DAT102x: Predicting Heart Disease Mortality

Reshma Patel July 2018

Executive Summary

This document presents a prediction of the rate of heart disease (per 100,000 individuals) across the United States at the county-level from other socioeconomic indicators. We also need to find how poverty and other factors influence health, which may help to understand what indicators are related to heart disease prevalence rates in counties across the United States.

The data is compiled from a wide range of sources and made publicly available by the United States Department of Agriculture Economic Research Service (USDA ERS). There are 33 variables in this data set. Each row in the data set represents a United States county, and the data set we are working with covers two particular years, denoted a, and b, having four types of categories- area, demographic, economic and health indicators which influence heart disease mortality rate.

From initial analysis and data exploration one can find that, The median mortality rate for counties considered metro is lower than that of counties considered non-metro. Prevalence of adult smoking (but not excessive drinking) is positively correlated mortality rate. Counties with relatively large older populations have a lower median mortality rate than counties with less than 20% older population. There is not a strong and obvious correlation between birth rate and mortality rate.

To perform feature selection and shrinking coefficients Lasso regression has been used. As observed in Lasso regression, some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model. As observed the series of coefficients of Lasso regression indicates following as the most influencing the heart disease mortality rate for given counties in USA.

the economic indicators such as

econ__pct_unemployment - Unemployment, annual average, as percent of population, having relatively low influence on heart disease mortality rate, according to survey the unemployed might likely have chance of not having insurance, which may cause in unaided health care. **demo__death_rate_per_1k** - Deaths per 1,000 of population, have also relatively low influence as heart disease mortality rate is part of this overall death rate in county. **health__pct_adult_obesity** - Percent of adults who meet clinical definition of obese, the persons with obesity are more likely to have diabetes and might have physical inacticity in later life which are the larger influencing indicators for the heart disease.

health__pct_adult_smoking- Percent of adults who smoke significant influence on heart disease.

demo__pct_adults_with_high_school_diploma- Percent of adult population that does not have a high school diploma have lower chance of good job and likely not to have insurance for the health care.

econ__pct_uninsured_adults - Percent of adults without health insurance, might not have good health care benefits in terms of economic condition.

demo__pct_female- Percent of population that is female, has low chance of family & social support and community safety which may cause in less access to insured health benefits. **demo__pct_adults_less_than_a_high_school_diploma** - Percent of adult population that does not have a high school diploma. Higher rates of educational achievement are linked to better jobs and higher incomes resulting in better health and insurance benefits.

health__pct_physical_inacticity- Percent of adult population that is physically inactive, the higher rate of population which are physical inactive is directly proportional to heart disease mortality.

health__pct_low_birthweight- Percent of babies born with low birth weight have significant effect on heart disease as child with low birth weight might have worse health outcome in later life.

health__pct_diabetes- Percent of population with diabetes, is the largest influence on heart disease.

Data exploration and analysis

The individual features and their statistics presents the minimum, maximum, Mean, Median, Standard deviation and Count for each independent variable.

For heart disease mortality rate, the following statistics have been emerged.

Table 1: Summary of statistics of each heart disease mortality rate.

count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**median**
3198.00	279.369	58.9533	109.00	237.000	275.000	317.000	512.000	275.0

The minimum mortality rate of heart disease id 109.00, while the maximum is 512.00. The median mortality rate 275.0 helps to determine the measure of central tendency and is quite helpful when imputing missing data. The standard deviation is the square root of sample variance (which is a measure of the variability (spread or dispersion) of data) indicates value of 58.9533. A small variance indicates it is clustered closely around the mean.

Since heart disease mortality rate in this analysis, it has been noted that the mean (279.369) and median(275.0) of this value are significantly closer and that the comparatively small standard deviation(58.9533) indicates that there is considerable small dispersion of data for the heart disease mortality rate.

The histogram for heart disease mortality rate indicates that the mortality rate has the normal distribution over the frequency spread.

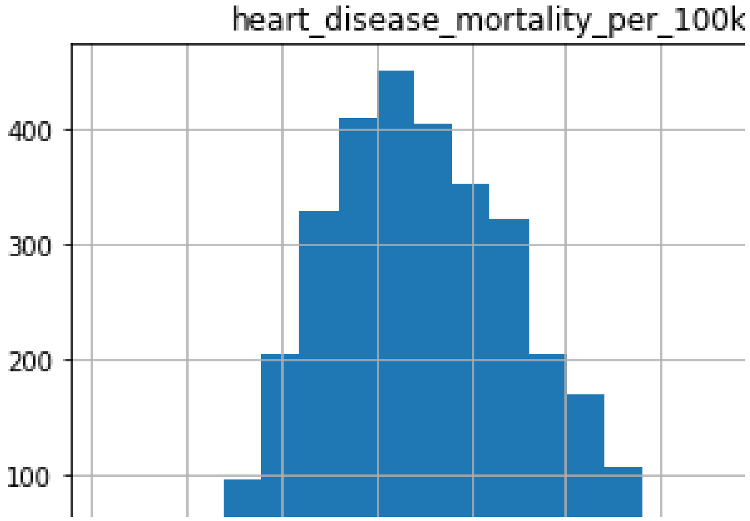

Figure 1: Histogram of heart disease mortality rate

In the following table the various statistics for other predictor variables are shown below.

Table 2: Summary of statistics of each feature.

	count	mean	std	min	max
row_id	3198	3116.98	1830.23	0	6276
econ__pct_civilian_labor	3198	0.46719	0.07439	0.207	1
econ__pct_unemployment	3198	0.05969	0.02294	0.01	0.248
econ__pct_uninsured_adults	3196	0.21746	0.06736	0.046	0.496
econ__pct_uninsured_children	3196	0.08606	0.03984	0.012	0.281
demo__pct_female	3196	0.49881	0.02439	0.278	0.573
demo__pct_below_18_years_of_age	3196	0.22771	0.03428	0.092	0.417
demo__pct_aged_65_years_and_older	3196	0.1700	0.04369	0.045	0.346
demo__pct_hispanic	3196	0.09020	0.14276	0	0.932
demo__pct_non_hispanic_african_american	3196	0.09104	0.14716	0	0.858
demo__pct_non_hispanic_white	3196	0.7699	0.20784	0.053	0.99
demo__pct_american_indian_or	3196	0.02468	0.08456	0	0.859

_alaskan_native					
demo__pct_asian	3196	0.01310	0.02543	0	0.341
demo__pct_adults_less_than_a_high_school_diploma	3198	0.14881	0.06820	0.01507	0.4735
demo__pct_adults_with_high_school_diploma	3198	0.35056	0.07055	0.06532	0.5589
demo__pct_adults_with_some_college	3198	0.30114	0.05231	0.10954	0.4739
demo__pct_adults_bachelors_or_higher	3198	0.19947	0.08930	0.01107	0.7989
demo__birth_rate_per_1k	3198	11.6769	2.73951	4	29
demo__death_rate_per_1k	3198	10.3011	2.78614	0	27
health__pct_adult_obesity	3196	0.30766	0.04322	0.131	0.471
health__pct_adult_smoking	2734	0.21362	0.06289	0.046	0.513
health__pct_diabetes	3196	0.10926	0.02321	0.032	0.203
health__pct_low_birthweight	3016	0.08389	0.02225	0.033	0.238
health__pct_excessive_drinking	2220	0.16484	0.05047	0.038	0.367
health__pct_physical_inacticity	3196	0.27716	0.05300	0.09	0.442
health__air_pollution_particulate_matter	3170	11.6258	1.55799	7	15
health__homicides_per_100k	1231	5.94749	5.03182	-0.4	50.49
health__motor_vehicle_crash_deaths_per_100k	2781	21.1326	10.4859	3.14	110.45
health__pop_per_dentist	2954	3431.43	2569.45	339	28130
health__pop_per_primary_care_physician	2968	2551.33	2100.45	189	23399
heart_disease_mortality_per_100k	3198	279.369	58.9533	109	512

The count of each feature indicates that, there significant amount of missing values in some of the features , and we need to impute the missing data with either most frequent values or median values.

The frequency distribution for each independent variable can be identified by following figure:

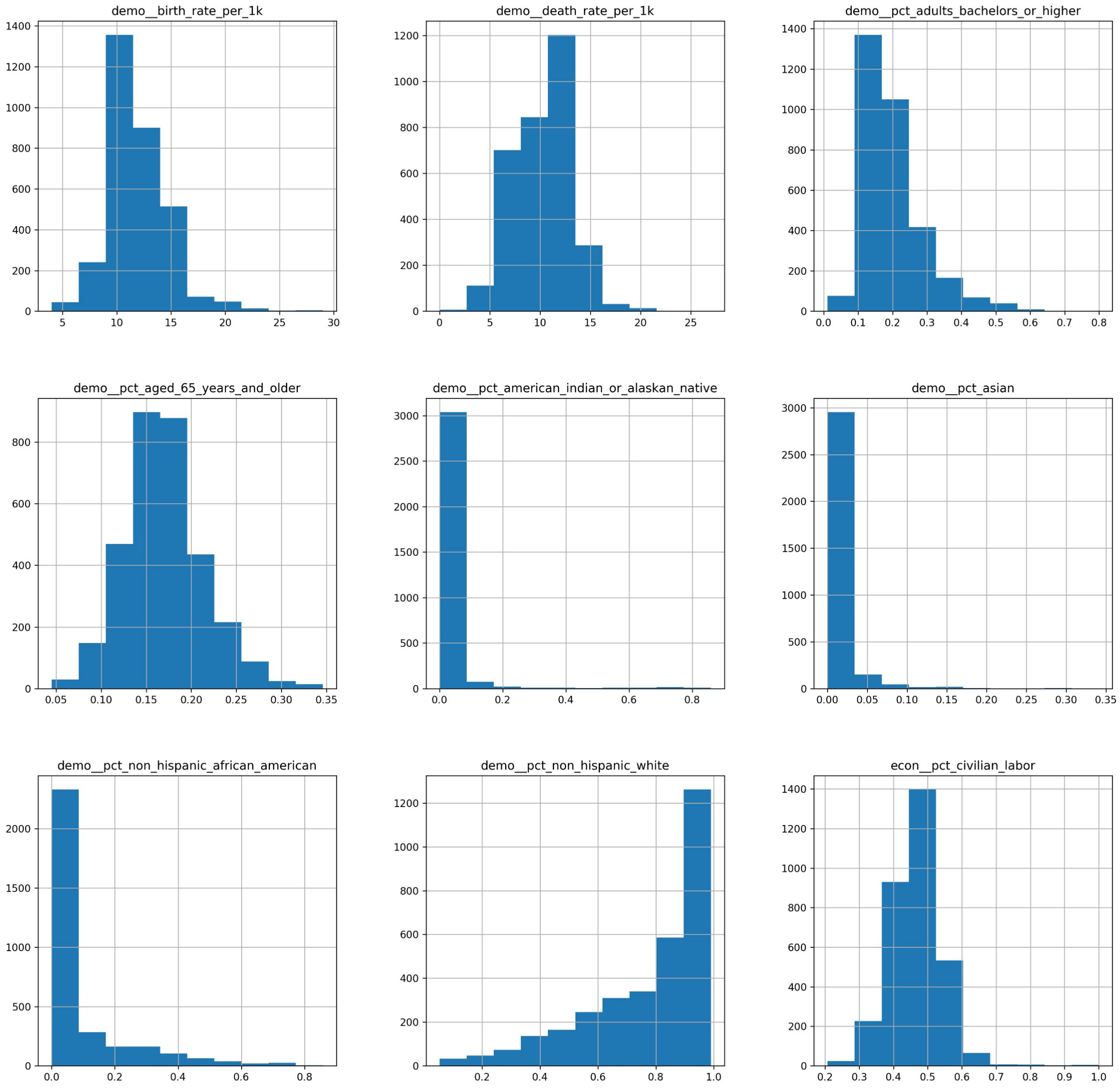

Figure 2: Histogram for birth rate, death rate, adults, aged over 65, and ethnicity.

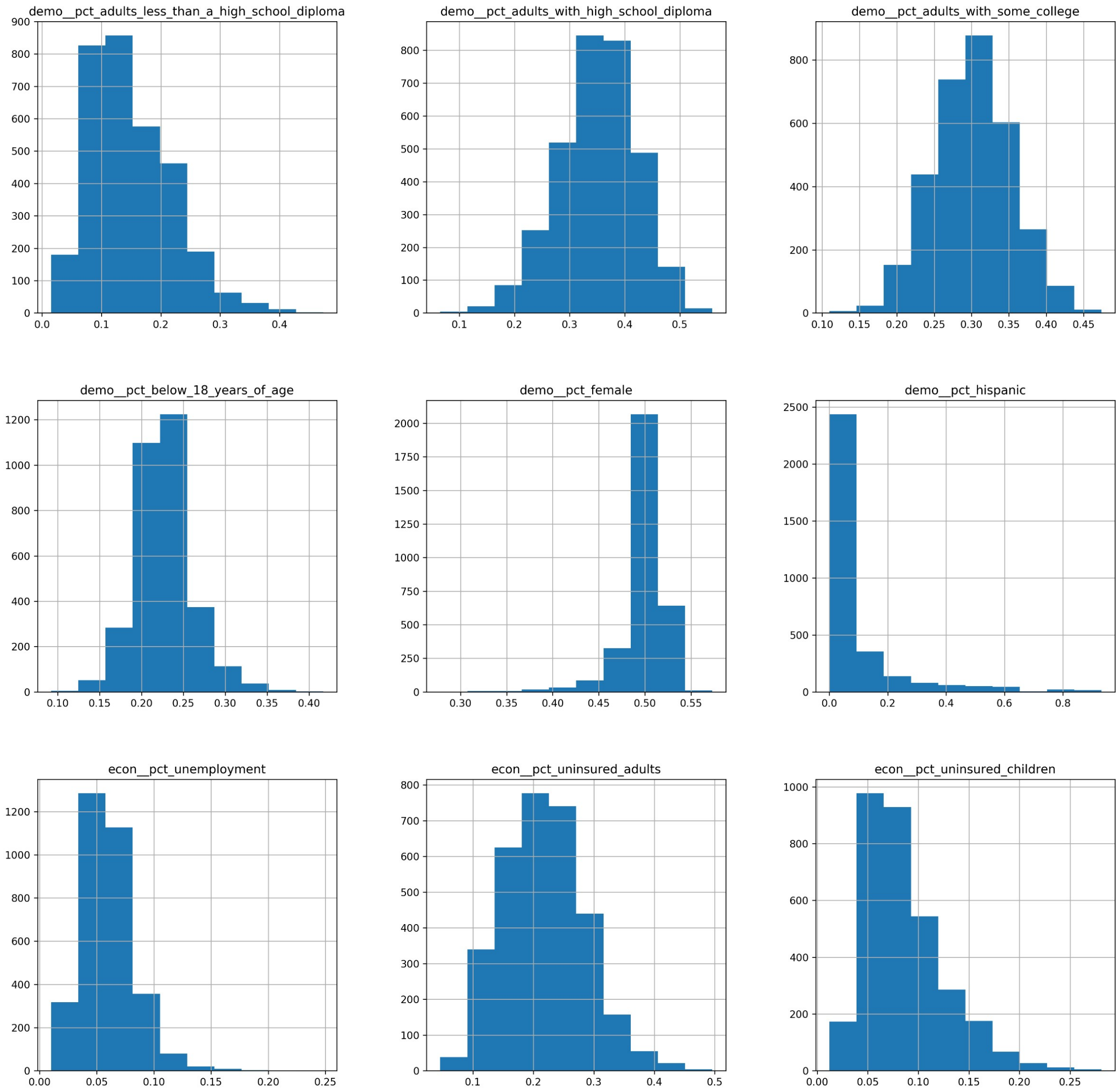

Figure 4: Histogram for economic factors and age, female and education factor

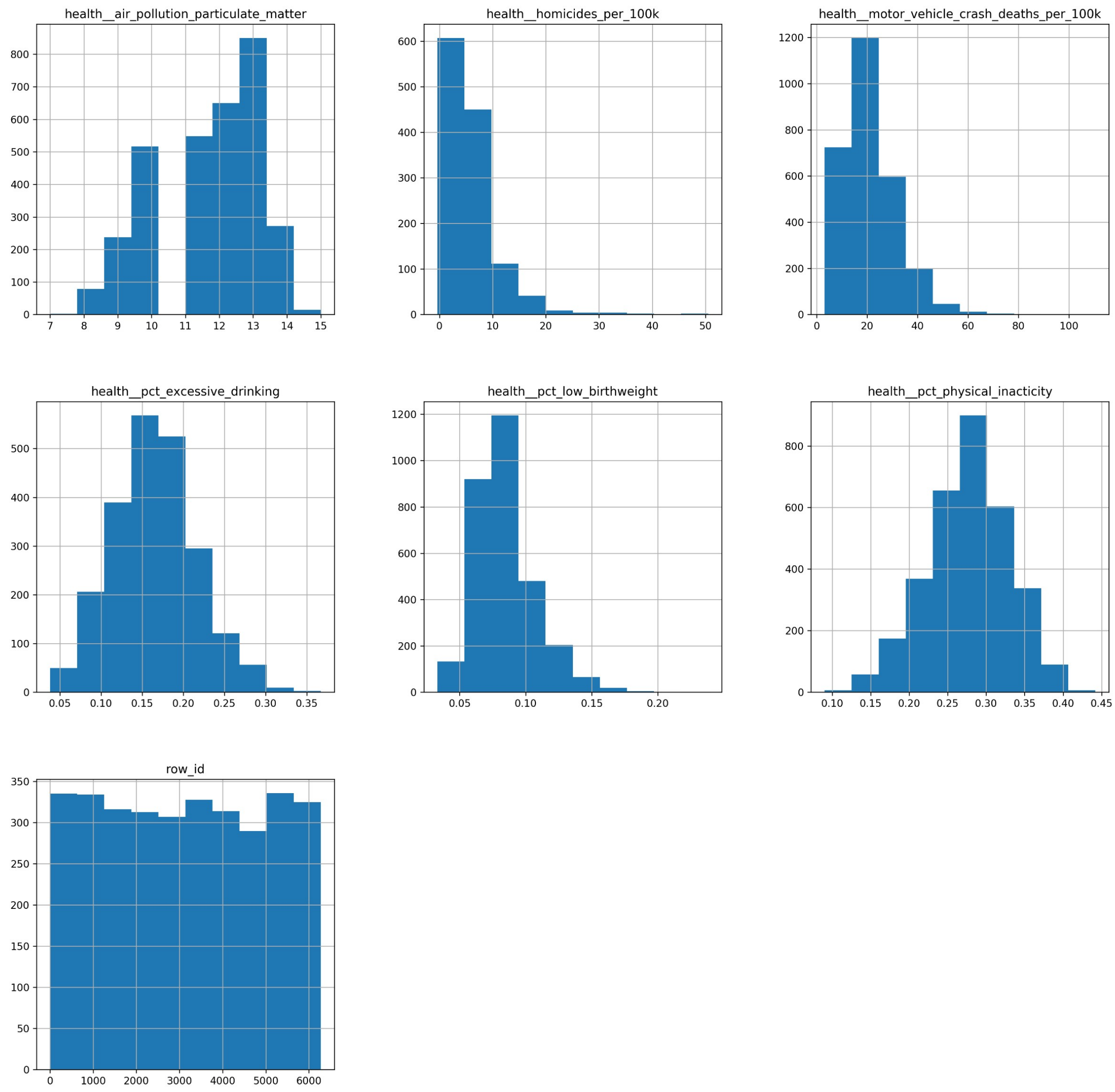

Figure 5: Histogram for health factors, air pollution, other death factors

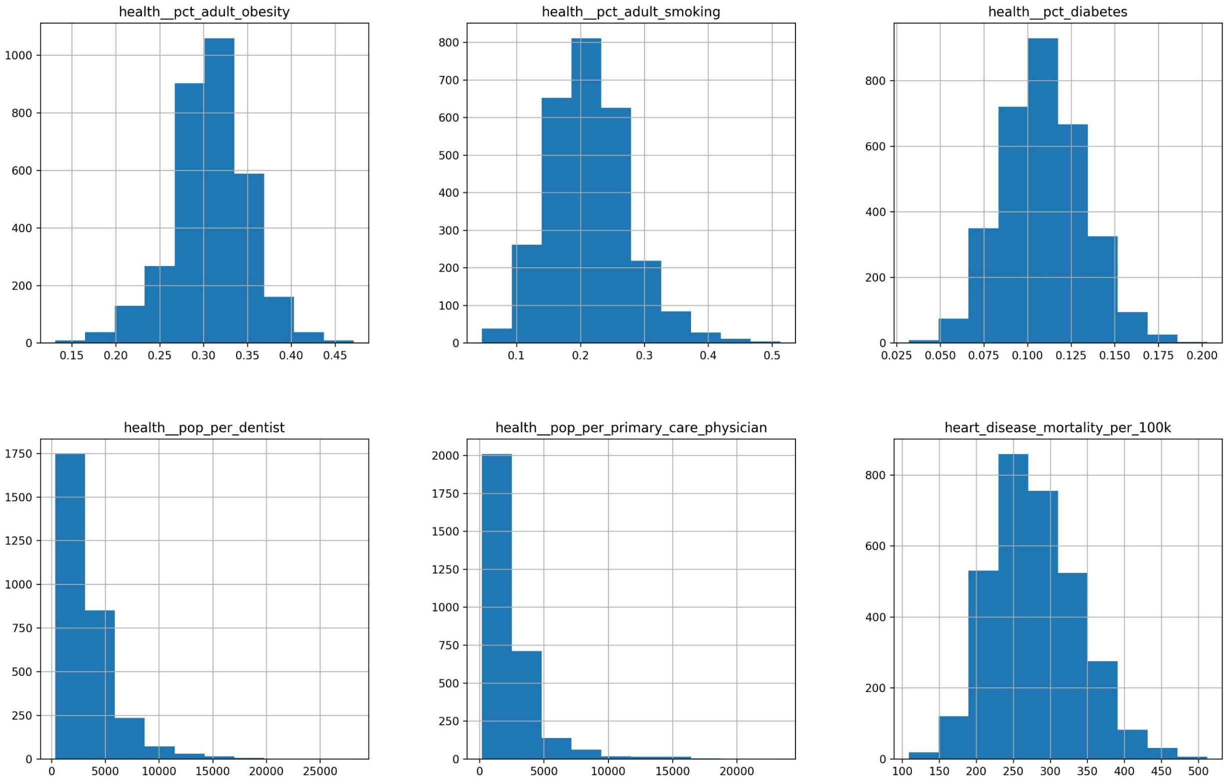

Figure 6: Histogram for health influencing factors such as dentist, primary care,diabetes and smoking

From the initial data exploration some other intuition came out as follows,

Metro-Nonmetro vs mortality rate:

- The median mortality rate for counties considered metro is lower than that of counties considered non-metro.
- The difference in median mortality rates between metro and non-metro counties is less than 100 deaths per 100,000.

These can be figured out from following chart plot and boxplot.

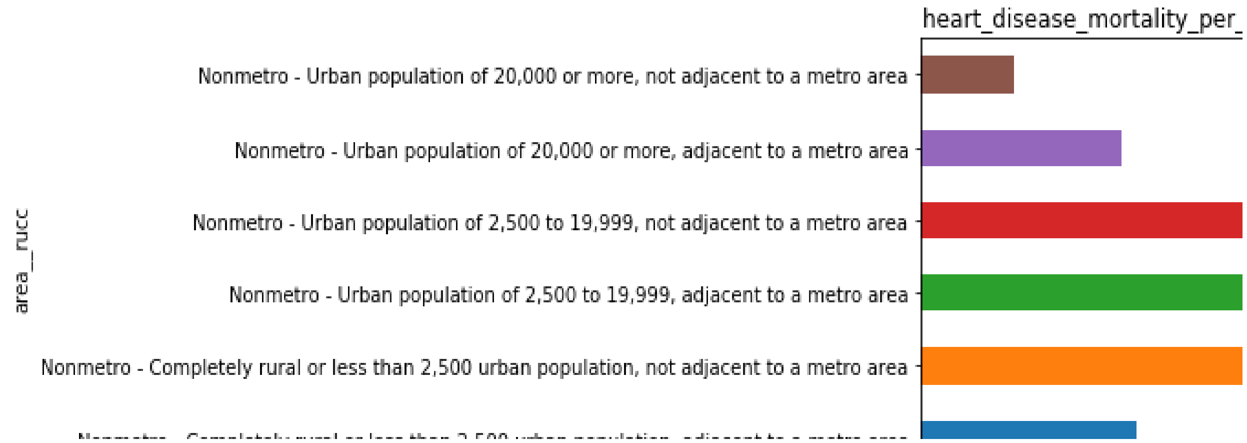

Figure 7: Heart disease mortality rate for Nonmetro counties

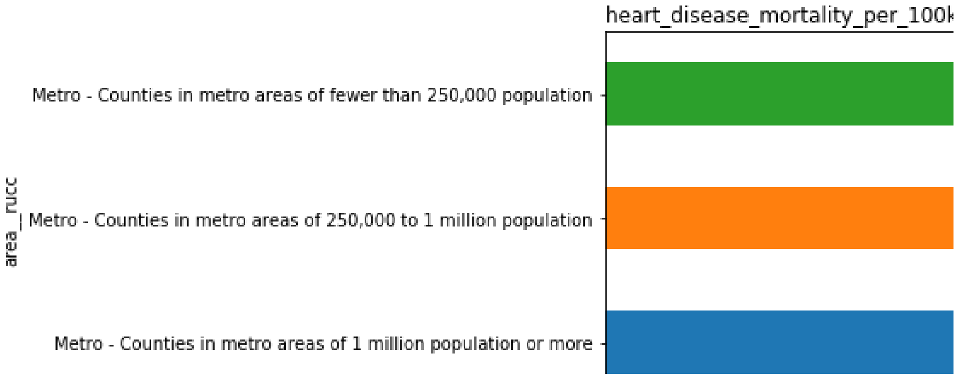

Figure 8: Heart disease mortality rate for Metro counties

The following figure shows the median mortality rate difference between Metro and Non-metro counties which is 21.0.

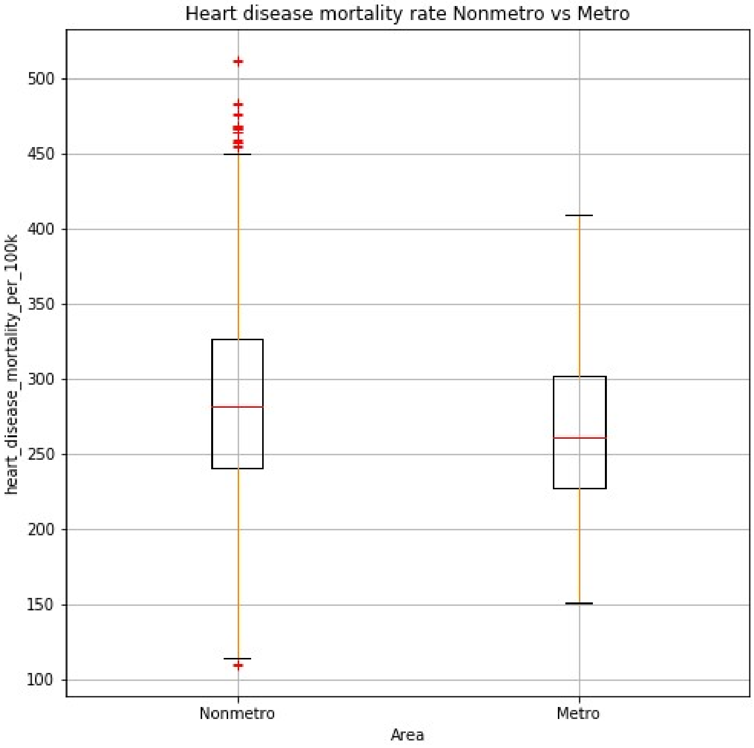

Figure 9: Mortality rate for Metro vs Non-metro

Birth rate vs mortality rate:

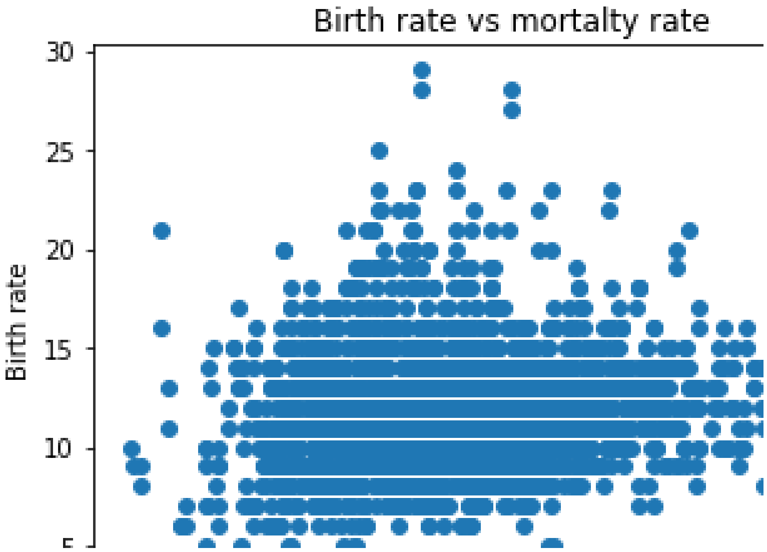

Figure 10: Birth rate vs Heart disease mortality rate

From the given scatter plot one can imply that, there is not a strong and obvious correlation between birth rate and mortality rate. (Reason: the correlation co-eff is nearer to 0 particularly 0.1421)

Tobacco and alcohol use vs mortality rate

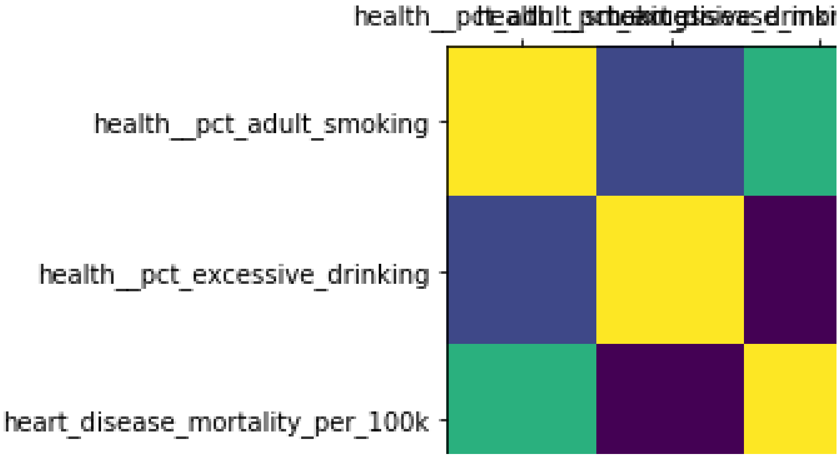

Figure 11: Tobacco, alcohol and mortality rate correlation

Table 3: Correlation among Tobacco, alcohol and heart disease mortality rate

	health__pct_adult_smoking	health__pct_excessive_drinking	heart_disease_mortality_per_100k
health__pct_adult_smoking	1.000000	-0.084902	0.497063
health__pct_excessive_drinking	-0.084902	1.000000	-0.382172
heart_disease_mortality_per_100k	0.497063	-0.382172	1.000000

Given correlation table and heatmap introduce the apparent relationship between adult smoking and excessive drinking within a county.

Prevalence of adult smoking (but not excessive drinking) is positively correlated with mortality rate.

Reason:	health__pct_adult_smoking	health__pct_excessive_drinking
heart_disease_mortality	0.497063(positive)	-0.382172 (negative)

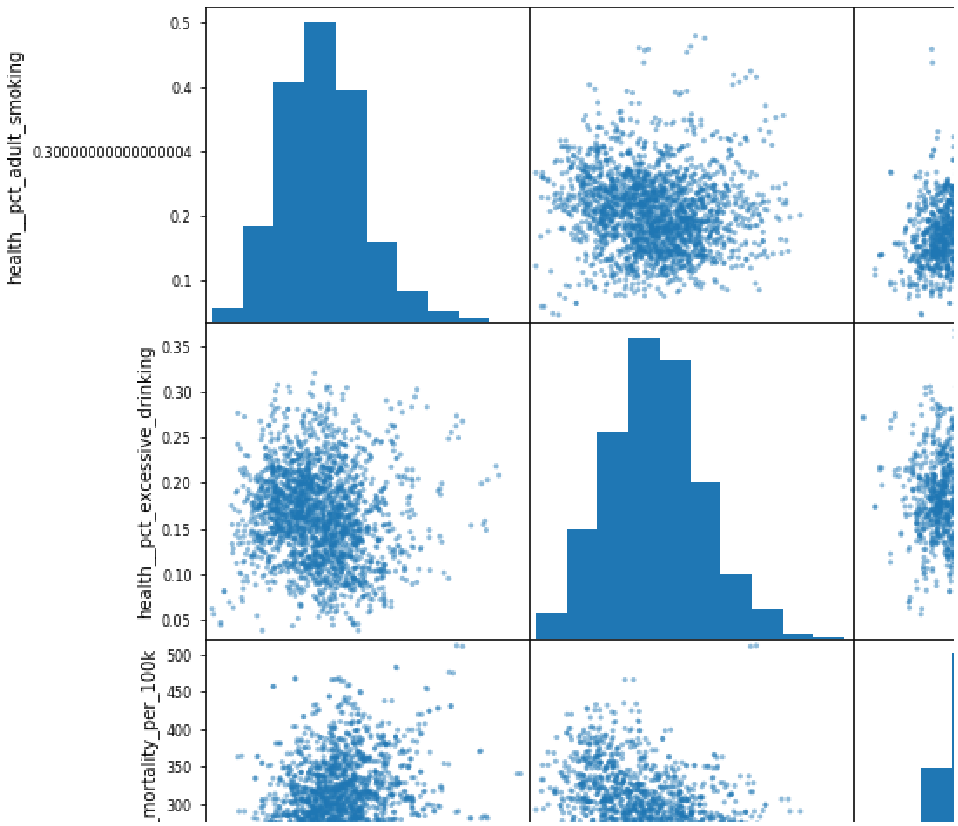

Figure 12: Scatter plot of Tobacco usage, Drinking habit and Heart disease

Older people and metro/non-metro vs mortality rate

Let's define counties for which the older population (aged 65 years and older) consitutes more than 20% of the population as having a "relatively large older population."

The following statements are true about the apparent relationship between a relatively large older population, metro/non-metro counties, and mortality rate.

Older people and metro/non-metro vs mortality rate-

1.Relatively large population having older people > 0.2

```
For all counties and old_population > 0.2 ( median )
demo__pct_aged_65_years_and_older      0.226
heart_disease_mortality_per_100k         254.000

For Metro and old_population > 0.2
demo__pct_aged_65_years_and_older      0.222
heart_disease_mortality_per_100k        250.000
```

For Nonmetro and old_population > 0.2
demo__pct_aged_65_years_and_older   0.227
heart_disease_mortality_per_100k   254.000

2. Relatively large population having older people < 0.2 - as 20%

For all counties and old_population < 0.2   (median)
demo__pct_aged_65_years_and_older   0.157
heart_disease_mortality_per_100k   281.000

For Metro and old_population < 0.2
demo__pct_aged_65_years_and_older   0.144
heart_disease_mortality_per_100k   263.000

For Nonmetro and old_population < 0.2
demo__pct_aged_65_years_and_older   0.164
heart_disease_mortality_per_100k   294.000

From given data one can easily indicate that:

- When narrowing down to only non-metro counties, there is a larger difference than for only metro counties when comparing median mortality rate of counties with relatively large older populations to those with less than 20% older population.
- Counties with relatively large older populations have a lower median mortality rate than counties with less than 20% older population.

Correlation between each feature with Heart disease mortality rate

Table 4: Correlation among Tobacco, alcohol and heart disease mortality rate

	heart_disease_mortality_per_100k
health__pct_adult_obesity	0.6568
health__pct_adult_smoking	0.5308
health__pct_diabetes	0.6876
health__pct_low_birthweight	0.5134
health__pct_excessive_drinking	-0.3711
health__pct_physical_inacticity	0.7293

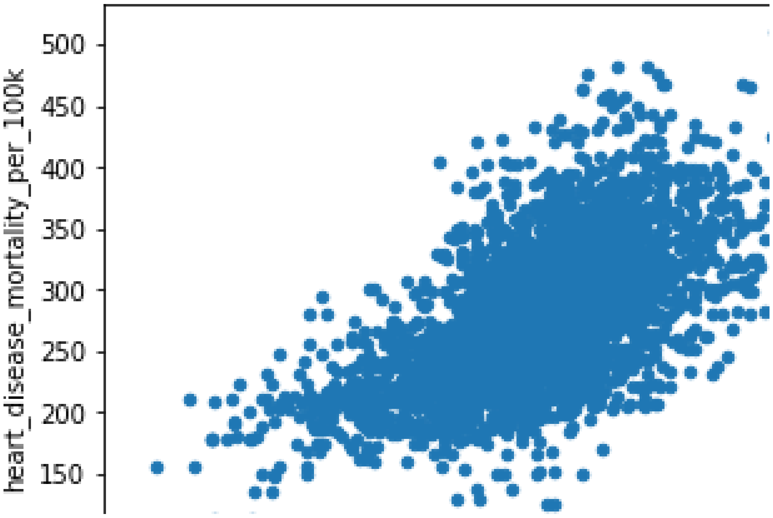

Figure 13: Population with obesity has positively correlated with heart disease mortality rate

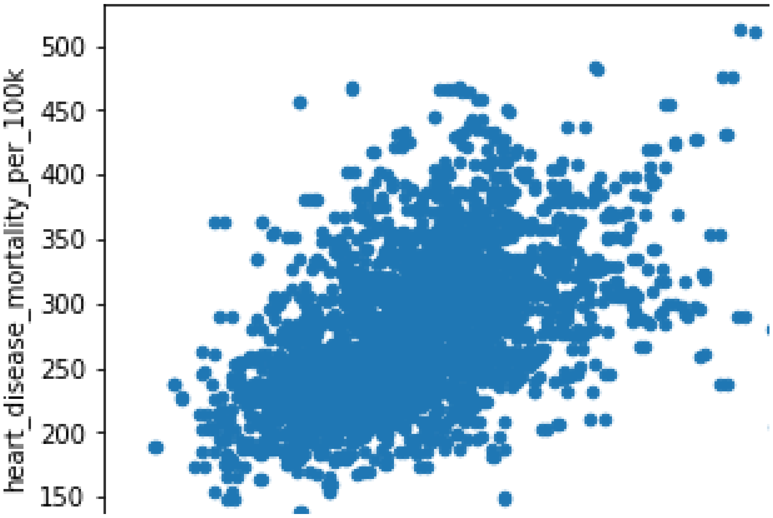

Figure 14: Population with smoking habit has positively correlated with heart disease mortality rate

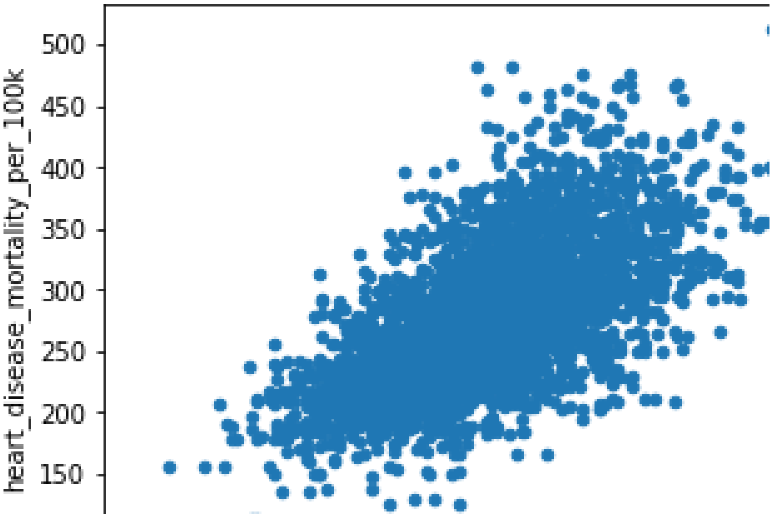

Figure 15: Population with diabetes is positively correlated with heart disease

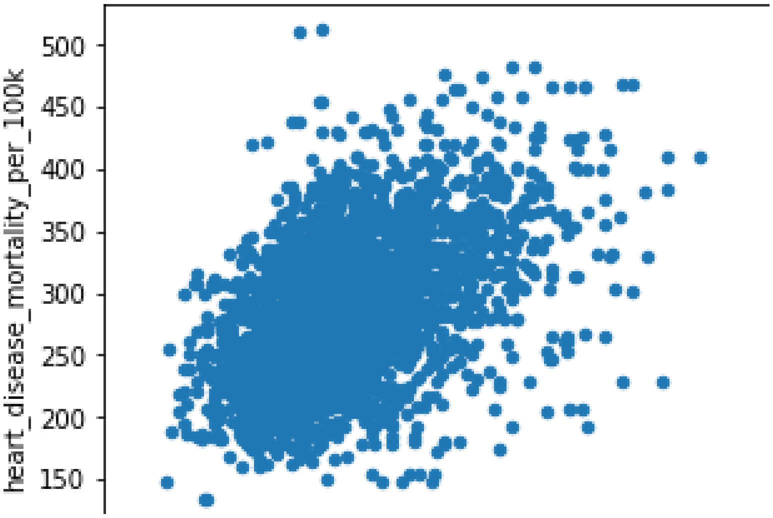

Figure 16: Population with lower birth weight is positively correlated with heart disease

Regression analysis and predictions

For the predictions of heart disease mortality rate, the Lasso regression technique has been used as the Lasso (Least Absolute Shrinkage Selector Operator) regression gives the better prediction in case of number of features is large and it automatically does feature selection.

It provides *sparse solutions*, it is generally the model of choice for modeling cases where the number of features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored.

It gives much **better output**, require **fewer tuning parameters** and can be **automated** to a large extend.

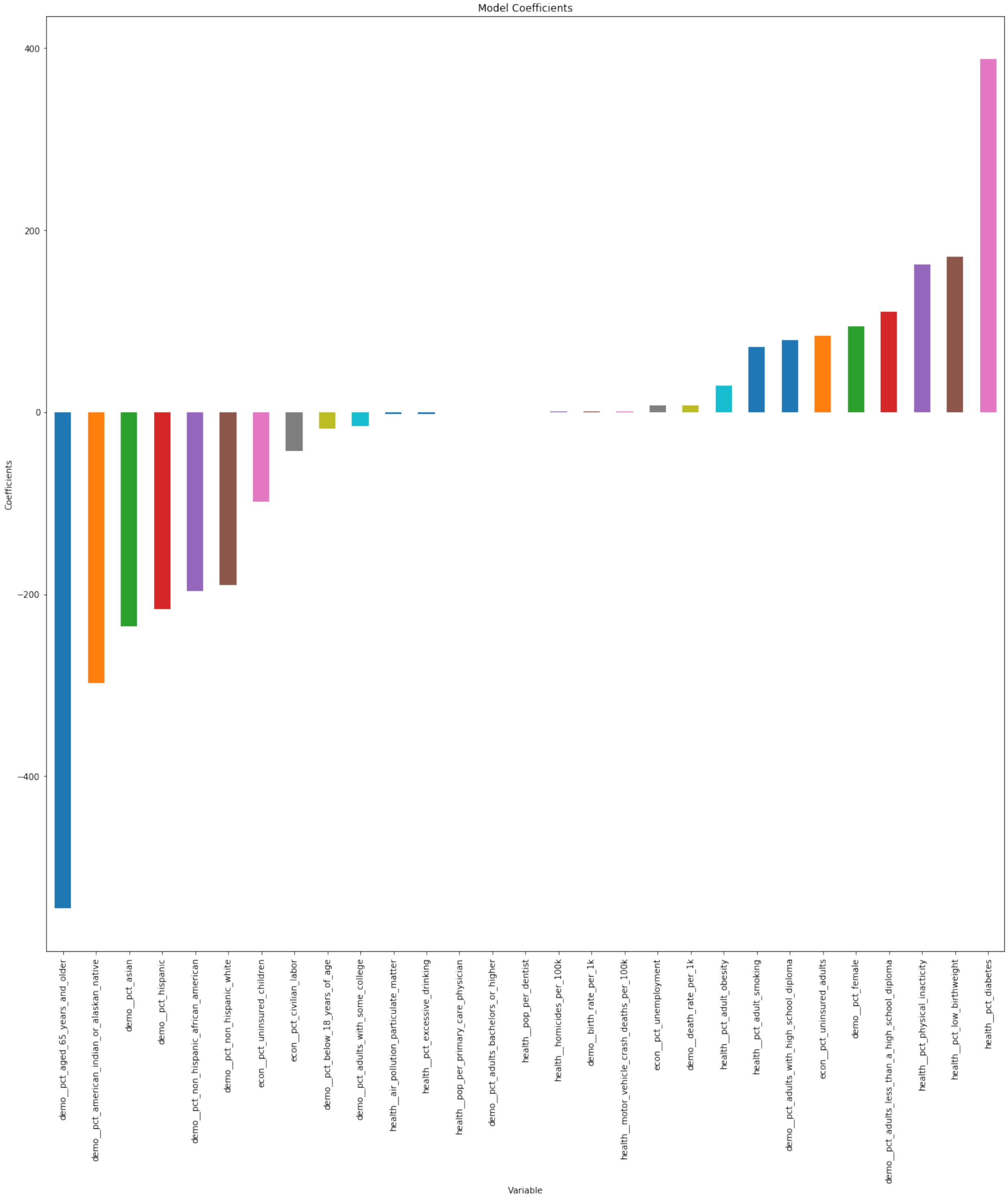

Figure 17 : Model coefficients of Lasso regression model

The lasso coefficients become zero in a certain range and are reduced by a constant factor, which explains their low magnitude in comparison to other techniques

Sorted list of coefficients of Lasso regression.

Table 5: Coefficients of Lasso regression

	Coefficient
demo__pct_aged_65_years_and_older	-545.1859364493
demo__pct_american_indian_or_alaskan_native	-297.4070660226
demo__pct_asian	-235.5948755722
demo__pct_hispanic	-216.1175144534
demo__pct_non_hispanic_african_american	-196.8891527563
demo__pct_non_hispanic_white	-189.6084776164
econ__pct_uninsured_children	-98.4731035127
econ__pct_civilian_labor	-42.7930728207
demo__pct_below_18_years_of_age	-18.0713152182
demo__pct_adults_with_some_college	-15.0240576446
health__air_pollution_particulate_matter	-2.2594949598
health__pct_excessive_drinking	-1.8493712819
health__pop_per_primary_care_physician	-0.0004680462
demo__pct_adults_bachelors_or_higher	-0
health__pop_per_dentist	0.0001135085
health__homicides_per_100k	0.2624483068
demo__birth_rate_per_1k	0.6119784638
health__motor_vehicle_crash_deaths_per_100k	0.7619784004
econ__pct_unemployment	7.2177056628
demo__death_rate_per_1k	7.6190843178
health__pct_adult_obesity	28.8478276138
health__pct_adult_smoking	71.8712878372
demo__pct_adults_with_high_school_diploma	78.6695441171
econ__pct_uninsured_adults	83.9681088389
demo__pct_female	93.8333846459
demo__pct_adults_less_than_a_high_school_diploma	109.9440859454
health__pct_physical_inacticity	161.6956795045
health__pct_low_birthweight	170.9366784241
health__pct_diabetes	387.7385406975

We can see that in case of lasso, even at smaller alpha's, our coefficients are reducing to absolute zeroes. Therefore, lasso selects the only some feature while reduces the coefficients of others to zero. This property is known as feature selection.

The statistics derived from regression model:

Coefficients:

```
[-4.27930728e+01    7.21770566e+00   8.39681088e+01   -9.84731035e+01
  9.38333846e+01   -1.80713152e+01  -5.45185936e+02  -2.16117514e+02
 -1.96889153e+02    -1.89608478e+02   -2.97407066e+02  -2.35594876e+02
  1.09944086e+02     7.86695441e+01  -1.50240576e+01   -0.00000000e+00
  6.11978464e-01     7.61908432e+00    2.88478276e+01   7.18712878e+01
  3.87738541e+02   1.70936678e+02     -1.84937128e+00   1.61695680e+02
 -2.25949496e+00      2.62448307e-01     7.61978400e-01     1.13508502e-04
 -4.68046196e-04]
```

Table 5: Error statistics of Lasso regression

Score	Value
Intercepts	292.3248299653
Root Mean squared error	77.2047852141
Lasso regression coefficient of determination R^2	1
Mean absolute error	61.3478388863

The main problem with lasso regression is when we have correlated variables, it retains only least number of variables and sets other correlated variables to zero. That will possibly lead to some loss of information resulting in lower accuracy in our model.

Scatter plot of predicted value to the original value indicates the linear estimation of the prediction from actual data.

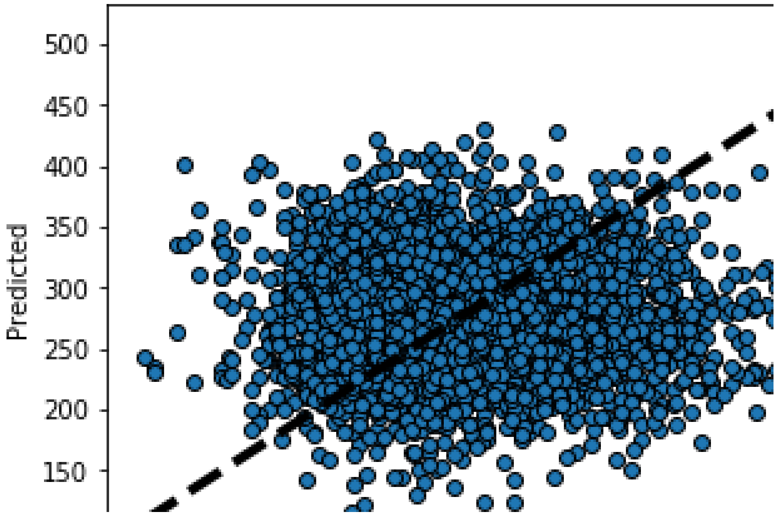

Figure 18 : Linear estimation of predicted values from actual values

Likewise the following line plot indicates the efficiency in terms of R^ error of Lasso regression over the original data.

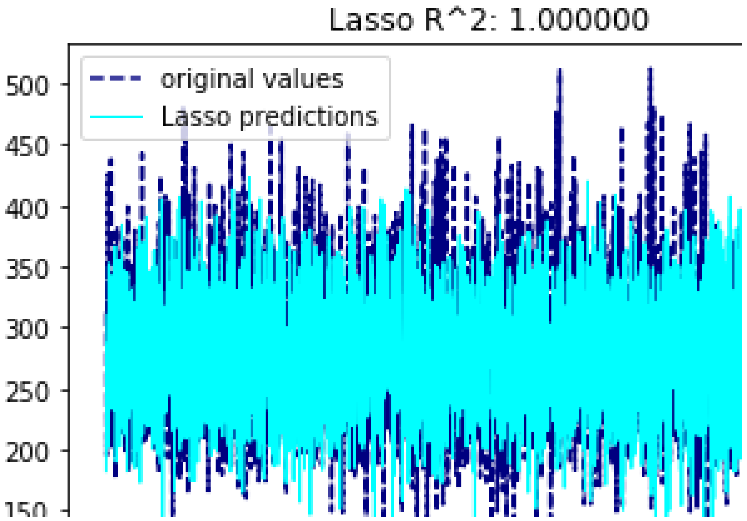

Figure 19 : Line plot of predicted values from actual values indicating

R^2 error almost equals to 1.00

Conclusion

This analysis has shown that the health indicators such as obesity, adult smoking, physical inacticity , lower birthweight ,diabetes has the significant influence on the heart disease mortality rate. In addition to that economic indicators such as unemployment, uninsured adults and the demographic features such as female, adults less than a high school diploma are need to be consider for the prediction of mortality rate.