Analysis of heart mortality rate by county

Murat Eliby

Executive Summary

In this capstone project, we are exploring features of counties in the US, to see how it affects the heart disease mortality rate. We have been provided with a training set, consisting of 3198 instances. The values dataset consists of the various features of demographic and economic types, which are to be analysed and used to make predictions about the target label, the 'heart mortality rate per 100k'.

Performing analysis on the features, some interesting conclusions were found such as:

- After mapping the data into metro/nonmetro areas for area_rucc, it is found that metropolitan areas have typically a smaller rate of heart diseases mortality
- The highest correlation between features was demo__pct_adults_bachelors_or_higher & demo__pct_adults_with_high_school_diploma at 0.74
- The lowest correlation between features was demo__pct_hispanic & health__pop_per_primary_care_physician at 0.000065
- The biggest correlation of heart disease mortality rate is physical inactivity, followed by diabetes and adult obesity, which all have a high correlation between each other and are listed in the top 10 of highest correlations among features.
- Most of the numeric features including the target column appear to resemble a normal distribution, with few features particularly from the demographic type having a strong skew to the right.
- A noticeable imbalance in classes can also be found among categorical features, except for 'Yr' which is perfectly symmetrical and found to be statistically irrelevant and removed.
- Further numeric features of row_id & demo__pct_american_indian_or_alaskan_native were also found to have no noticeable correlation for prediction and removed.
- Non-metro areas were found to have a higher ratio of people aged 65 and over, however there was no correlation between higher age and heart disease mortality rate.

The dataset consists of 35 columns, 4 categorical datatypes and the remaining numeric. As the target label for prediction is of numeric type, we explore regression models with the aim of minimising the Root Mean Squared Error. Going through three potential models of linear regression, decision forest and boosted decision tree, the best performing model is found to be boosted decision tree regression scoring a 27.42 for the internal validation and 32.8 on unseen data. A conclusion is reached that a better model can be achieved with less overfitting of data, if a higher number of training data is provided with less missing data. Nevertheless, given Root Mean Squared Error of 32.8 on unseen data is still almost half of the standard deviation of the heart diseases mortality rate at 58.95, the achievement can be deemed as satisfactory.

Exploratory analysis

After joining the values and labels into one dataset, by the common key identifier 'row id', we look at the brief overview of our dataset features by total count and data type:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3198 entries, 0 to 3197
Data columns (total 35 columns):
row_id                                          3198 non-null int64
area__rucc                                      3198 non-null object
area__urban_influence                           3198 non-null object
econ__economic_typology                         3198 non-null object
econ__pct_civilian_labor                        3198 non-null float64
econ__pct_unemployment                          3198 non-null float64
econ__pct_uninsured_adults                      3196 non-null float64
econ__pct_uninsured_children                    3196 non-null float64
demo__pct_female                                3196 non-null float64
demo__pct_below_18_years_of_age                 3196 non-null float64
demo__pct_aged_65_years_and_older               3196 non-null float64
demo__pct_hispanic                              3196 non-null float64
demo__pct_non_hispanic_african_american         3196 non-null float64
demo__pct_non_hispanic_white                    3196 non-null float64
demo__pct_american_indian_or_alaskan_native     3196 non-null float64
demo__pct_asian                                 3196 non-null float64
demo__pct_adults_less_than_a_high_school_diploma 3198 non-null float64
demo__pct_adults_with_high_school_diploma       3198 non-null float64
demo__pct_adults_with_some_college              3198 non-null float64
demo__pct_adults_bachelors_or_higher            3198 non-null float64
demo__birth_rate_per_1k                         3198 non-null float64
demo__death_rate_per_1k                         3198 non-null float64
health__pct_adult_obesity                       3196 non-null float64
health__pct_adult_smoking                       2734 non-null float64
health__pct_diabetes                            3196 non-null float64
health__pct_low_birthweight                     3016 non-null float64
health__pct_excessive_drinking                  2220 non-null float64
health__pct_physical_inacticity                 3196 non-null float64
health__air_pollution_particulate_matter        3170 non-null float64
health__homicides_per_100k                      1231 non-null float64
health__motor_vehicle_crash_deaths_per_100k     2781 non-null float64
health__pop_per_dentist                         2954 non-null float64
health__pop_per_primary_care_physician          2968 non-null float64
yr                                              3198 non-null object
heart_disease_mortality_per_100k                3198 non-null int64
dtypes: float64(29), int64(2), object(4)
```

As we can see there are a total of 35 columns, consisting of 34 features and 1 target column. As we can see from the index count, there are total of 3198 entries, which matches the count of target label 'heart_disease_mortality_per_100k', providing confirmation that there are no missing values for the target label however we can see there are missing data for several features, which we will look at closer shortly.

Looking over the datatypes, we can see that there are 4 object types, which in this case would indicate a categorical feature. It is also clear that none of the categorical features are missing from the dataset.

The count of the columns provides an insight on how many rows are missing data, majority of columns have the full set of 3198 counts, few columns missing data for 2 instances/rows with 3196 counts in the demographic section mostly, while the features for health data have more missing data. Worth noting is that those features with 2 nulls happen to belong to 2 unique instances across these features, dropping these 2 rows out of the dataset is a reasonable option.

Individual feature statistics:

Summary statistics have been computed on the dataset for numeric features, as per below. The available data percentage is obtained by dividing individual feature count over total count of 3198.

	count	mean	std	min	median	max
row_id	3198	3116.986	1830.237	0	3113.5	6276
econ__pct_civilian_labor	3198	0.467191	0.0744	0.207	0.468	1
econ__pct_unemployment	3198	0.059696	0.022947	0.01	0.057	0.248
econ__pct_uninsured_adults	3196	0.217463	0.067362	0.046	0.216	0.496
econ__pct_uninsured_children	3196	0.086067	0.039849	0.012	0.077	0.281
demo__pct_female	3196	0.498811	0.024399	0.278	0.503	0.573
demo__pct_below_18_years_of_age	3196	0.227715	0.034282	0.092	0.226	0.417
demo__pct_aged_65_years_and_older	3196	0.170043	0.043694	0.045	0.167	0.346
demo__pct_hispanic	3196	0.090207	0.142763	0	0.035	0.932
demo__pct_non_hispanic_african_american	3196	0.091046	0.147165	0	0.022	0.858
demo__pct_non_hispanic_white	3196	0.769989	0.20785	0.053	0.853	0.99
demo__pct_american_indian_or_alaskan_native	3196	0.02468	0.084563	0	0.007	0.859
demo__pct_asian	3196	0.013109	0.025431	0	0.007	0.341
demo__pct_adults_less_than_a_high_school_diploma	3198	0.148815	0.068208	0.015075	0.133234	0.473526
demo__pct_adults_with_high_school_diploma	3198	0.350567	0.070554	0.065327	0.355015	0.558912
demo__pct_adults_with_some_college	3198	0.301143	0.052318	0.109548	0.301587	0.473953

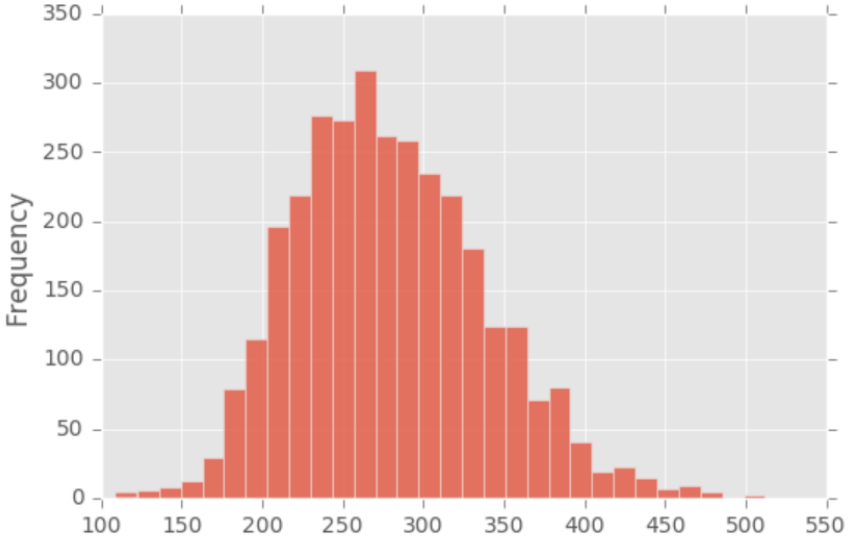
demo__pct_adults_bachelors_or_higher	3198	0.199475	0.089308	0.011078	0.176471	0.798995
demo__birth_rate_per_1k	3198	11.67699	2.739516	4	11	29
demo__death_rate_per_1k	3198	10.30113	2.786143	0	10	27
health__pct_adult_obesity	3196	0.307668	0.043228	0.131	0.309	0.471
health__pct_adult_smoking	2734	0.213628	0.062895	0.046	0.21	0.513
health__pct_diabetes	3196	0.10926	0.023216	0.032	0.109	0.203
health__pct_low_birthweight	3016	0.083896	0.022251	0.033	0.081	0.238
health__pct_excessive_drinking	2220	0.164841	0.050474	0.038	0.164	0.367
health__pct_physical_inacticity	3196	0.277161	0.053003	0.09	0.28	0.442
health__air_pollution_particulate_matter	3170	11.62587	1.557996	7	12	15
health__homicides_per_100k	1231	5.947498	5.031822	-0.4	4.7	50.49
health__motor_vehicle_crash_deaths_per_100k	2781	21.13262	10.48592	3.14	19.63	110.45
health__pop_per_dentist	2954	3431.434	2569.451	339	2690	28130
health__pop_per_primary_care_physician	2968	2551.339	2100.459	189	1999	23399
heart_disease_mortality_per_100k	3198	279.3693	58.95334	109	275	512

Analysing the table for any major discrepancies between the median and mean, we can see that the features for demographic information stand out, while all other features have comparatively similar figures. The biggest discrepancies of features are highlighted in yellow, consisting of demo__pct_hispanic, demo__pct_non_hispanic_african_american & demo__pct_american_indian_or_alaskan_native.

Our target column is the variable we are looking to predict, which is the heart mortality rate. The mean value of 279 a median value 275 would indicate that there is no significant skewness to be expected, just a minor skew to the right as the median is slightly above the median.

Distributions of features:

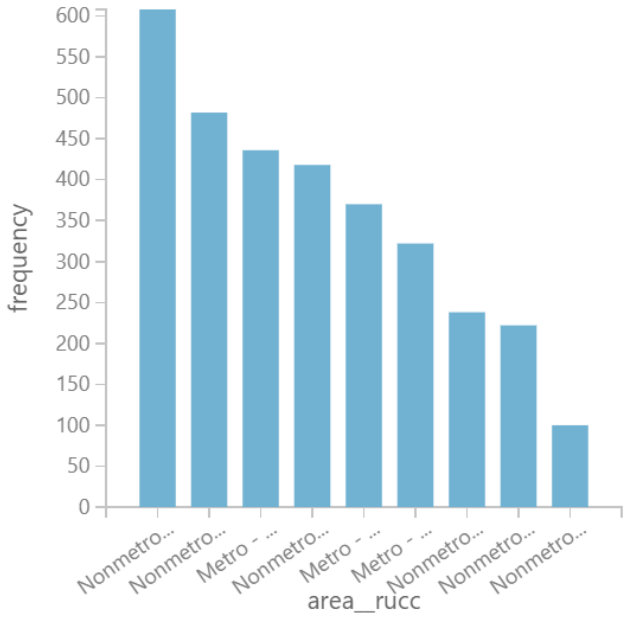
Target column: heart_disease_mortality_per_100k



Plotting the target variable, we confirm that indeed the distribution resembles a normal distribution, with a small skew to the right as expected. With a mean of 279.37 and standard deviation of 58.95, using the 68–95–99.7 rule we can conclude 95% of heart mortality rates lie between 161.47 and 397.37 per 100k.

Categorical features:

area__rucc, dividing into 9 categories of metropolitan and non-metropolitan areas:



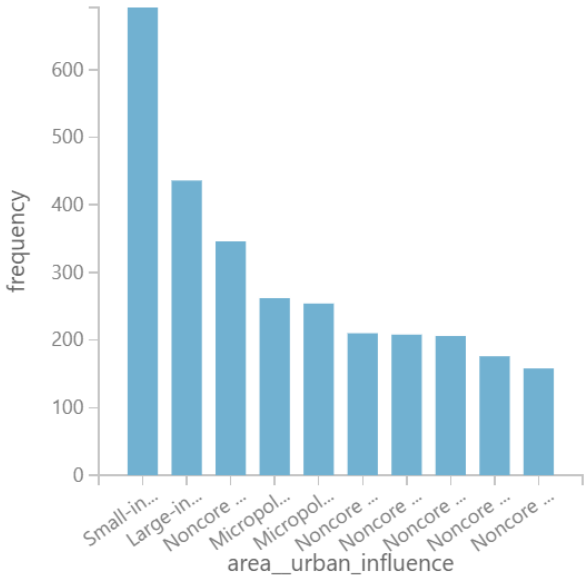
Dividing the categories into 2 of metro and non-metro, we get the following count:

Non-Metro: 2070

Metro: 1128

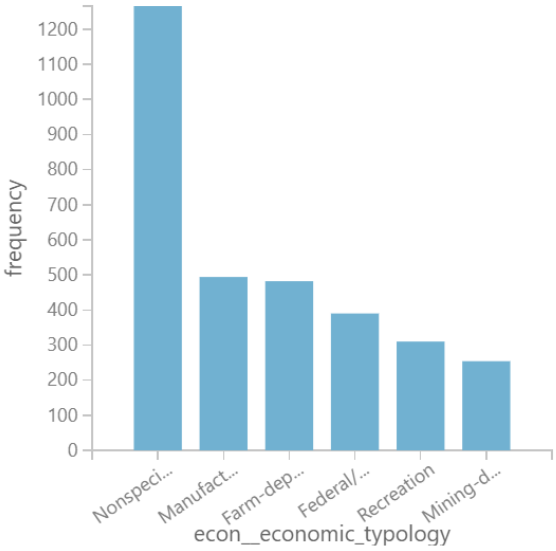
There is almost twice the amount of non-metropolitan areas in our dataset compared to metropolitan, indicating a noticeable imbalance in the classes of the feature.

area__urban_influence, breaking down the locations into more detailed major and minor areas:



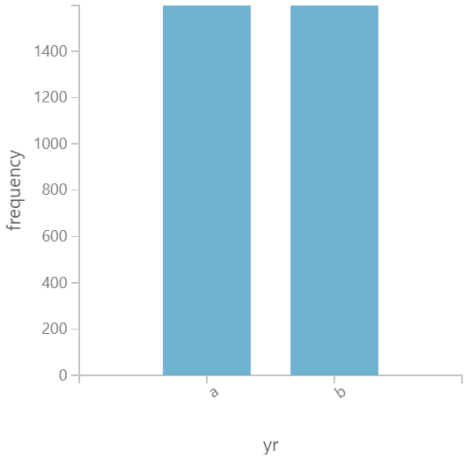
Given that metropolitan areas are under-represented in relation to non-metropolitan areas, as expected there is a skew in this feature as well, having a derivative relationship with area_rucc.

econ__economic_typology, grouping the areas into 6 economic types:



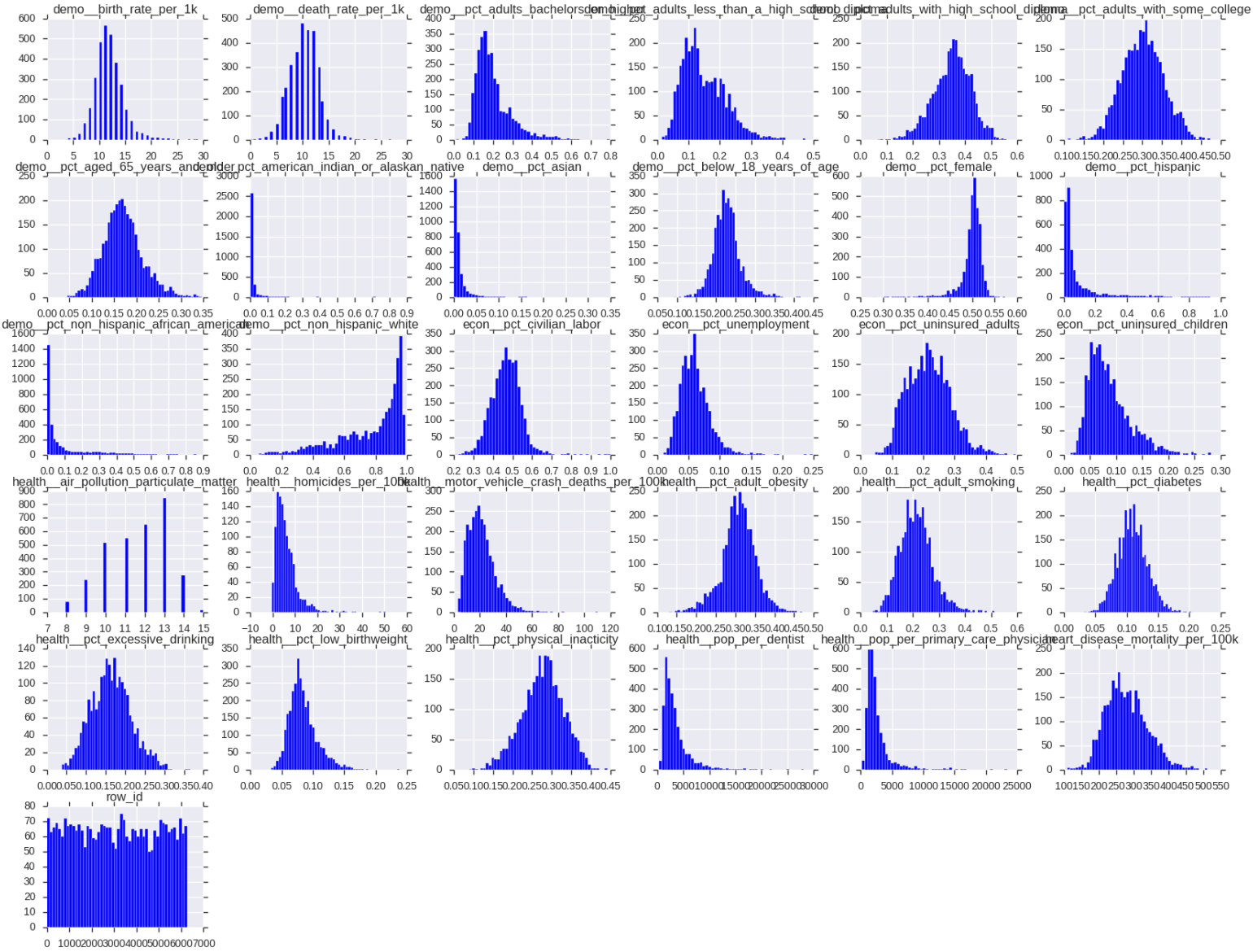
The class 'Nonspecialized' is over-represented in this feature, with 39.6% frequency from the full set, again establishing an imbalance in classes. However, among the other classes the variance between classes are less extreme, with manufacturing counting the second-most at 15.4% and mining coming last at 7.9%.

Yr – simple division into 2 categories of a & b:



Being the only categorical feature that is symmetrical in our dataset, both classes count exactly at 1599 instances.

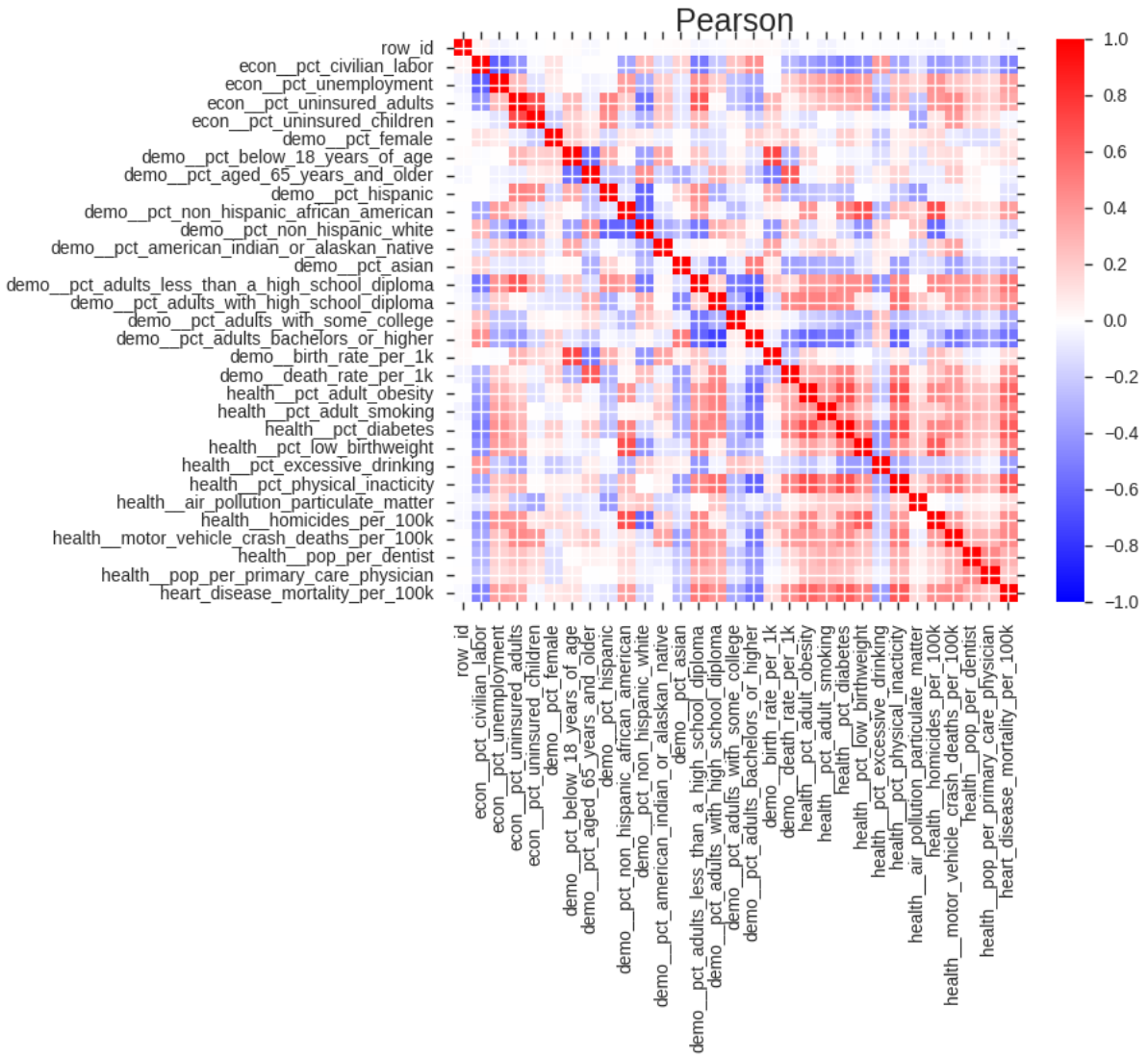
Overview of numeric feature distributions:



Looking over the matrix of histograms for numeric features, we confirm our previous exploratory analysis about the significant skewness of the demographic features, particularly demo__pct_hispanic, demo__pct_non_hispanic_african_american & demo__pct_american_indian_or_alaskan_native. We also find that there are more variables that show significant skewness in the health categories such as health__pop_per_dentist & health__pop_per_primary_care_physician. Otherwise, the features tend to resemble a bell-shaped curve, with no significant outliers that skew the whole distribution either side to a major degree that stands out.

Correlations

As there are many features in our dataset, we need to filter out relevant features. A step in understanding the significances between features is by plotting the correlation:



Given the high number of features, we plot the correlation matrix and sort for the top 10 highest absolute correlations between features and the bottom 10:

Top 10:

demo__pct_adults_bachelors_or_higher	demo__pct_adults_with_high_school_diploma	0.739895
demo__birth_rate_per_1k	demo__pct_below_18_years_of_age	0.729499
econ__pct_uninsured_children	econ__pct_uninsured_adults	0.717686
health__pct_diabetes	health__pct_adult_obesity	0.70116
health__homicides_per_100k	demo__pct_non_hispanic_african_american	0.694183
health__pct_physical_inacticity	health__pct_adult_obesity	0.683851

demo__pct_non_hispanic_african_american	health__pct_low_birthweight	0.68141
health__pct_diabetes	health__pct_physical_inacticity	0.67447
econ__pct_uninsured_adults	demo__pct_adults_less_than_a_high_school_diploma	0.66328
health__pct_physical_inacticity	heart_disease_mortality_per_100k	0.650305

The highest correlation is between demo__pct_adults_bachelors_or_higher & demo__pct_adults_with_high_school_diploma at 0.74, which is not surprising given the nature of the individual features. Further correlations also depict an expected result such as a high correlation between adult obesity, or the highest correlation between the target label heart_disease_mortality_per_100k against physical inactivity at 0.65.

Bottom 10, excluding row_id:

econ__pct_uninsured_children	health__pct_adult_smoking	0.005351
econ__pct_civilian_labor	demo__pct_aged_65_years_and_older	0.005105
health__pct_low_birthweight	demo__pct_hispanic	0.005086
demo__pct_below_18_years_of_age	econ__pct_unemployment	0.004807
heart_disease_mortality_per_100k	demo__pct_american_indian_or_alaskan_native	0.004626
econ__pct_civilian_labor	demo__birth_rate_per_1k	0.00448
health__pop_per_primary_care_physician	demo__pct_aged_65_years_and_older	0.004279
health__pct_physical_inacticity	demo__pct_non_hispanic_white	0.00175
demo__pct_below_18_years_of_age	health__pct_adult_smoking	0.00105
demo__pct_hispanic	health__pop_per_primary_care_physician	0.000065

Analysing the figures for the bottom 10 correlations between features, there are some interesting observations that may come as a surprise. For instance, there is no correlation between percentage of civilian labour against birth rate, or percentage of unemployment against aged below 18.

As we are interested in the heart mortality rate, let's look at the correlations against it:

health__pct_physical_inacticity	heart_disease_mortality_per_100k	0.650305
health__pct_diabetes	heart_disease_mortality_per_100k	0.631765
health__pct_adult_obesity	heart_disease_mortality_per_100k	0.593775
demo__pct_adults_bachelors_or_higher	heart_disease_mortality_per_100k	0.541385
heart_disease_mortality_per_100k	demo__pct_adults_less_than_a_high_school_diploma	0.527382
health__pct_adult_smoking	heart_disease_mortality_per_100k	0.497063
health__pct_low_birthweight	heart_disease_mortality_per_100k	0.476757
econ__pct_civilian_labor	heart_disease_mortality_per_100k	0.476644
health__motor_vehicle_crash_deaths_per_100k	heart_disease_mortality_per_100k	0.459803
heart_disease_mortality_per_100k	demo__death_rate_per_1k	0.444757
health__homicides_per_100k	heart_disease_mortality_per_100k	0.441164

demo__pct_adults_with_high_school_diploma	heart_disease_mortality_per_100k	0.428137
health__pct_excessive_drinking	heart_disease_mortality_per_100k	0.382172
heart_disease_mortality_per_100k	demo__pct_non_hispanic_african_american	0.375385
heart_disease_mortality_per_100k	econ__pct_unemployment	0.37162
demo__pct_adults_with_some_college	heart_disease_mortality_per_100k	0.340764
heart_disease_mortality_per_100k	econ__pct_uninsured_adults	0.334217
heart_disease_mortality_per_100k	health__pop_per_dentist	0.301232
demo__pct_asian	heart_disease_mortality_per_100k	0.267458
heart_disease_mortality_per_100k	health__pop_per_primary_care_physician	0.219111
heart_disease_mortality_per_100k	demo__pct_non_hispanic_white	0.157544
heart_disease_mortality_per_100k	health__air_pollution_particulate_matter	0.150019
demo__birth_rate_per_1k	heart_disease_mortality_per_100k	0.142176
demo__pct_below_18_years_of_age	heart_disease_mortality_per_100k	0.121956
heart_disease_mortality_per_100k	demo__pct_hispanic	0.112437
demo__pct_female	heart_disease_mortality_per_100k	0.08704
heart_disease_mortality_per_100k	demo__pct_aged_65_years_and_older	0.056203
heart_disease_mortality_per_100k	econ__pct_uninsured_children	0.034482
row_id	heart_disease_mortality_per_100k	0.019117
heart_disease_mortality_per_100k	demo__pct_american_indian_or_alaskan_native	0.004626
yr	heart_disease_mortality_per_100k	0.000032

There are numerous features with high correlations ranging above 0.5, the top 5 being:

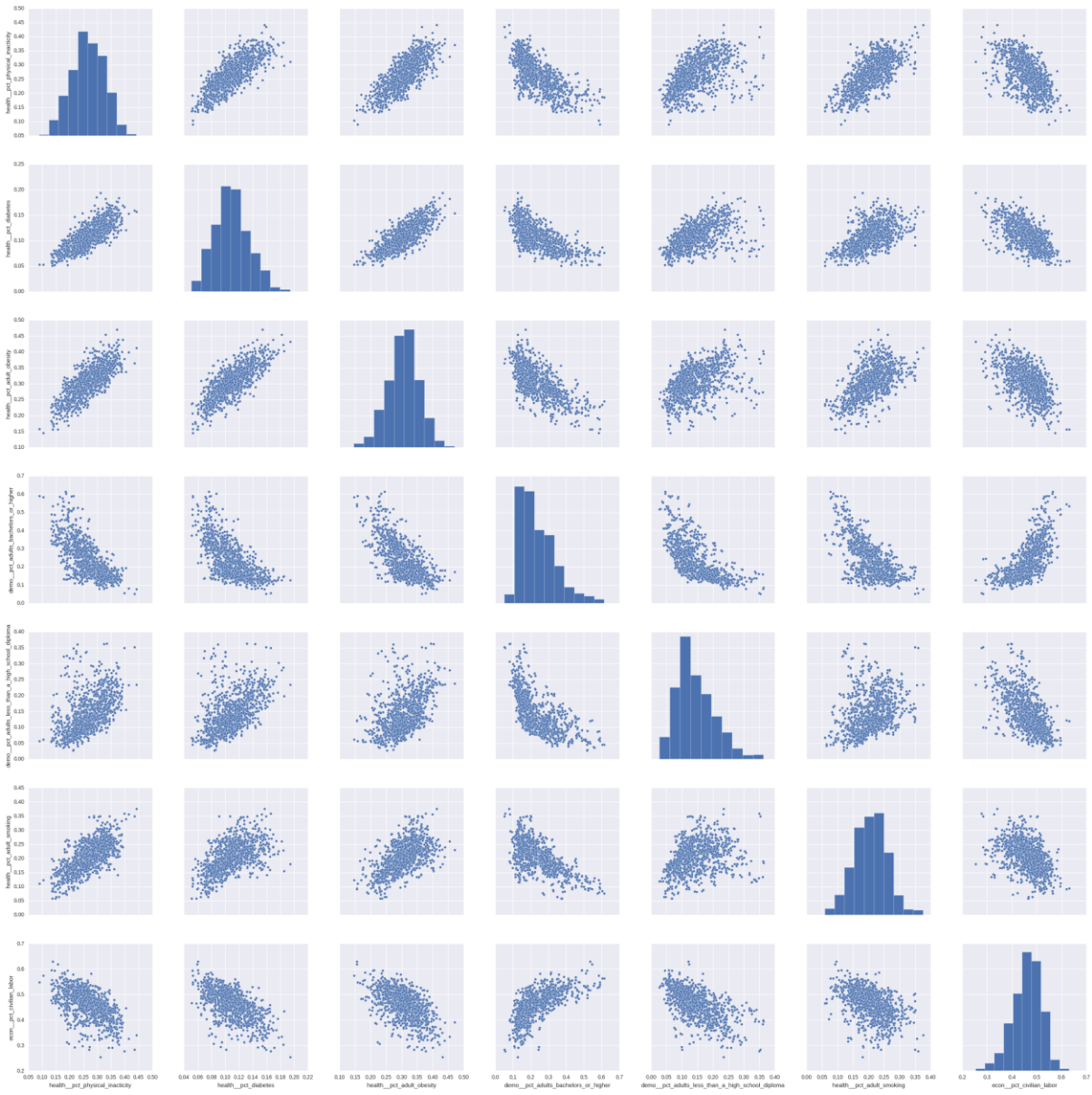
1. health__pct_physical_inacticity
2. health__pct_diabetes
3. health__pct_adult_obesity
4. demo__pct_adults_bachelors_or_higher
5. demo__pct_adults_less_than_a_high_school_diploma

The top 5 features (excl. row_id) that seem to have fewest effect on heart mortality rate are found to be:

1. demo__pct_american_indian_or_alaskan_native
2. econ__pct_uninsured_children
3. demo__pct_aged_65_years_and_older
4. demo__pct_female
5. demo__pct_hispanic

To visualise the relationship between the key features, a scatter plot matrix is used in the order of:

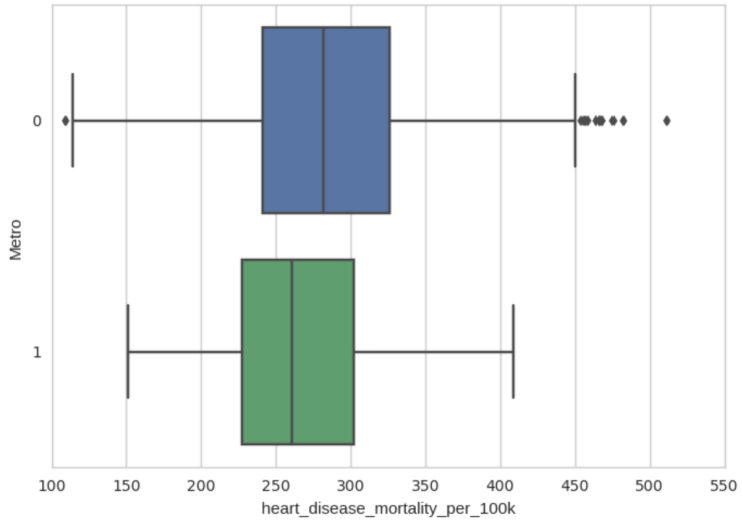
["health__pct_physical_inacticity", "health__pct_diabetes", "health__pct_adult_obesity", "demo__pct_adults_bachelors_or_higher", "demo__pct_adults_less_than_a_high_school_diploma", "health__pct_adult_smoking", "econ__pct_civilian_labor"]



Looking over the relationships between the features, we can observe many positive and negative correlations, with some clearly linear relationships such as between physical inactivity and obesity and diabetes and obesity. On the other hand, negative correlations were found between numerous features as well, notably reduced obesity with higher civilian labour or reduced smoking with increased ratio of bachelor's degree or higher.

Categorical relationships:

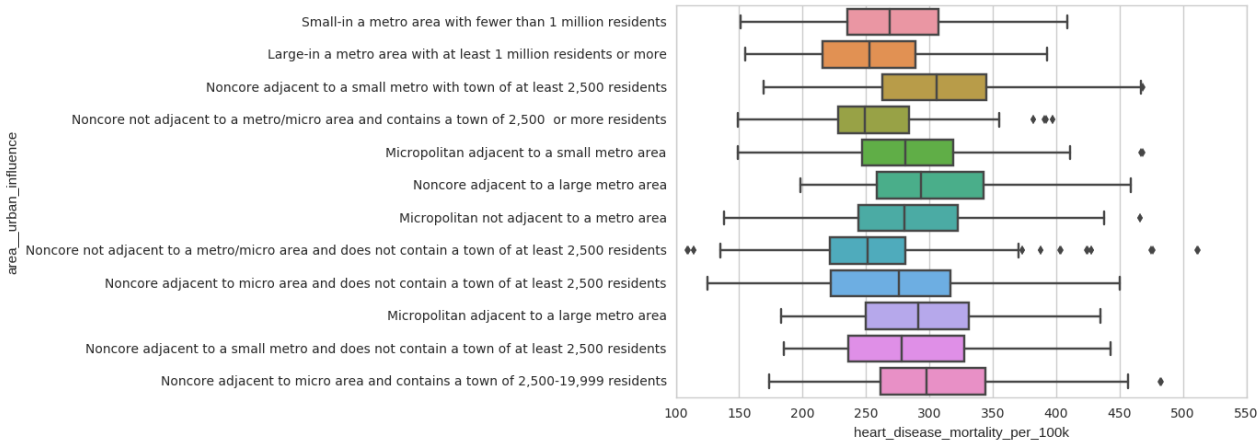
area__rucc grouped by 1- Metro, 0 – Non-metro:



From the above boxplot the average rate for heart disease mortality is higher for non-metropolitan areas, with the below summary statistics which shows a difference of 20 based on the mean. However, the range of the non-metropolitan class is wider, surpassing both the minimum and maximum of the metropolitan class.

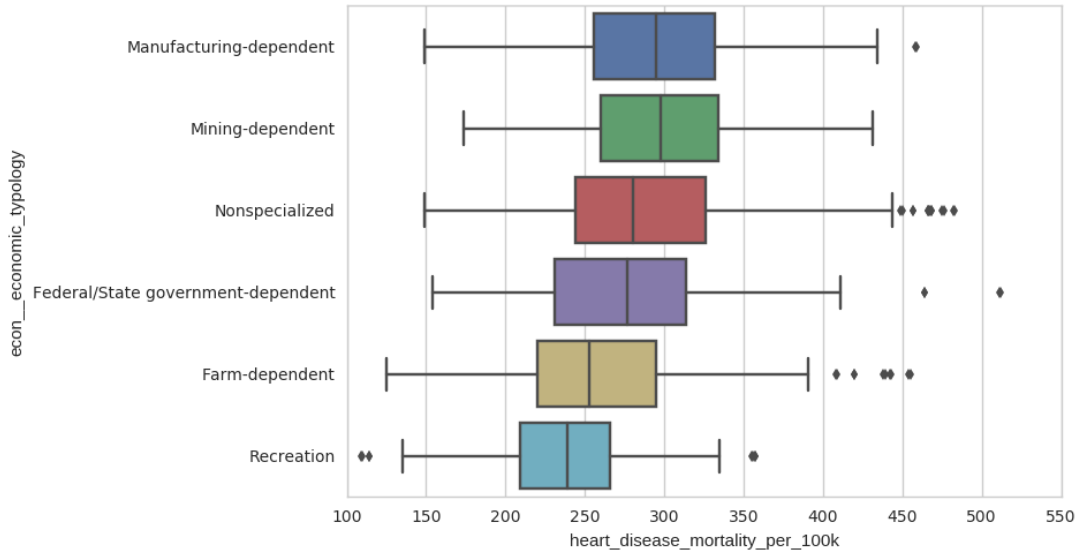
```
Metro:
Count    1128.000000
mean      266.191489
std        50.429501
min       151.000000
25%       227.000000
50%       261.000000
75%       302.000000
max       409.000000

Non-Metro:
count    2070.000000
mean      286.550242
std        61.957844
min       109.000000
25%       241.000000
50%       282.000000
75%       326.000000
max       512.000000
```

Area_urban_influence:



Among the 12 different classes, we can see a variety of ranges including some classes with many outliers, such as the 'Noncore not adjacent to a metro/micro...'. The median of all classes ranges around 250 to 300, with areas of larger populations showing an increased amount as expected given the previous look on metro/non-metro boxplot.

Economic_typology:

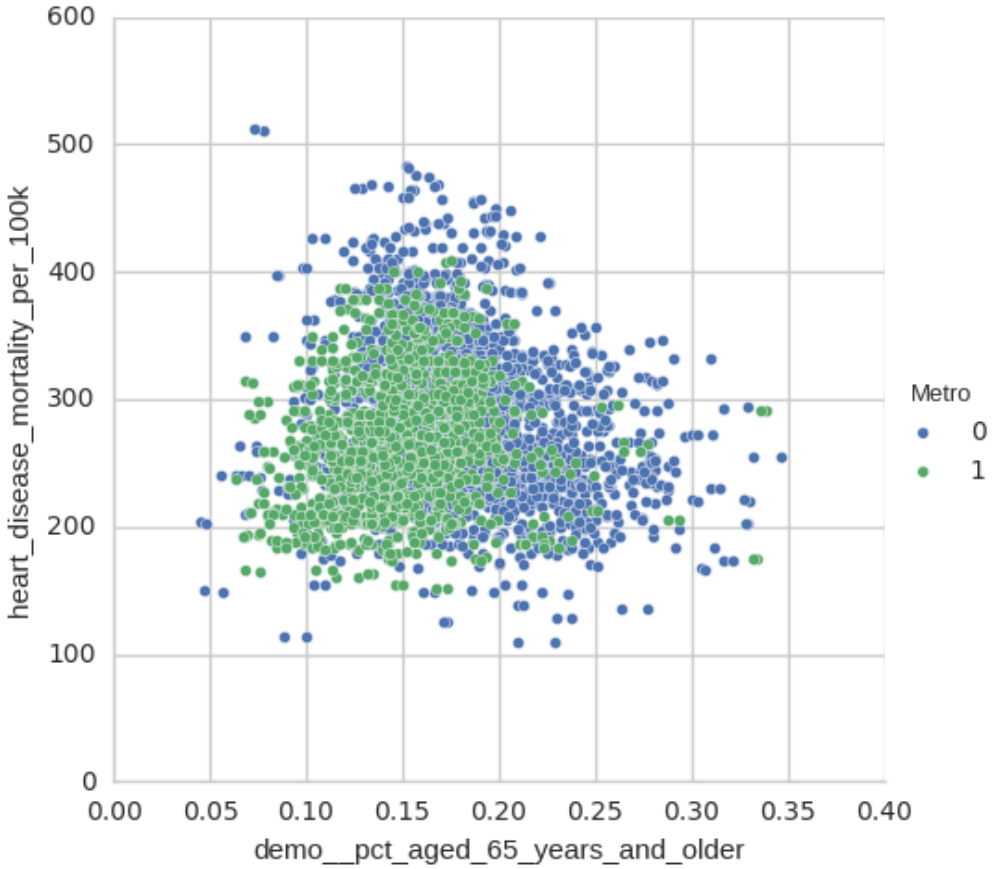


For the economic typology classes, mining shows the highest median close to 300, while recreation-dependent areas hold a median of close 240.

Some conclusions that can be made from the above is that metro areas have overall less heart diseases mortality rates compared to non-metro areas. This observation goes in line with the fact that areas that are more focused on labour intensive work such as manufactoruing and mining have higher heart mortality rates, while white-collar work such as recreation tends to have a lesser value.

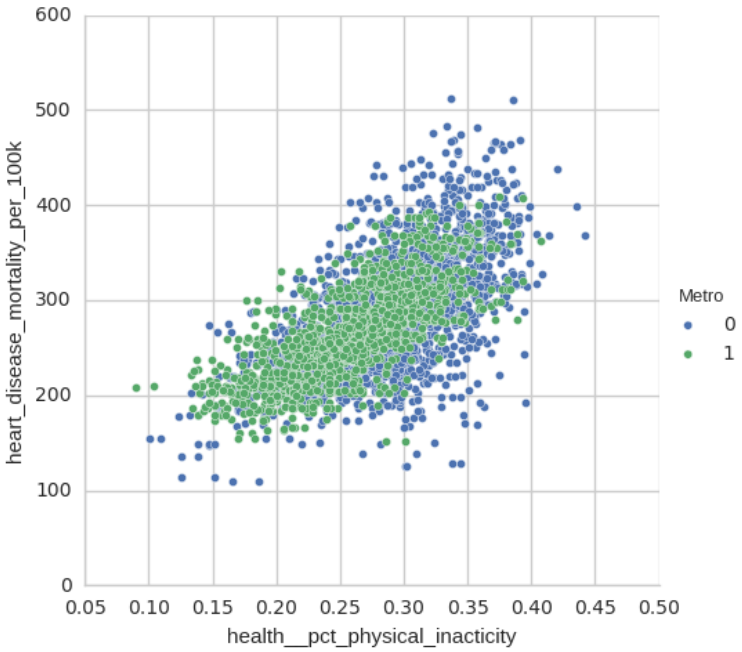
Multi-faceted relationships:

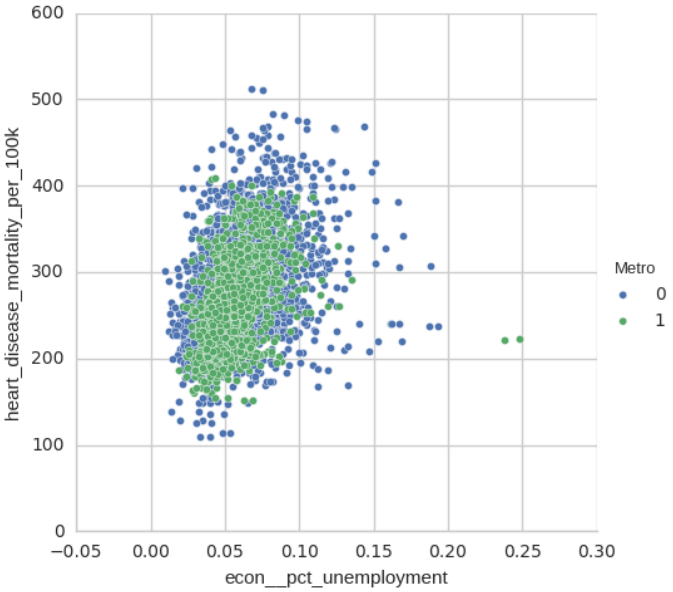
To understand the relationships between features as well as the target variable, we plot a conditioned scatter plot. An interesting exploration is to see whether there is a relationship in heart disease mortality rate and population with a higher ratio of elderly, aged over 65. At the same time, we will class the points by metro (1) or non-metro (0). What we discover is that firstly non-metro areas have a higher ratio of aged over 65. However, there is no visible correlation between heart mortality rate and age group of over 65. Thus, while we can conclude that non-metro areas have a higher ration of aged over 65, it has no significant impact on predicting heart disease mortality rate.



We continue to look at the relationship between physical inactivity, heart mortality and metro areas.

We have already established previously that physical inactivity is the highest predictor of heart diseases mortality, when looking at the correlations. This linear relationship is evident on the scatter plot, at the same time it is revealed that physical inactivity tends to be a little higher in non-metro areas. A different outcome is found when plotting against percentage of unemployment, where there is no apparent correlation between unemployment and heart mortality, while non-metro areas exhibit both higher occurrences of unemployment for certain areas, but also have instances of less unemployment compared to metro areas.





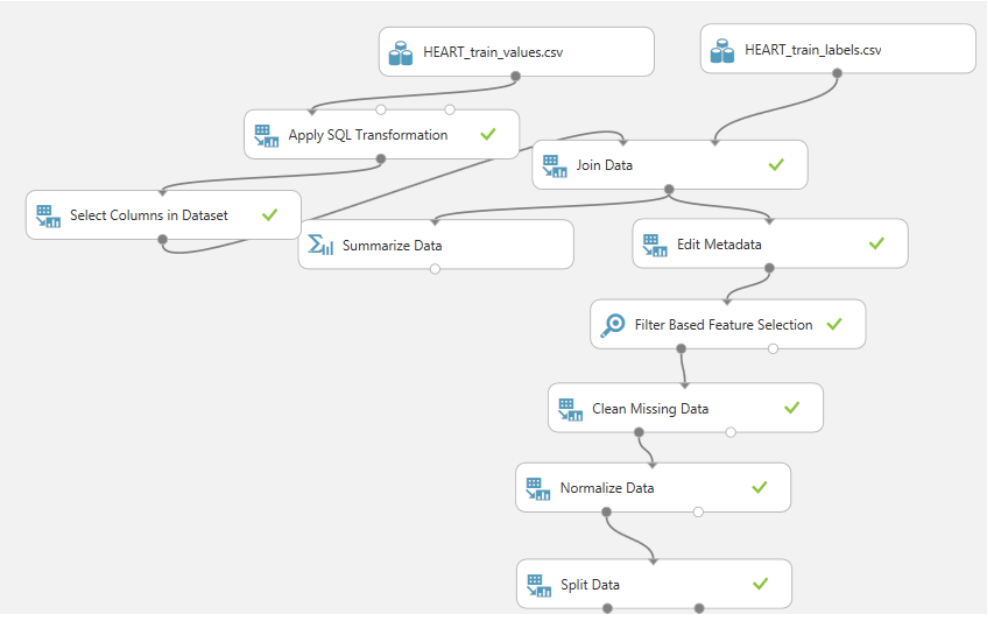
Regression of heart mortality rate based on demographic and economic factors – Azure Machine Learning Studio

To begin with the task of predicting heart disease mortality rate, the initial step is to understand the underlying problem to choose the correct model. As the target label is not a categorical datatype, but a numeric one we can deduce that this is a regression problem. Initially three models were tested in parallel to determine the best performer consisting of decision forest regression, boosted decision tree regression and linear regression. As the linear regression model consistently underperformed against the latter, it was removed.

An analysis on the data was performed as part of the data cleaning process, which found the below:

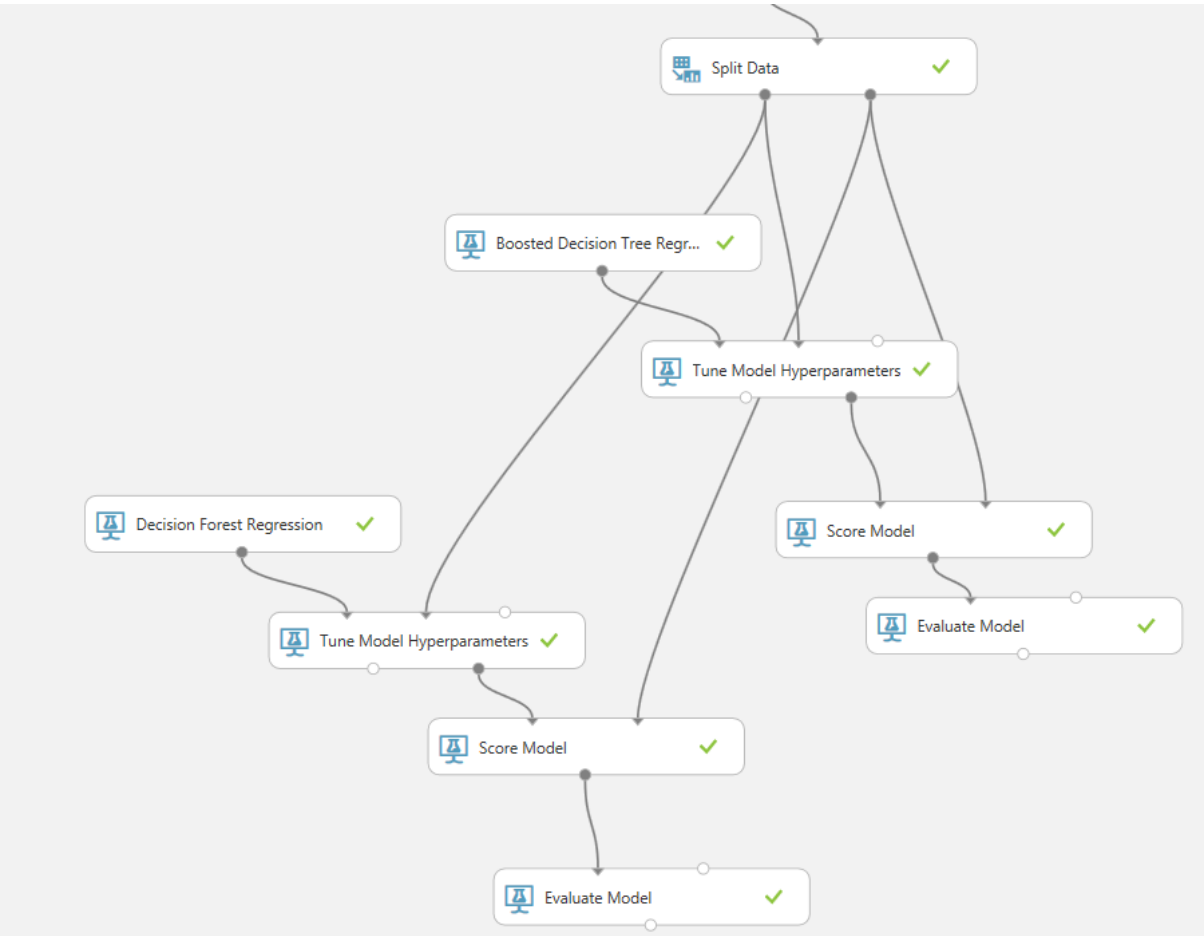
- `demo__pct_american_indian_or_alaskan_native` has 544 / 17.0% zeros **Zeros**
- `demo__pct_asian` has 517 / 16.2% zeros **Zeros**
- `demo__pct_non_hispanic_african_american` has 299 / 9.3% zeros **Zeros**
- `health__homicides_per_100k` has 1967 / 61.5% missing values **Missing**
- `health__motor_vehicle_crash_deaths_per_100k` has 417 / 13.0% missing values **Missing**
- `health__pct_adult_smoking` has 464 / 14.5% missing values **Missing**
- `health__pct_excessive_drinking` has 978 / 30.6% missing values **Missing**
- `health__pct_low_birthweight` has 182 / 5.7% missing values **Missing**
- `health__pop_per_dentist` has 244 / 7.6% missing values **Missing**
- `health__pop_per_primary_care_physician` has 230 / 7.2% missing values **Missing**

As a first step, using SQL transformation, 2 rows were dropped, that was mentioned previously that had numerous NULLs across features. Also, a given the high count of missing data for health__pct_excessive_drinking & health__homicides_per_100k, it is decided to drop these 2 features completely. Data has been joined between train values and train labels on row_id, and using Edit Metadata, we specify 4 columns to be categorical.



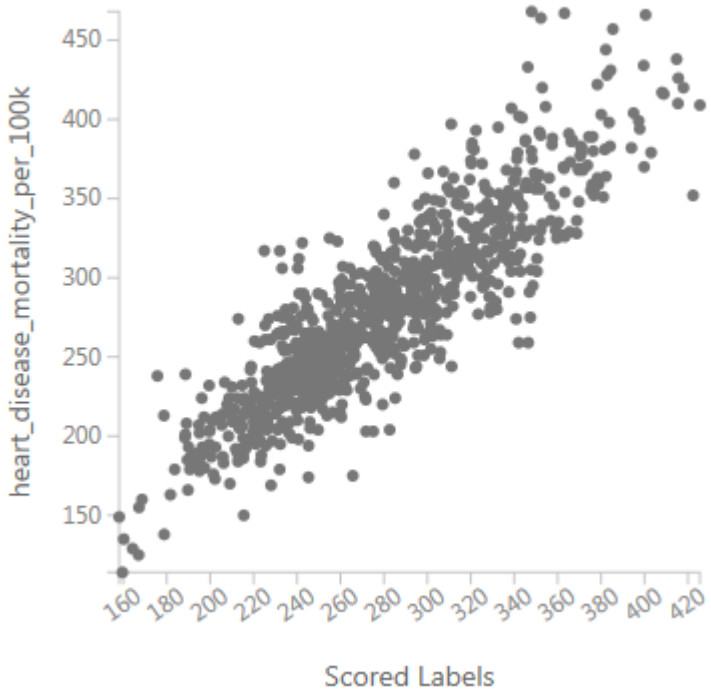
To lessen the curse of dimensionality, given the high count of features, a feature selection process is conducted based on the Pearson correlation of target column *heart disease mortality rate* against other features using *Filter Based Feature Selection*. 3 features from the bottom of the correlation list are removed which have a correlation of less than 3%: Yr, Row_id, and demo__pct_american_indian_or_alaskan_native.

As a next step, we fill the missing data using probabilistic PCA, as opposed to MICE, due to rather larger ratio of missing data for certain columns such as percentage of smokers and vehicle crash deaths. Lastly the numeric columns (with the exception of the target column) is normalized to ensure that all features are weighted equally in their representation, which will hopefully yield a more accurate prediction model.



The normalized data is split 70/30 and used to train in parallel a decision forest and decision tree regression model, with hyperparameters for both being tuned on the entire grid of default parameter ranges. Using the evaluation tool, we determine the Root Mean Square Error for decision forest to be 31.29, and the decision tree model to be 27.42, subsequently we choose the latter model as the final pick.

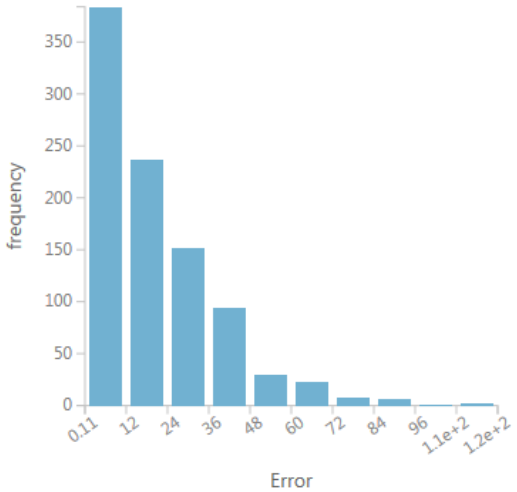
Conclusion:



Metrics

Mean Absolute Error	20.520499
Root Mean Squared Error	27.423988
Relative Absolute Error	0.440278
Relative Squared Error	0.22322
Coefficient of Determination	0.77678

Error Histogram



A clear linear relationship is obvious when plotting the predicted labels against the actual values of heart disease mortality rate. The figure of the Root Mean Squared Error in this scenario represents the discrepancy between the predicted and actual values, which is 27.42. As this figure is well below the standard deviation of the target column, which is at 58.95, we can conclude that the model performs reasonably well. This is supported by the Coefficient of Determination that scored at 0.77, indicating that 77% of the variability in the dataset can be explained by the model and is a common measure for accuracy.

In saying that, we need to keep in mind that the training dataset in our project consisted of 3198 instances, which is considered very small in the world of data science. Furthermore, splitting the dataset further into 70/30 for validation brings the number down. This is a strong limitation in training a reliable model, as there is insufficient data. Furthermore, the chosen model of decision tree regression is known for tendency to overfit data, which means it will not perform as well on new unseen data when tested. This is proven by the fact that the RMSE score on the test set indeed comes out higher, at approximately 32.8. To ensure a

better result, ideally, we would like to see significantly higher amount of training data. Additionally, a more complete data per feature is also desired, to avoid the removal of features due to missing data, and reduction of imputations of missing data during data cleaning.