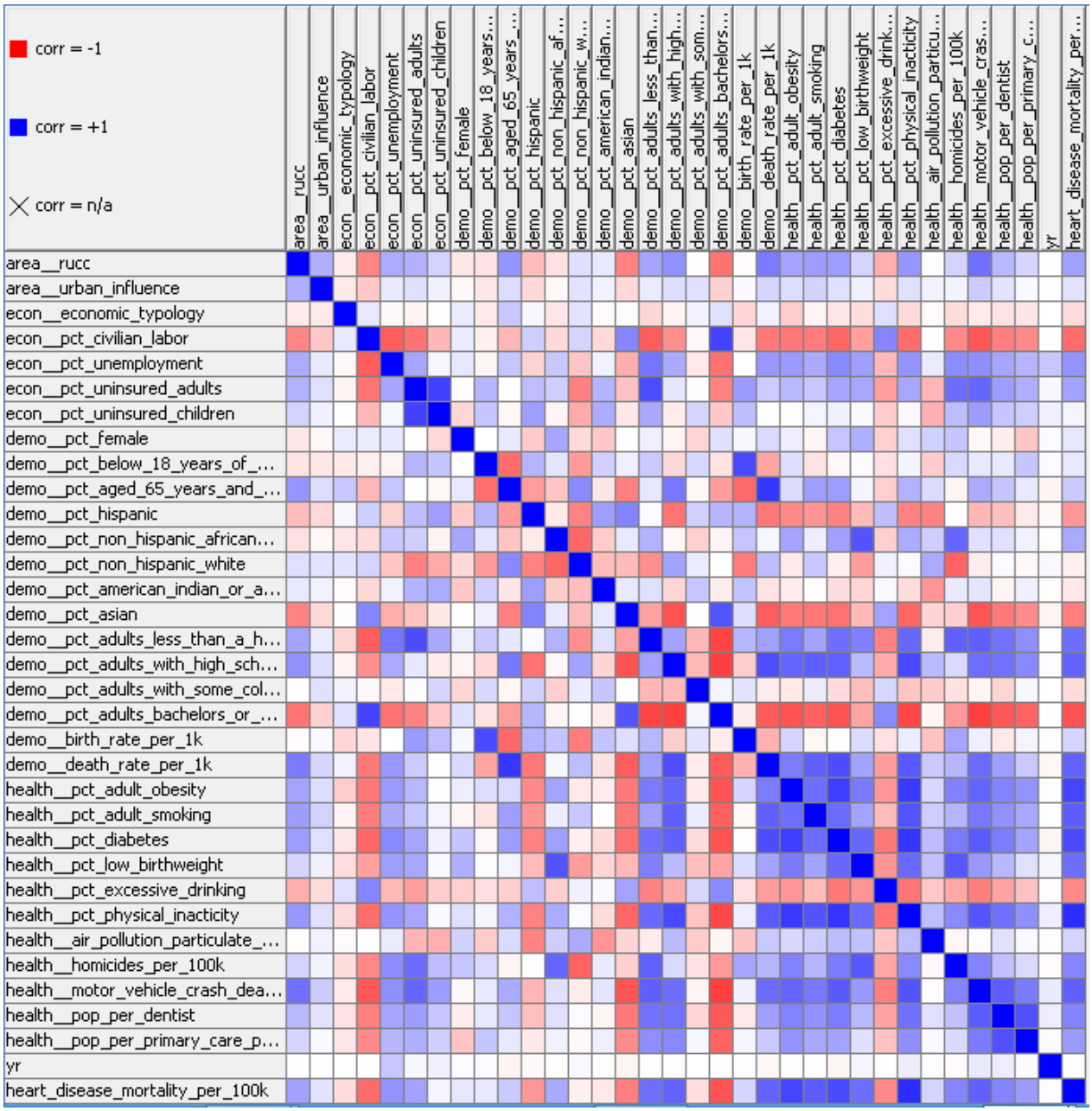
3. Correlation Analysis

Predicting Heart Disease Mortality with Machine Learning Techniques
Lorenzo Negri, July 2018

Overview of Relationships

The scope of this analysis is to identify relationships between features in the data – in particular, between **target variable** and the other features. The following calculated heat-map correlation matrix compares all features with one another.



The heat-map indicates that white or very pale colours are not indicating correlation; blue indicates positive relationship and red negative correlation, more intense means more correlation. The matrix shows apparent relationship between heart disease mortality and other features. Especially with, physical inacticity (higher percentage increase mortality), bachelor or higher education (higher percentage decrease mortality), and other health features that increase heart disease mortality rate.

It can be also noticed that:

- There is a strong positive correlation with percentage of bachelor or higher education and of civilian labor. Which is quite self-explanatory.
- Negative correlations with percentage of bachelor or higher education and most of health features and uninsured/unemployed ones. Meaning that better education probably leads to better culture of health and more healthy care possibilities.
- Positive correlation between percentage of age over 65 and death rate per 1k. Which is quite self-explanatory.

Some of these correlations will be managed for the feature selection and engineering of the prediction model.

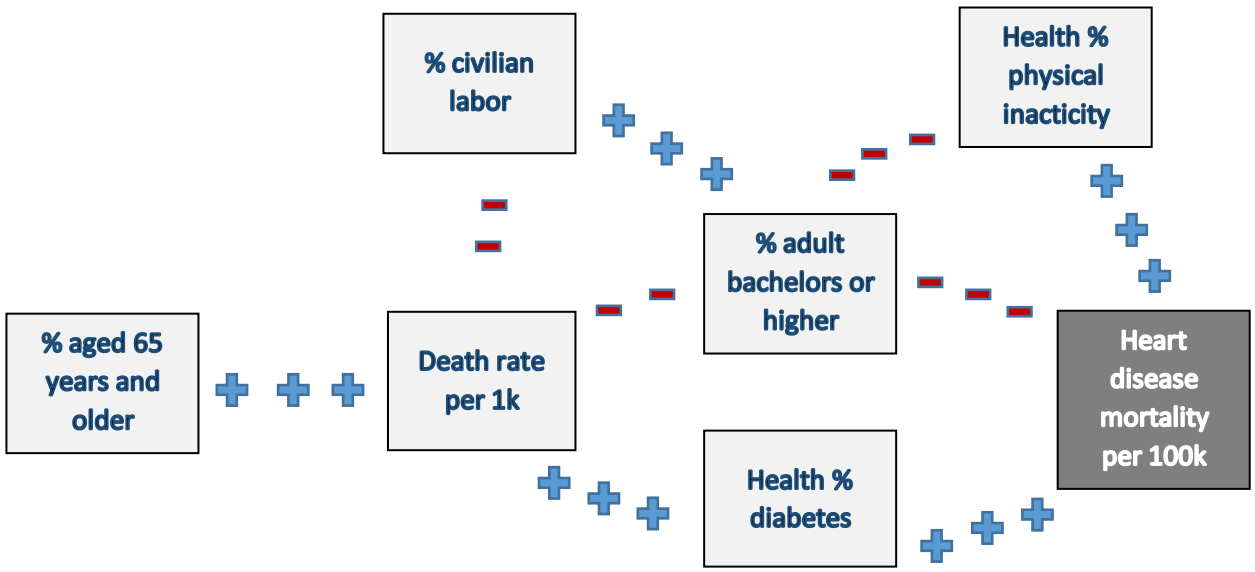
Numeric Relationships

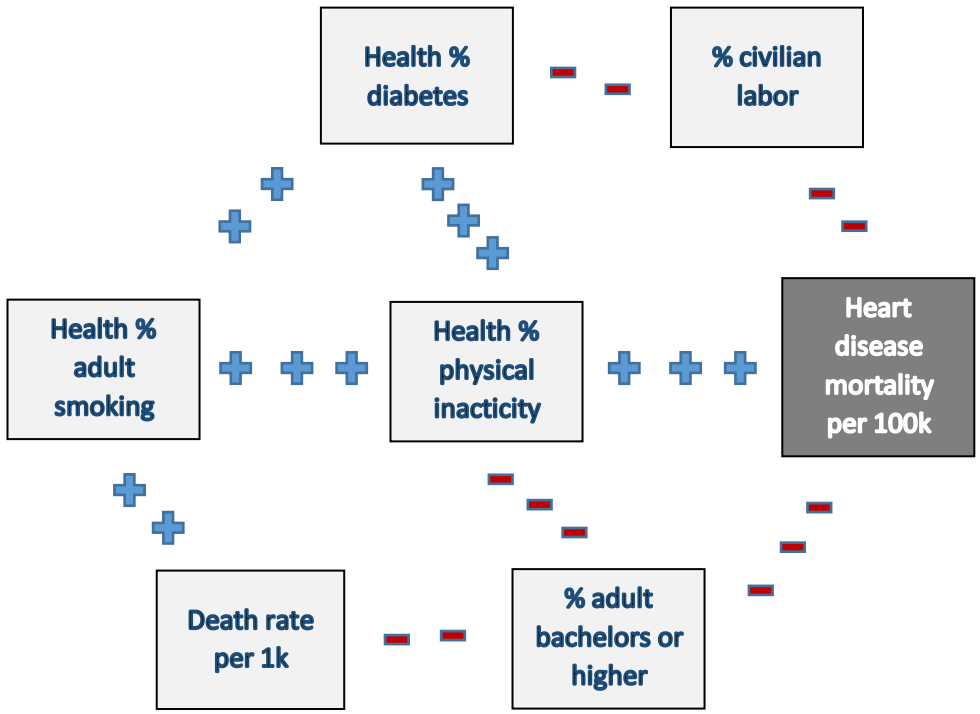
The correlation measure matrix between some relevant numeric columns (selected by high Spearman's rank correlation coefficient between high volume of missing observations columns and cross-measured correlation with the target variable:

	pct_civilian_labor	pct_aged_65_years_and_older	pct_adults_bachelors_or_higher	demo__death_rate_per_1k	health__pct_adult_smoking	health__pct_physical_inacticity	health__pct_diabetes	heart_disease_mortality_per_100k
pct_civilian_labor	1	-0.29	0.734	-0.53	-0.51	-0.57	-0.6	-0.58
pct_aged_65_years_and_older	-0.29	1	-0.38	0.787	0.382	0.302	0.383	0.207
pct_adults_bachelors_or_higher	0.73	-0.38	1	-0.64	-0.62	-0.72	-0.68	-0.68
demo__death_rate_per_1k	-0.53	0.787	-0.64	1	0.622	0.651	0.699	0.615
health__pct_adult_smokin	-0.51	0.382	-0.62	0.622	1	0.671	0.58	0.635
health__pct_physical_inacticit	-0.57	0.302	-0.72	0.651	0.671	1	0.793	0.828
health__pct_diabete	-0.6	0.383	-0.68	0.699	0.58	0.793	1	0.711
heart_disease_mortality_per_100	-0.58	0.207	-0.68	0.615	0.635	0.828	0.711	1

The table validate the heat-map by showing correlations for the numeric features has mentioned. It is interesting to see some concatenated correlation effect that could let us exclude in some cases the apparent relationships.

Some of the correlations effect schemes

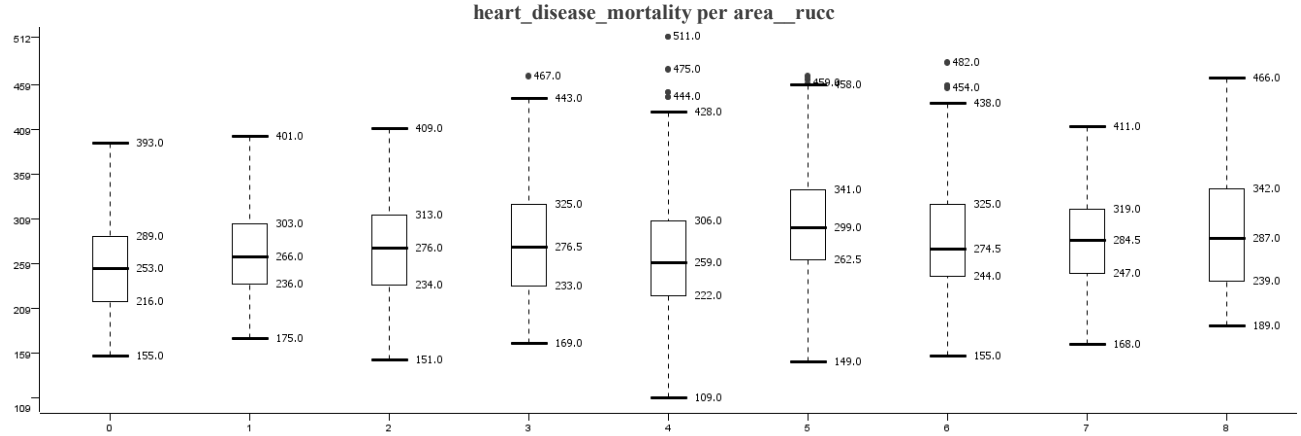




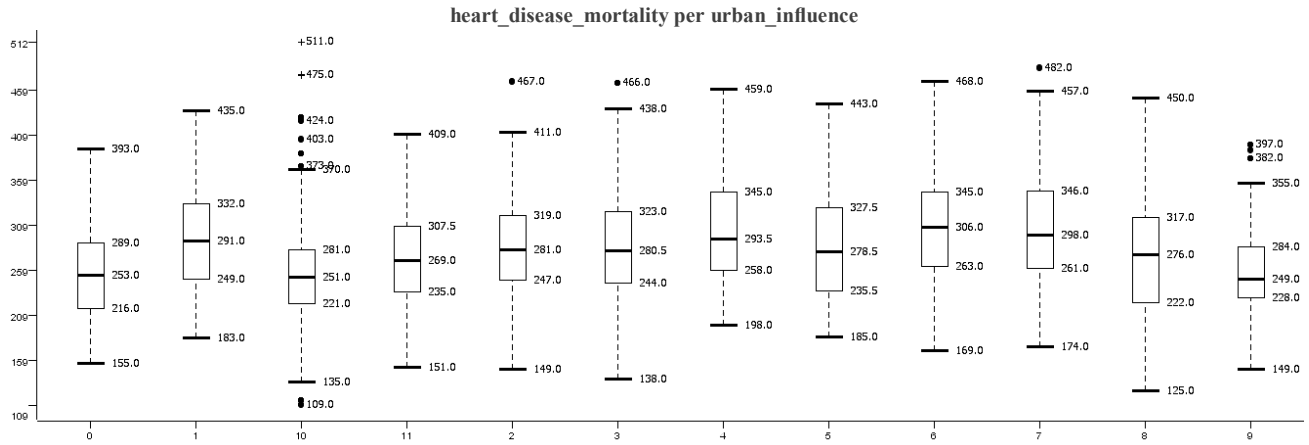
It appears from the schemes above, that *pct_aged_65_years_and_older* is showing more correlation to *heart_disease_mortality_per_100k* than how the matrix was measuring. In addition, the correlation measure from the matrix for *health__pct_adult_smoking* with the target variable could maybe be too high, caused by directly and self-explanatory relationship with *health__pct_physical_inacticity*.

Categorical Relationships

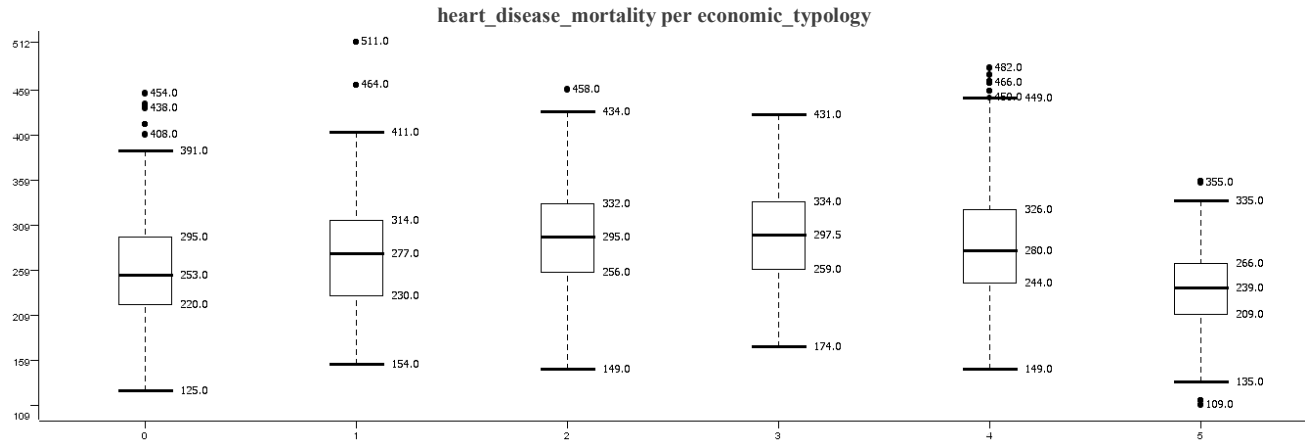
Some pale apparent relationship between categorical feature values and heart disease mortality rate can be analysed. The following boxplots show the categorical columns compared with *heart_disease_mortality_per_100k*:



Nominal values (area__rucc): 0=Metro - Counties in metro areas of 1 million population or more, **1=Metro** - Counties in metro areas of 250,000 to 1 million population, **2=Metro** - Counties in metro areas of fewer than 250,000 population, **3=Nonmetro** - Completely rural or less than 2,500 urban population, adjacent to a metro area, **4=Nonmetro** - Completely rural or less than 2,500 urban population, not adjacent to a metro area, **5=Nonmetro** - Urban population of 2,500 to 19,999, adjacent to a metro area, **6=Nonmetro** - Urban population of 2,500 to 19,999, not adjacent to a metro area, **7=Nonmetro** - Urban population of 20,000 or more, adjacent to a metro area, **8=Nonmetro** - Urban population of 20,000 or more, not adjacent to a metro area.



Nominal values (urban_influence): 0=Large-in a metro area with at least 1 million residents or more, **1=Micropolitan** adjacent to a large metro area, **10=Micropolitan** adjacent to a small metro area, **11=Micropolitan** not adjacent to a metro area, **2=Noncore** adjacent to a large metro area, **3=Noncore** adjacent to a small metro and does not contain a town of at least 2,500 residents, **4=Noncore** adjacent to a small metro with town of at least 2,500 residents, **5=Noncore** adjacent to micro area and contains a town of 2,500-19,999 residents, **6=Noncore** adjacent to micro area and does not contain a town of at least 2,500 residents, **7=Noncore** not adjacent to a metro/micro area and contains a town of 2,500 or more residents, **8=Noncore** not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents, **9=Small-**in a metro area with fewer than 1 million residents.



Nominal values (economic_typology): 0=Farm-dependent, 1=Federal/State government-dependent, 2=Manufacturing-dependent, 3=Mining-dependent, 4=Nonspecialized, 5=Recreation

The box plots show some differences in terms of the median and range of *heart_disease_mortality* rate for some categorical values, they are not significantly important but interesting. For example:

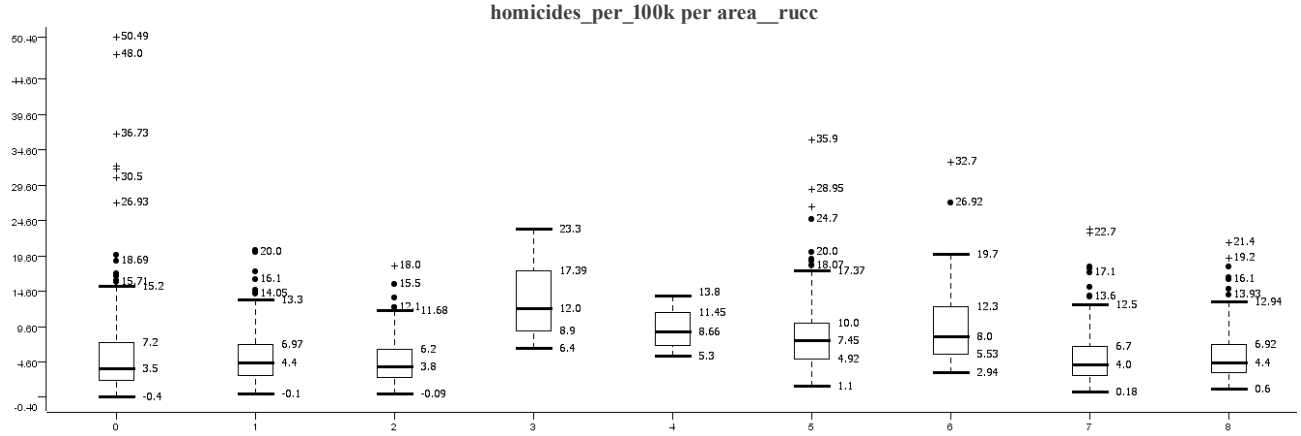
- There is some difference between different types of micropolitan *urban_influence*.
- There is less wide range of *heart_disease_mortality* rate for large and small *urban_influence* than for metropolitan and noncore, and the median of *heart_disease_mortality* is lower for large and small *urban_influence*.
- There is less wide range of *heart_disease_mortality* rate for Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area in *area__rucc* but the median is similar for all.
- *Economic_typology* in Recreation type has the smallest range of *heart_disease_mortality* rate and the smallest median.
- Some hypothetic outliers are clearly visible especially in Micropolitan adjacent to a small metro area on *urban_influence*.

There could be minimal correlations from *heart_disease_mortality* rate and some of the categorical values, and in this problem a conversion to indicator values (binary features for each nominal value) can be done to experiment the best combination that suites the performance of the machine-learning algorithm.

The outliers found in these analysis reminds that, in all the dataset, if treated more efficiently can obtain better ML results. In the experiments: clipping all numeric data at 90% shown some good performance on predictions.

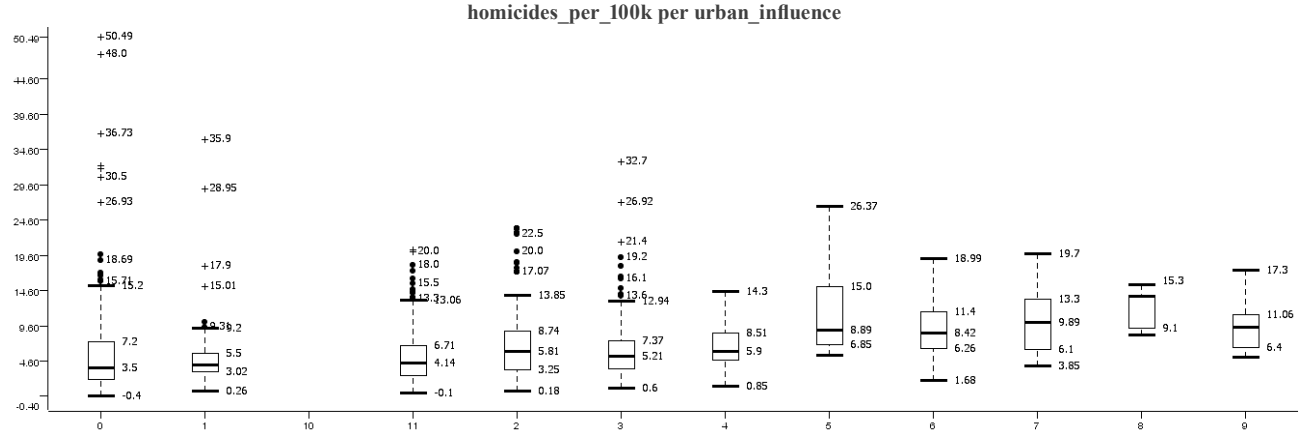
Some relationship between categorical feature values and columns with the most missing data values can be analysed with the purpose to obtain a better prediction of the missing data. The following boxplots show the more interesting categorical columns compared with *health__homicides_per_100k* rate and *health__pct_excessive_drinking* percentage:

homicides_per_100k per area__rucc



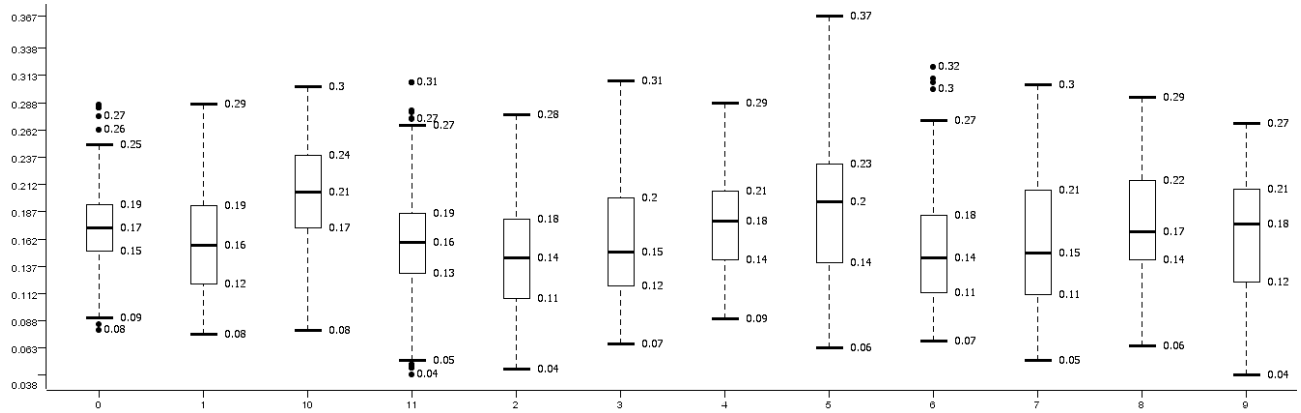
Nominal values (area__rucc): 0=Metro - Counties in metro areas of 1 million population or more, **1=Metro** - Counties in metro areas of 250,000 to 1 million population, **2=Metro** - Counties in metro areas of fewer than 250,000 population, **3=Nonmetro** - Completely rural or less than 2,500 urban population, adjacent to a metro area, **4=Nonmetro** - Completely rural or less than 2,500 urban population, not adjacent to a metro area, **5=Nonmetro** - Urban population of 2,500 to 19,999, adjacent to a metro area, **6=Nonmetro** - Urban population of 2,500 to 19,999, not adjacent to a metro area, **7=Nonmetro** - Urban population of 20,000 or more, adjacent to a metro area, **8=Nonmetro** - Urban population of 20,000 or more, not adjacent to a metro area.

homicides_per_100k per urban_influence



Nominal values (urban_influence): 0=Large-in a metro area with at least 1 million residents or more, **1=Micropolitan** adjacent to a large metro area, **10=Micropolitan** adjacent to a small metro area, **11=Micropolitan** not adjacent to a metro area, **2=Noncore** adjacent to a large metro area, **3=Noncore** adjacent to a small metro and does not contain a town of at least 2,500 residents, **4=Noncore** adjacent to a small metro with town of at least 2,500 residents, **5=Noncore** adjacent to micro area and contains a town of 2,500-19,999 residents, **6=Noncore** adjacent to micro area and does not contain a town of at least 2,500 residents, **7=Noncore** not adjacent to a metro/micro area and contains a town of 2,500 or more residents, **8=Noncore** not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents, **9=Small-**in a metro area with fewer than 1 million residents.

pct_excessive_drinking per urban_influence



Nominal values (urban_influence): 0=Large-in a metro area with at least 1 million residents or more, **1=Micropolitan** adjacent to a large metro area, **10=Micropolitan** adjacent to a small metro area, **11=Micropolitan** not adjacent to a metro area, **2=Noncore** adjacent to a large metro area, **3=Noncore** adjacent to a small metro and does not contain a town of at least 2,500 residents, **4=Noncore** adjacent to a small metro with town of at least 2,500 residents, **5=Noncore** adjacent to micro area and contains a town of 2,500-19,999 residents, **6=Noncore** adjacent to micro area and does not contain a

town of at least 2,500 residents, **7=Noncore** not adjacent to a metro/micro area and contains a town of 2,500 or more residents, **8=Noncore** not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents, **9=Small-**in a metro area with fewer than 1 million residents.

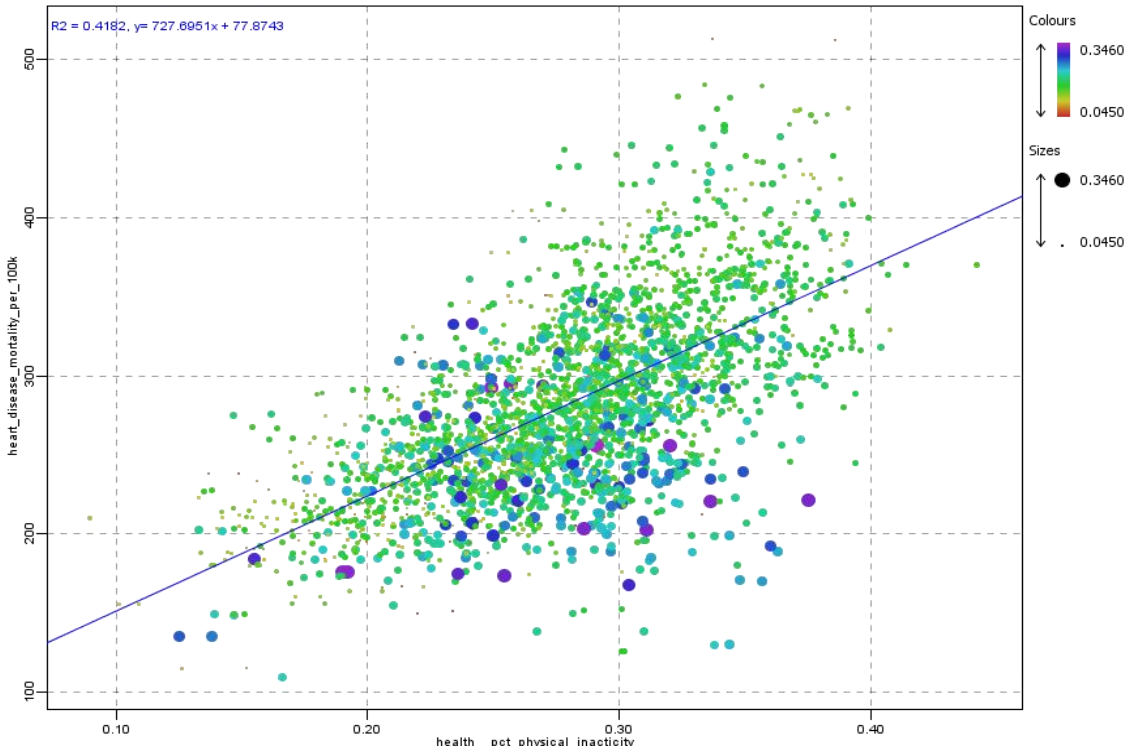
To obtain better result when managing and predicting missing values, the apparent relationships with these categorical data, suggests there is need to take into account the following:

- The difference in median and range between single or groups of nominal values for *homicides_per_100k* and *pct_excessive_drinking*.
- The need to select the best nominal values to contribute to increase performance when predicting and managing missing values.

Multi-faceted Relationships

Hidden and apparent relationships between features will help ML model to predict the label (the target in our problem: *heart_disease_mortality_per_100k*). Relationships are often complex or hidden, and may only become apparent considering multiple features in combination with one another. We need multi-faceted plots to help identify these relationships, with some revealed and other may need more investigations.

The plot bellow shows some interesting aspects of *pct_aged_65_years_and_older*. Although the correlation with *heart_disease_mortality* is relatively low (+0.20 corr. Measure), when we plot *pct_aged_65_years_and_older* (as colour and size) into a scatter plot of *heart_disease_mortality* and *health__pct_physical_inacticity*, it appears that the more higher the percentage of *aged_65_years_and_older* is more plotted on the bottom part. Following a different trend between others that are more linearly distributed. Separating it into two parts can help increase ML predictive performance.

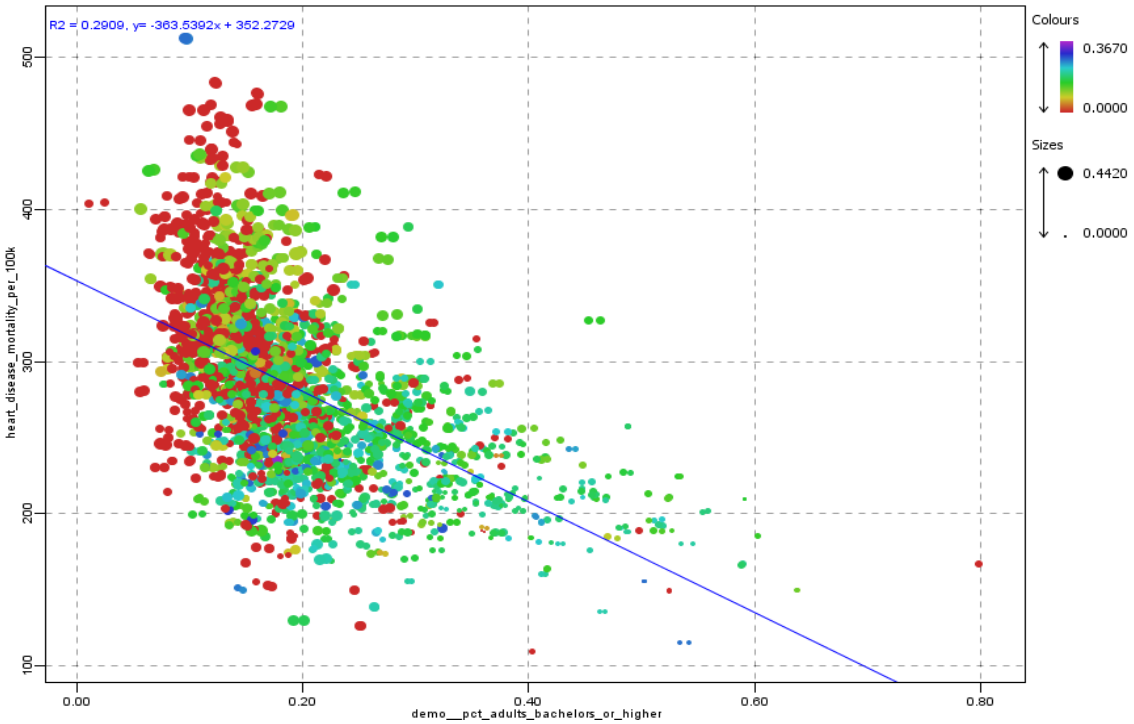


Colour and Size values (legend)

Colours from lowest value (red) to highest value (blue/purple) = **pct_aged_65_years_and_older**

Size from lowest value (small) to highest value (big) = **pct_aged_65_years_and_older**

The following multi-plot is with *heart_disease_mortality, pct_adults_bachelors_or_higher, health__pct_excessive_drinking,* and *pct_physical_inacticity.*

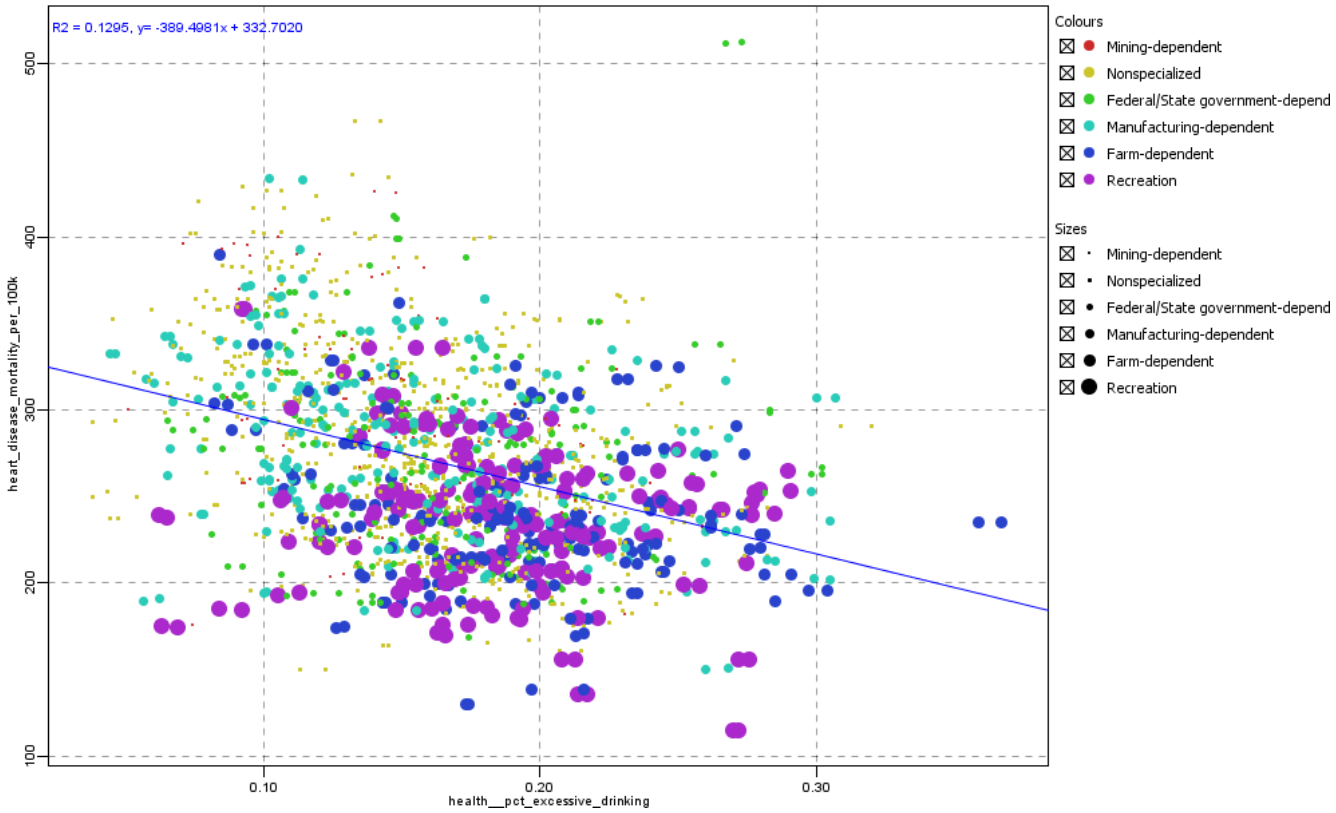


Colour and Size values (legend)
Colours from lowest value (red) to highest value (blue/purple) = **pct_excessive_drinking**
Size from lowest value (small) to highest value (big) = **pct_physical_inacticity**

We can see interesting things in the plot above like percentage of excessive drinking where the warm colours are more on the left side of the plot where we have less percentage of bachelor education and mild and good drinkers tend to ne on the right side where percentage *adults_bachelors_or_higher* is greater. *Pct_physical_inacticity* is bigger in size on the plot when it is greater and smaller the other side, in this plot we can see that it is getting smaller when *pct_ adults_bachelors_or_higher* is bigger.

The plot above maybe can also explain why it seems like more drinking percentage decrease heart disease mortality. However, it needs more investigations.

Below we can see in the plot that some type of economic_typologies tend to have lower rate of heart disease mortality. For example, Recreation and Farm-dependent are more on the bottom part of the plot.



In the experiments, all considerations about relationships hidden, apparent and real, are managed and manipulated to obtain better predicting results. Some techniques used in this problem:

- Clipping *pct_aged_65_years_and_older* from a value and above, consisting in making all above values the same as the clipped one (after some experiments better results obtained with the value of 0.22)
- Selecting features starting with the redundant ones correlated to each other. For example, *pct_civilian_labor* has nearly the same correlation with heart disease mortality as *pct_adults_bachelors_or_higher* and both are highly correlated to each other. After doing some experiments, the use of Permutation Feature Importance scoring or similar technique helps to select other features and score the ones that have been unselected before to have or not confirm.
- Converting categorical features into to indicator values (binary features for each unique values) increase prediction performance when correctly selected. Experiments and Permutation Feature Importance scoring or similar automated technique helps to get the best set of selected nominal values.