CASE STUDY FOR
GHAMUT CORPORATION

HEALTH OUTCOMES ANALYSIS IN USA

March 2023

Maslenkova Svetlana

OVERVIEW

1. Data Sources
2. Data Analysis & Insights
3. Target variable
4. Predictive features
5. Predicting model
6. Results
7. Feature importance analysis

DATA SOURCES

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

Health related data
Social Vulnerability
Mortality

United States® Census Bureau

Demographics

USDA Economic Research Service
U.S. DEPARTMENT OF AGRICULTURE

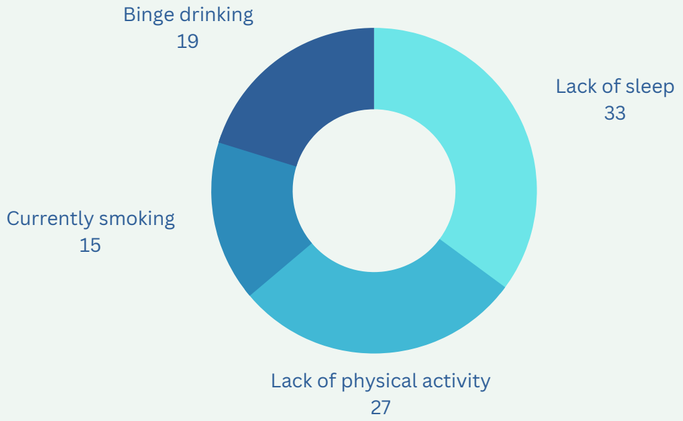
Poverty, Unemployment, Education

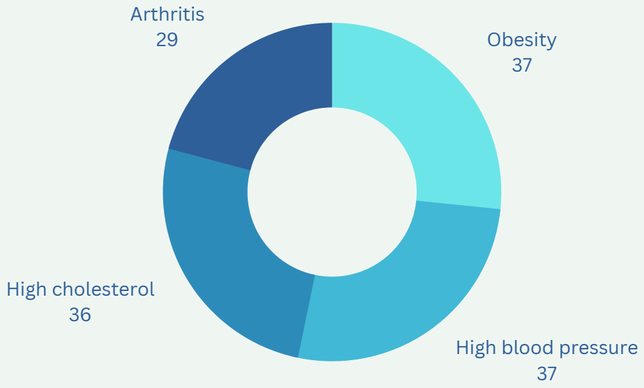
Current Results
weather and science facts

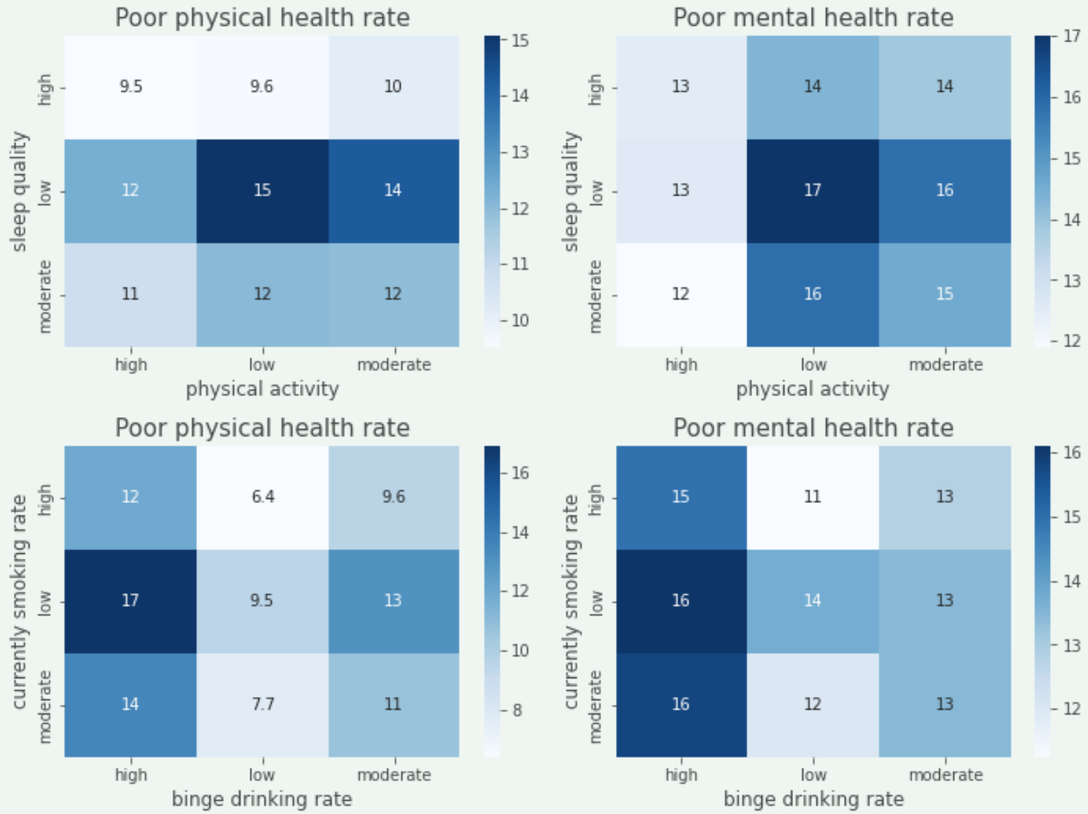
Amount of sunshine

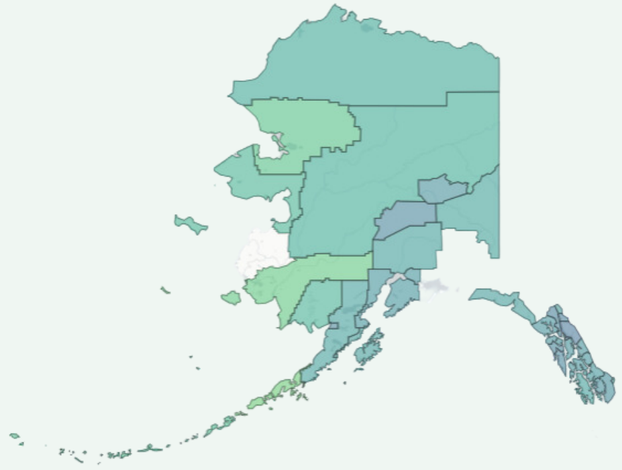
HEALTH RISK BEHAVIOURS & PREVENTIVE MEASURES

Ehe effect of health risk behaviours on health rates

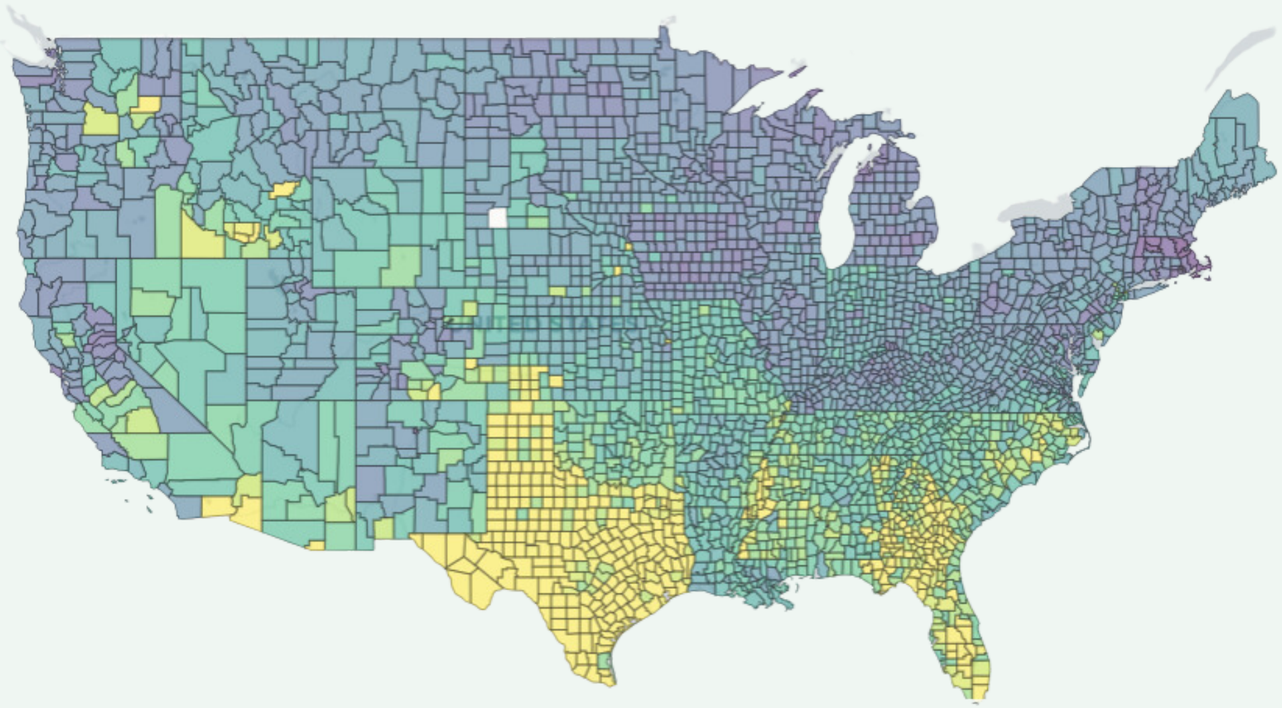




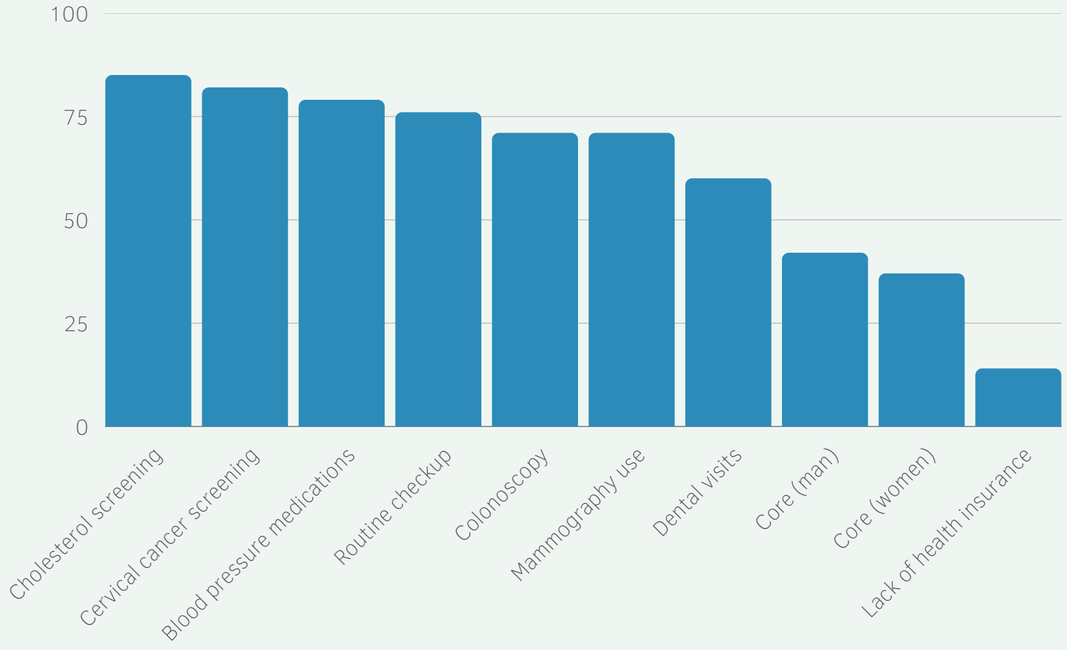




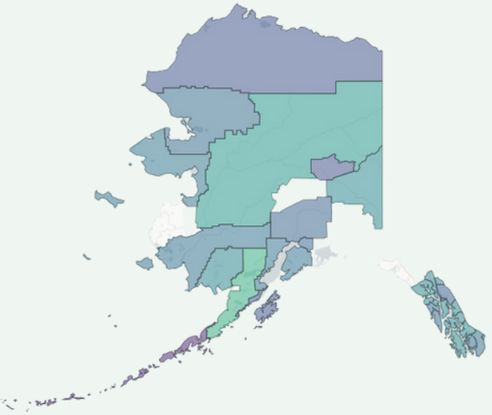
Current lack of health insurance



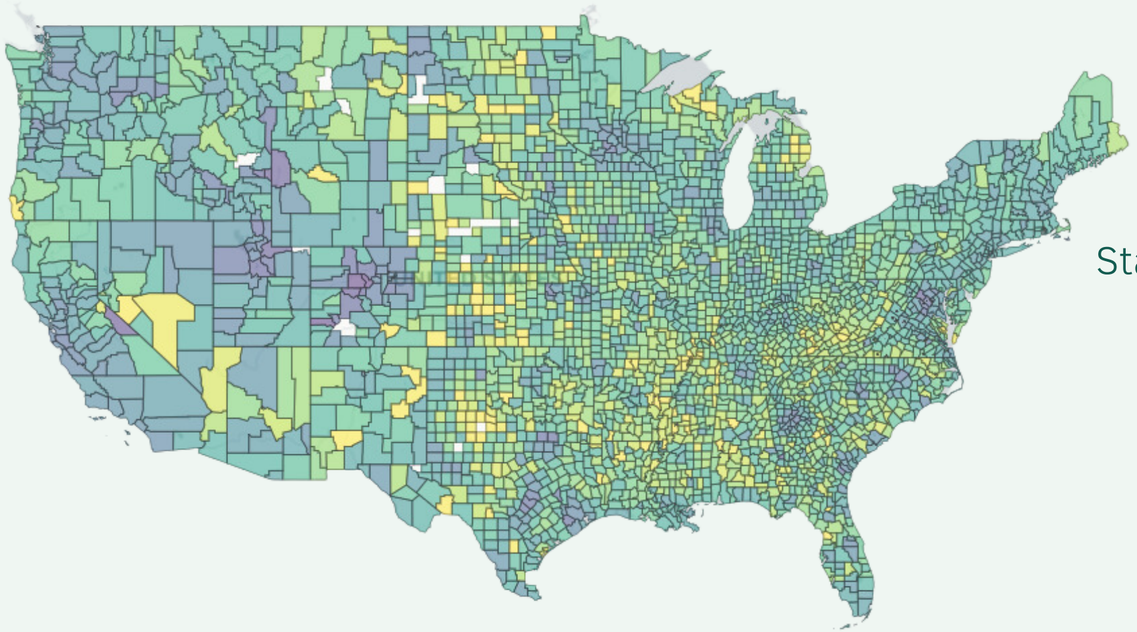
Preventive measures crude prevalence rates (median)

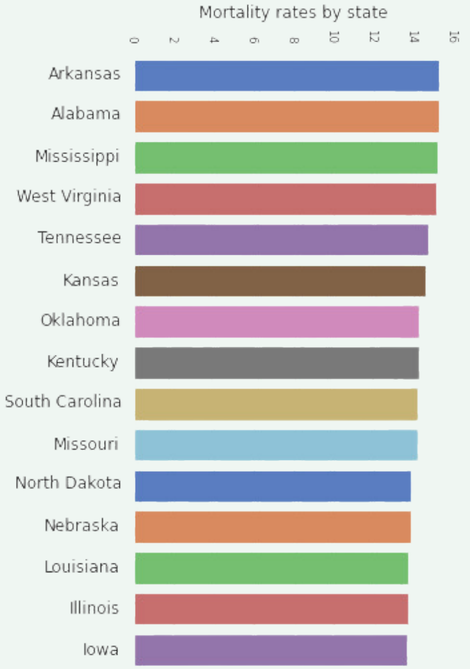


MORTALITY



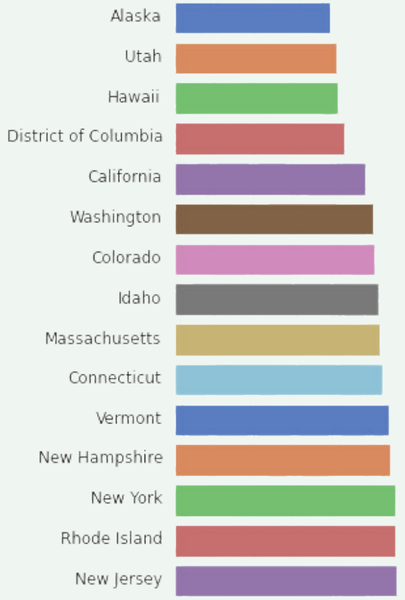
Majority of US counties have mortality rates between 11 and 15 per 1000 population





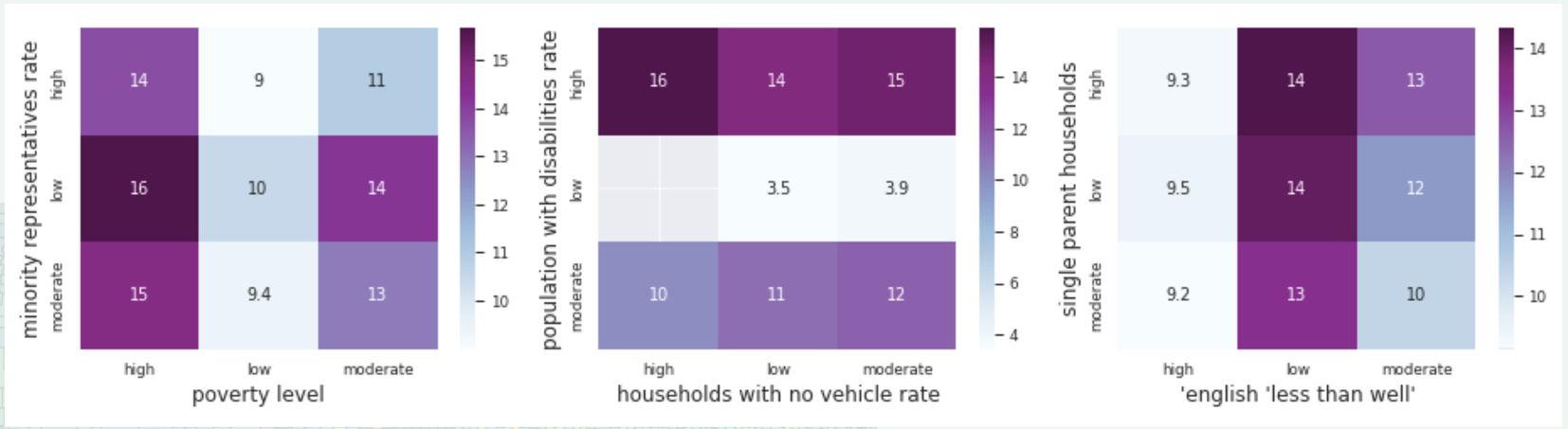
States with the highest mortality rates

States with the lowest mortality rates



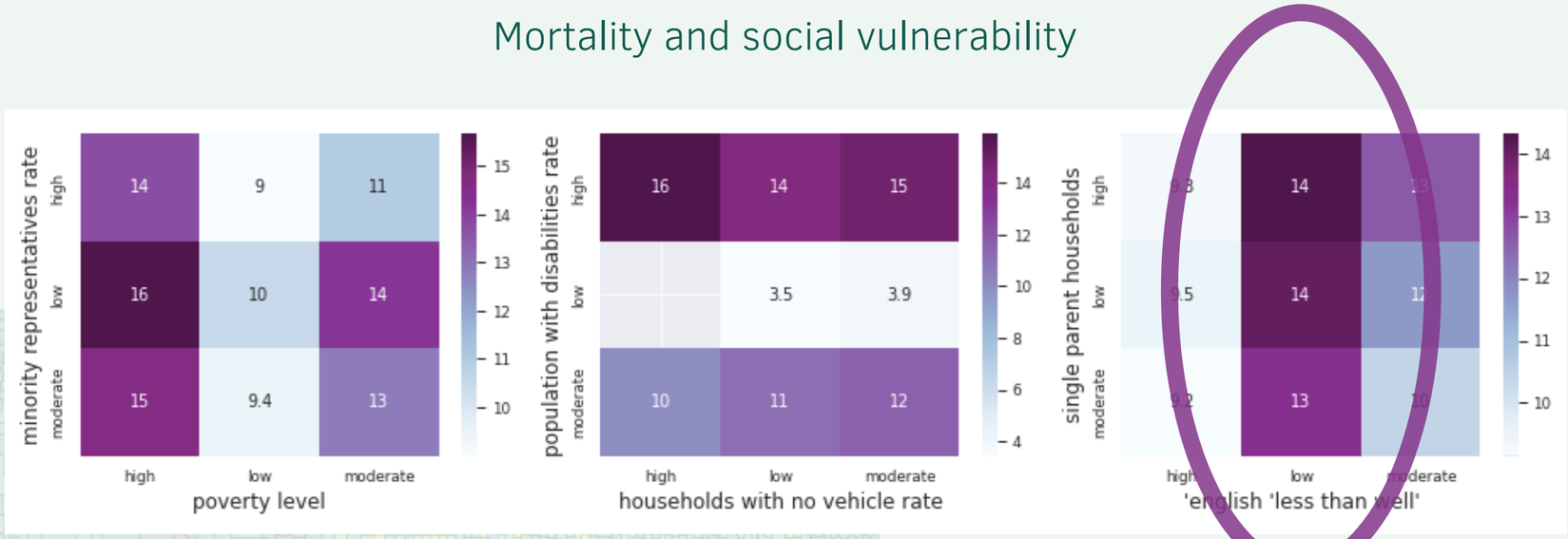
MORTALITY

Mortality and social vulnerability



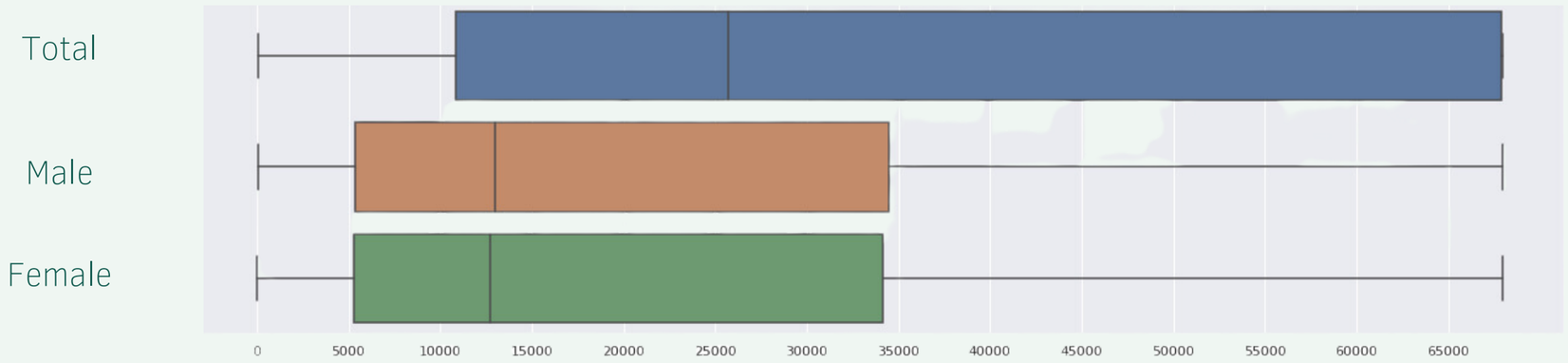
MORTALITY

Mortality and social vulnerability

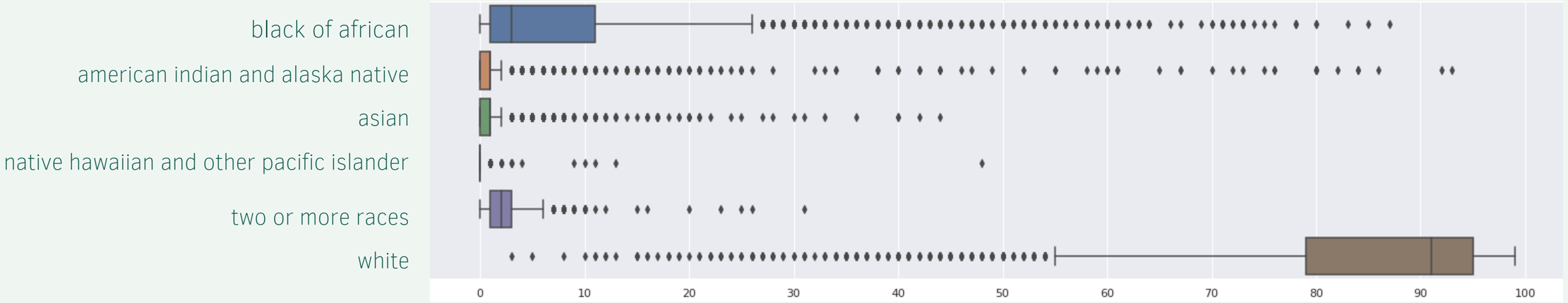


DEMOGRAPHICS

Population by gender

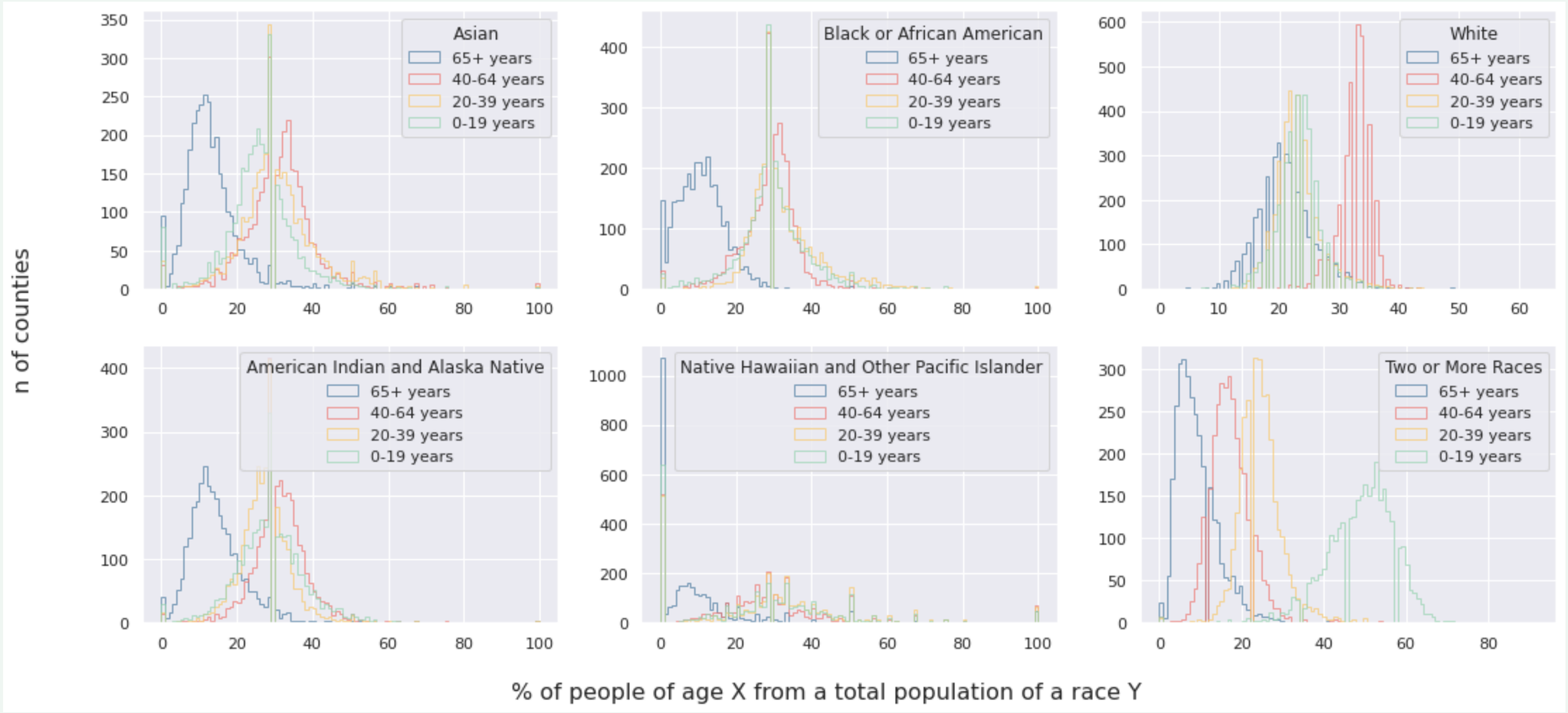


Population by race, % of total county population



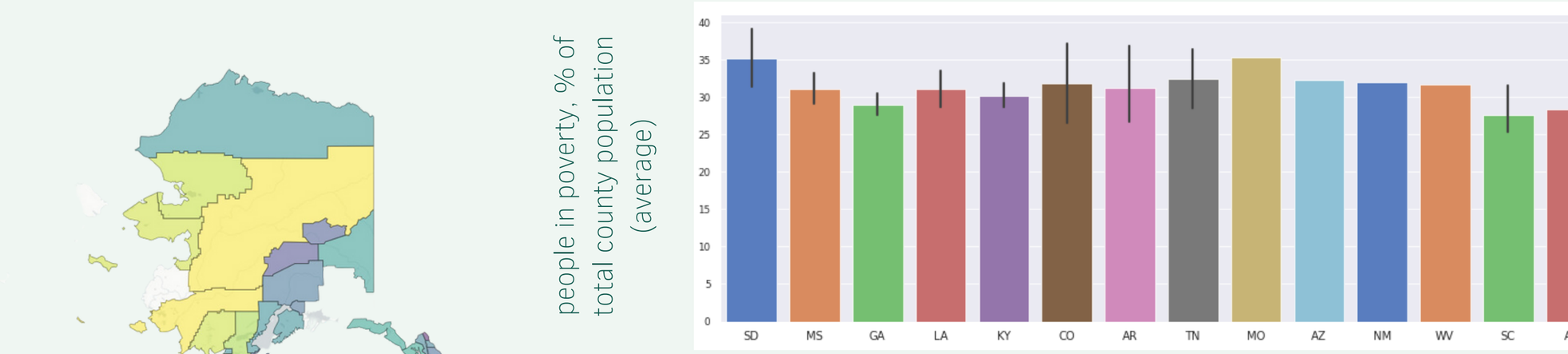
DEMOGRAPHICS

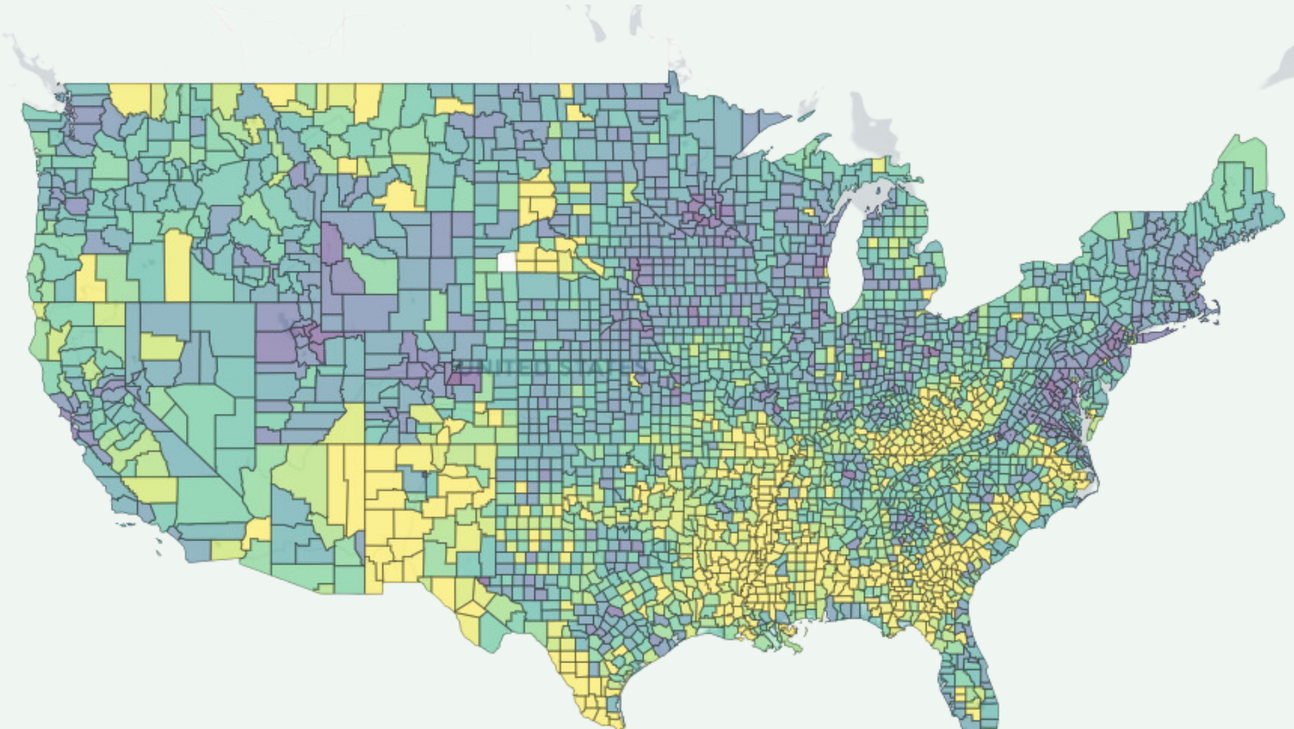
Race and length of life



POVERTY

Counties with the highest poverty levels, by state

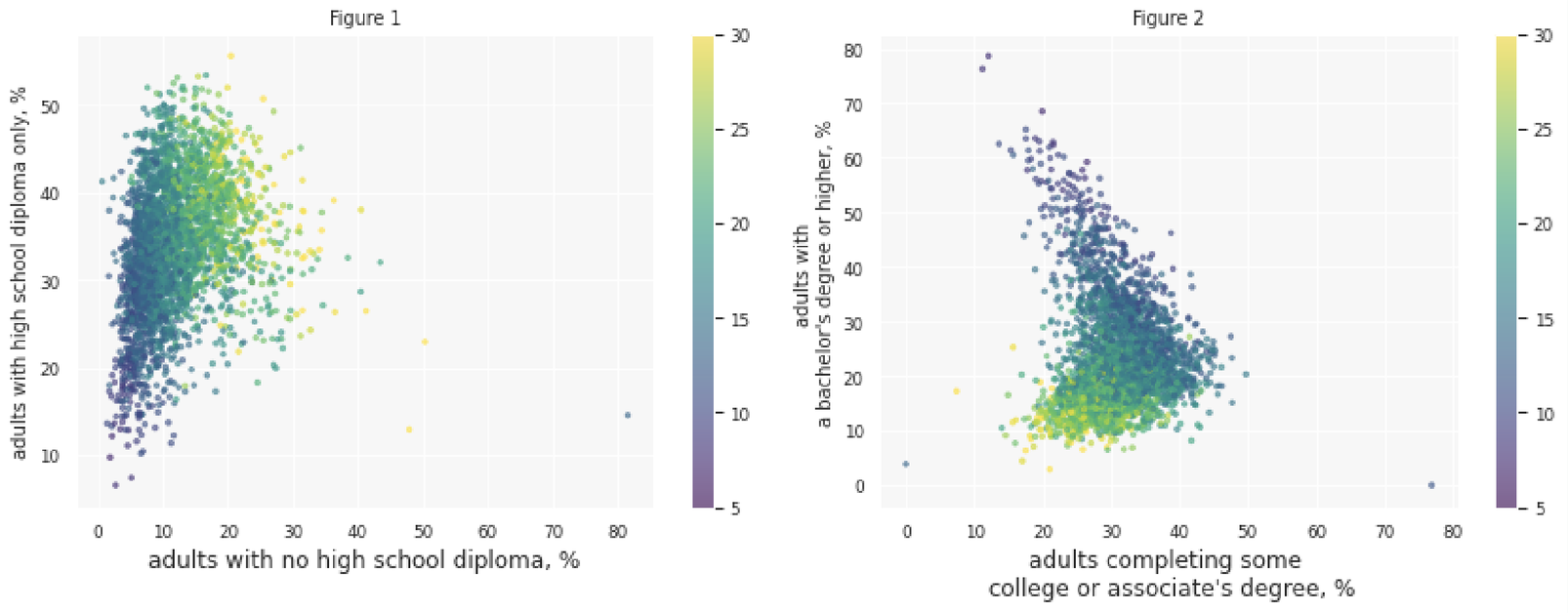




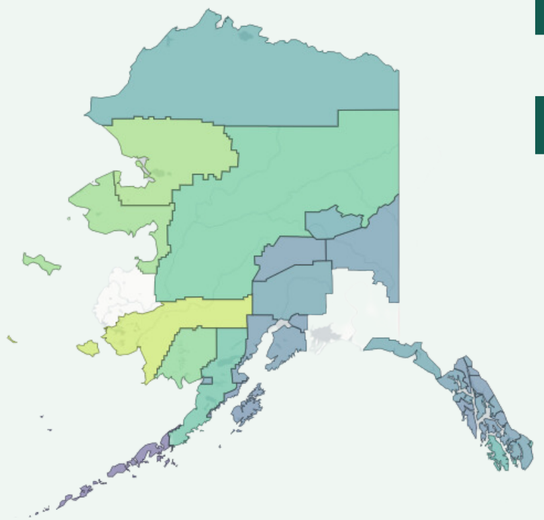
EDUCATION AND HEALTH

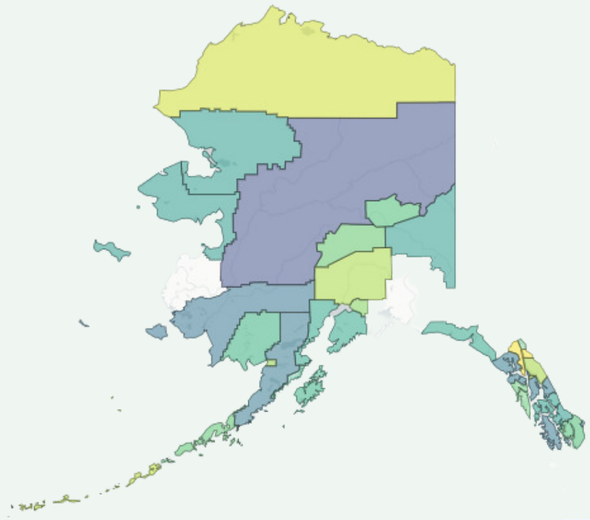
In the majority of US counties around 33% of adults have only a high school diploma, around 31% completed some college or associate's degree, 21% have bachelor's degree or higher, and around 11% of adults do not have a high school diploma.

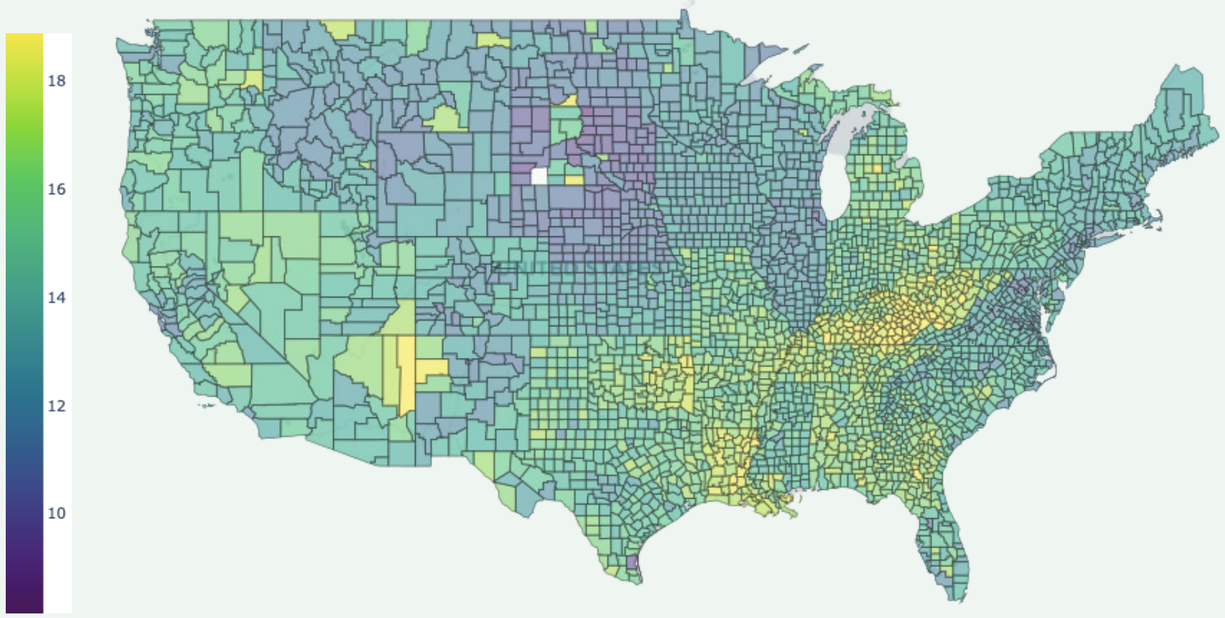
Education and 'poor general health' rates

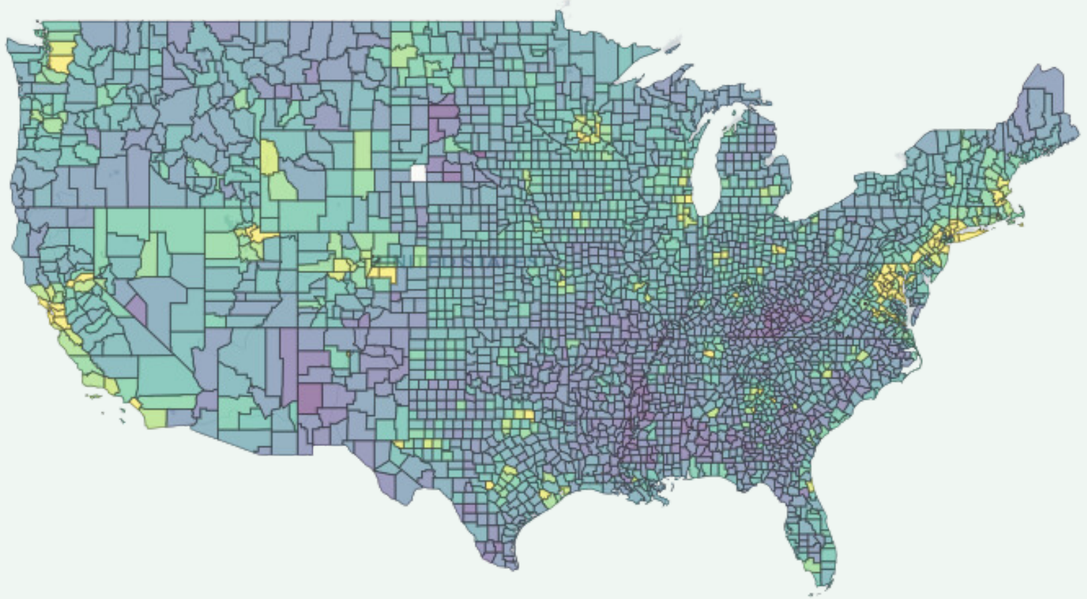


MENTAL HEALTH AND MEDIAN HOUSEHOLD INCOME



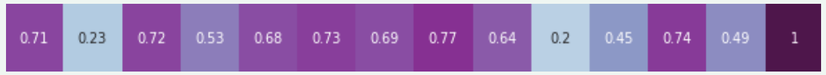






TARGET VARIABLE

1. Each health outcome column multiplied by the correlation coefficient between this feature and mortality rate.
2. All feature columns are summed up, resulting into one new feature column.
3. The resulting variables are scaled between 0 and 1 using the min max scaling method.
4. Substract the resulting variable from 1 to change the direction from 0 to 1, where 0.0 corresponds to poor population health and 1..0 corresponds to good population health.

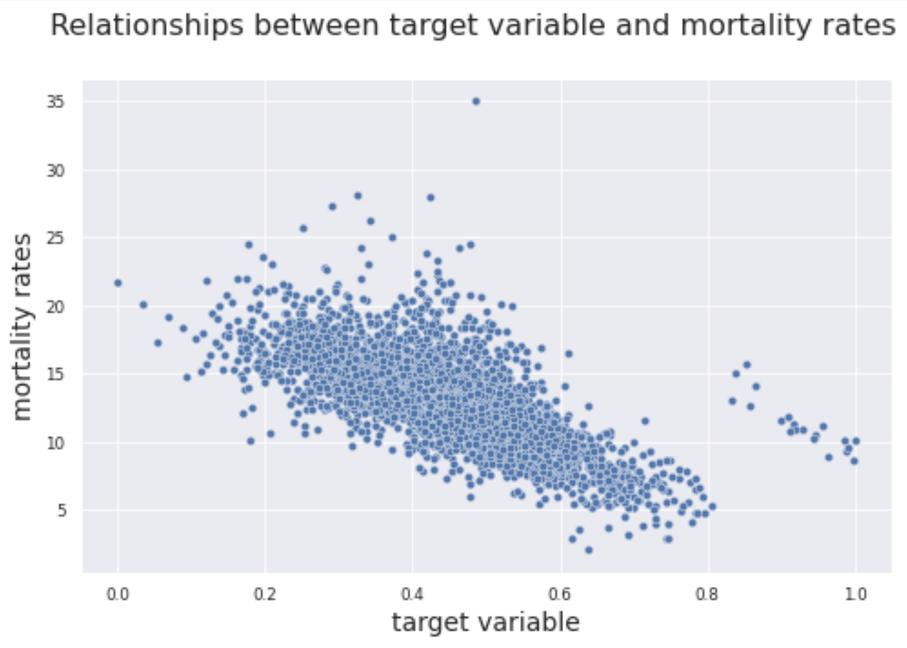


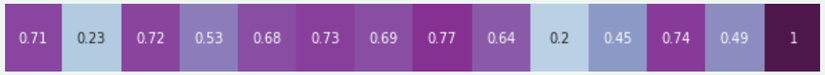
$$C = [c_1 \quad c_2 \quad \dots \quad c_n]$$

$$O = \begin{bmatrix} o_{11} & o_{12} & \dots & o_{1n} \\ o_{21} & o_{22} & \dots & o_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ o_{m1} & o_{m2} & \dots & o_{mn} \end{bmatrix}$$

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix}$$

TARGET VARIABLE





$$C = \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix}$$

$$O = \begin{bmatrix} o_{11} & o_{12} & \dots & o_{1n} \\ o_{21} & o_{22} & \dots & o_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ o_{m1} & o_{m2} & \dots & o_{mn} \end{bmatrix}$$

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix}$$

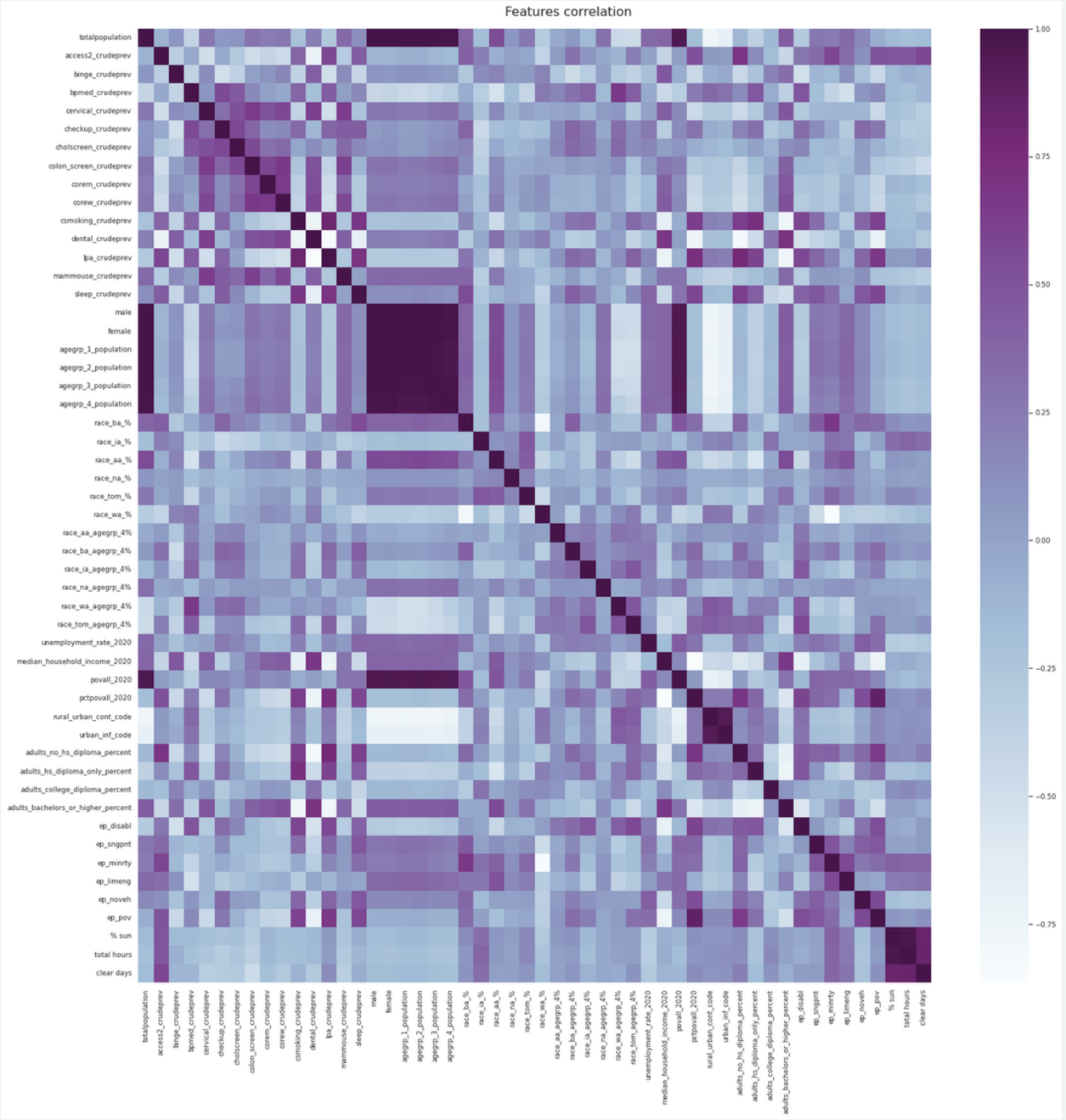
FEATURE MATRIX

Three sets of features were used in the prediction model:

- all 80 features
- 40 features
- 30 features

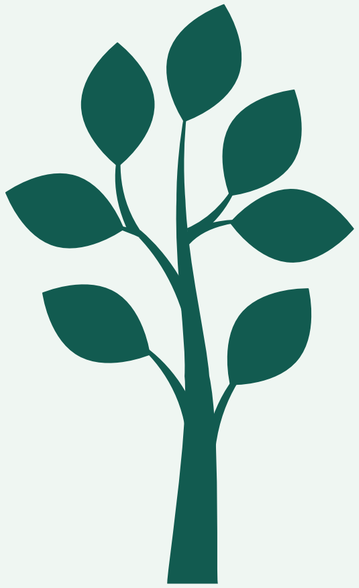
Majority of features are crude prevalence rates on **county level**, however, some features are of state-level.

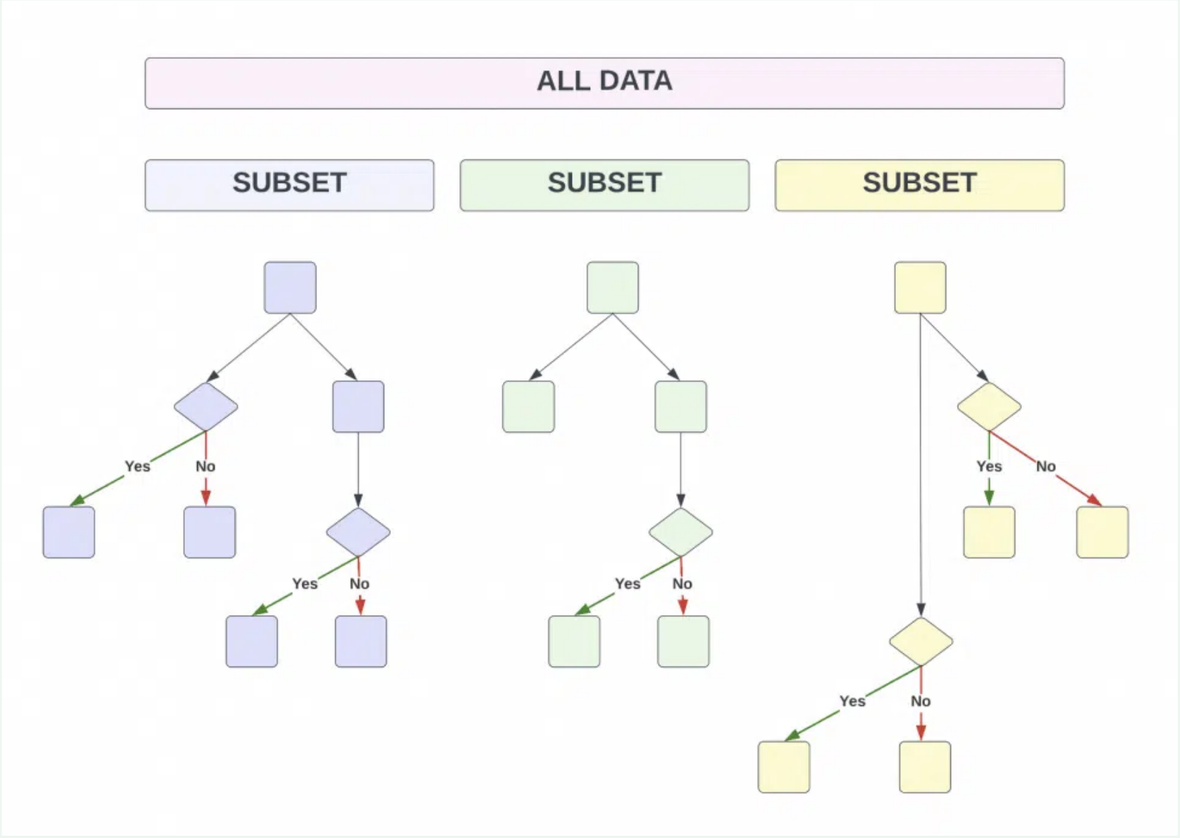
Each sample is. a set of features corresponding to **one county**.



PREDICTION MODEL

The model used in this study is XGBoost regressor, since it is one of the state-of-the-art models for tabular data problems, it is fast to train, and it is well interpretable.

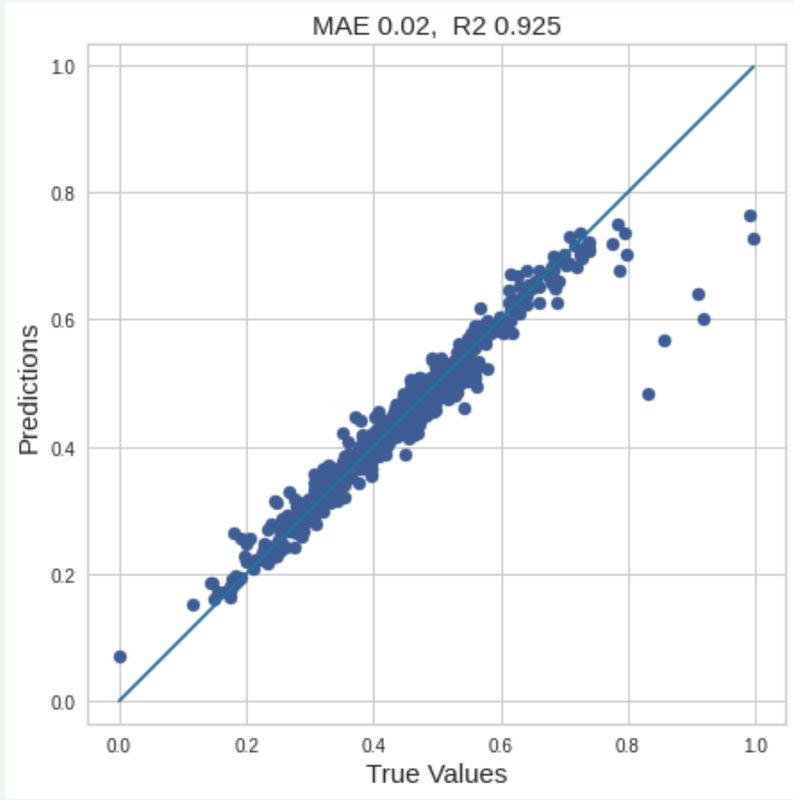




https://www.aiplusinfo.com/blog/introduction-to-xgboost-and-its-uses-in-machine-learning/

RESULTS

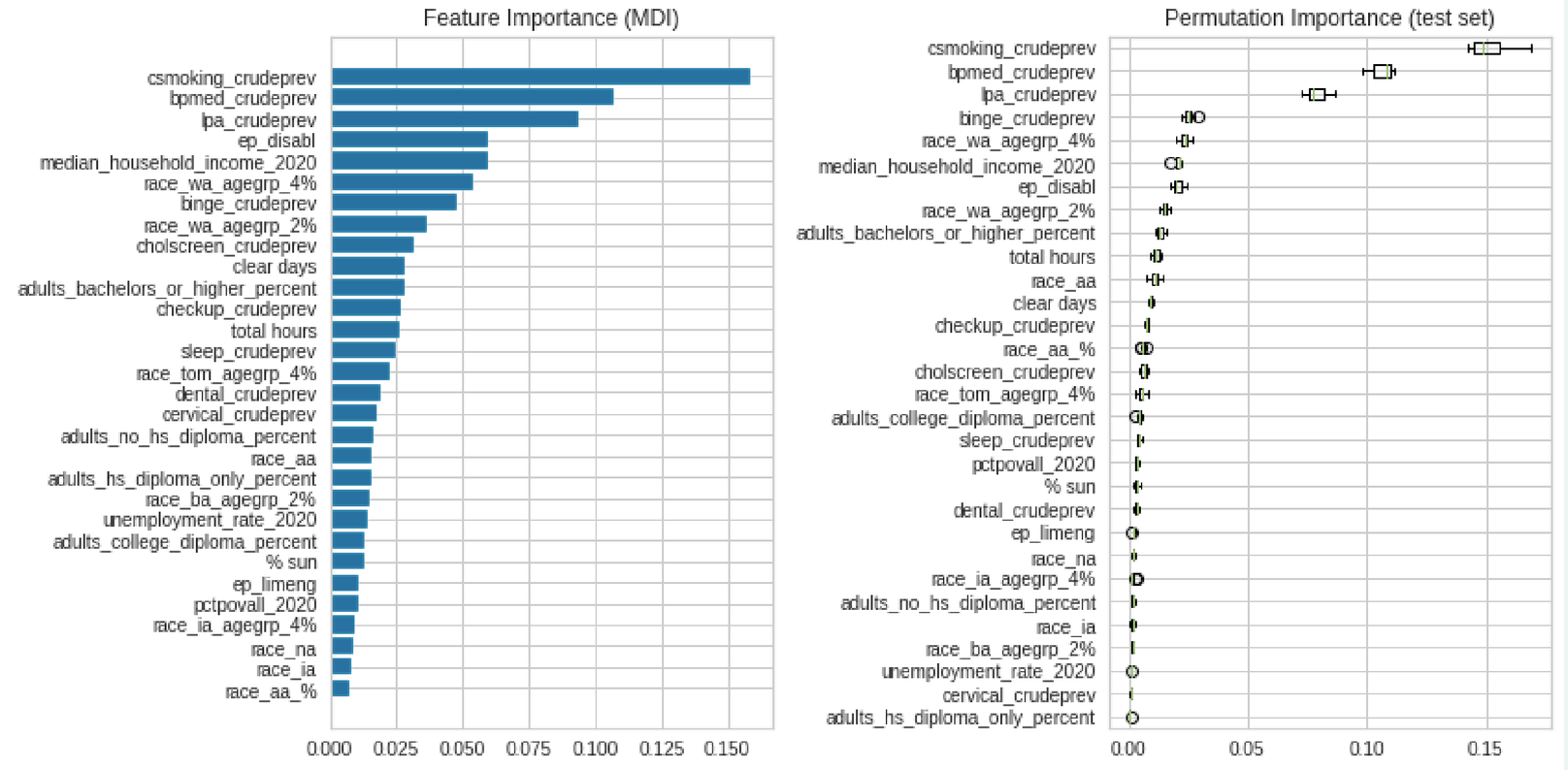
n of features	MAE	R2
80	0.021(std 0.0004)	0.912 (std .0061)
40	0.020 (std 0.0005)	0.916 (std 0.0023)
30	0.020 (std 0.0004)	0.923 (std 0.052)



FEATURE IMPORTANCE

The most important predictors:

- Health risk behaviour: "currently smoking", "low physical activity", "sleep less than 7 hours", "binge drinking"
- Preventive measures: "taking blood pressure medicines", "cholesterol screening", "routine checkup visits", "cervical cancer screening", "dentist visits".
- Social vulnerability status: such as percent of people with disabilities, household income, unemployment rate, poverty, limited english, and education level.

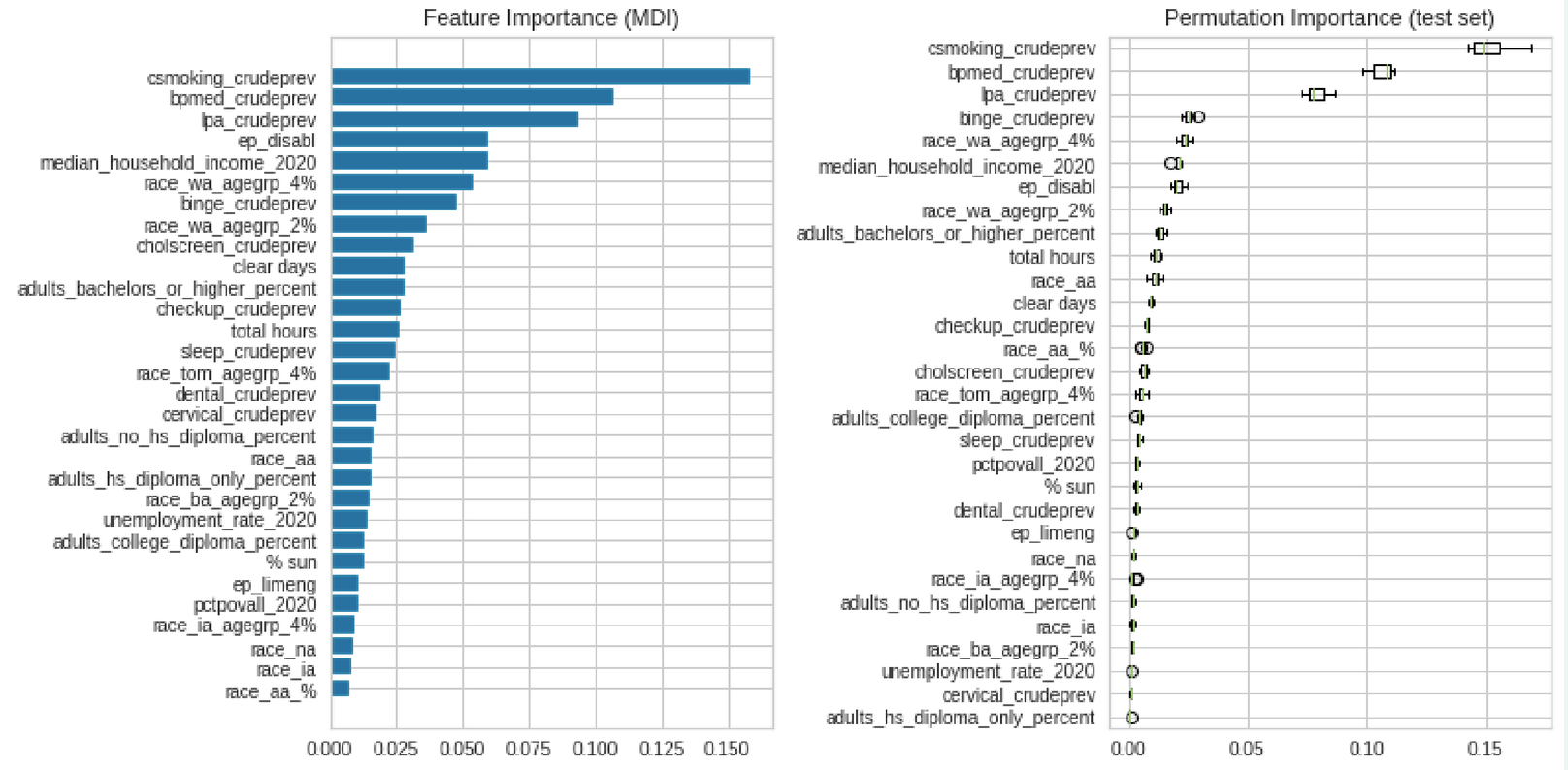


FEATURE IMPORTANCE

- Demographics: percent of people of asian race, elderly people of two or more races, percent of young people of black or African American race, elderly and young American Indian and Alaska native people, as well as percent of native Hawaiian people in the county.

white race people in elderly and between 20 and 39 years old

average annual sunshine



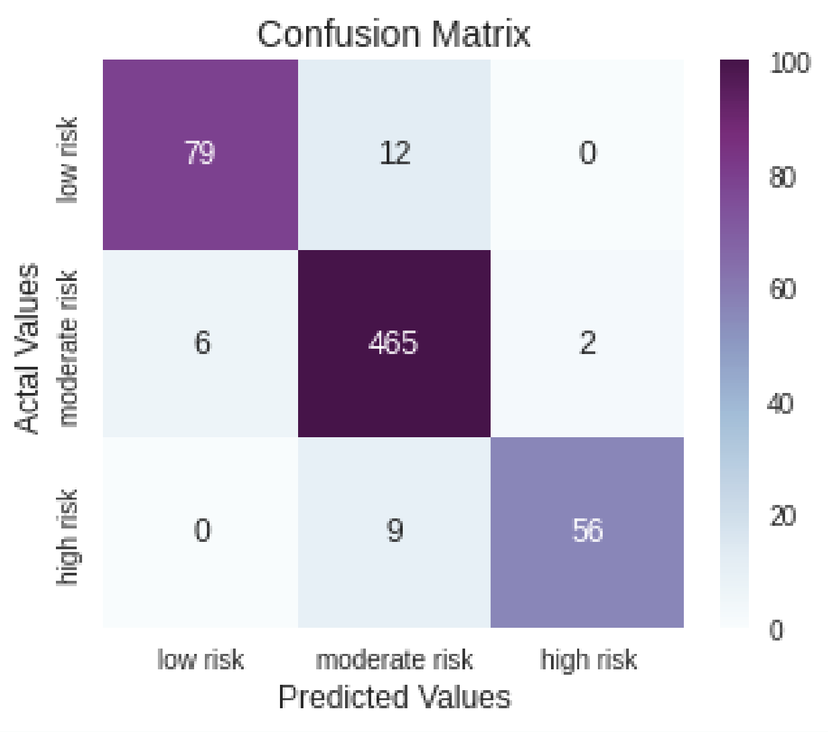
THANK YOU

FOR LISTENING

March 2023

Maslenkova Svetlana

CLASSES ANALYSIS



	precision	recall	f1-score	support
0	0.93	0.87	0.90	91
1	0.96	0.98	0.97	473
2	0.97	0.86	0.91	65
accuracy			0.95	629
macro avg	0.95	0.90	0.93	629
weighted avg	0.95	0.95	0.95	629