Analysis of Heart Disease Mortality

Laura Gaffney | Data Science Capstone | July 19, 2018

Executive Summary

This document presents an analysis of heart disease mortality related to four categories of data: information about the county, economic indicators, health indicators and demographic information. The data comes from a variety of sources and is published by the United States Department of Agriculture Economic Research Service. Heart disease mortality is presented as a rate of heart disease (per 100,000 individuals) across the United States at the county level.

Summary statistics and descriptive statistics were calculated from the data. Visualizations are used to show relationships between socioeconomic factors and the associated heart disease mortality. A machine learning model was created to predict heart disease mortality rate based on the available data.

Some of the most significant factors are:

- **Area**: Metro counties tend to have lower rates of heart disease than nonmetro counties. Counties with recreation-based economies had the lowest rate of heart disease (and the lowest standard deviation between counties).

- **Economic:** Counties with a higher percent of employed residents tended to have lower rates of heart disease mortality. A lower percentage of uninsured adults also correlated with lower heart disease.

- **Demographic:** The age of the population had a slight correlation with heart disease, and younger populations had lower rates. Counties with higher percentages of African Americans had higher rates of heart disease, while counties with higher percentages of Asians had lower rates of heart disease. Education level was also impactful, with higher levels of education associated with lower rates of heart disease.

- **Health:** Obesity, smoking, diabetes, physical inactivity and low birthweight were all associated with higher rates of heart disease.

Data Description

The dataset included information from 3198 counties. The data was described by categorical information about the area, plus numeric data about the population. An analysis of the data started with observing the min, max, median, standard deviation and count of each data point across all categories and then drilled down into specific areas. Scatter plots were created to visualize the data.

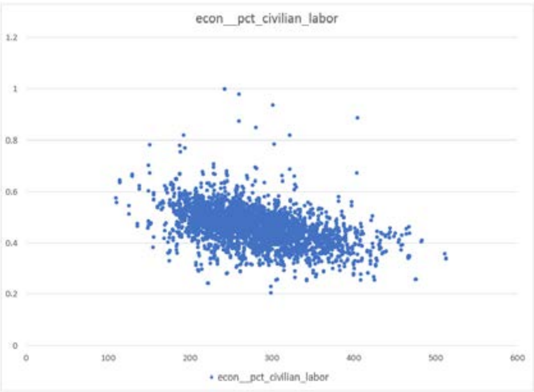
Numeric features are summarized below. Most are percentages. The values that are not percentages were normalized in the predictive model described later.

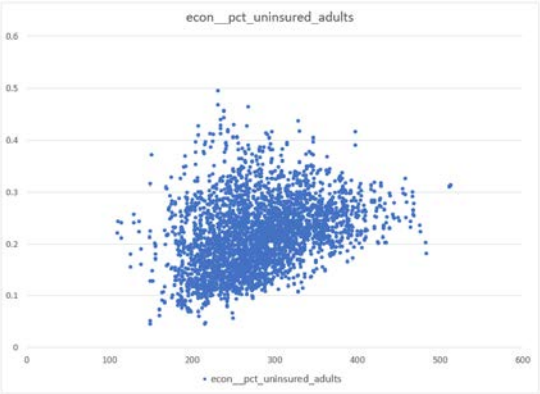
Feature	Min	Max	Mean	Median	Std Dev
econ__pct_civilian_labor	0.21	1.00	0.47	0.47	0.07
econ__pct_unemployment	0.01	0.25	0.06	0.06	0.02
econ__pct_uninsured_adults	0.05	0.50	0.22	0.22	0.07

econ__pct_uninsured_children	0.01	0.28	0.09	0.08	0.04
demo__pct_female	0.28	0.57	0.50	0.50	0.02
demo__pct_below_18_years_of_age	0.09	0.42	0.23	0.23	0.03
demo__pct_aged_65_years_and_older	0.05	0.35	0.17	0.17	0.04
demo__pct_hispanic	0.00	0.93	0.09	0.04	0.14
demo__pct_non_hispanic_african_american	0.00	0.86	0.09	0.02	0.15
demo__pct_non_hispanic_white	0.05	0.99	0.77	0.85	0.21
demo__pct_american_indian_or_alaskan_native	0.00	0.86	0.02	0.01	0.08
demo__pct_asian	0.00	0.34	0.01	0.01	0.03
demo__pct_adults_less_than_a_high_school_diploma	0.02	0.47	0.15	0.13	0.07
demo__pct_adults_with_high_school_diploma	0.07	0.56	0.35	0.36	0.07
demo__pct_adults_with_some_college	0.11	0.47	0.30	0.30	0.05
demo__pct_adults_bachelors_or_higher	0.01	0.80	0.20	0.18	0.09
demo__birth_rate_per_1k	4.00	29.00	11.68	11.00	2.74
demo__death_rate_per_1k	0.00	27.00	10.30	10.00	2.79
health__pct_adult_obesity	0.13	0.47	0.31	0.31	0.04
health__pct_adult_smoking	0.05	0.51	0.21	0.21	0.06
health__pct_diabetes	0.03	0.20	0.11	0.11	0.02
health__pct_low_birthweight	0.03	0.24	0.08	0.08	0.02
health__pct_excessive_drinking	0.04	0.37	0.16	0.16	0.05
health__pct_physical_inacticity	0.09	0.44	0.28	0.28	0.05
health__air_pollution_particulate_matter	7.00	15.00	11.63	12.00	1.56
health__homicides_per_100k	-0.40	50.49	5.95	4.70	5.03
health__motor_vehicle_crash_deaths_per_100k	3.14	110.45	21.13	19.63	10.49
health__pop_per_dentist	339	28130	3431	2690	2569
health__pop_per_primary_care_physician	189	23399	2551	1999	2100

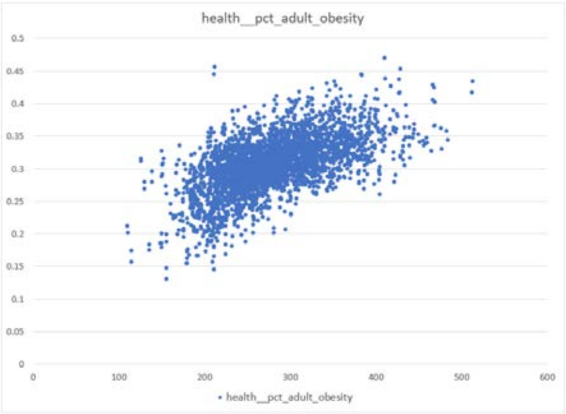
Several features were found to significantly impact the heart disease mortality rate, and are displayed below by category.

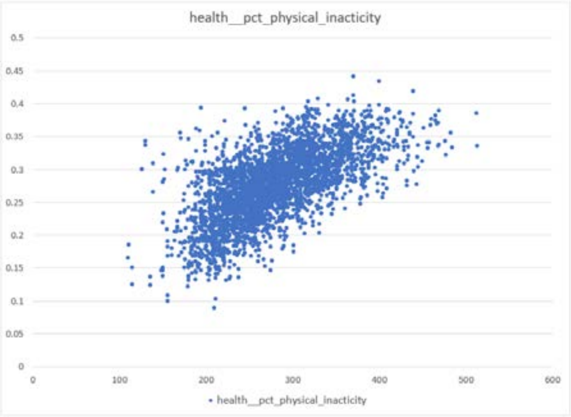
First, in the economic category, the rate of heart disease mortality decreased as the civilian labor force increased (a similar relationship exists as unemployment decreases). Heart disease mortality increased as the rate of uninsured adults increased:



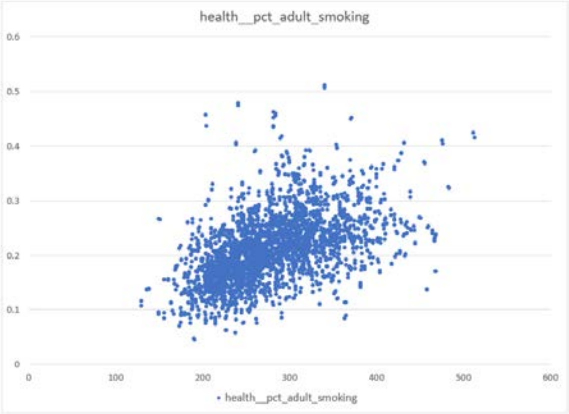


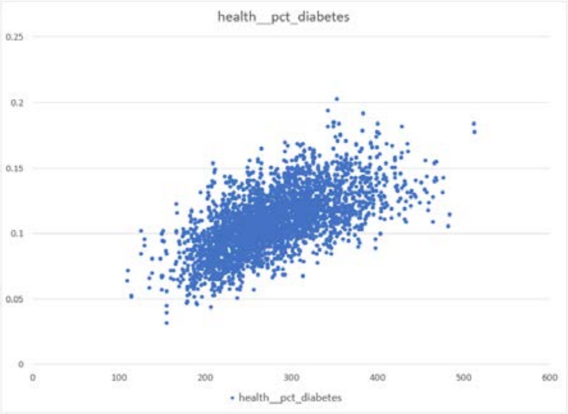
In the health category, obesity and physical inactivity were strongly correlated with each other and with higher rates of heart disease mortality:



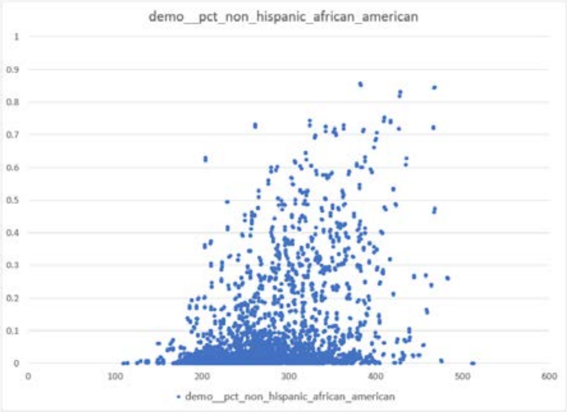


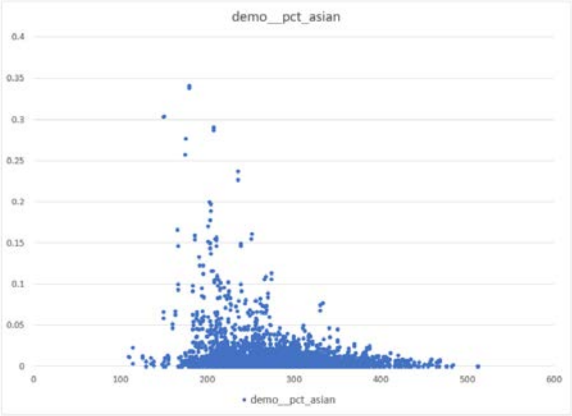
Higher rates of smoking and diabetes were also higher in counties with higher heart disease mortality:



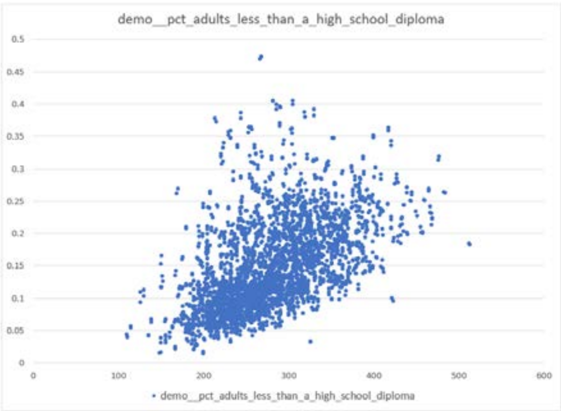


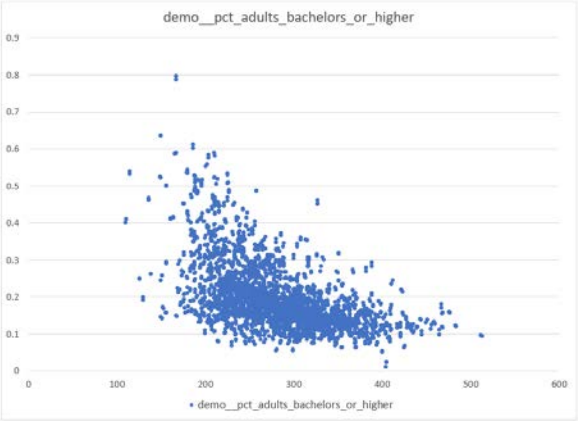
In the demographic area, two races had interesting tails on their scatter plots. Not many counties had a high percentage of African Americans, but of those that did, they tended to have a higher rate of heart disease mortality. The opposite was true for the few counties with higher percentages of Asians – they tended to have a lower rate of heart disease mortality:



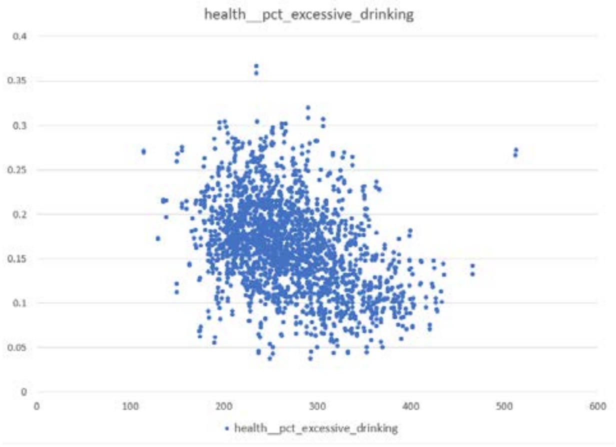


Also in the demographic category, higher education levels were associated with lower rates of heart disease. The plots below show the highest level of education. Counties with a higher portion of residents with bachelors degrees or higher had lower mortality rates that those with high school diplomas. Supporting correlations also existed for education levels of a high school diploma and some college.





One somewhat surprising finding in the data was that counties with a higher percent of excessive drinking were associated with a lower rate of heart disease mortality. While excessive drinking is typically associated with lesser health, in this case it might actually reduce a health risk.



Of the features in the scatter plots above, most higher-risk features are stronger in nonmetro areas, as shown in the table below. This is not surprising, since as discussed later in this section, the heart disease mortality rates are generally higher in nonmetro counties.

econ__economic_typology	Metro Mean	Nonmetro Mean
econ__pct_civilian_labor	0.48	0.46
econ__pct_uninsured_adults	0.20	0.23
demo__pct_adults_less_than_a_high_school_diploma	0.13	0.16
demo__pct_adults_bachelors_or_higher	0.25	0.17
health__pct_adult_obesity	0.30	0.31
health__pct_adult_smoking	0.19	0.18
health__pct_diabetes	0.10	0.11
health__pct_physical_inactivity	0.26	0.29

Categorical data included Rural-Urban Continuum Codes, which were further classified as metro or nonmetro. In the dataset, 1128 counties were considered metro and 2070 counties were considered nonmetro. The mean, standard deviation and count of heart disease mortality rates in each code are summarized below. The codes are listed in ascending order of mean. Overall, metro areas have lower rates than nonmetro areas, with the exception of isolated completely rural areas.

Rural-Urban Continuum Code	Mean	StdDev	Count
Metro - Counties in metro areas of 1 million population or more	258	50.45	436
Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area	267	66.12	484
Metro - Counties in metro areas of 250,000 to 1 million population	270	46.76	370
Metro - Counties in metro areas of fewer than 250,000 population	273	52.96	322
Nonmetro - Urban population of 20,000 or more, adjacent to a metro area	283	50.36	222
Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area	284	60.71	238

Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area	288	62.56	418
Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area	293	62.74	100
Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area	302	57.82	608

For economic typology, counties with a recreation economy had the lowest rate of heart disease mortality. Farm-dependent counties were second lowest. About half of the "Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area" counties have farming economies, so it seems that while nonmetro counties usually have higher rates of heart disease than metro counties, farming communities are an exception.

Row Labels	Mean of heart_disease_mortality_per_100k
Recreation	239
Farm-dependent	259
Federal/State government-dependent	278
Nonspecialized	287
Manufacturing-dependent	295
Mining-dependent	301

Machine Learning Predictive Model

To create a predictive model, the data was imported into the Azure Machine Learning Studio. The data was cleansed, replacing missing values with the median (across all counties) for that feature. For columns with values that were not percentages, the data was normalized to values between zero and one. Additionally, features that were less impactful were removed from the model.

The model was trained using a Boosted Decision Tree Regression algorithm. This type of model performs a sequence of simple evaluations for each data row, moving along a tree structure from trunk to leaf. The model had a root-mean-square-error of 33.7, when scored against existing data.

Summary and Next Steps

In summary, the data shows potential indicators of heart disease spanning area information, demographics and health. Many of the risk factors were stronger in nonmetro areas. Health outreach to the specific high risk counties could help reduce the rate of heart disease.

To refine the data analysis, future steps could include:

- More feature engineering within the machine learning model
- Additional drill-down into categorical features to find relationships with high risk factors
- Visualizations comparing features between areas and economic typologies

This additional analysis could help identify deeper relationships within the data and help target the message and counties for health outreach.