ANALYSIS OF HEART DISEASE MORTALITY

EXECUTIVE SUMMARY

This report presents an analysis of heart disease mortality rate among US counties based on various economic, health and demographic features.

The data provided consists of 2 sets - Training (3198 rows, 33 features) along with actual mortality rate, and Test (3080 rows) for which the rate must be predicted.

Following sections describe a systematic approach to tackling this data-science problem: Data Exploration, Feature Representation & Selection, Missing Data, Model Creation, Prediction & Evaluation

To summarize, a regression model was created to predict mortality rate from features, most notable among them being, unsurprisingly, the **lack of physical activity** in the population.

TOOLS : Excel , SQL Server , Azure ML, Kaggle , Jupyter

DATA EXPLORATION

Summary statistics for all features were generated using pandas package describe() and Excel heatmap.
Numerical features are shown below:

	count	mean	std	min	25%	median	75%	max
row_id	3198							
econ__pct_civilian_labor	3198	0.47	0.07	0.21	0.42	0.47	0.51	1.00
econ__pct_unemployment	3198	0.06	0.02	0.01	0.04	0.06	0.07	0.25
econ__pct_uninsured_adults	3196	0.22	0.07	0.05	0.17	0.22	0.26	0.50
econ__pct_uninsured_children	3196	0.09	0.04	0.01	0.06	0.08	0.11	0.28
demo__pct_female	3196	0.50	0.02	0.28	0.49	0.50	0.51	0.57
demo__pct_below_18_years_of_age	3196	0.23	0.03	0.09	0.21	0.23	0.25	0.42
demo__pct_aged_65_years_and_older	3196	0.17	0.04	0.05	0.14	0.17	0.20	0.35
demo__pct_hispanic	3196	0.09	0.14	0.00	0.02	0.04	0.09	0.93
demo__pct_non_hispanic_african_american	3196	0.09	0.15	0.00	0.01	0.02	0.10	0.86
demo__pct_non_hispanic_white	3196	0.77	0.21	0.05	0.65	0.85	0.94	0.99
demo__pct_american_indian_or_alaskan_native	3196	0.02	0.08	0.00	0.00	0.01	0.01	0.86
demo__pct_asian	3196	0.01	0.03	0.00	0.00	0.01	0.01	0.34
demo__pct_adults_less_than_a_high_school_diploma	3198	0.15	0.07	0.02	0.10	0.13	0.19	0.47
demo__pct_adults_with_high_school_diploma	3198	0.35	0.07	0.07	0.31	0.36	0.40	0.56
demo__pct_adults_with_some_college	3198	0.30	0.05	0.11	0.26	0.30	0.34	0.47
demo__pct_adults_bachelors_or_higher	3198	0.20	0.09	0.01	0.14	0.18	0.23	0.80
demo__birth_rate_per_1k	3198	11.68	2.74	4	10	11	13	29
demo__death_rate_per_1k	3198	10.30	2.79	0	8	10	12	27
health__pct_adult_obesity	3196	0.31	0.04	0.13	0.28	0.31	0.33	0.47
health__pct_adult_smoking	2734	0.21	0.06	0.05	0.17	0.21	0.25	0.51
health__pct_diabetes	3196	0.11	0.02	0.03	0.09	0.11	0.12	0.20
health__pct_low_birthweight	3016	0.08	0.02	0.03	0.07	0.08	0.10	0.24
health__pct_excessive_drinking	2220	0.16	0.05	0.04	0.13	0.16	0.20	0.37
health__pct_physical_inacticity	3196	0.28	0.05	0.09	0.24	0.28	0.31	0.44
health__air_pollution_particulate_matter	3170	11.63	1.56	7	10	12	13	15
health__homicides_per_100k	1231	5.95	5.03	-0.40	2.62	4.70	7.89	50.49
health__motor_vehicle_crash_deaths_per_100k	2781	21.13	10.49	3.14	13.49	19.63	26.49	110.45
health__pop_per_dentist	2954	3431.43	2569.45	339	1812	2690	4090	28130
health__pop_per_primary_care_physician	2968	2551.34	2100.46	189	1420	1999	2859	23399
heart_disease_mortality_per_100k	3198	279.37	58.95	109	237	275	317	512

This quickly shows us columns with missing data - 13 features have just 2 ‘bad’ rows, while one has almost 2/3rd’s missing. We can also spot outliers, which in this case are also ‘errors’ – negative homicide rate.

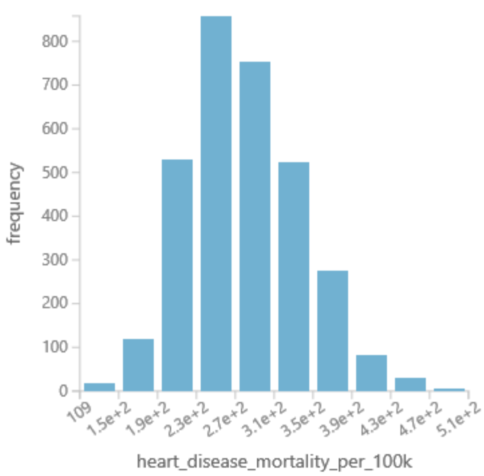
Since heart_disease_mortality_per_100k is our target column of interest, its distribution was visualized in Azure ML studio, through histogram and boxplot.

Statistics

Mean	279.3693
Median	275
Min	109
Max	512
Standard Deviation	58.9533
Unique Values	301
Missing Values	0
Feature Type	Numeric Feature

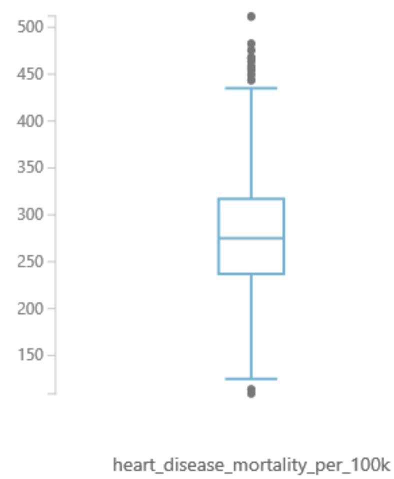
Visualizations

heart_disease_mortality_per_100k
Histogram



Visualizations

heart_disease_mortality_per_100k
BoxPlot



A slightly right-skewed normal distribution can be seen from bell-curve i.e. in the given data, the counties mostly have an even distribution of mortality rate with a few of them having high rates.

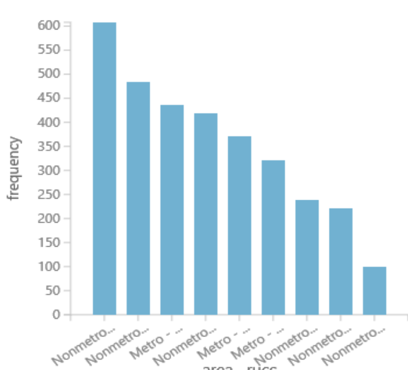
Other than numerical features, categorical features such as region and typology can be plotted as such:

Statistics

Unique Values	9
Missing Values	0
Feature Type	String Feature

Visualizations

area__rucc
Histogram

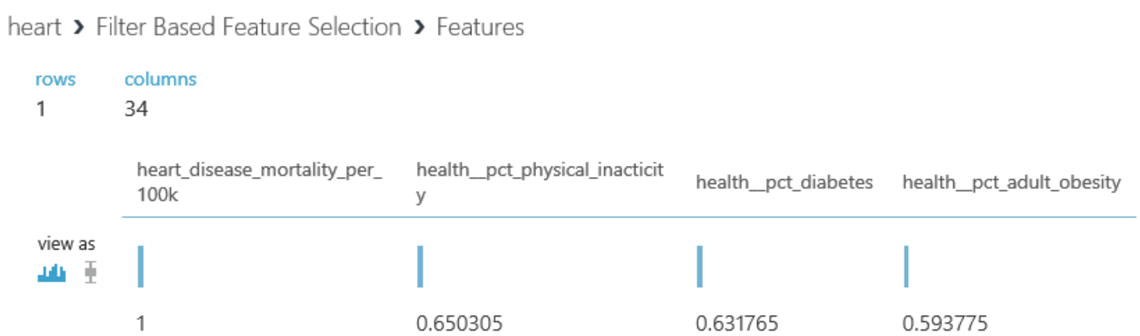


FEATURE REPRESENTATION & SELECTION

Numerical features need to be normalized – standardized to a value between -1 & 1 to bring them all on same scale e.g. percentage rates are < 1 while population rate is in the thousands.

Categorical features need to be categorized, either using a numerical encoding – setting value 1 to metro, 2 to non-metro etc., or one-hot encoding – setting a binary value of 1 or 0 to each type of area, which expands the number of features to as many areas.

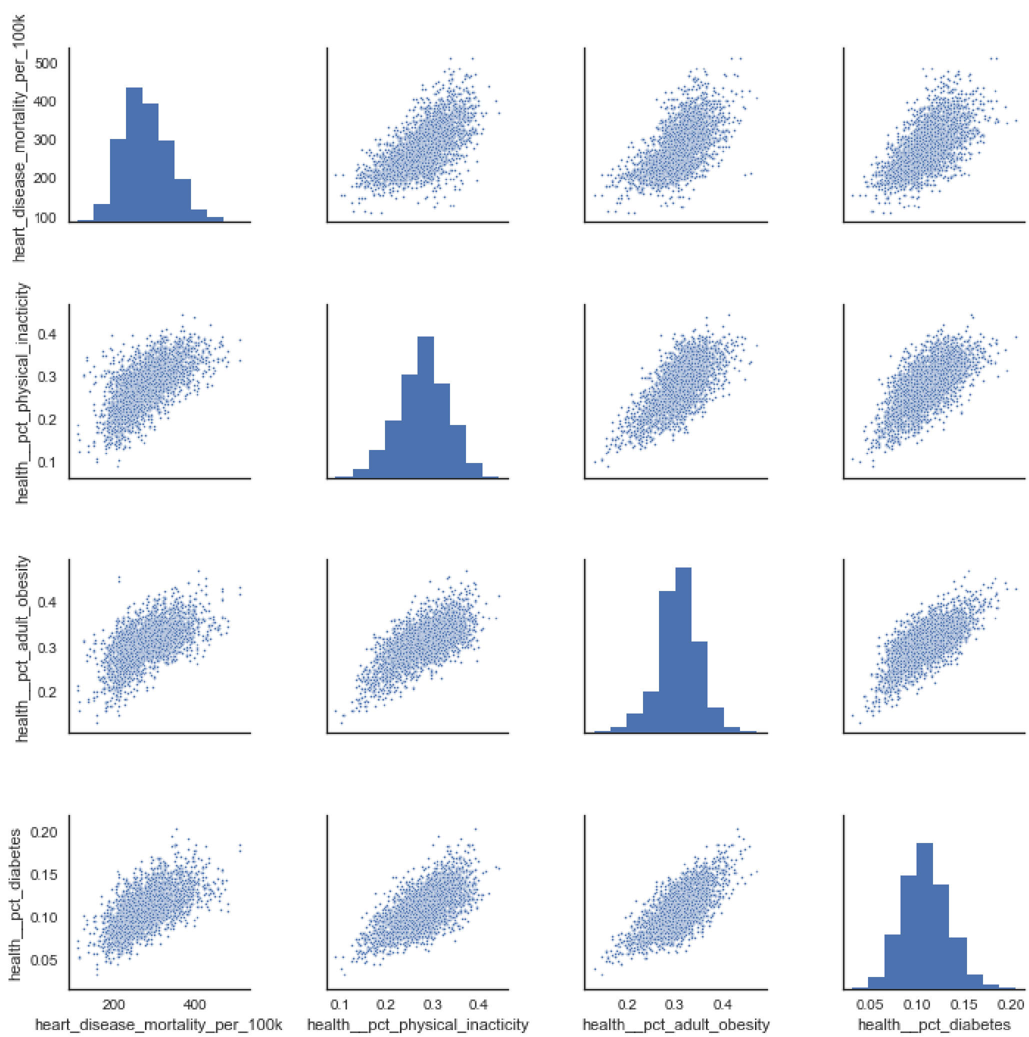
Out of 33 features, we can use the Pearson correlation coefficient, to find out the most important ones that might affect the target variable (mortality rate). The top 3 are shown here:


heart › Filter Based Feature Selection › Features

rows	columns
1	34

heart_disease_mortality_per_100k	health__pct_physical_inacticity	health__pct_diabetes	health__pct_adult_obesity
1	0.650305	0.631765	0.593775

As expected, physical inactivity and related features like obesity & diabetes seem to be the main culprits. This is also proven using pair-wise scatter plot using seaborn package, which clearly shows almost-linear correlation.



A correlation heat-map, also from seaborn, shows the interaction between every pair of features as well as target.

Red means positive correlation (as one feature increases, so does other). Blue means negative correlation.

Those who have a ‘bachelors degree or higher’ will obviously not say they just have ‘some college degree’ – hence the dark blue negative correlation between the two. And they also tend to take of their health better, probably due to better earnings, hence its negative correlation to all health-related features. While we can see a positive correlation between these features and the less-educated ‘high school diploma’ population. Similarly, the higher the working ‘civilian labor’ population, the better any health-related issues. While, the higher the vices (drinking and smoking), the worse any health-related issues.

This analysis can be used to narrow down the list of features presented to the model, to improve accuracy and duration.

MISSING DATA

Before we can feed the data to any machine learning model, the principle 'GARBAGE-IN, GARBAGE-OUT' must be kept in mind.
We can deal with missing in 2 ways:
- Remove : column / row
- Replace : constant / computed

Removing the entire column, or entire row whenever we encounter missing data, can be detrimental if there are many missing values, because our training data set gets vastly reduced. In this dataset, we would only be left with ~1000 rows if we did so.

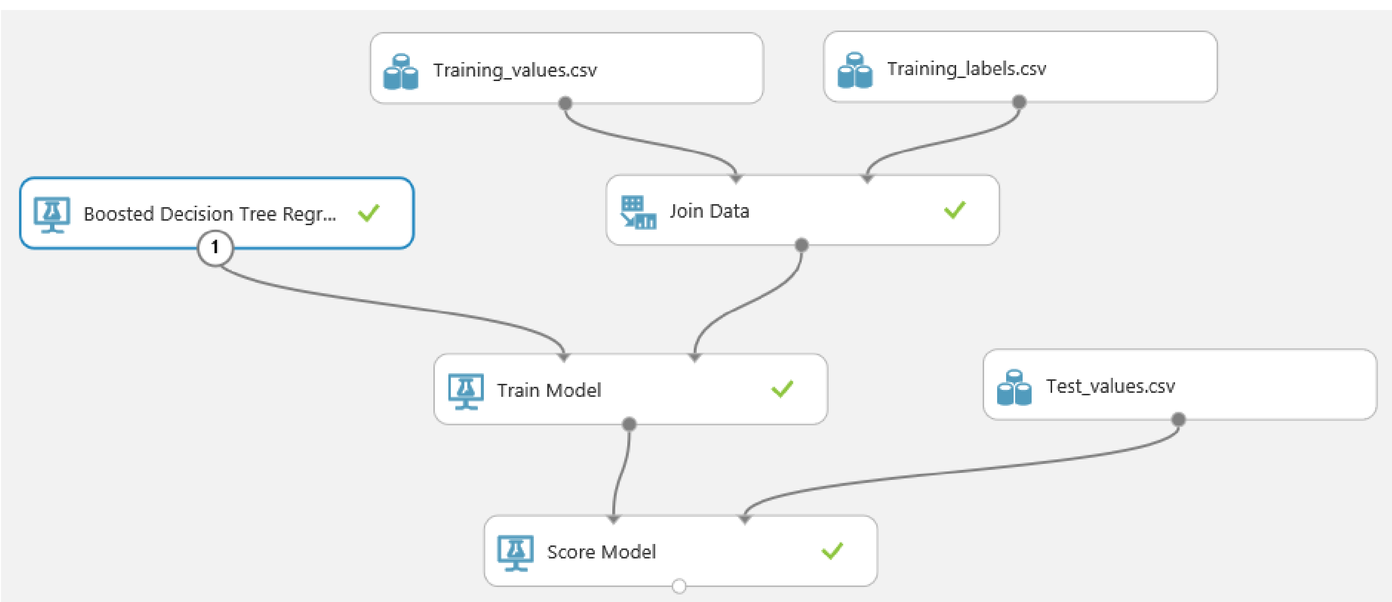
Replacing the missing value with a calculated guess is a better option. Constant values like the Mean, Median, Mode of that column are good starting points. Going further, statistical methods like MICE and PCA can be used to fill-in-the blanks with better estimates drawn from distribution of that column as well as correlation with other columns.

MODEL CREATION-PREDICTION-EVALUATION

The training data and labels are first passed to a machine learning model, which 'learns' from it, and then predicts target when given test data.

Azure ML studio offers several in-built machine learning models to train and predict data : Linear Regression, Decision, Trees, Neural Networks etc.

A simplistic model-creation-prediction flowchart is shown below :



Since we do not know the Test labels, how do we check if the model is good or not ? We use the training set itself for 'testing'. By splitting training set into 75%-25% , we reserve 25% of data for testing the error (the metric used here is RMSE) of our model. We can compare the performance of several models this way, and choose the best one. After trying various algorithms, the one with lowest RMSE was found to be Boosted Decision Tree Regression.

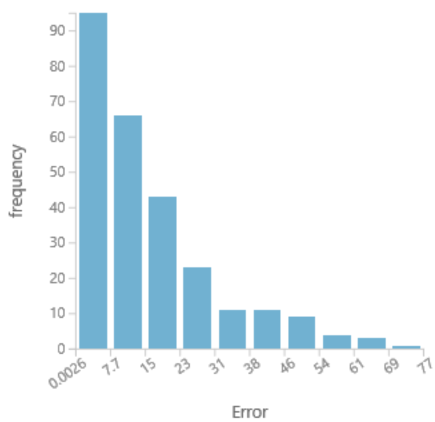
Charts below show the error-rate, and comparison of actual-vs-predicted rate, which shows a pretty-strong correlation between the two, which means the model performs well.

heartdisease › Evaluate Model › Evaluation results

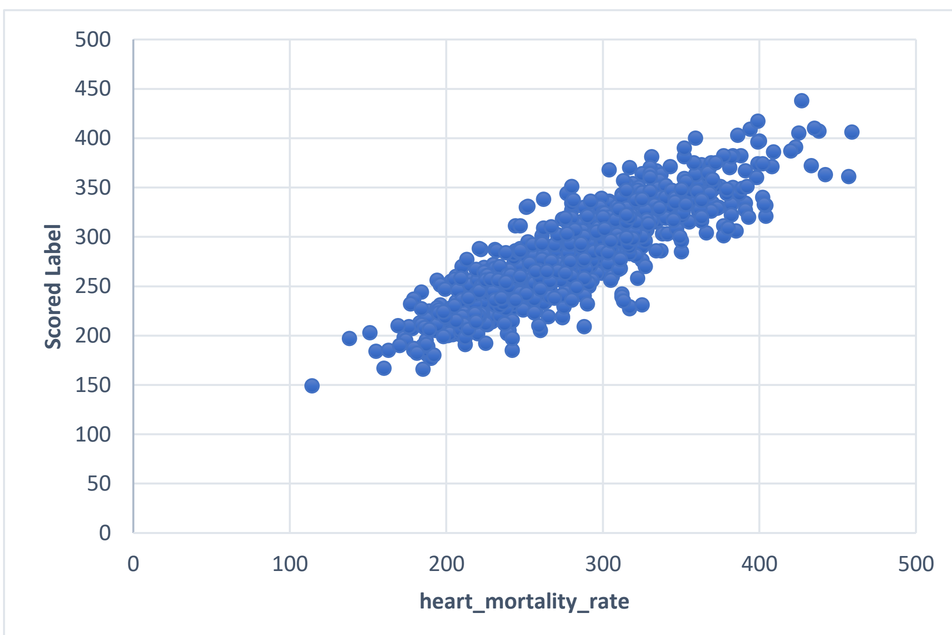
Metrics

Mean Absolute Error	16.123813
Root Mean Squared Error	21.996947
Relative Absolute Error	0.36834
Relative Squared Error	0.164066
Coefficient of Determination	0.835934

Error Histogram



heart_disease_mortality_per_100k	Scored Labels
274	262.182159
319	344.943115
363	373.188263
269	264.733063



CONCLUSION

Regional, economic, demographic and health factors of a county can be used to reliably predict (and hopefully prevent) its heart disease mortality rate, using a regression model with boosted decision trees. Analysis indicated that increasing physical activity among population could be the single most effective method to reducing the occurrence of heart disease, or at least its fatality rate. Other features such as education and vices also play determining roles.