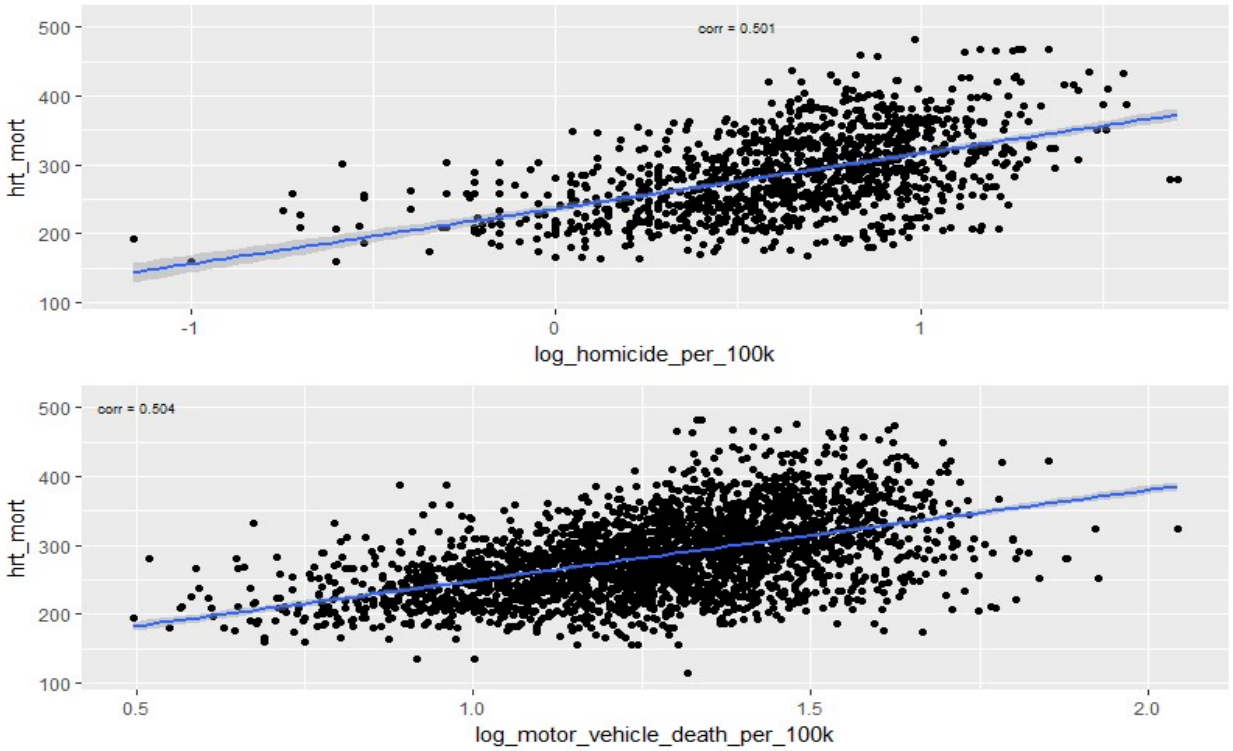

Figure 18 - Correlation between $\log_{10}$(Death Rates) and Heart Disease Mortality

Each version was used in separate iterations of the predictive model, and surprisingly, the "base" version (prior to $\log_{10}$ conversion) showed marginally better predictive power.

PREDICTIVE MODEL

In order to predict heart disease mortality rate, we need a predictive model that provides a numerical answer, implying that the model will be a regression model. There are many metrics that can show the quality of a regression models:

- Least Square Error (LSE, aka Mean Square Error, aka Sum of Squares Error (SSE))
- Root Mean Square Error (square root of LSE)
- Total Variation (aka Total Sum of Squares - SST)
- Regression Sum of Squares (SSR)
- R^2

For this model we will be using the Root Mean Square Error (RMSE) statistic to determine the model's quality. RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}(\hat{\text{y}}_{\text{i}} - \text{y}_{\text{i}})^2}$$

Where N is the number of data records in the sample, $\hat{y}_i$ is the predicted value of the *i*th member of the sample, and y_i is the actual value of the *i*th member of the sample. By minimizing this value, we are essentially minimizing the absolute error of the model.

Three types of regression models were tested

- Linear Regression
- Boosted Decision Tree
- Decision Forest Regression

Each of the models was given the same set of features, and were trained using 50% of the model data and scored with the remaining 50%. Comparison between the 3 different models showed that the Decision Forest Regression model (RMSE ≈ 32) performed the better than Linear Regression (RMSE ≈ 36) and Boosted Tree (RMSE ≈ 34).

Once the Decision Forest was chosen, further data steps were undertaken to further refine the predictive power of the model.

- Filling in missing data values by substituting with the mean value of the feature
- Normalizing the data using the "Z-Score" method: $Z = \frac{x - mean(x)}{stdev(x)}$ where mean(x) and stdev(x) are the mean and standard deviation of the feature in question

After the RMSE was minimized against the training set of data, scoring of the model was performed against a separate test set of data, comprising an additional 3080 rows. Various permutations of feature selections were performed, with the best model predicting values for the testing set (with unknown values) with a RMSE of **36.6206** as calculated by DataScienceCapstone.org

CONCLUSION

Predicting heart disease mortality rate is a challenging endeavor, as there are many "environmental" (i.e., not necessarily specific to an individual such as genetics) factors that can affect the outcome. Within the given data set, we see that heart disease mortality can be predicted reasonably well based on the following features:

- Economic Typology (categorical)
- Population (categorical)
- Metropolitan Type (categorical)
- Air Pollution concentration (categorical)
- Percent of Civilian Labor (numeric)
- Percent non-Hispanic African American (numeric)
- Percent of Adult Physical Inactivity (numeric)
- Percent of adults with less than a High School diploma (numeric)
- Percent of adults with Bachelor's Degree or higher (numeric)
- Percent of adults that are obese and have diabetes (numeric)
- Percent of adults that smoke and have diabetes (numeric)
- Percent of adults that are obese, smoke, and have diabetes (numeric)
- Homicides per 100,000 individuals (numeric)
- Motor vehicle crash deaths per 100,000 individuals (numeric)