DAT102x: Predicting Heart Disease Mortality

Executive Summary

This document presents an analysis of data concerning the rate of heart disease (per 100,000 individuals) across the United States at the county-level from other socioeconomic indicators. The analysis is based on 3,198 observations which is compiled from a wide range of sources and made publicly available by the United States Department of Agriculture Economic Research Service (USDA ERS).

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between county-level characteristics and rate of heart disease were identified. After exploring the data, a regression model to predict the rate of heart disease from its features was created.

After performing the random forest model, significant features were found in this analysis were:

- **Does not have a high school diploma** - % of population does not have a high school diploma (US Census, American Community Survey)
- **Adult that is physically inactive** - % adult population that is physically inactive (National Center for Chronic Disease Prevention and Health Promotion)
- **Adults who obese** - % adults who meet clinical definition of obese (National Center for Chronic Disease Prevention and Health Promotion)
- **Population with diabetes** - % population with diabetes (National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation)
- **Babies born with low birth weight** - % babies born with low birth weight (National Center for Health Statistics)
- **Deaths per 1,000 of population** - % Deaths per 1,000 of population (US Census Population Estimates)
- **High school diploma** - % adult population which has a high school diploma as highest level of education achieved (US Census, American Community Survey)
- **Bachelor's degree or higher** - % adult population which has a bachelor's degree or higher as highest level of education achieved (US Census, American Community Survey)
- **Civilian labor force rate** - Civilian labor force, annual average, as percent of population (Bureau of Labor Statistics, http://www.bls.gov/lau/)
- **Adults who smoke** - % adults who smoke (Behavioral Risk Factor Surveillance System)

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics, while the features can be group into 4 groups.

1. AREA — Information about the county
2. ECON — economic indicators
3. HEALTH — health indicators
4. DEMO — demographics information

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 3,198 observations are shown here:

ECON

As % of population	Count	Mean	Median	std	Min	Max
Civilian labor force rate	3,198	46.72%	46.80%	7.44%	20.70%	100.00%
Unemployment rate	3,198	5.97%	5.70%	2.29%	1.00%	24.80%
Adults without health insurance	3,196	21.75%	21.60%	6.74%	4.60%	49.60%
Children without health insurance	3,196	8.61%	7.70%	3.98%	1.20%	28.10%

HEALTH – PART 1

% of	Count	Mean	Median	std	Min	Max
Adults who obese	3,196	30.77%	30.90%	4.32%	13.10%	47.10%
Adults who smoke	2,734	21.36%	21.00%	6.29%	4.60%	51.30%
Population with diabetes	3,196	10.93%	10.90%	2.32%	3.20%	20.30%
Babies born with low birth weight	3,016	8.39%	8.10%	2.23%	3.30%	23.80%
Adult engages in excessive consumption of alcohol	2,220	16.48%	16.40%	5.05%	3.80%	36.70%
Adult that is physically inactive	3,196	27.72%	28.00%	5.30%	9.00%	44.20%

HEALTH – PART 2

	Count	Mean	Median	std	Min	Max
Fine particulate matter in µg/m³	3,170	11.63	12.00	1.56	7.00	15.00
Deaths by homicide per 100,000 population	1,231	5.95	4.70	5.03	- 0.40	50.49
Deaths by motor vehicle crash per 100,000 population	2,781	21.13	19.63	10.49	3.14	110.45
Population per dentist	2,954	3,431	2,690	2,569	339	28,130
Population per Primary Care Physician	2,968	2,551	1,999	2,100	189	23,399

DEMO - BASIC

As % of population	Count	Mean	Median	std	Min	Max
Female	3,196	49.88%	50.30%	2.44%	27.80%	57.30%
< 18 years of age	3,196	22.77%	22.60%	3.43%	9.20%	41.70%
>=65 years of age	3,196	17.00%	16.70%	4.37%	4.50%	34.60%
Hispanic	3,196	9.02%	3.50%	14.28%	0.00%	93.20%
African American	3,196	9.10%	2.20%	14.72%	0.00%	85.80%
Hispanic and White	3,196	77.00%	85.30%	20.79%	5.30%	99.00%
Native American	3,196	2.47%	0.70%	8.46%	0.00%	85.90%
Asian	3,196	1.31%	0.70%	2.54%	0.00%	34.10%

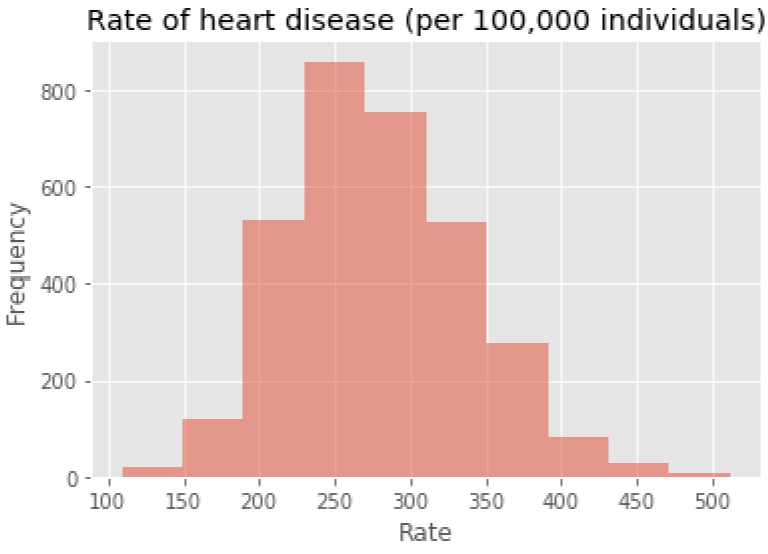
DEMO - HIGHEST LEVEL OF EDUCATION ACHIEVED FOR ADULT POPULATION

As % of population	Count	Mean	Median	std	Min	Max
Does not have a high school diploma	3,198	14.88%	13.32%	6.82%	1.51%	47.35%
High school diploma	3,198	35.06%	35.50%	7.06%	6.53%	55.89%
Some college	3,198	30.11%	30.16%	5.23%	10.95%	47.40%
Bachelor's degree or higher	3,198	19.95%	17.65%	8.93%	1.11%	79.90%

DEMO – BIRTHS/DEATHS

	Count	Mean	Median	std	Min	Max
Births per 1,000 of population	3,198	11.68	11.00	2.74	4.00	29.00
Deaths per 1,000 of population	3,198	10.30	10.00	2.79	-	27.00

A histogram of the **Rate of heart disease (per 100,000 individuals)** shows that the rate are bell-shaped but slightly left-skewed.



In addition to the numeric values, the observations include categorical features, including:

- **Rural-Urban Continuum Code**
  'Metro - Counties in metro areas of 1 million population or more'
  'Metro - Counties in metro areas of 250,000 to 1 million population'
  'Metro - Counties in metro areas of fewer than 250,000 population'
  'Nonmetro - Urban population of 20,000 or more, adjacent to a metro area'
  'Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area'
  'Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area'
  'Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area'
  'Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area'
  'Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area'

- **Urban Influence Codes**
  'Large-in a metro area with at least 1 million residents or more'
  'Small-in a metro area with fewer than 1 million residents'
  'Micropolitan adjacent to a large metro area'
  'Micropolitan adjacent to a small metro area'
  'Micropolitan not adjacent to a metro area'
  'Noncore adjacent to a large metro area'
  'Noncore adjacent to a small metro with town of at least 2,500 residents'
  'Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents'
  'Noncore adjacent to micro area and contains a town of 2,500-19,999 residents'
  'Noncore adjacent to micro area and does not contain a town of at least 2,500 residents'
  'Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents'
  'Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents'

- **County Typology Codes**
  'Manufacturing-dependent', 'Mining-dependent', 'Nonspecialized', 'Farm-dependent', 'Recreation’, 'Federal/State government-dependent'

- **Air Pollution Particulate Matter**
  - Fine particulate matter in µg/m³ (label 7 to 15)

- **Year**
  a or b

Bar charts were created to show frequency of these features, and indicate the following:

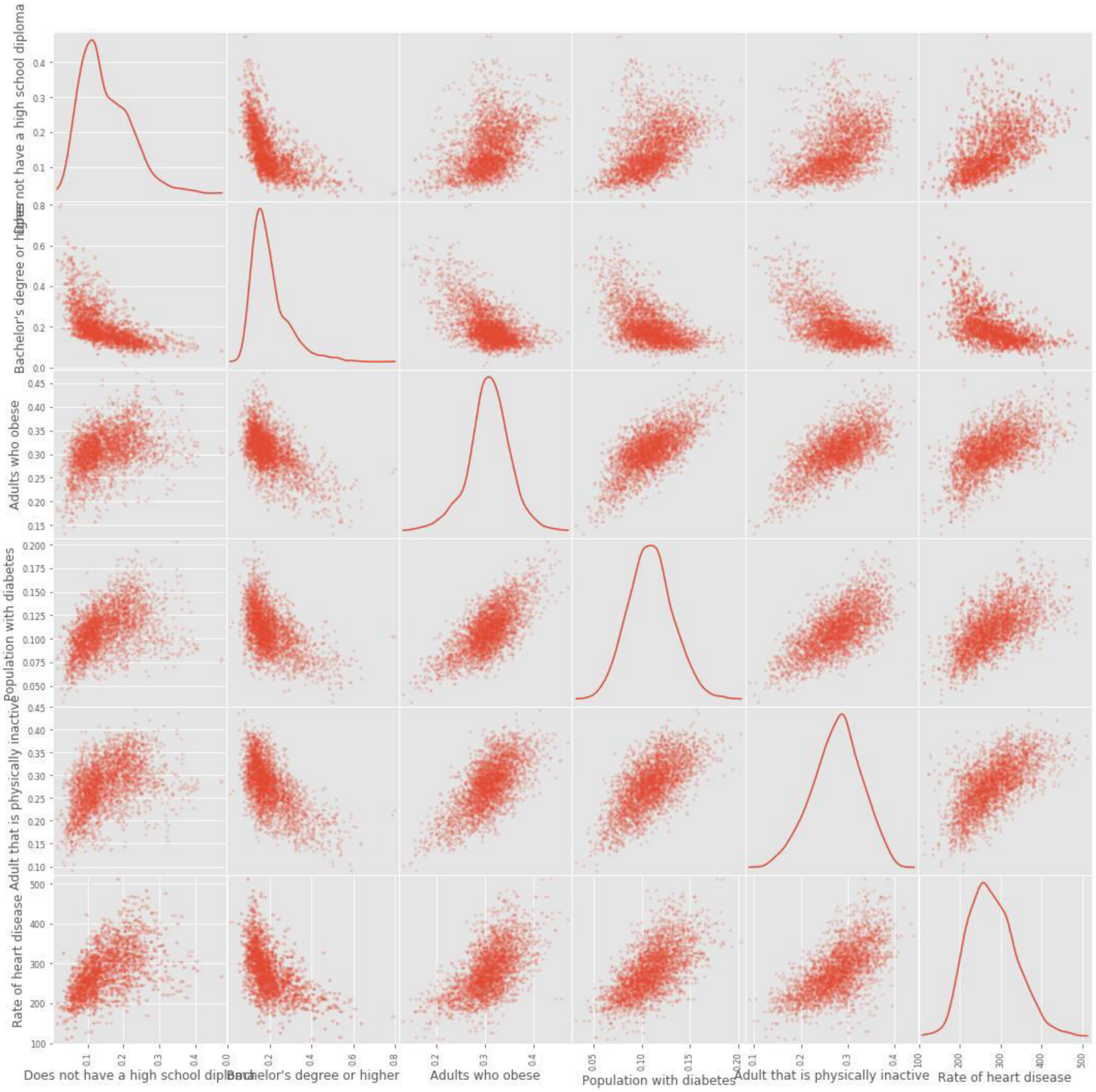
- 'Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area' are the most common group, followed by 'Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area' and 'Metro - Counties in metro areas of 1 million population or more'; 'Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area' is very uncommon.
- The 'Small-in a metro area with fewer than 1 million residents'
- 'Nonspecialized' is the dominant county typology, besides this group, 'Farm-dependent' and 'Manufacturing-dependent' are most common and having similar count.
- Samples are mainly in the range of Air Pollution Particulate Matter 10-13.

Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between **rate of heart disease** and the other features.

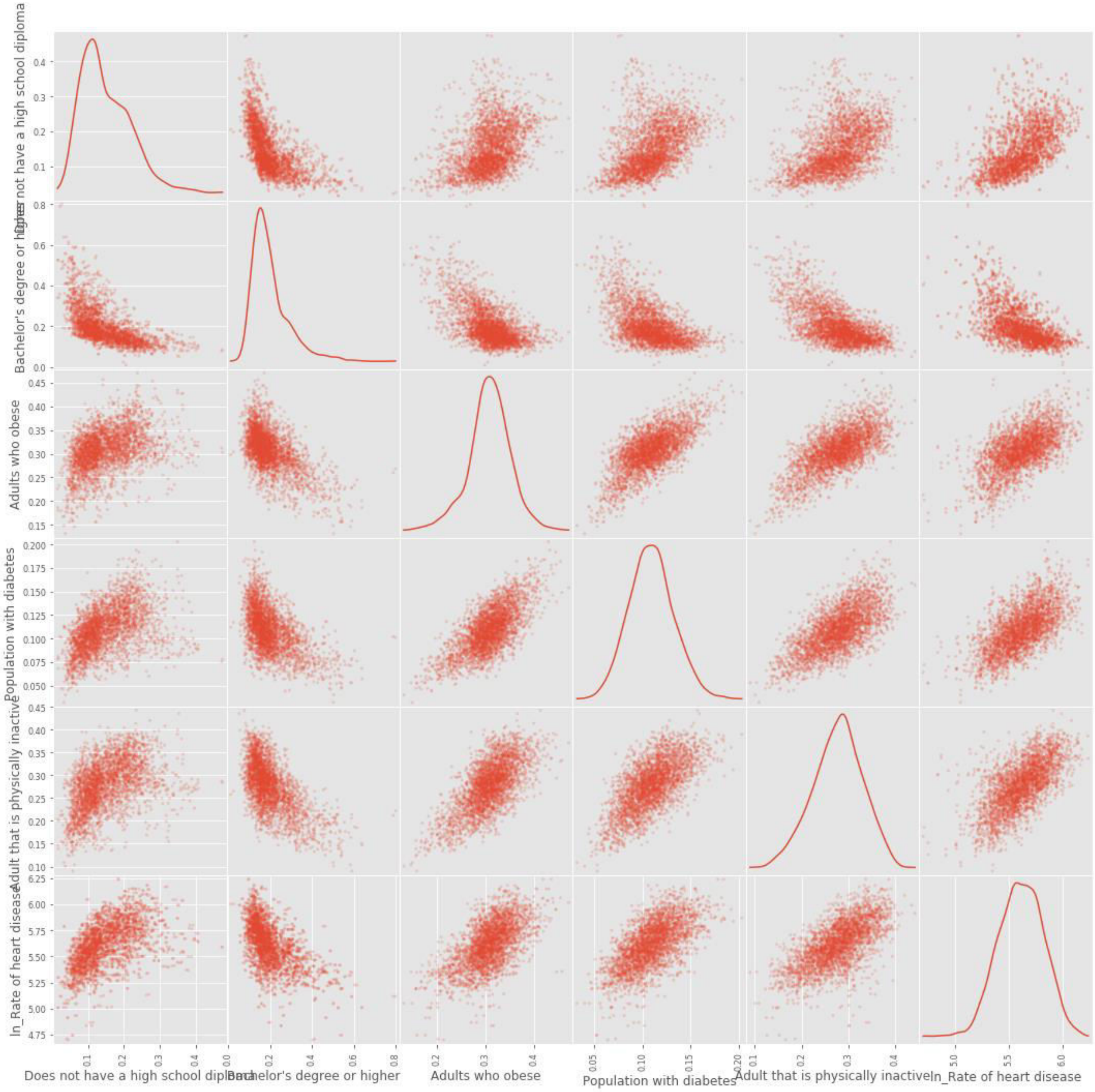
Numeric Relationships

The following scatter-plot matrix was generated initially to compare numeric features with one another. The key features in this matrix are shown here:

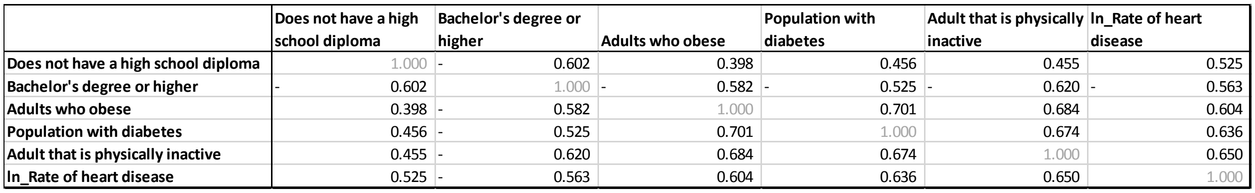


Viewing plots in the bottom row or the right-most column of this matrix shows an apparent relationship between rate of heart disease and other numeric features. Specifically, as percentage of adult without a high school diploma, obese, physically inactive and population with diabetes increase, so does rate of heart disease; and as percentage of adult with Bachelor's degree or higher increases, rate of heart disease reduces.

It can be seen from these plots that the relationships between numeric features and rate of heart disease often exhibits a “curved” nature that is not quite linear. In an attempt to improve the fit of the features to rate, the log normal value for rate was calculated. The resulting scatter-plot matrix shows increased linearity in the relationships between log- rate of heart disease and the other numeric features:

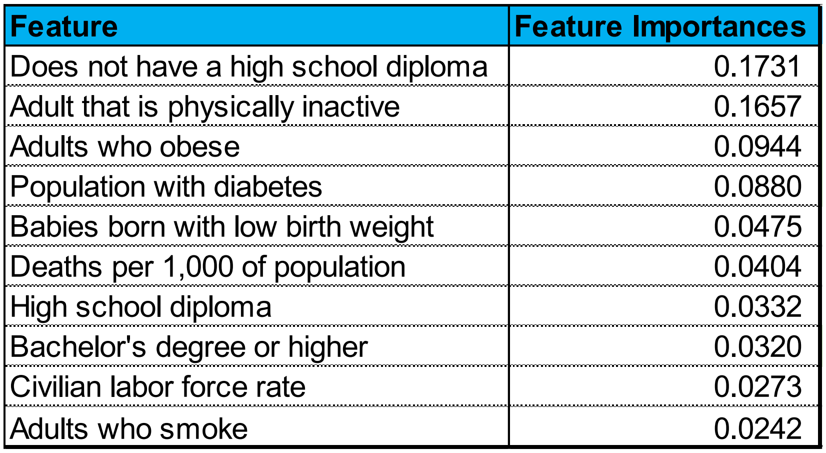


The correlation between the numeric columns was then calculated with the following results:



	Does not have a high school diploma	Bachelor's degree or higher	Adults who obese	Population with diabetes	Adult that is physically inactive	ln_Rate of heart disease
Does not have a high school diploma	1.000	- 0.602	0.398	0.456	0.455	0.525
Bachelor's degree or higher	- 0.602	1.000	- 0.582	- 0.525	- 0.620	- 0.563
Adults who obese	0.398	- 0.582	1.000	0.701	0.684	0.604
Population with diabetes	0.456	- 0.525	0.701	1.000	0.674	0.636
Adult that is physically inactive	0.455	- 0.620	0.684	0.674	1.000	0.650
ln_Rate of heart disease	0.525	- 0.563	0.604	0.636	0.650	1.000

These correlations validate the plots by showing a negative correlation between ‘percentage of adult having Bachelor's degree or higher’ and ln(rate of heart disease), and moderate to strong positive correlations for the other numeric features.

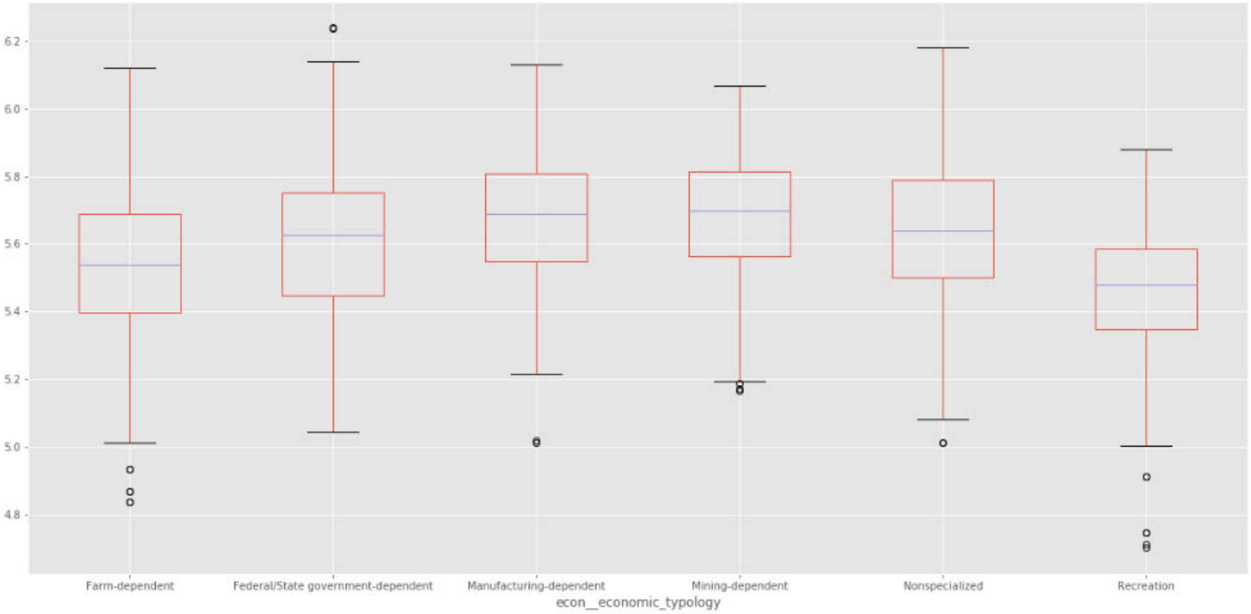


Feature	Feature Importances
Does not have a high school diploma	0.1731
Adult that is physically inactive	0.1657
Adults who obese	0.0944
Population with diabetes	0.0880
Babies born with low birth weight	0.0475
Deaths per 1,000 of population	0.0404
High school diploma	0.0332
Bachelor's degree or higher	0.0320
Civilian labor force rate	0.0273
Adults who smoke	0.0242

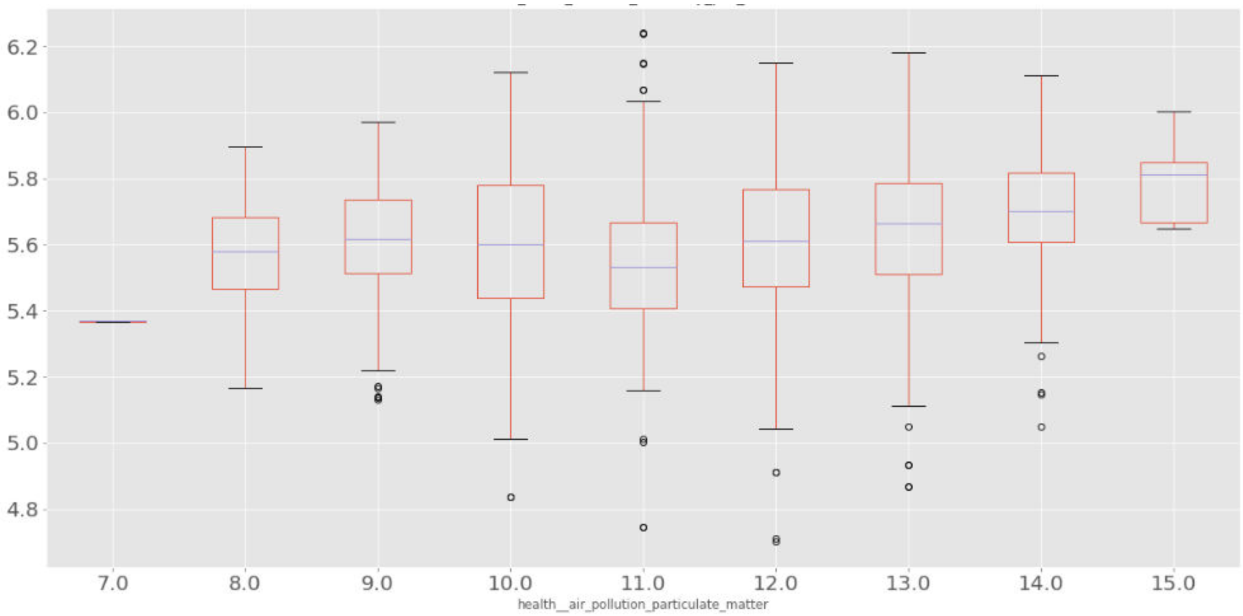
Categorical Relationships

Having explored the relationship between rate of heart disease and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and rate. The following box-plots show the categorical columns that seem to exhibit a relationship with the log of heart disease rate:

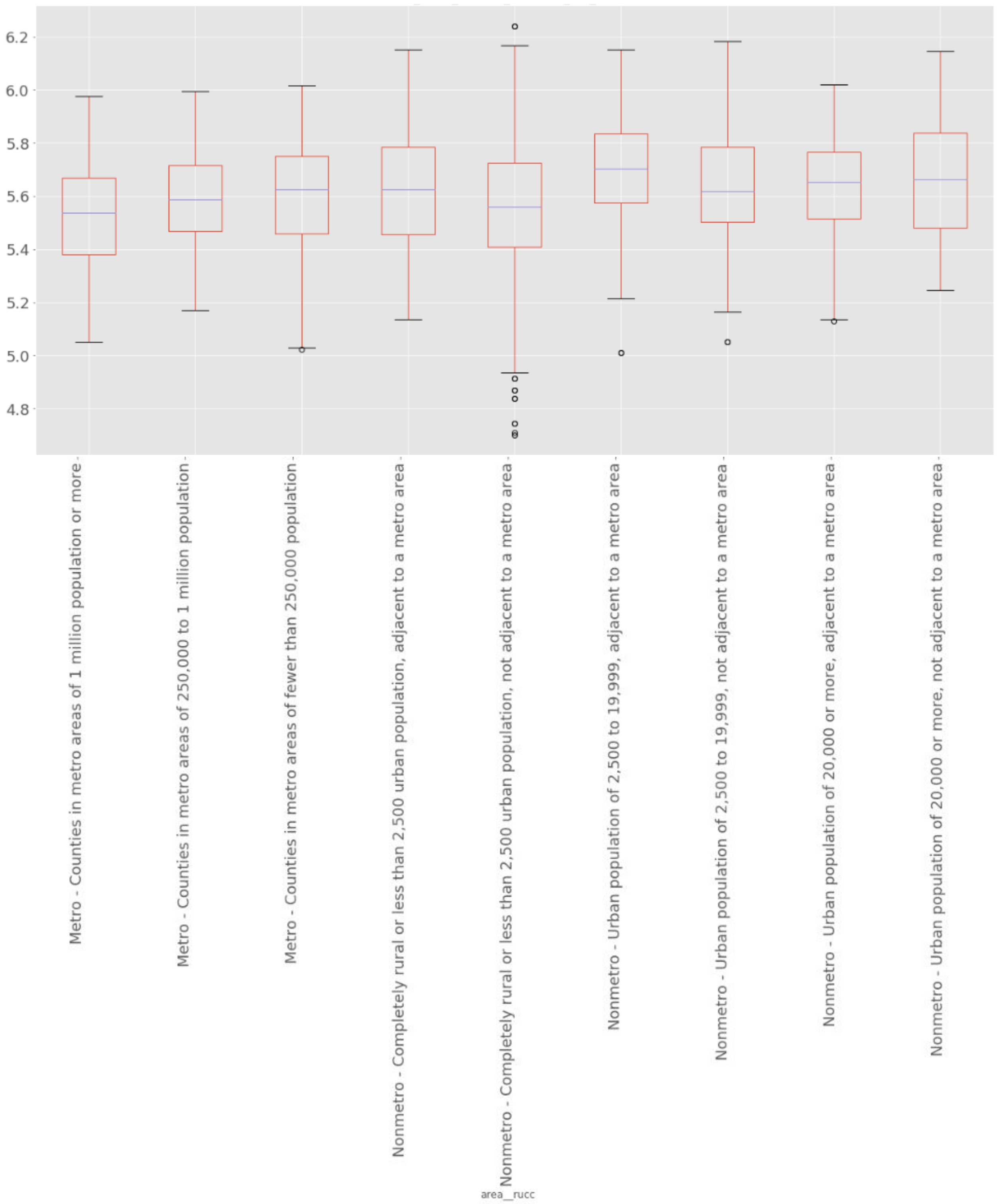
ln heart disease rate by County Typology Codes



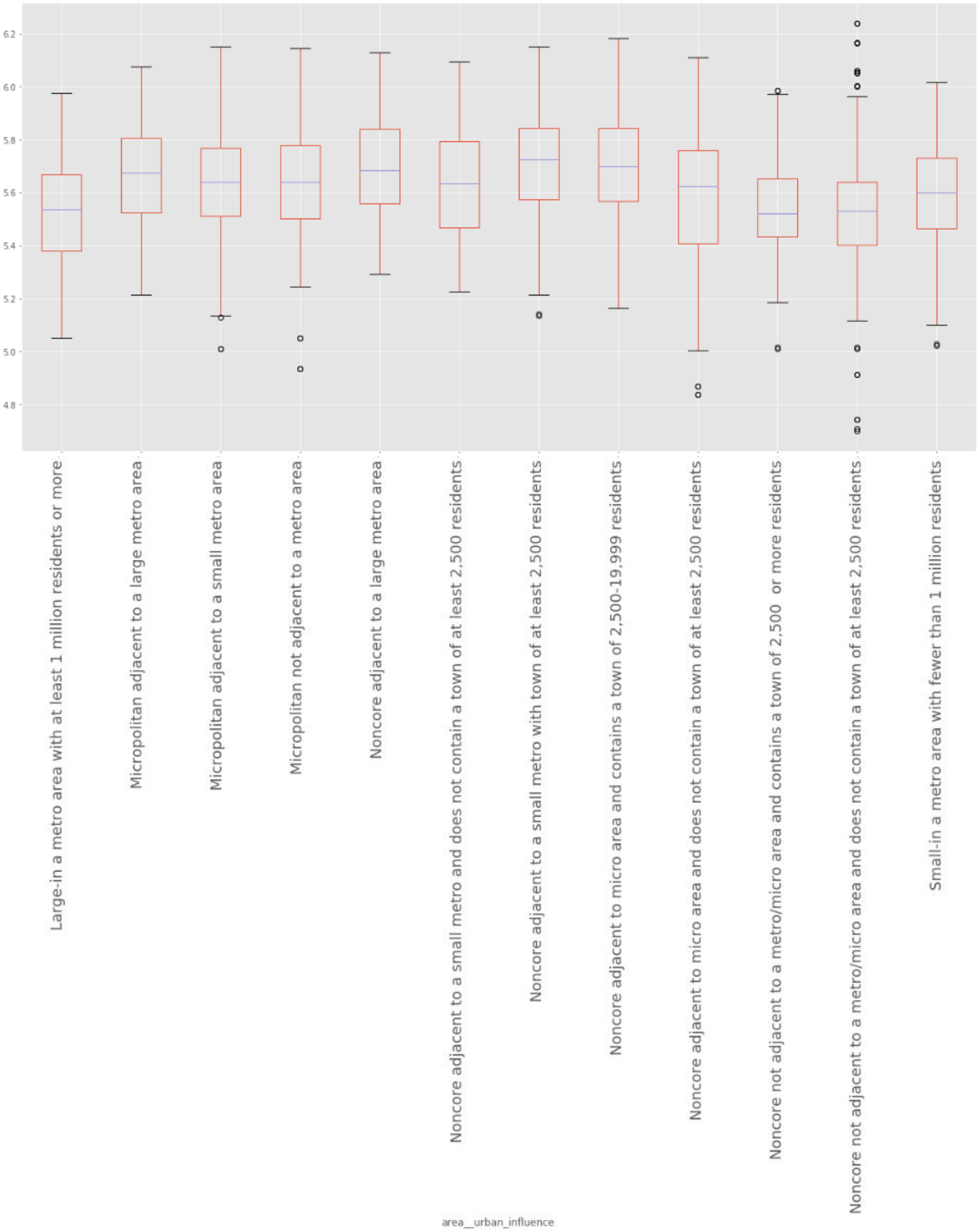
ln heart disease rate by Air Pollution Particulate Matter



ln heart disease rate by Rural-Urban Continuum Code



ln heart disease rate by Urban Influence Codes



The box plots show some clear differences in terms of the median and range of rate for different categorical features. For example:

- Citizen in 'Manufacturing-dependent' and 'Mining-dependent' countries has higher heart disease rate, while 'Farm-dependent' and 'Recreation' has lower.
- Countries with lower fine particulate matter are significantly having lower heart disease rate, they could be combined into 4 groups, 7, 8-12, 13-14 and 15 μg/m³
- Heart disease rate in metro area have negative correlation with their population size; for countries in non-metro area with small and median urban population, both of them will have a significant higher heart disease rate if they are adjacent to a metro area.
- Heart disease rate of 'noncore not adjacent to a metro/micro area' and 'Small/ Large' countries are lower than 'Noncore adjacent' and 'Micropolitan' group.

Classification of heart disease rate

Based on the analysis of heart disease rate data, a predictive model to classify the rate into two categories: *Low* (heart disease rate less than or equal to 280) and *Premium* (heart disease rate 280 or more). The model was created using the Random Forest algorithm and trained with 80% of the data. Testing the model with the remaining 20% of the data yielded the following results:

- True Positives: 239
- True Negatives: 310
- False Positives: 37
- False Negatives: 54

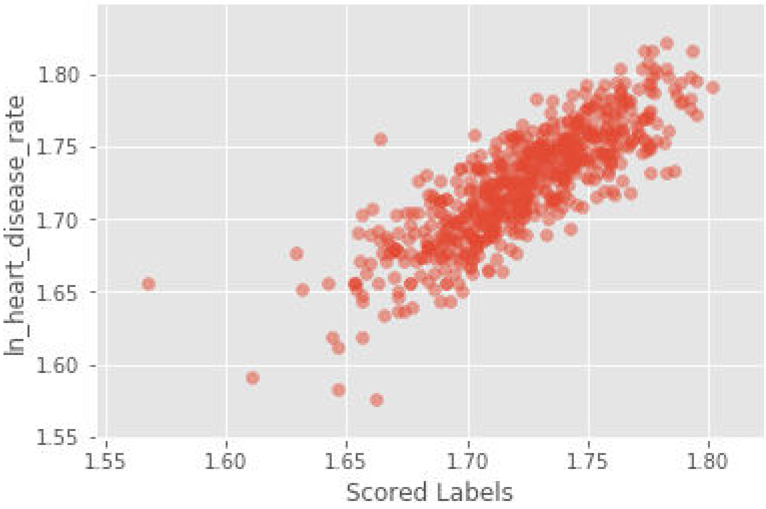
And here are the standard performance metrics for classification:

- Accuracy: 86.4%
- Precision: 86.6%
- Recall: 81.6%
- F1 Score: 84.0%

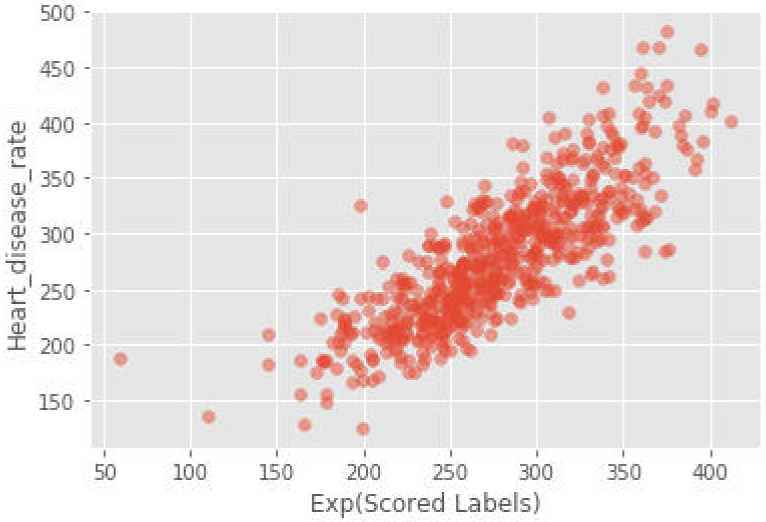
Regression

After creating a classification model to predict heart disease rate categories, a regression model to predict the actual heart disease rate was created. Based on the apparent relationships identified when analyzing the data, a lasso regression model was created to predict the heart disease rate, with the numeric feature (instead of heart disease rate) being taken ln and normalized.

The model was trained with 80% of the data and tested with the remaining 20%. A scatter plot showing the predicted log heart disease and the actual log rate is shown below:



This plot shows a clear linear relationship between predicted and actual values in the test dataset. The Root Mean Square Error (RMSE) for the test results is 0.02. Since the model predicts the log of heart disease, and not the heart disease itself, this figure does not represent the monetary amount by which the predicted value varies from the actual value. When the predicted log rate is converted back to its exponential value, the following scatter plot shows the results, with RMSE = 32.



Conclusion

This analysis has shown that the rate of heart disease (per 100,000 individuals) can be confidently predicted from the features 'Does not have a high school diploma', 'Bachelor's degree or higher', 'Adults who obese', 'Population with diabetes', 'Adult that is physically inactive', and 'Rate of heart disease'.