Impact of Demographic Factors on Cancer Mortality Rates

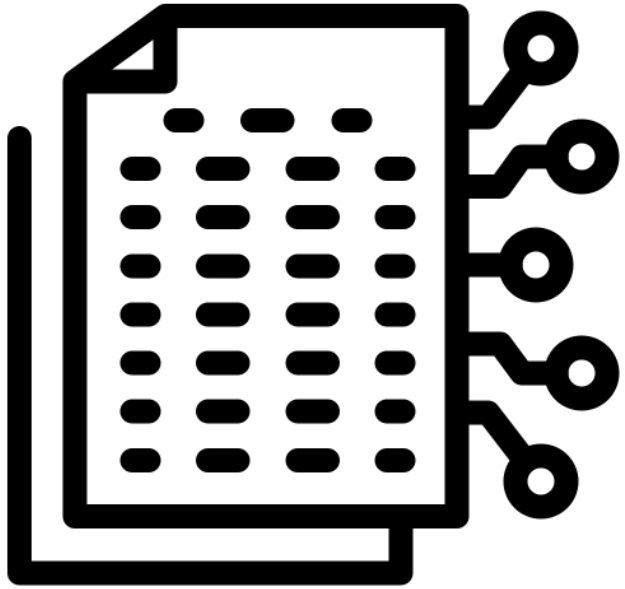
Szymon Abramczyk

Sonia Bogdańska

Table of Contents:

Data

- Table *avg_household_size*: Average household size in regions of the USA.
- Table *cancer_reg*: Data on cancer incidence and mortality in US counties.

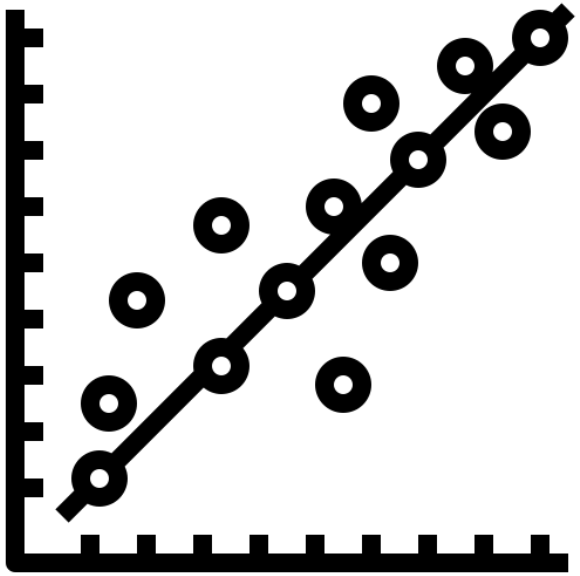


Example variables:

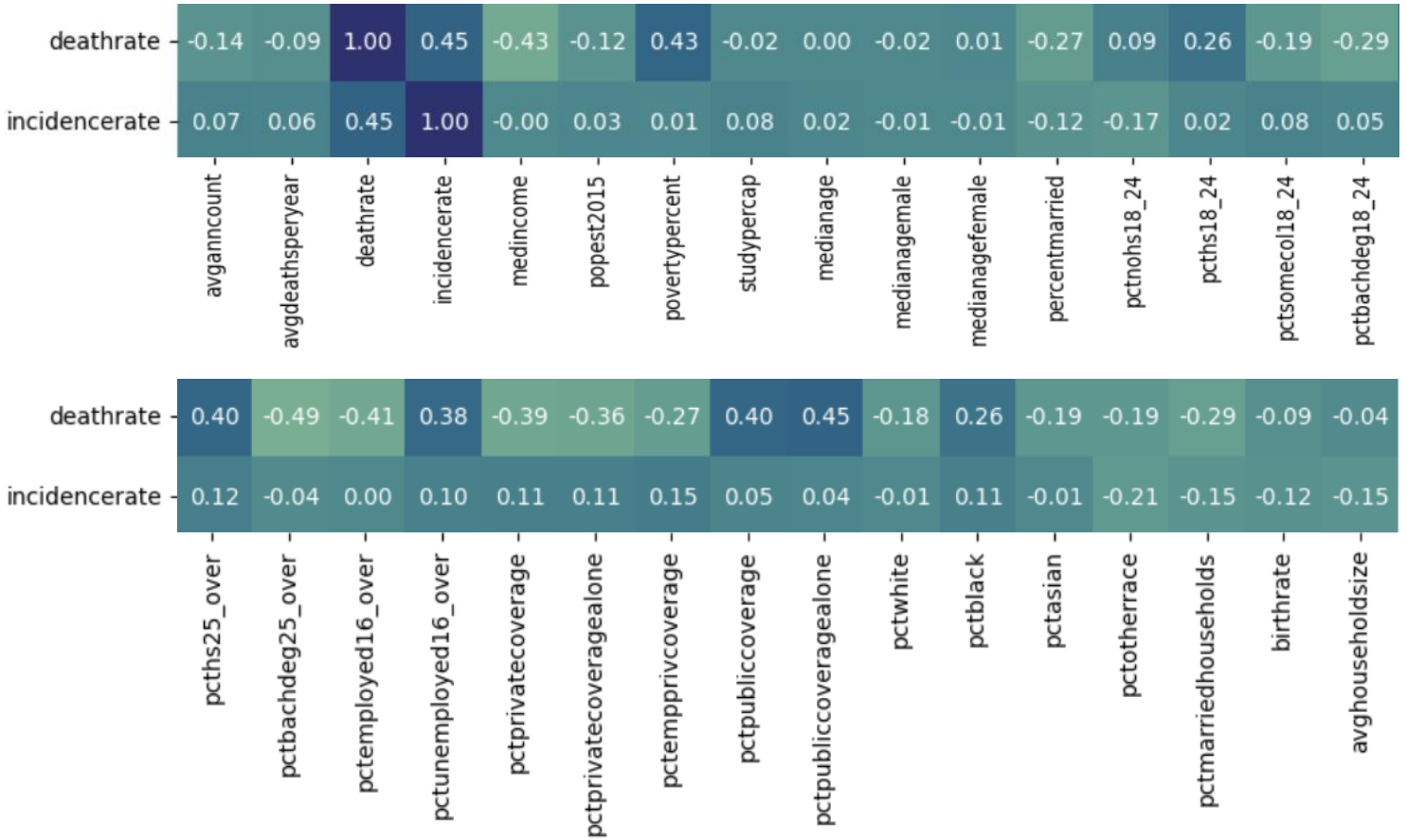
- Average number of cancer diagnoses per 100,000 residents in a county.
- Average number of cancer deaths per 100,000 residents in a county.
- Average household income in a county.
- Percentage of residents over 25 years old with at most a high school education.
- Percentage of residents over 25 years old with a bachelor's degree.
- Percentage of county residents who are married.

Highly correlated features

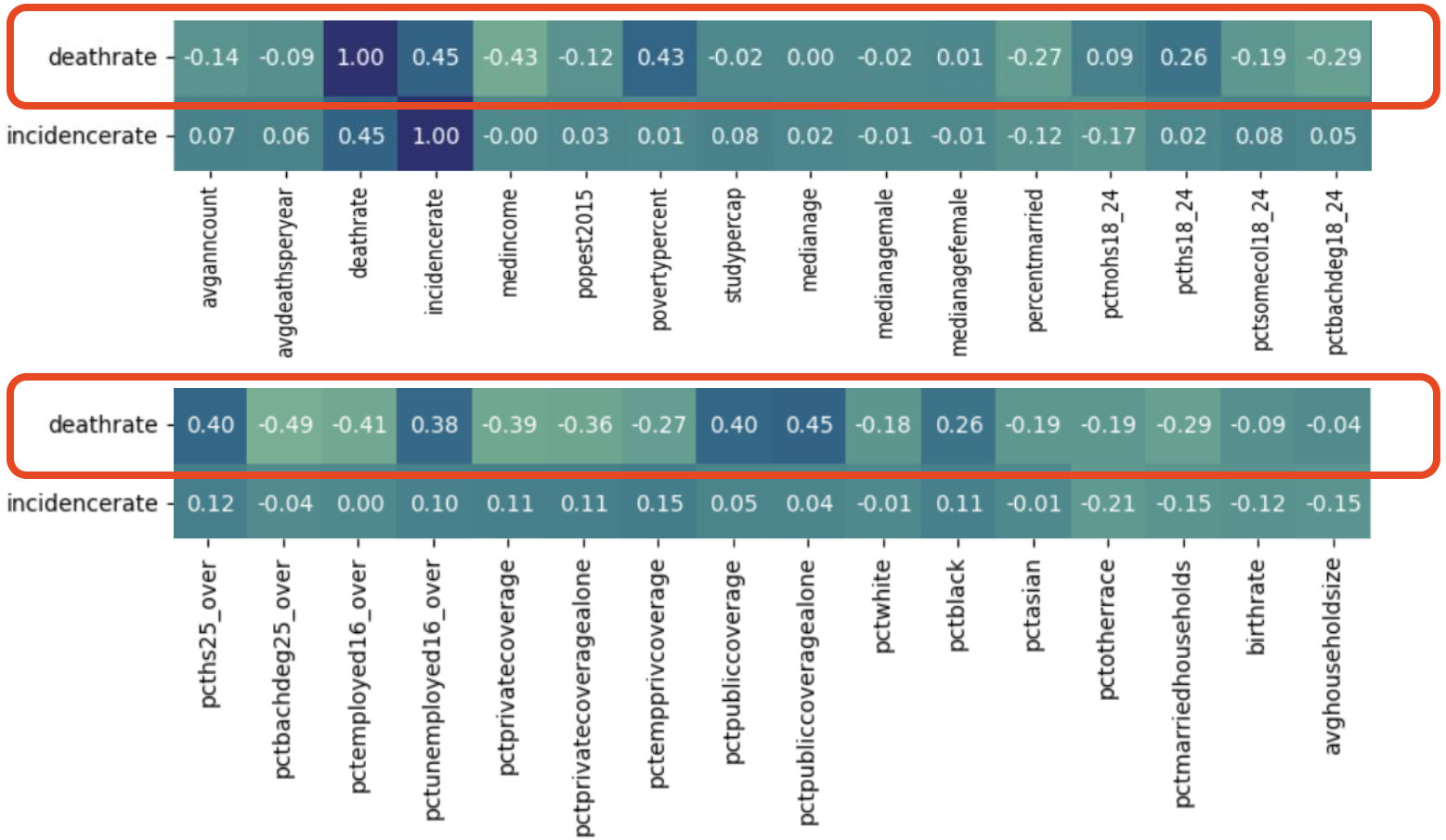
- Population size and average number of deaths **(0.98)**.
- Average number of reported cancer diagnoses and average number of deaths **(0.94)**.
- Percentage of people with only public insurance and percentage of the population living below the poverty line **(0.8)**.
- Average income and percentage of people with private health insurance **(0.75)**.



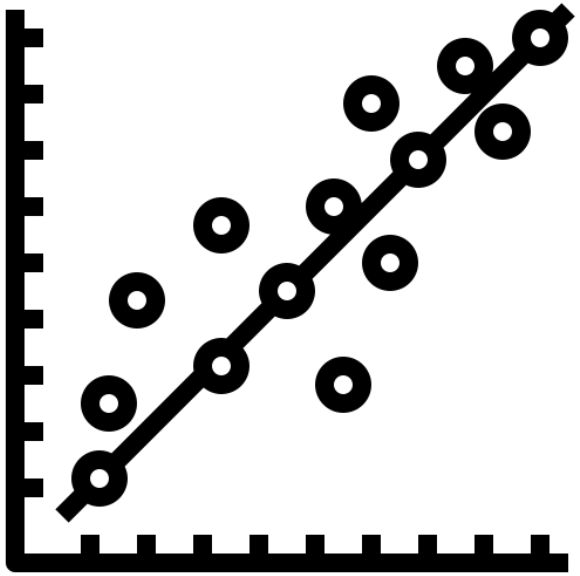
Selection of the dependent variable



Selection of the dependent variable

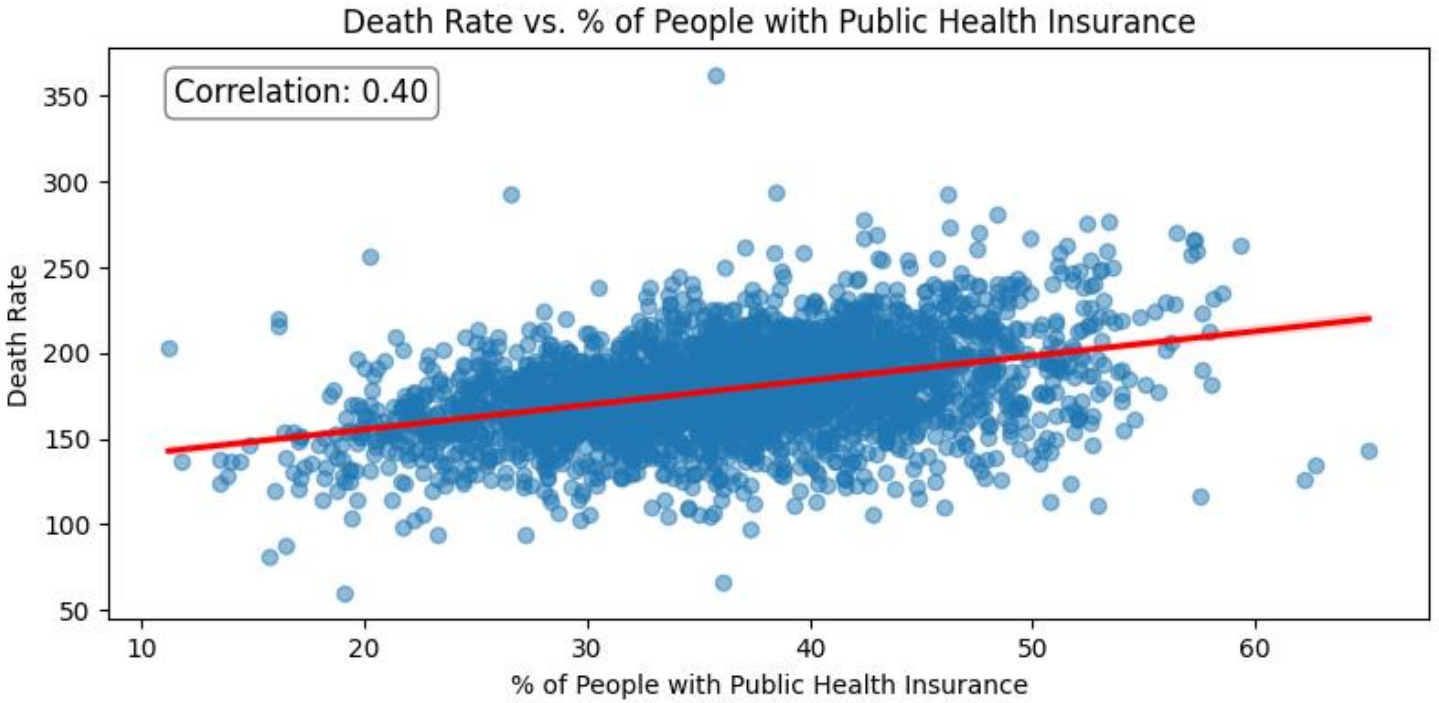


Features correlated with the dependent variable

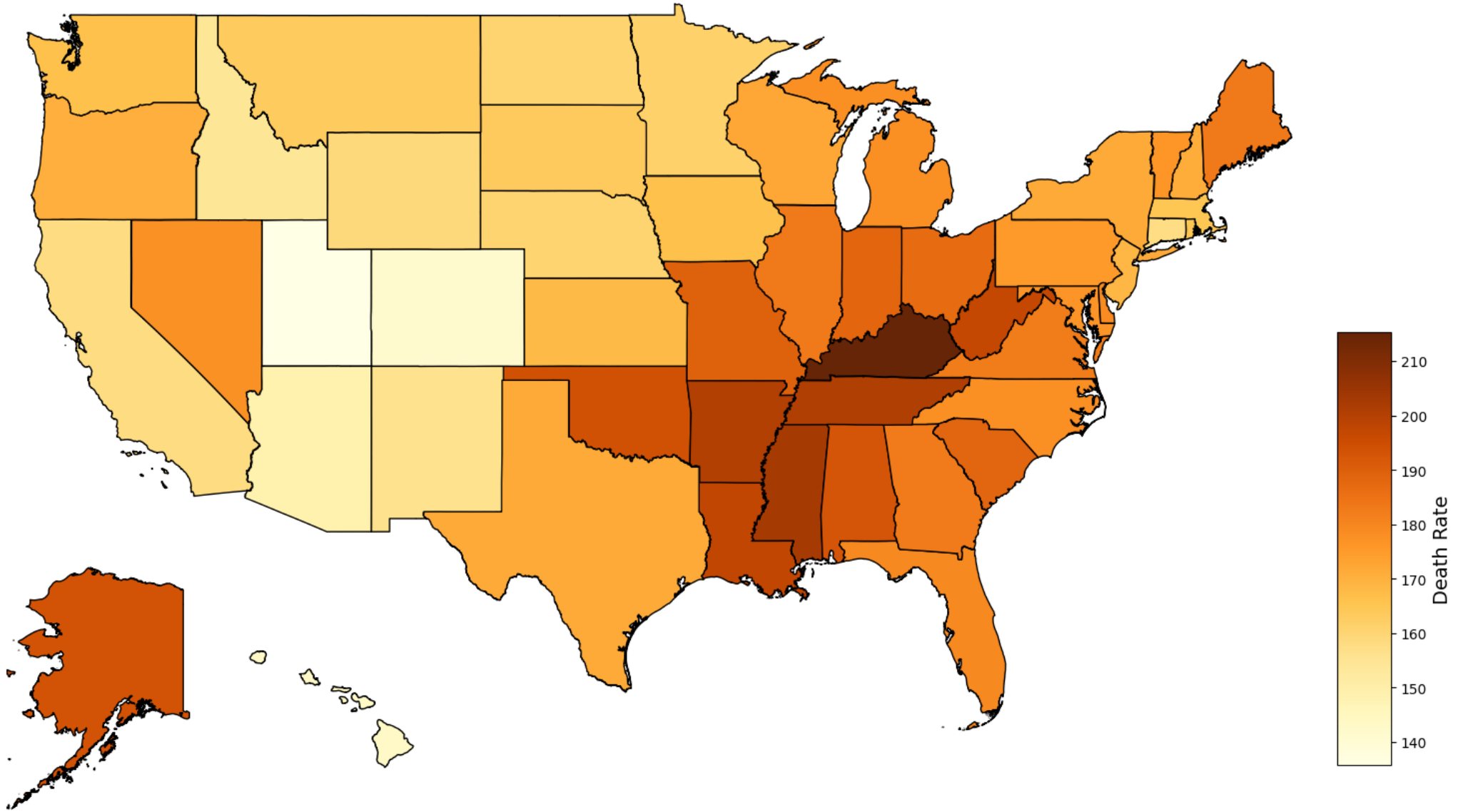


- Number of people with at least a bachelor's degree over 25 years old **(-0.49)**.
- Average number of cancer diagnoses per 100,000 residents **(0.45)**.
- Percentage of people with only public health insurance **(0.45)**.
- Percentage of the county population living below the poverty line **(0.43)**.
- Average household income in the county **(-0.43)**.
- Percentage of employed individuals over 16 years old **(-0.41)**.

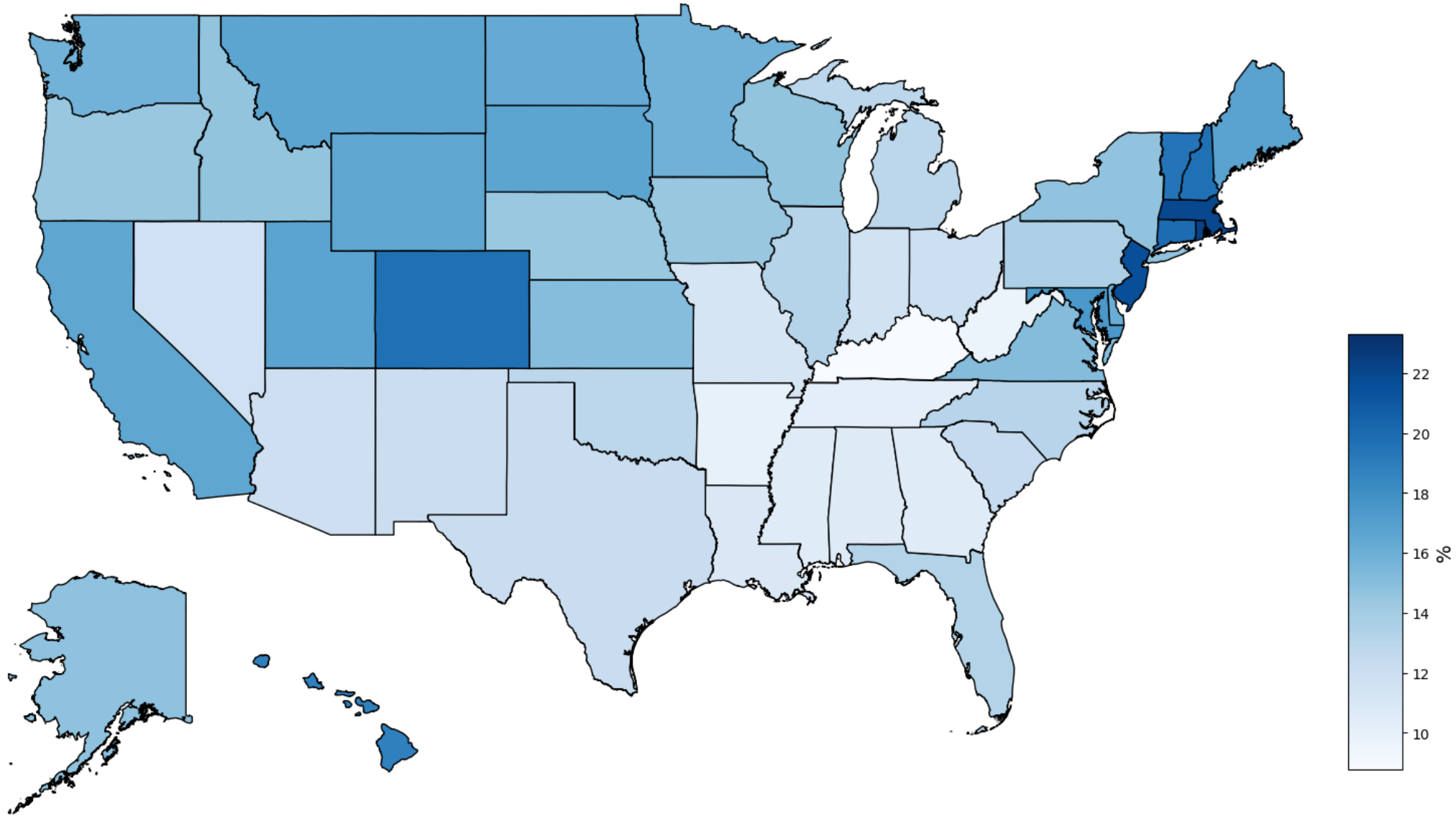
Example visualizations



Average Death Rate in the USA



Percentage of Residents Aged 25 and Over with a Bachelor's Degree



Conclusions from exploratory analysis

Regions with higher incomes have lower mortality rates.

Type of insurance (private vs. public) plays a significant role in mortality levels.

Higher education levels are correlated with lower mortality rates.

Research hypothesis

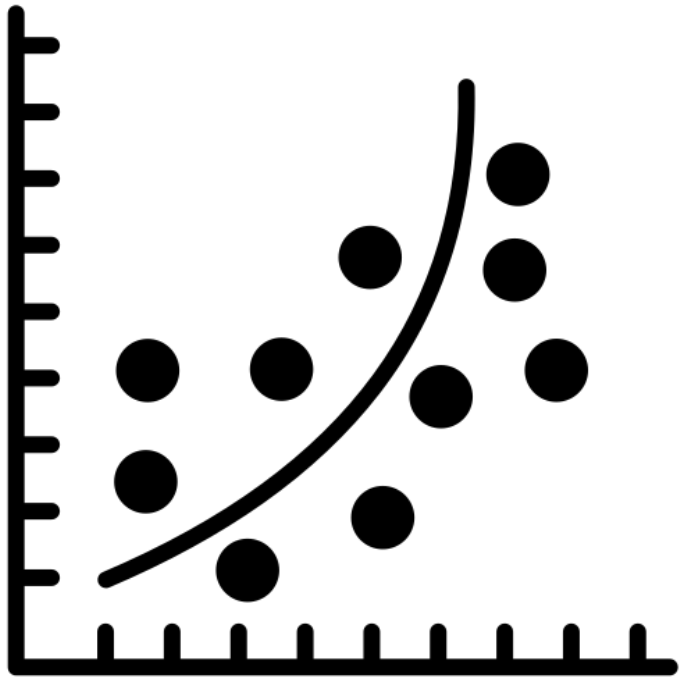
Demographic factors such as income, education level, and type of insurance have a significant impact on cancer mortality rates.

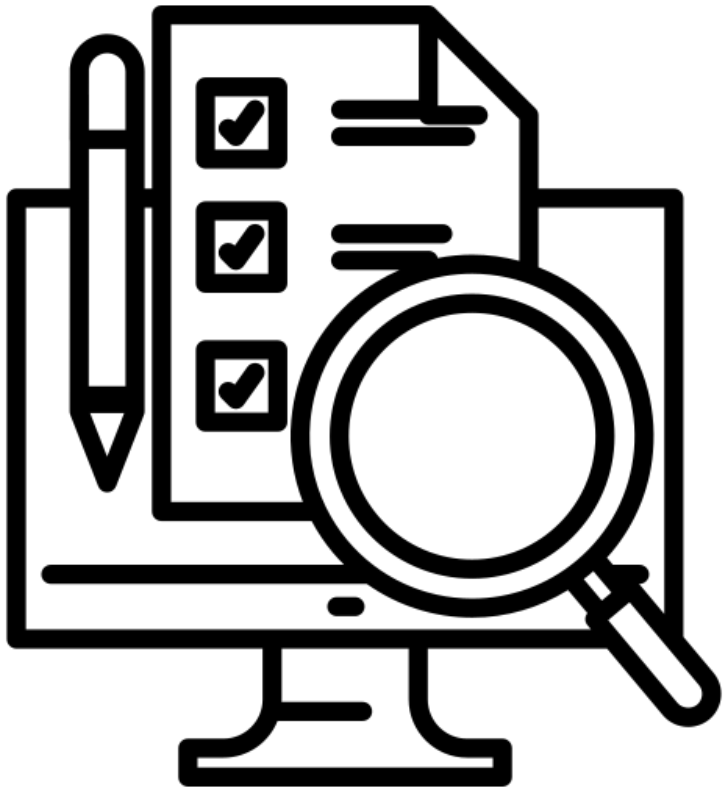
Development of the regression model

Goal: A predictive model to identify key factors influencing mortality rates.

Stages of work:

1. Data preparation
2. Feature selection
3. Cross-validation
4. Evaluation of models with the best results





Feature selection

4 sets of features

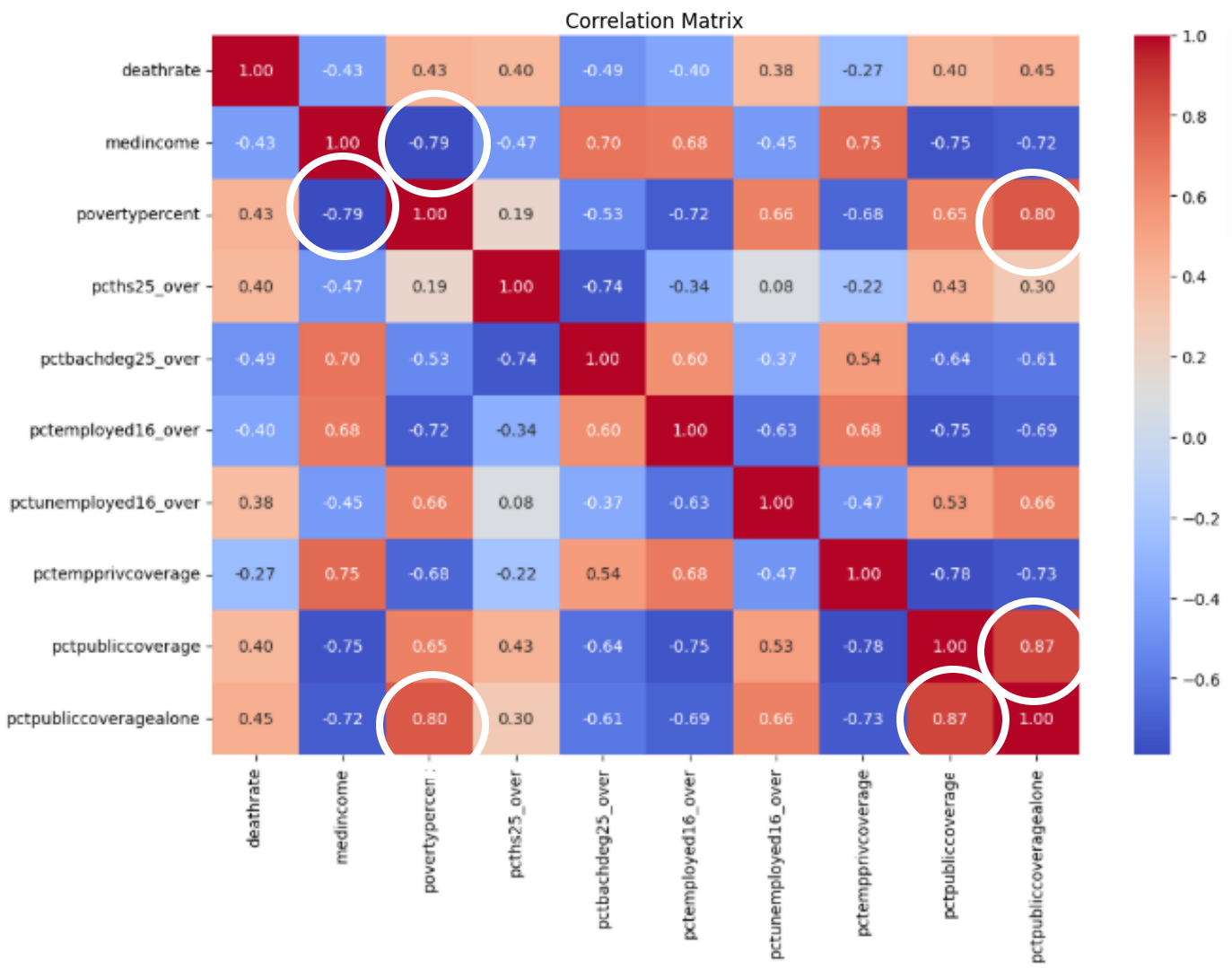
1. Selection based on the hypothesis
2. Set of all features
3. Greedy method
4. Selection using the SelectKBest method

Features from the hypothesis

Multicollinearity

⇩

Removal of redundant data

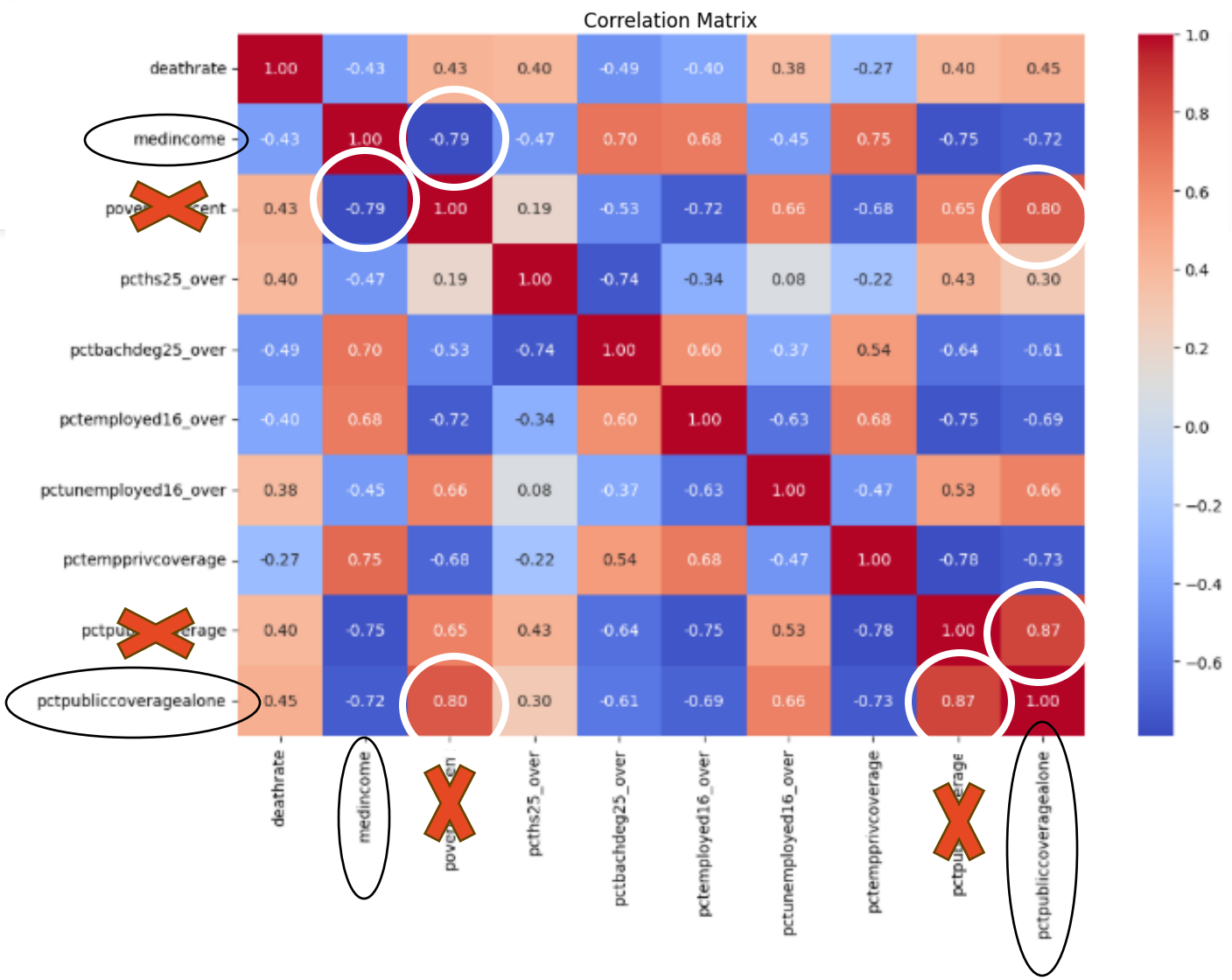


Features from the hypothesis

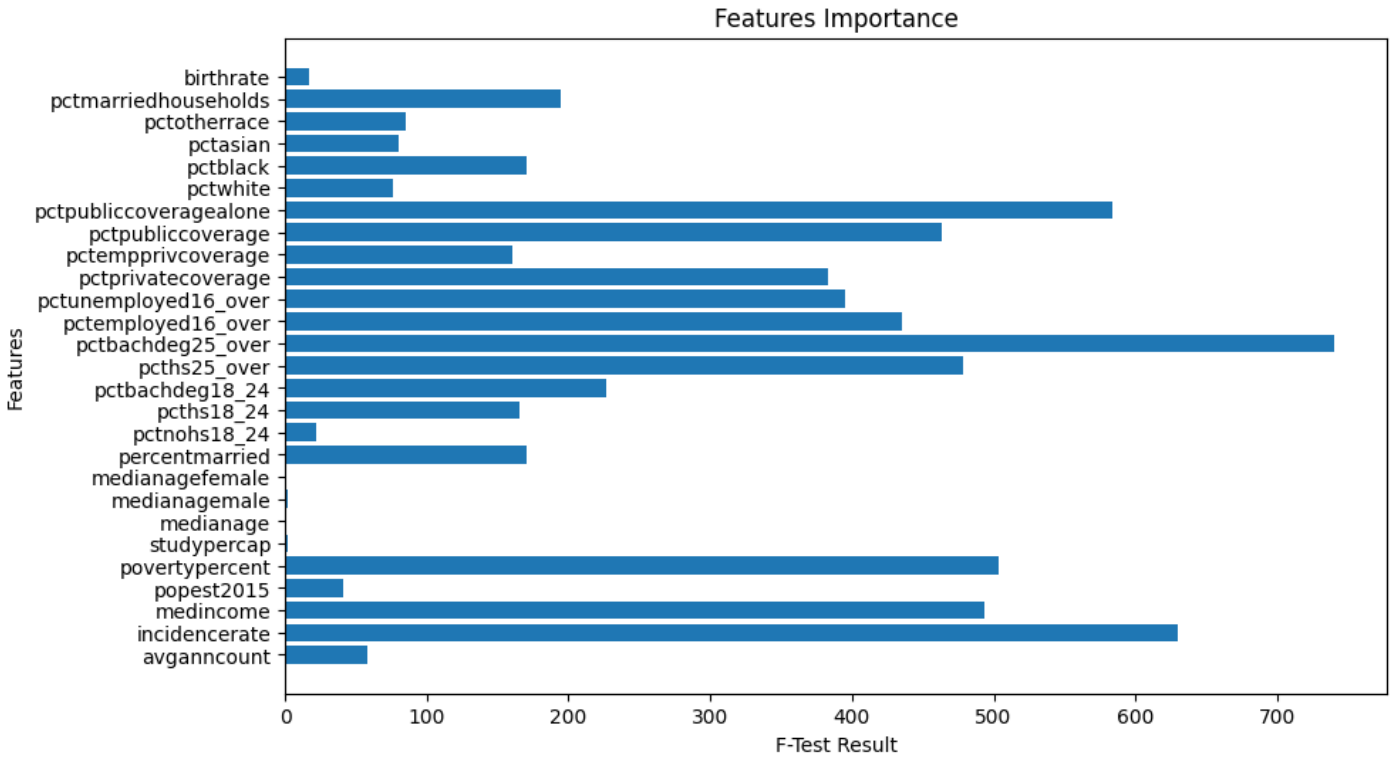
Multicollinearity

⇩

Removal of redundant data



Feature selection – other methods



Example features selected by SelectKBest

Modeling – used algorithms

Polynomial regression (2nd degree)	**Linear regression**
All features	Features from the hypothesis
Features selected by SelectKBest	Features selected greedily
Ridge regularization	Elastic Net regularization
Lasso regularization	

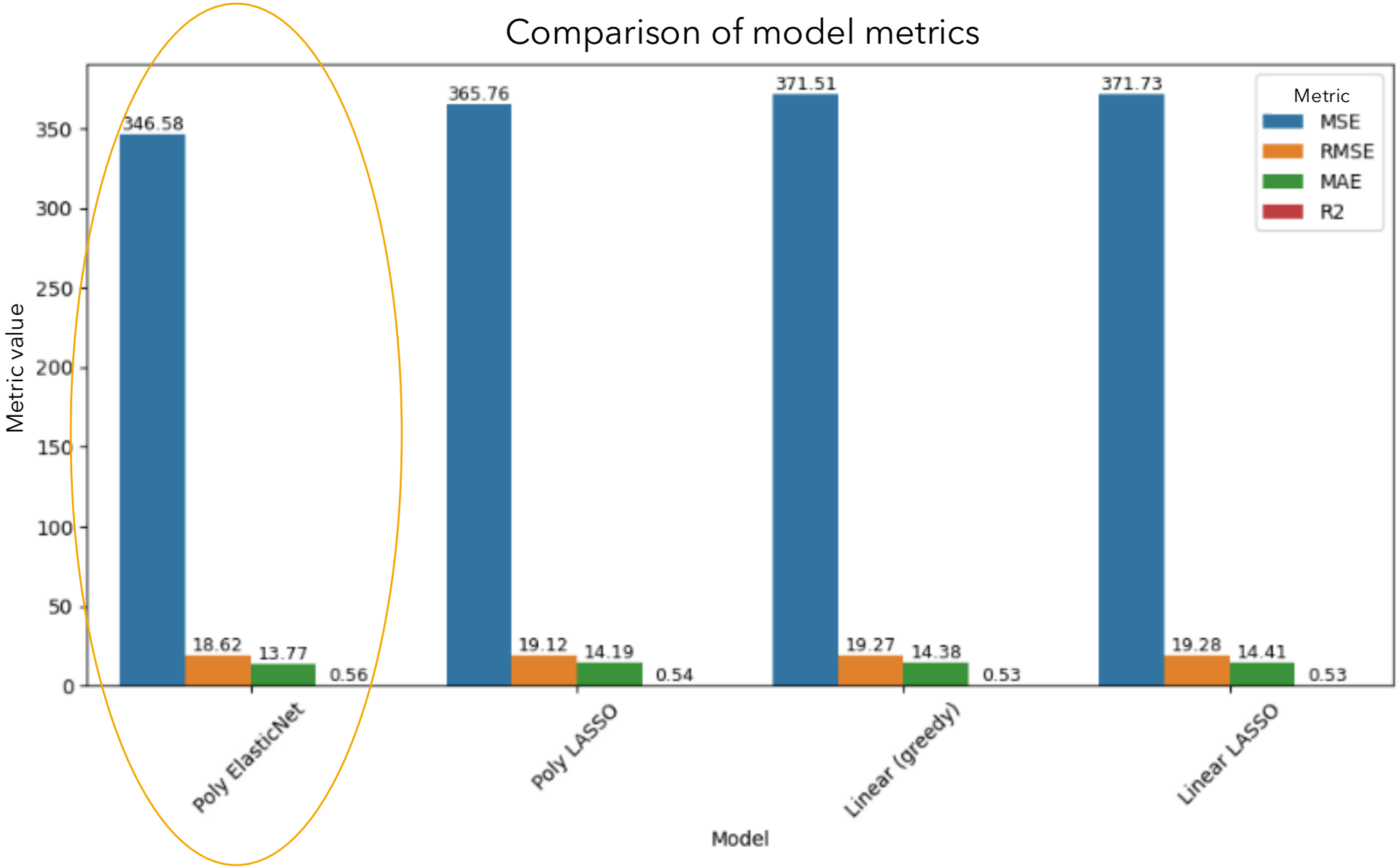
Comparison of results

	Feature_Set	Average CV Score
11	features_poly_EN	0.511380
9	features_poly_lasso	0.511253
10	features_poly_ridge	0.502823
2	features_greedy	0.502594
4	features_lasso	0.497069
6	features_EN	0.497019
5	features_ridge	0.496921
3	features_kbest	0.493294
8	features_poly_kbest	0.403964
7	features_poly_hypothesis	0.352660
1	features_hypothesis_ridge	0.324670
0	features_hypothesis	0.324575

Comparison of results

	Feature_Set	Average CV Score
11	features_poly_EN	0.511380
9	features_poly_lasso	0.511253
10	features_poly_ridge	0.502823
2	features_greedy	0.502594
4	features_lasso	0.497069
6	features_EN	0.497019
5	features_ridge	0.496921
3	features_kbest	0.493294
8	features_poly_kbest	0.403964
7	features_poly_hypothesis	0.352660
1	features_hypothesis_ridge	0.324670
0	features_hypothesis	0.324575

Comparison of model metrics



Interpretation of models – features with the greatest impact on predictions:

- Average number of cancer diagnoses per 100,000 residents in a county
- Percentage of residents over 25 years old with a high school education
- Percentage of the county population living below the poverty line

- Percentage of residents over 25 years old with a bachelor's degree
- Percentage of households in the county where residents are married

Sources

- https://commons.wikimedia.org/wiki/File:Data_icon.svg
- https://www.flaticon.com/free-icon/linear-regression_2103601
- https://icon-library.com/icon/dollar-icon-png-0.html.html>Dollar Icon Png # 9041
- https://www.pngwing.com/en/free-png-nnypv
- https://www.flaticon.com/freeicon/taskmanagement_13072224?term=checklist+computer&page=1&position=34&origin=search&related_id=13072224
- https://www.flaticon.com/freeicon/analysis_14639578?term=regression&page=1&position=5&origin=search&related_id=14639578
- https://www.flaticon.com/free-icon/trend_3121571?related_id=3121571&origin=pack
- https://www.flaticon.com/free-icon/trend_3121574?related_id=3121574&origin=pack