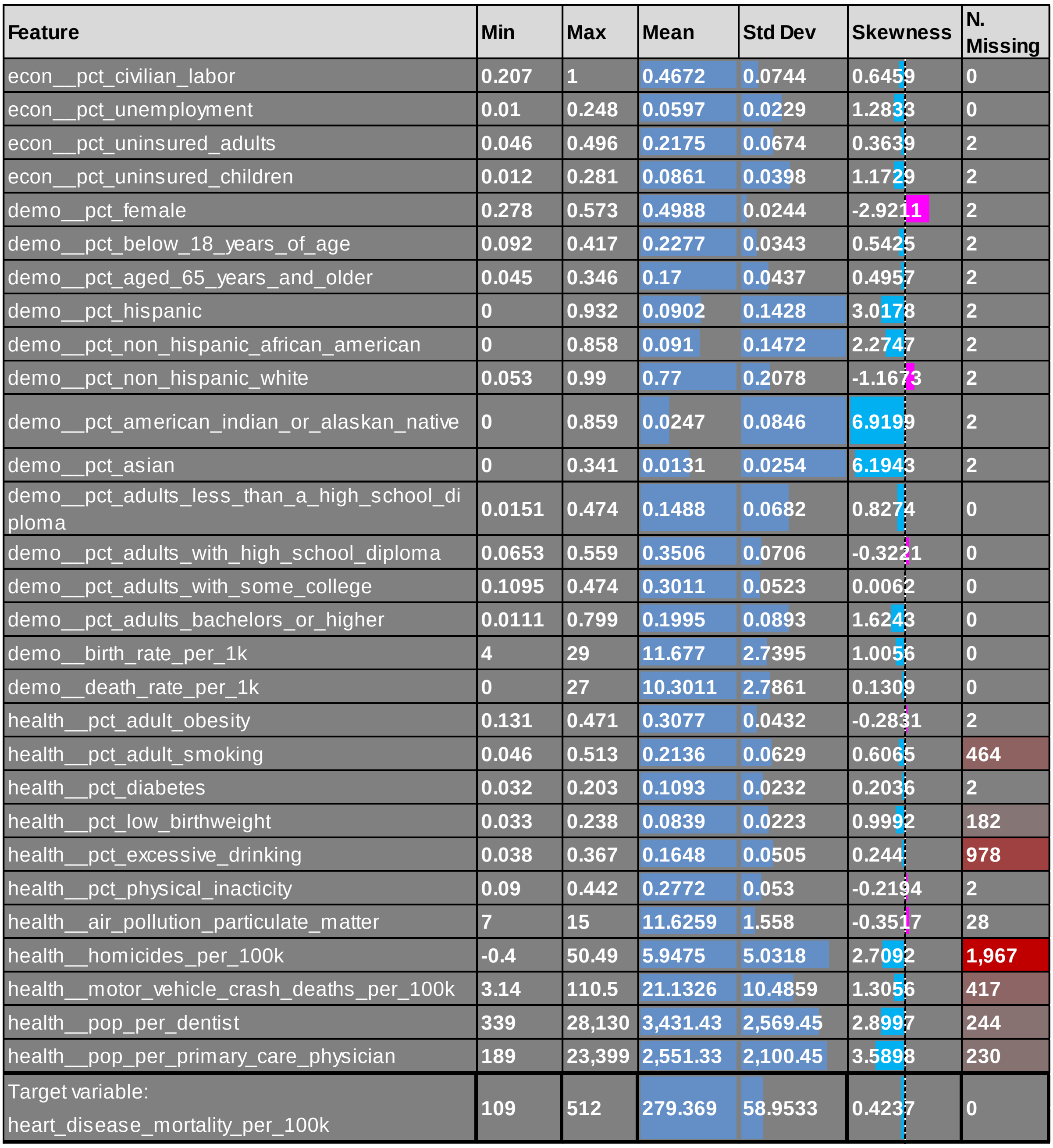
# 2. Data Exploration

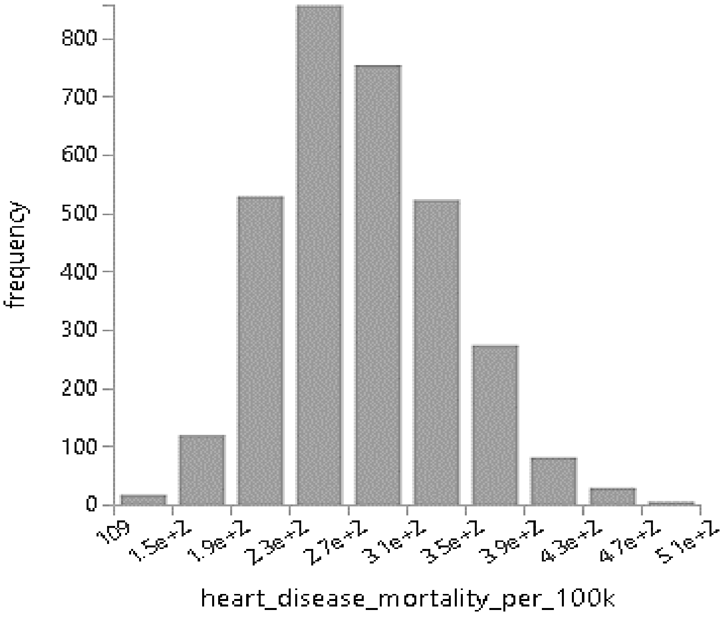
Predicting Heart Disease Mortality with Machine Learning Techniques
Lorenzo Negri, July 2018

### Numeric Feature Statistics

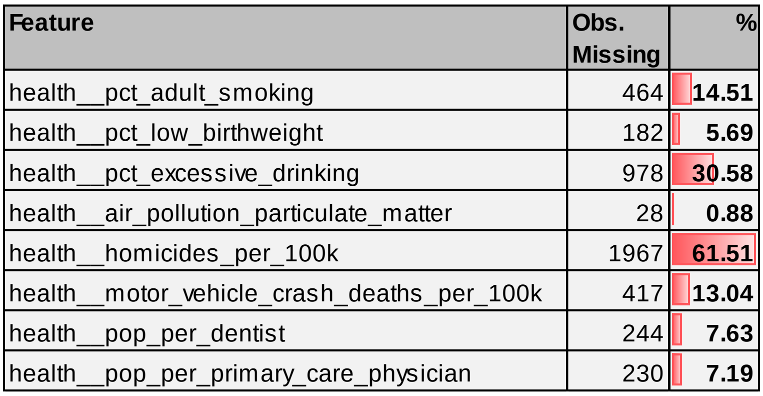
Summary statistics for minimum, maximum, mean, standard deviation, skewness and number of missing values for numeric columns, and the results calculation taken from all the observations are visible in table:



| Feature | Min | Max | Mean | Std Dev | Skewness | N. Missing |
|---|---|---|---|---|---|---|
| econ__pct_civilian_labor | 0.207 | 1 | 0.4672 | 0.0744 | 0.6459 | 0 |
| econ__pct_unemployment | 0.01 | 0.248 | 0.0597 | 0.0229 | 1.2833 | 0 |
| econ__pct_uninsured_adults | 0.046 | 0.496 | 0.2175 | 0.0674 | 0.3639 | 2 |
| econ__pct_uninsured_children | 0.012 | 0.281 | 0.0861 | 0.0398 | 1.1729 | 2 |
| demo__pct_female | 0.278 | 0.573 | 0.4988 | 0.0244 | -2.9211 | 2 |
| demo__pct_below_18_years_of_age | 0.092 | 0.417 | 0.2277 | 0.0343 | 0.5425 | 2 |
| demo__pct_aged_65_years_and_older | 0.045 | 0.346 | 0.17 | 0.0437 | 0.4957 | 2 |
| demo__pct_hispanic | 0 | 0.932 | 0.0902 | 0.1428 | 3.0178 | 2 |
| demo__pct_non_hispanic_african_american | 0 | 0.858 | 0.091 | 0.1472 | 2.2747 | 2 |
| demo__pct_non_hispanic_white | 0.053 | 0.99 | 0.77 | 0.2078 | -1.1673 | 2 |
| demo__pct_american_indian_or_alaskan_native | 0 | 0.859 | 0.0247 | 0.0846 | 6.9199 | 2 |
| demo__pct_asian | 0 | 0.341 | 0.0131 | 0.0254 | 6.1943 | 2 |
| demo__pct_adults_less_than_a_high_school_diploma | 0.0151 | 0.474 | 0.1488 | 0.0682 | 0.8274 | 0 |
| demo__pct_adults_with_high_school_diploma | 0.0653 | 0.559 | 0.3506 | 0.0706 | -0.3221 | 0 |
| demo__pct_adults_with_some_college | 0.1095 | 0.474 | 0.3011 | 0.0523 | 0.0062 | 0 |
| demo__pct_adults_bachelors_or_higher | 0.0111 | 0.799 | 0.1995 | 0.0893 | 1.6243 | 0 |
| demo__birth_rate_per_1k | 4 | 29 | 11.677 | 2.7395 | 1.0056 | 0 |
| demo__death_rate_per_1k | 0 | 27 | 10.3011 | 2.7861 | 0.1309 | 0 |
| health__pct_adult_obesity | 0.131 | 0.471 | 0.3077 | 0.0432 | -0.2831 | 2 |
| health__pct_adult_smoking | 0.046 | 0.513 | 0.2136 | 0.0629 | 0.6065 | 464 |
| health__pct_diabetes | 0.032 | 0.203 | 0.1093 | 0.0232 | 0.2036 | 2 |
| health__pct_low_birthweight | 0.033 | 0.238 | 0.0839 | 0.0223 | 0.9992 | 182 |
| health__pct_excessive_drinking | 0.038 | 0.367 | 0.1648 | 0.0505 | 0.244 | 978 |
| health__pct_physical_inacticity | 0.09 | 0.442 | 0.2772 | 0.053 | -0.2194 | 2 |
| health__air_pollution_particulate_matter | 7 | 15 | 11.6259 | 1.558 | -0.3517 | 28 |
| health__homicides_per_100k | -0.4 | 50.49 | 5.9475 | 5.0318 | 2.7092 | 1,967 |
| health__motor_vehicle_crash_deaths_per_100k | 3.14 | 110.5 | 21.1326 | 10.4859 | 1.3056 | 417 |
| health__pop_per_dentist | 339 | 28,130 | 3,431.43 | 2,569.45 | 2.8997 | 244 |
| health__pop_per_primary_care_physician | 189 | 23,399 | 2,551.33 | 2,100.45 | 3.5898 | 230 |
| Target variable:<br>heart_disease_mortality_per_100k | 109 | 512 | 279.369 | 58.9533 | 0.4237 | 0 |



The **target variable** of our interest for the training of the machine-learning algorithm is nearly normal distributed and the skewness indicates a slight right-skewed values. On the left, the frequency chart of the **target variable** column.



| Feature | Obs. Missing | % |
|---|---|---|
| health__pct_adult_smoking | 464 | 14.51 |
| health__pct_low_birthweight | 182 | 5.69 |
| health__pct_excessive_drinking | 978 | 30.58 |
| health__air_pollution_particulate_matter | 28 | 0.88 |
| health__homicides_per_100k | 1967 | 61.51 |
| health__motor_vehicle_crash_deaths_per_100k | 417 | 13.04 |
| health__pop_per_dentist | 244 | 7.63 |
| health__pop_per_primary_care_physician | 230 | 7.19 |

The **missing values** of columns with greater number appears in the table on the left, with also percentages. As we can see, *health_homicides_per_100k* has more than 60% of missing observations of the 3198 total. *Health_pct_excessive_drinking* is at second place with 30% of missing values.

## Categorical Features

The categorical features of the observations, includes:

- **area__rucc** – classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. There are three metro and six nonmetro categories.
- **area__urban_influence** – classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas.
- **econ__economic_typology** – classify all U.S. counties according to six mutually exclusive categories of economic dependence and six overlapping categories of policy-

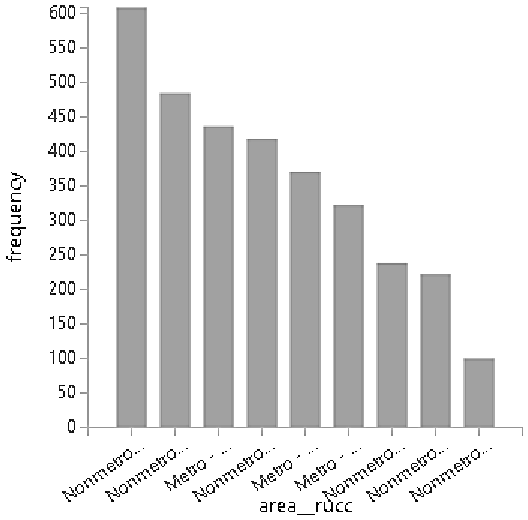
relevant themes. The economic dependence types include farming, mining, manufacturing, Federal/State government, recreation, and nonspecialized counties. The policy-relevant types include low education, low employment, persistent poverty, persistent child poverty, population loss, and retirement destination.

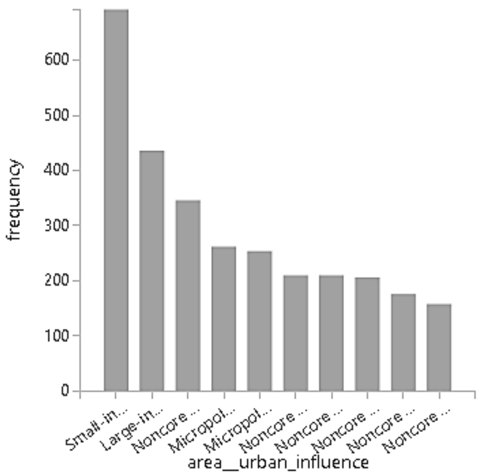
- **yr** – two year reference.

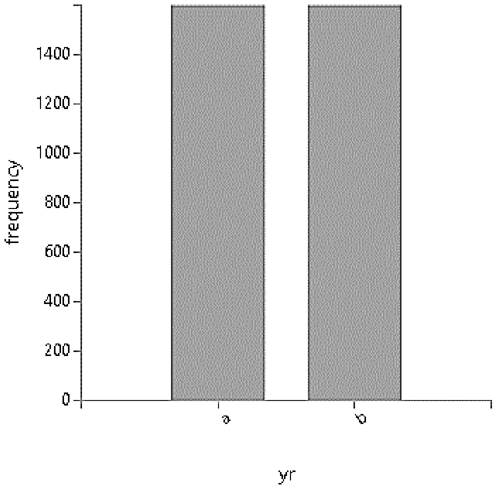
Bar charts visualization of these features, indicate the following relevant notes:

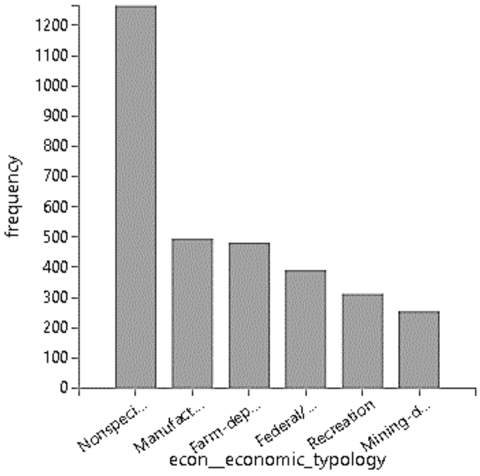
- The years of reference has the same frequency for both.
- Nonspecialized economic typology is 40% of all others in the same category.

It appears that "year" is not a categorical feature of some interest and that would not help getting better results for our problem.


frequency
Nonmetro...
Nonmetro...
Metro - ...
Nonmetro...
Metro - ...
Metro - ...
Nonmetro...
Nonmetro...
Nonmetro...
area__rucc


frequency
Small-in...
Large-in...
Noncore ...
Micropol...
Micropol...
Noncore ...
Noncore ...
Noncore ...
Noncore ...
Noncore ...
area__urban_influence


frequency
a
b
yr


frequency
Nonspeci...
Manufact...
Farm-dep...
Federal/...
Recreation
Mining-d...
econ__economic_typology