DS2500 - FALL 23

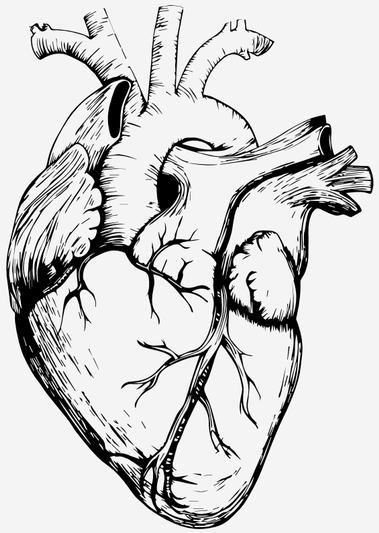
THE RELATIONSHIP BETWEEN AMBIENT OZONE CONCENTRATION AND HEART DISEASE MORTALITY RATE

Prediction model and correlation

Anh Nguyen

Background

- Ozone - a reactive gas molecule in the Earth’s atmosphere.
- Ambient ozone is a risk factor for cardiovascular disease.
- Ambient ozone levels today are 30-70% higher than they were 100 years ago.



- Heart disease is among the leading causes of death among all demographic groups.
- Cause of 1 in 5 deaths in the U.S.

Problem Statement and Goals

- It is important to be aware of possible relationships between ambient ozone and cardiovascular disease.

→ Build an ozone concentration - heart disease risk k-nearest neighbor prediction model specific to each state.

→ Analyze for correlation through linear regression.

Data Sources

Ozone concentration Data (ppb)

- Environmental Protection Agency (EPA)
- Collected at CASTNET sites across the country - collect air quality data
- 2015 - 2020

3-year averaged heart disease mortality rate (per 100,000 population)

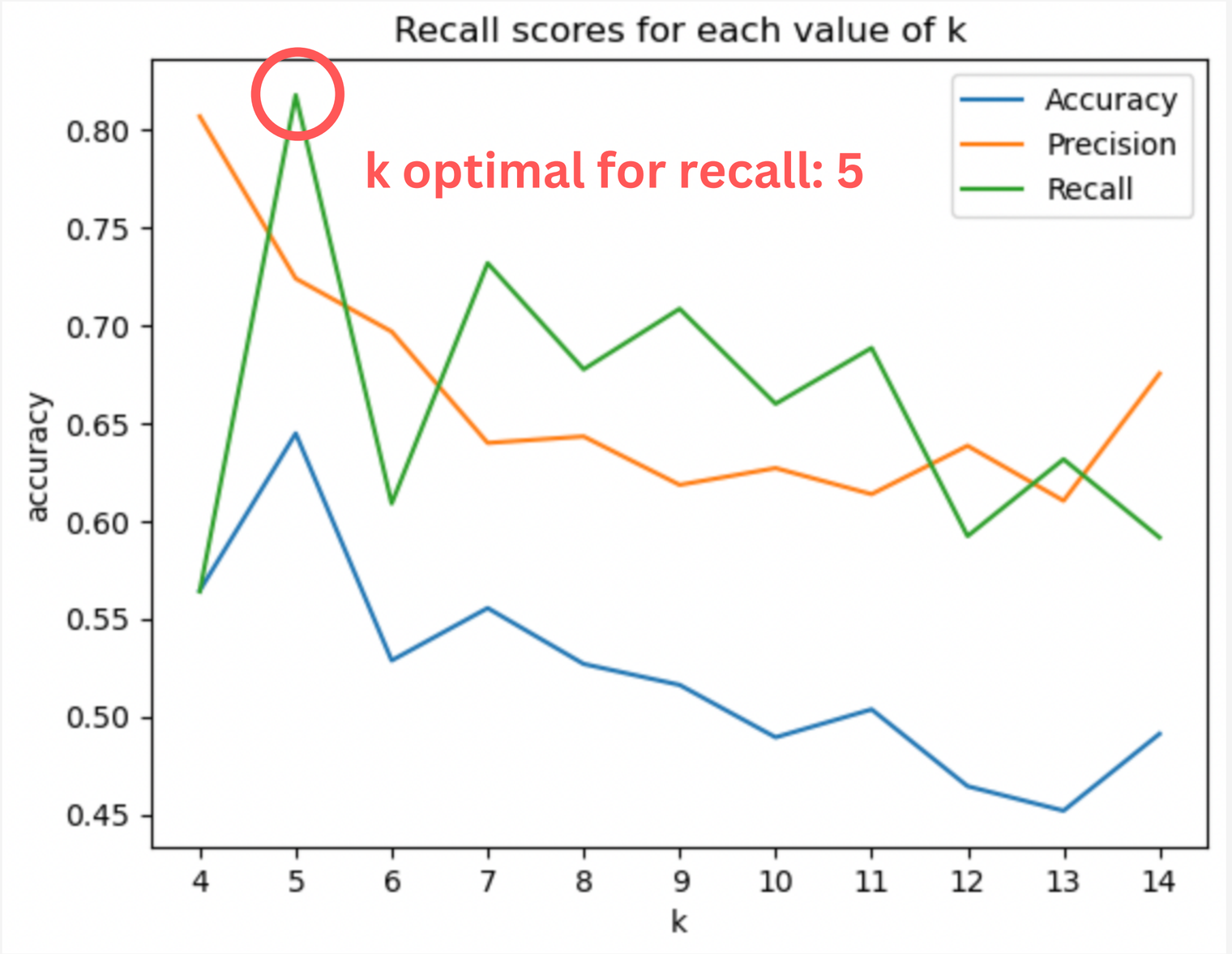
- Centers for Disease Control and Prevention (CDC)
- Averaged for 2015-2017 and 2018-2020

K-Nearest Neighbor Classification

Train on data about the average ozone concentration (feature) and heart disease risk (target) between 2015 and 2020 (High risk - above average, low risk - below average)

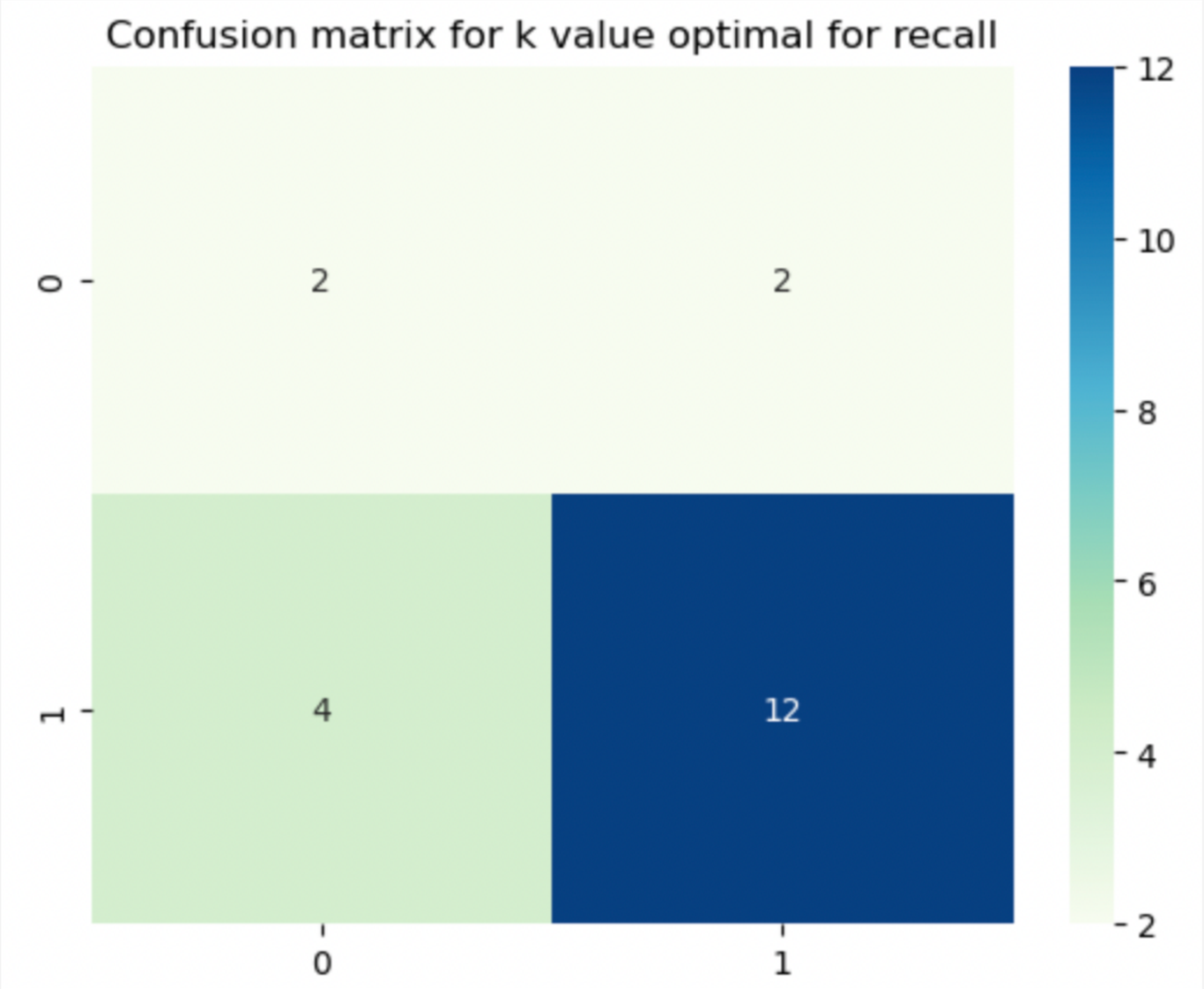
Choosing the value of k

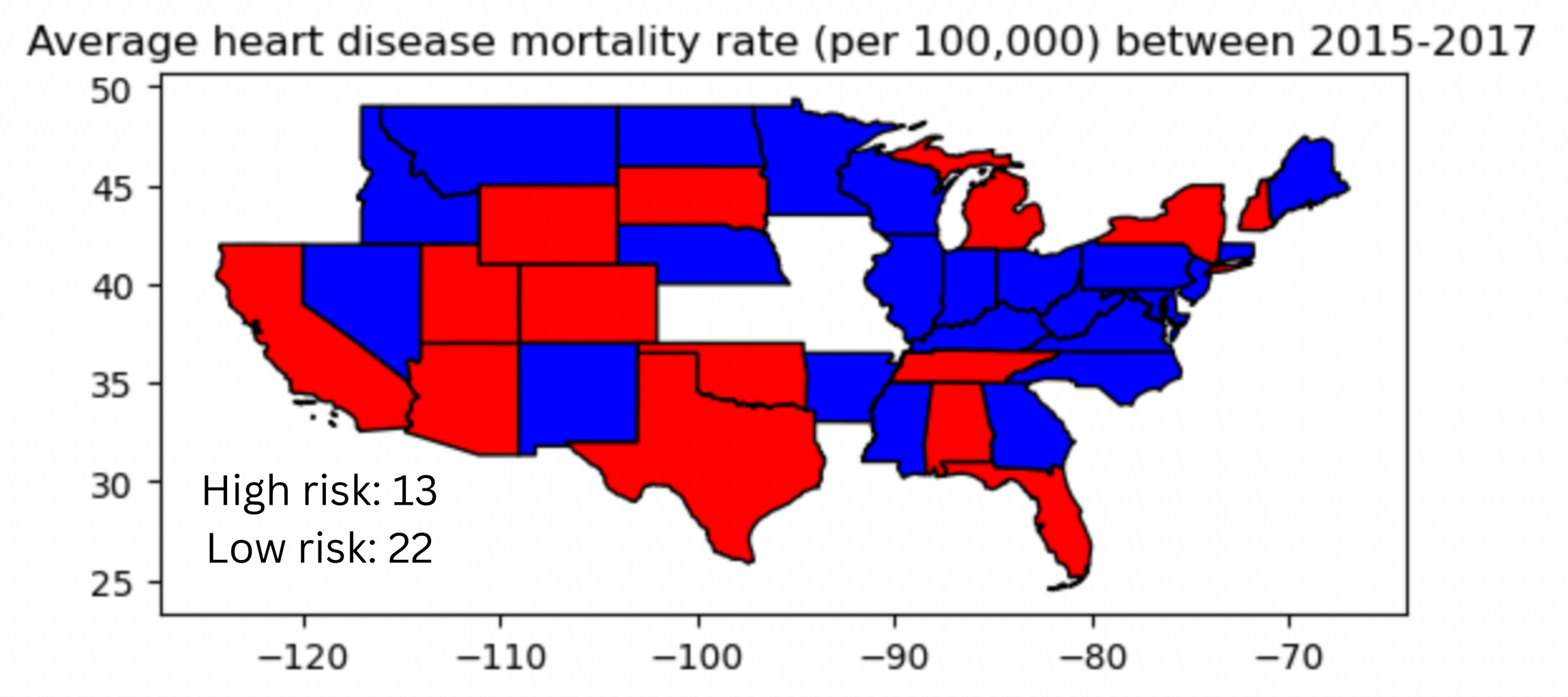
	k	accuracy	precision	recall
0	4	0.564286	0.806667	0.563810
1	5	0.644643	0.723810	0.817619
2	6	0.528571	0.696667	0.608810
3	7	0.555357	0.639762	0.731667
4	8	0.526786	0.643095	0.677381
5	9	0.516071	0.618333	0.708333
6	10	0.489286	0.626905	0.659762
7	11	0.503571	0.613571	0.688333
8	12	0.464286	0.638333	0.592143
9	13	0.451786	0.610238	0.631429
10	14	0.491071	0.675238	0.591429



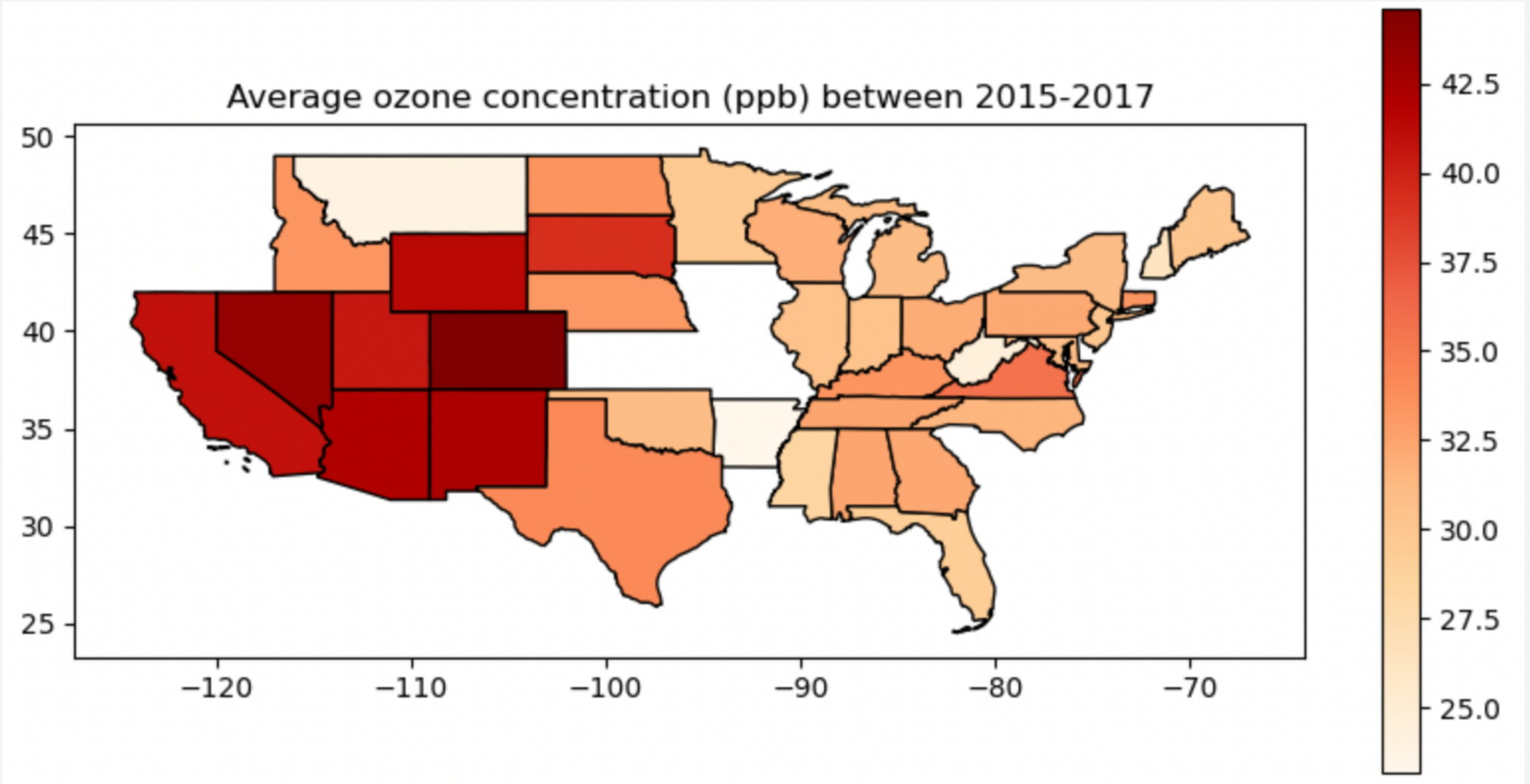
Report and confusion matrix

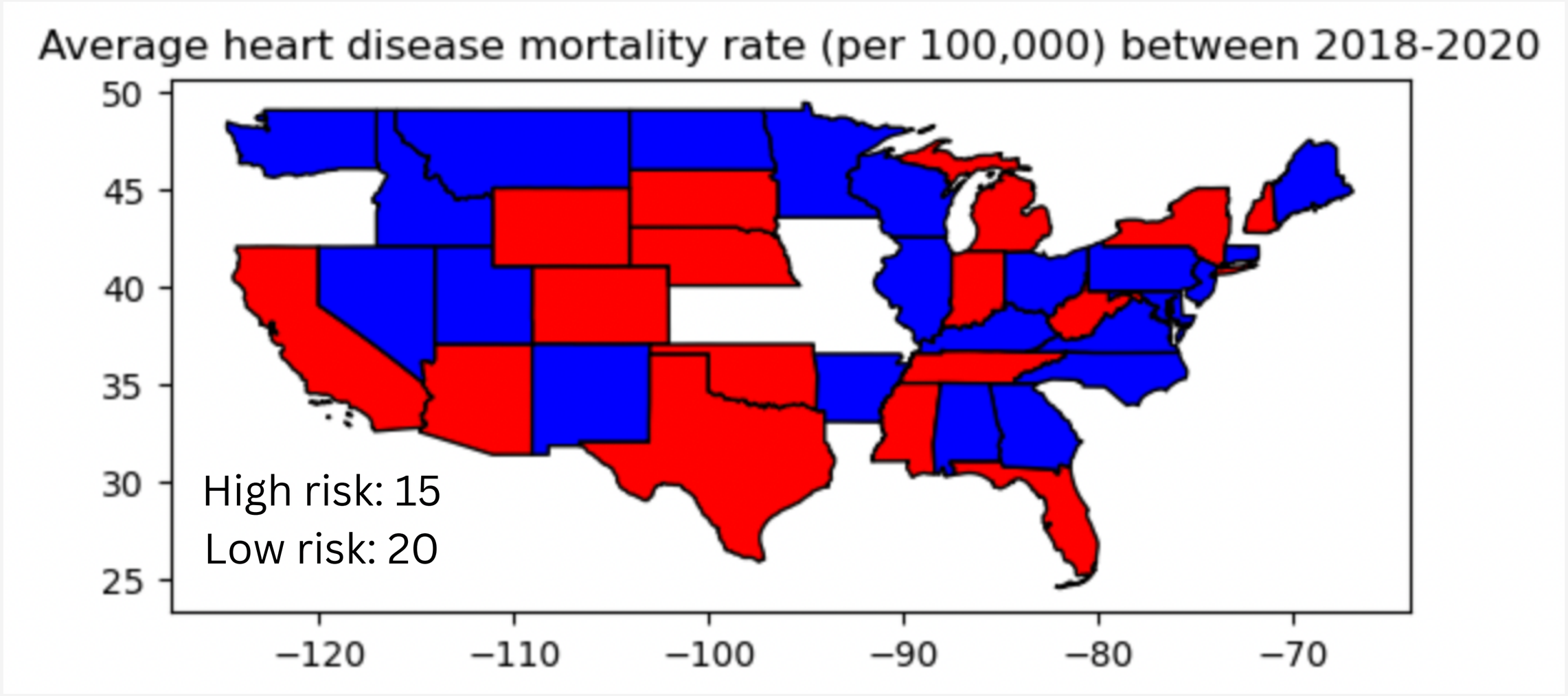
	precision	recall	f1-score	support
0	0.33	0.50	0.40	4
1	0.86	0.75	0.80	16
accuracy			0.70	20
macro avg	0.60	0.62	0.60	20
weighted avg	0.75	0.70	0.72	20

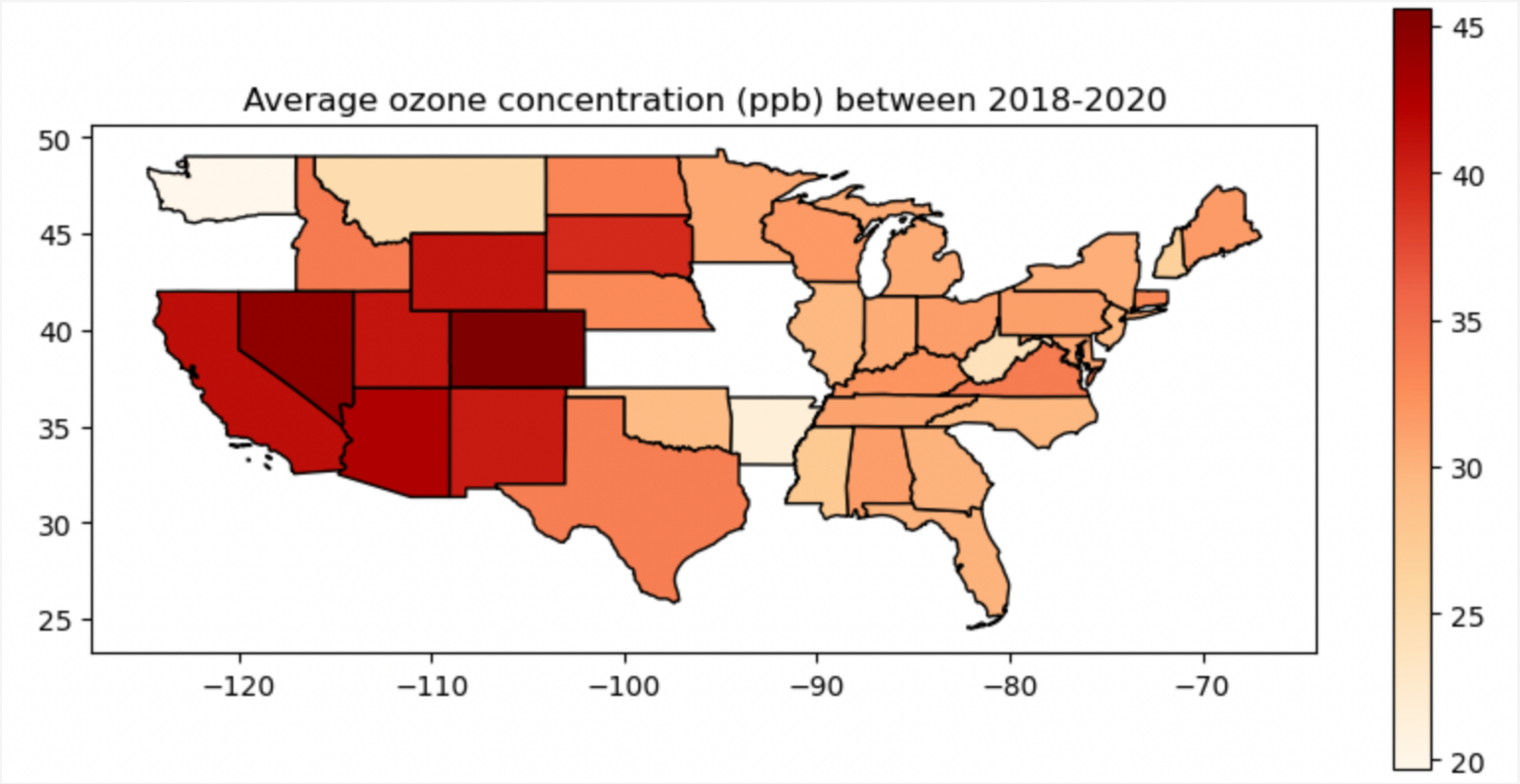


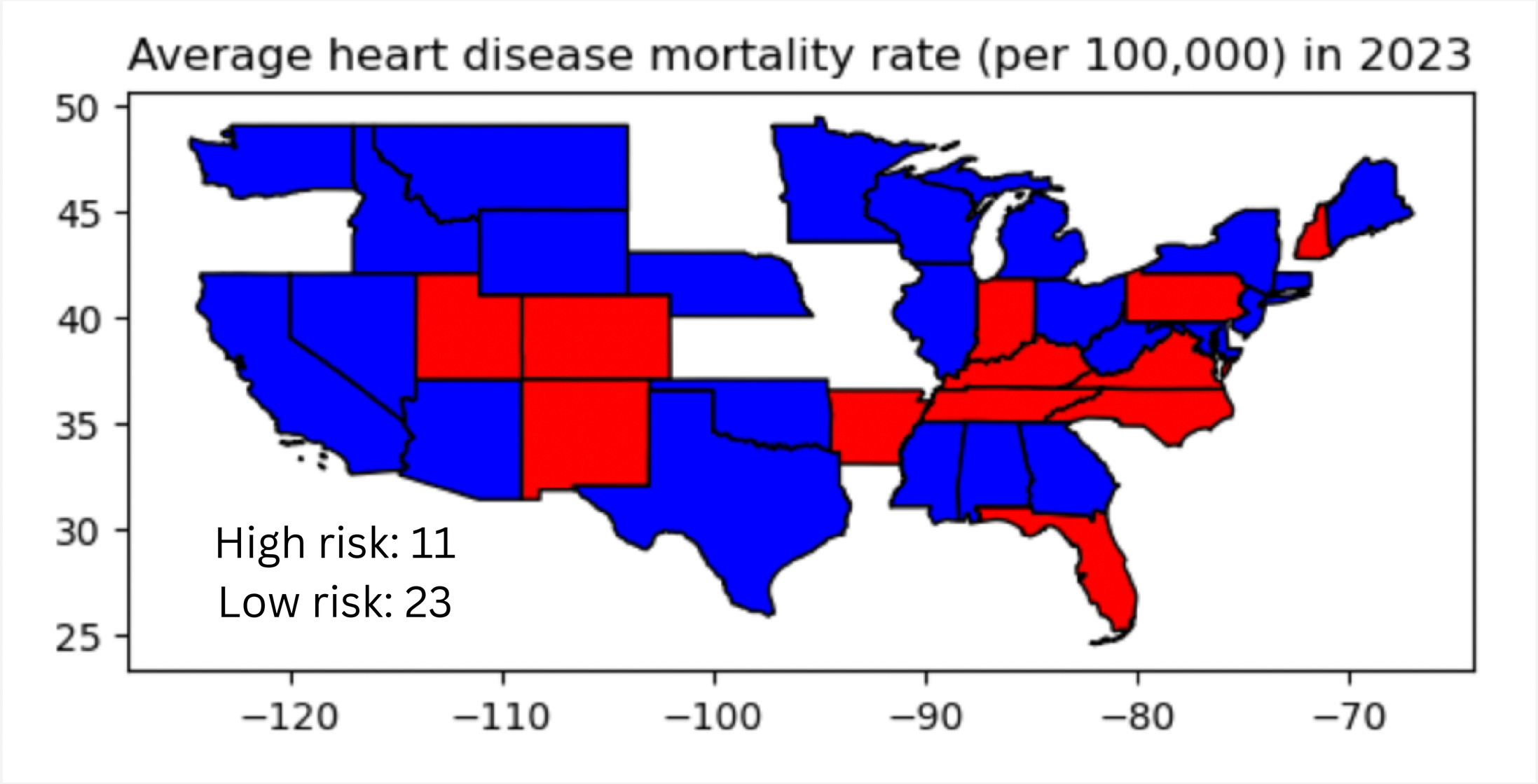


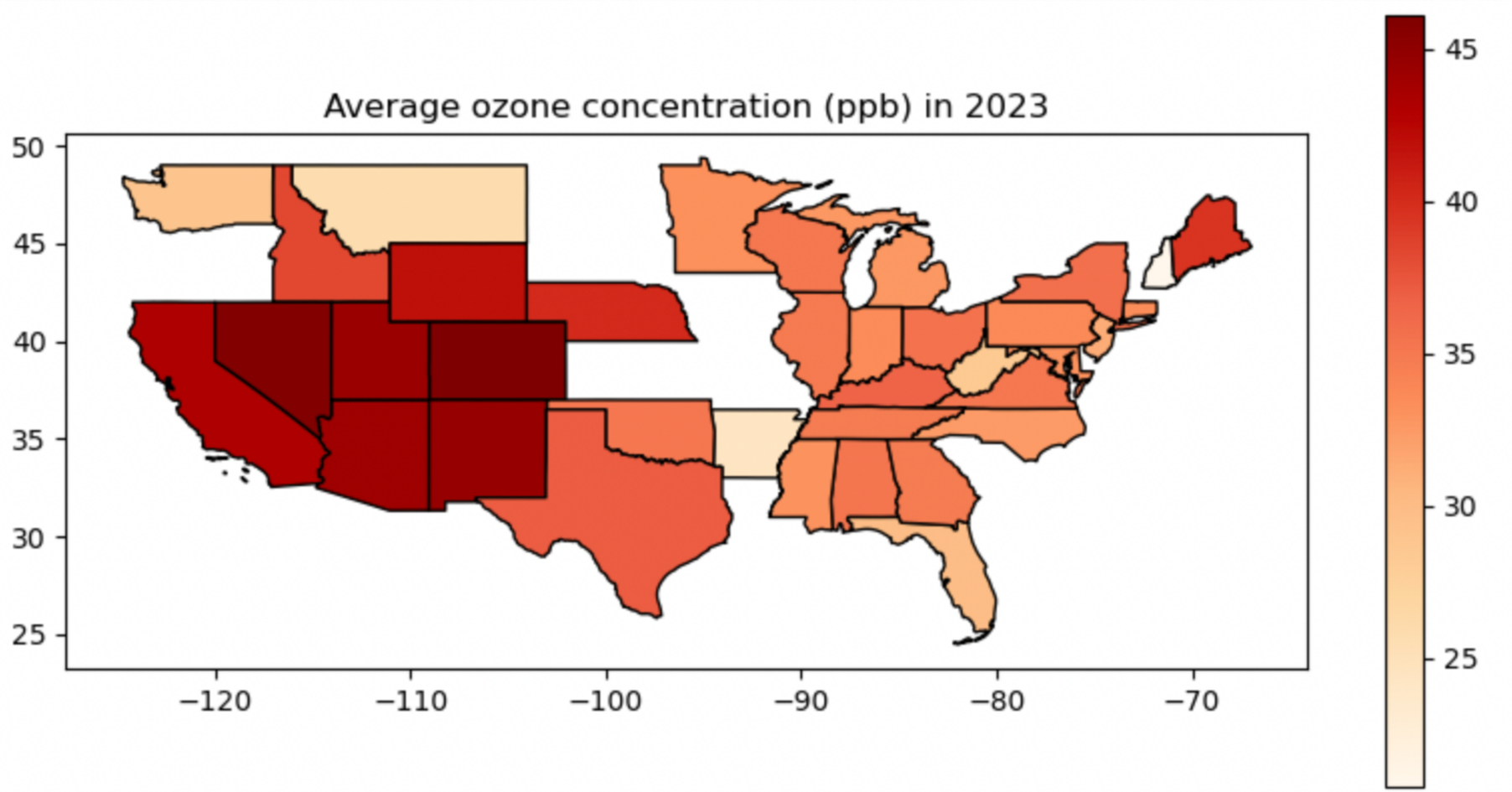
High - above average, Low - below average

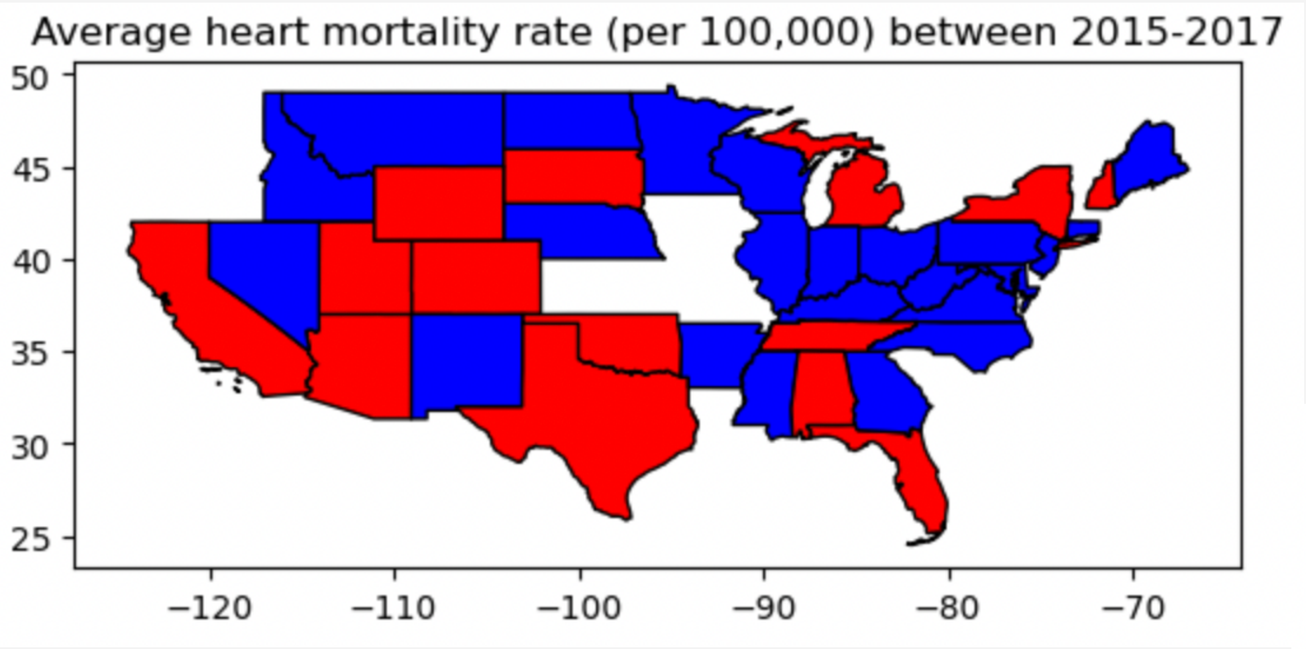

Average ozone concentration (ppb) between 2015-2017
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
42.5
40.0
37.5
35.0
32.5
30.0
27.5
25.0

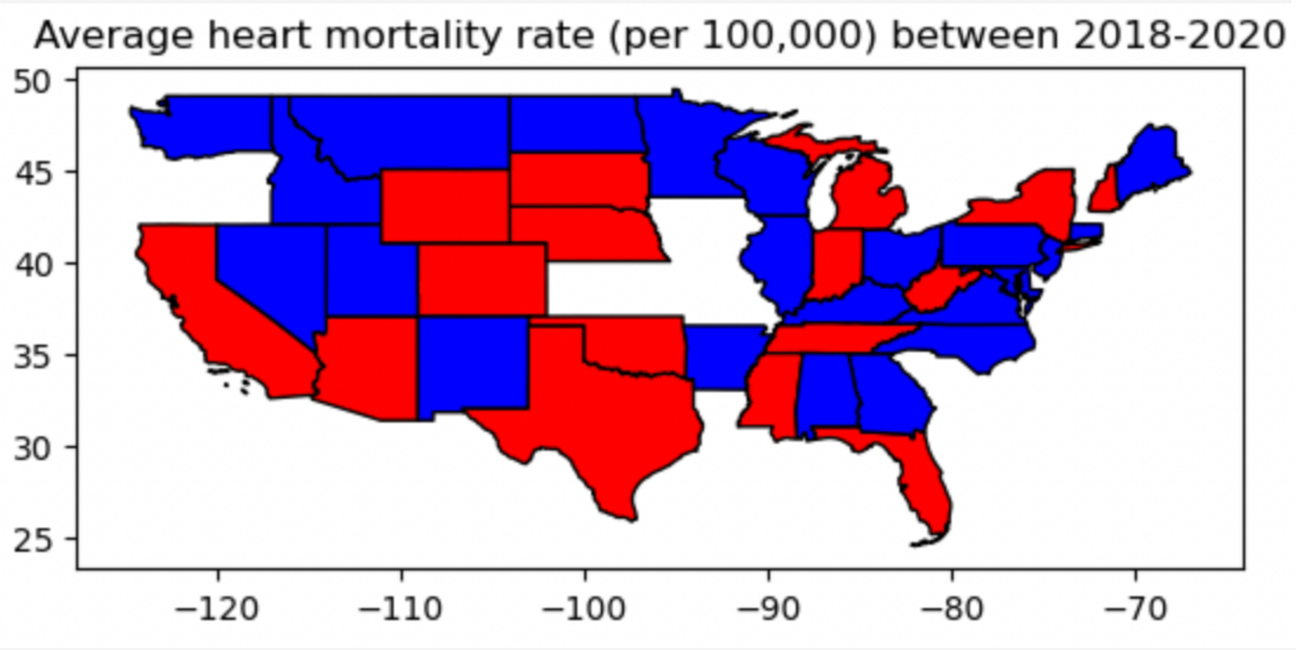

Average heart disease mortality rate (per 100,000) between 2018-2020
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
High risk: 15
Low risk: 20

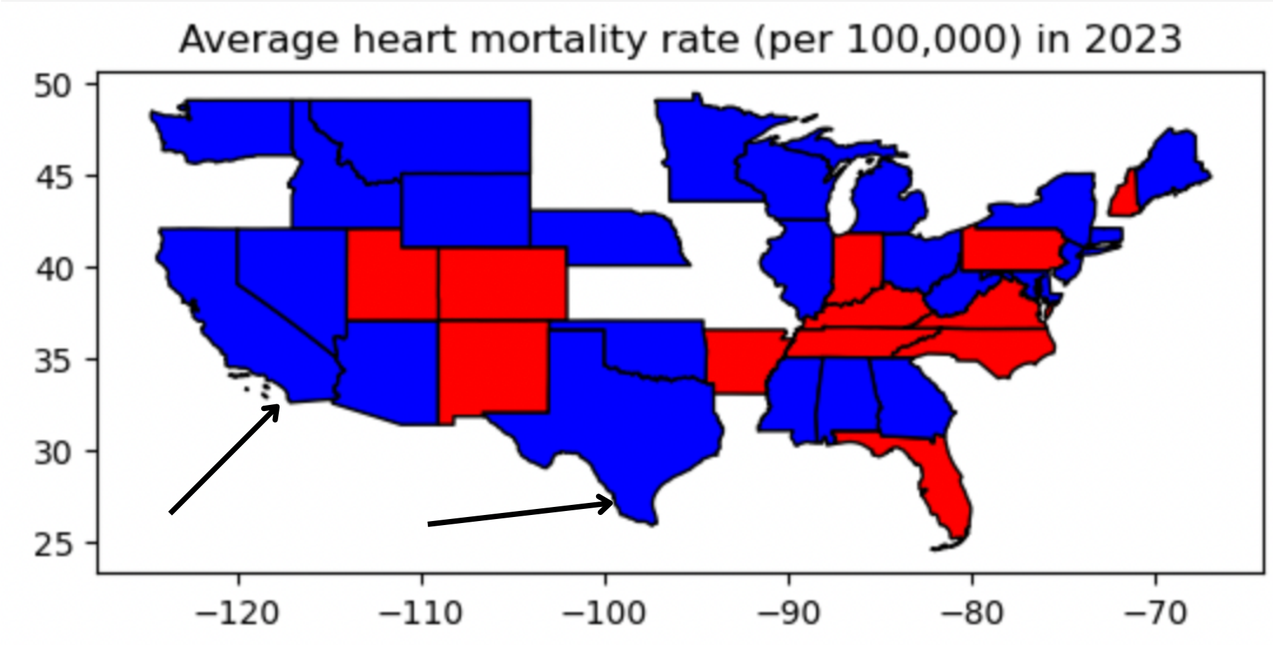

Average ozone concentration (ppb) between 2018-2020
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
45
40
35
30
25
20

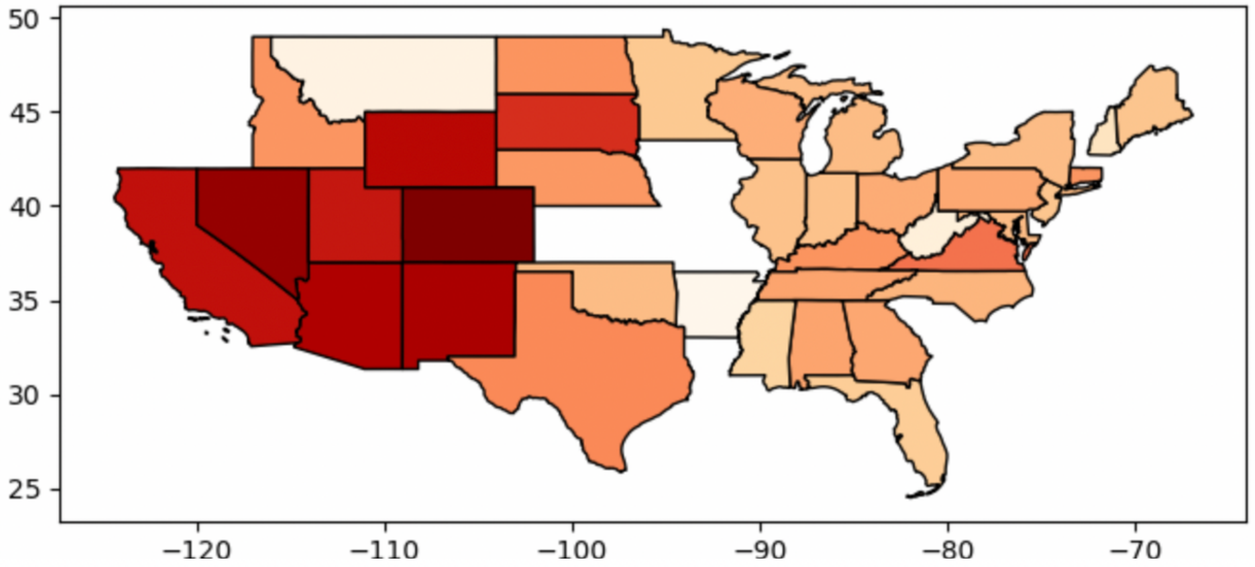

Average heart disease mortality rate (per 100,000) in 2023
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
High risk: 11
Low risk: 23

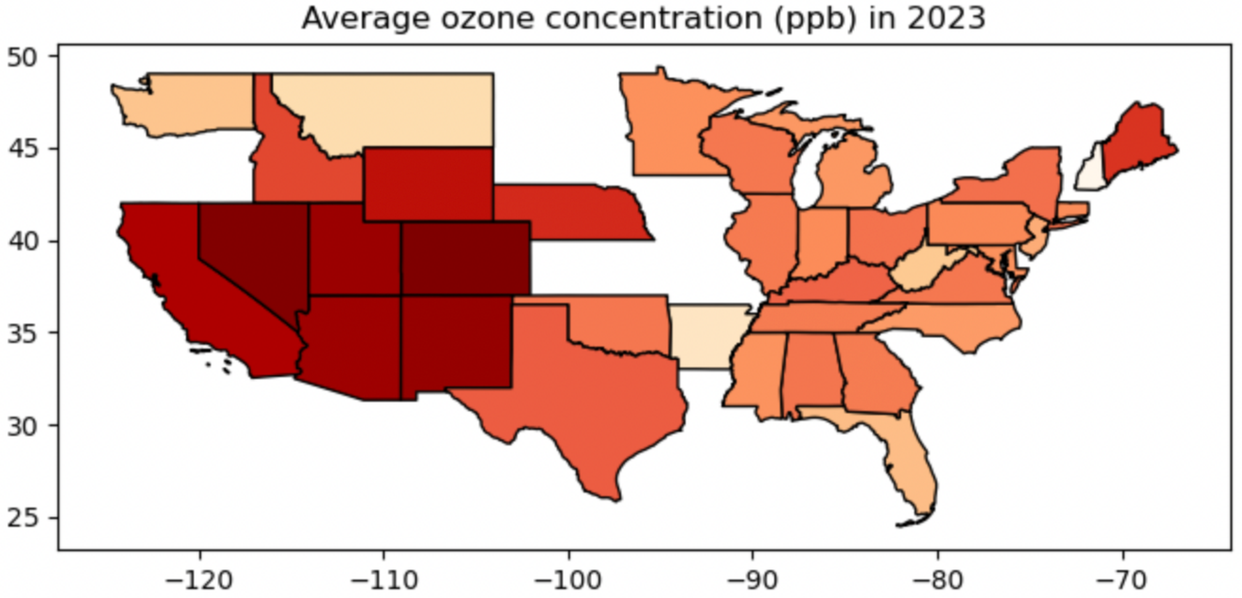

Average ozone concentration (ppb) in 2023
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
45
40
35
30
25

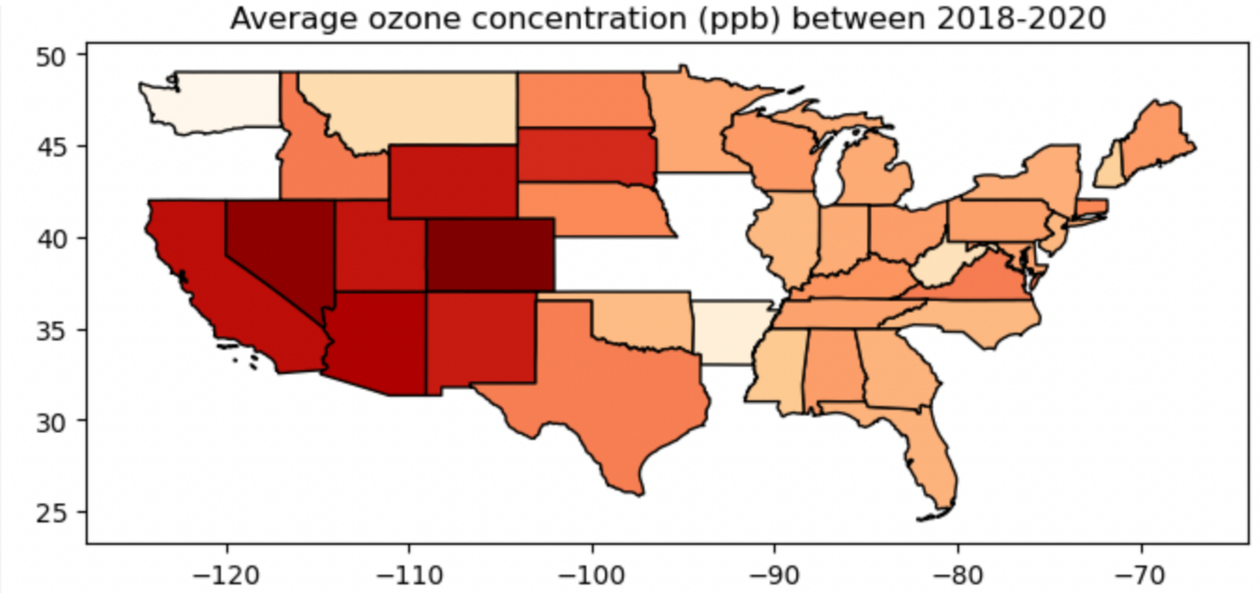

Average heart mortality rate (per 100,000) between 2015-2017
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

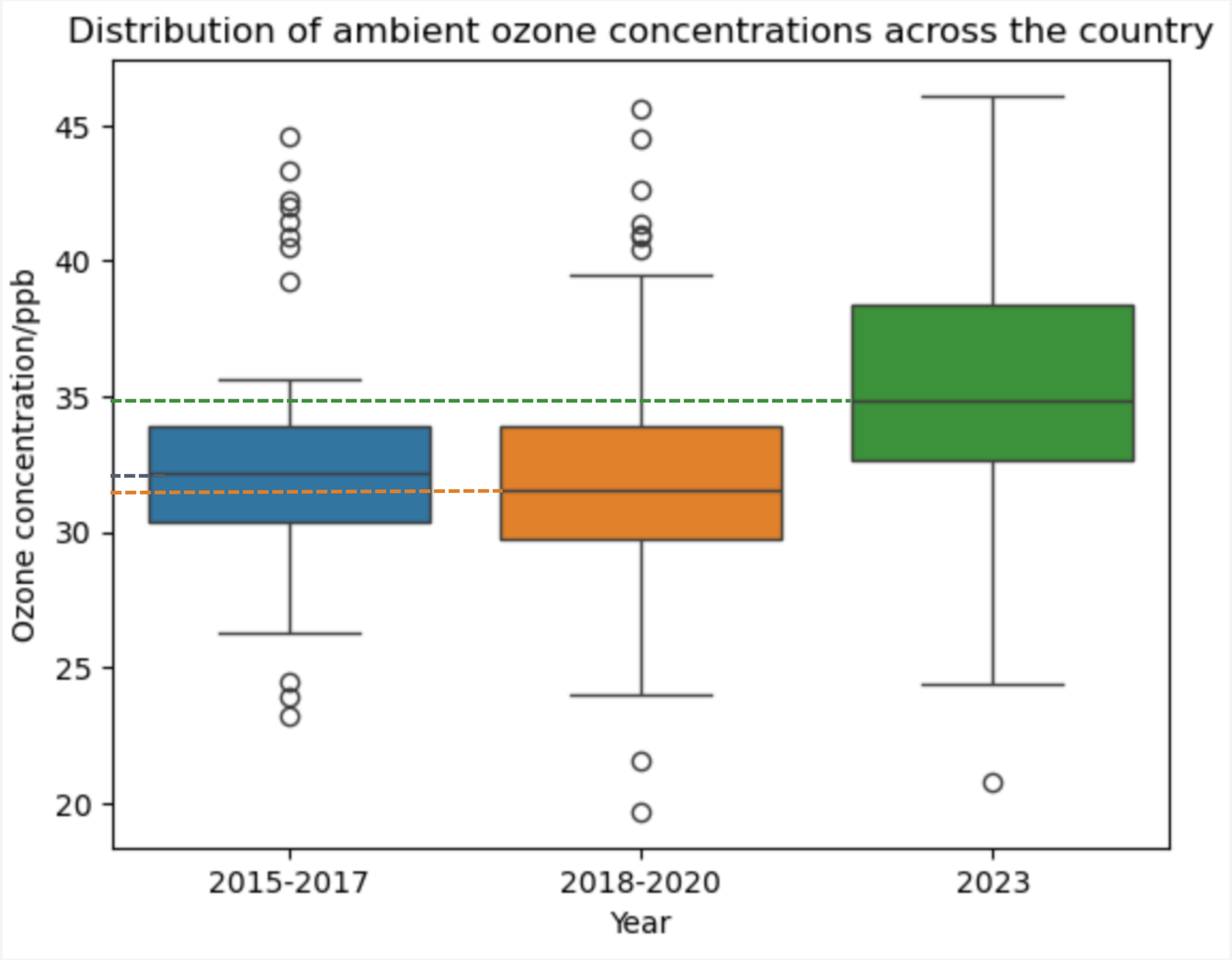

Average heart mortality rate (per 100,000) between 2018-2020
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70


Average heart mortality rate (per 100,000) in 2023
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

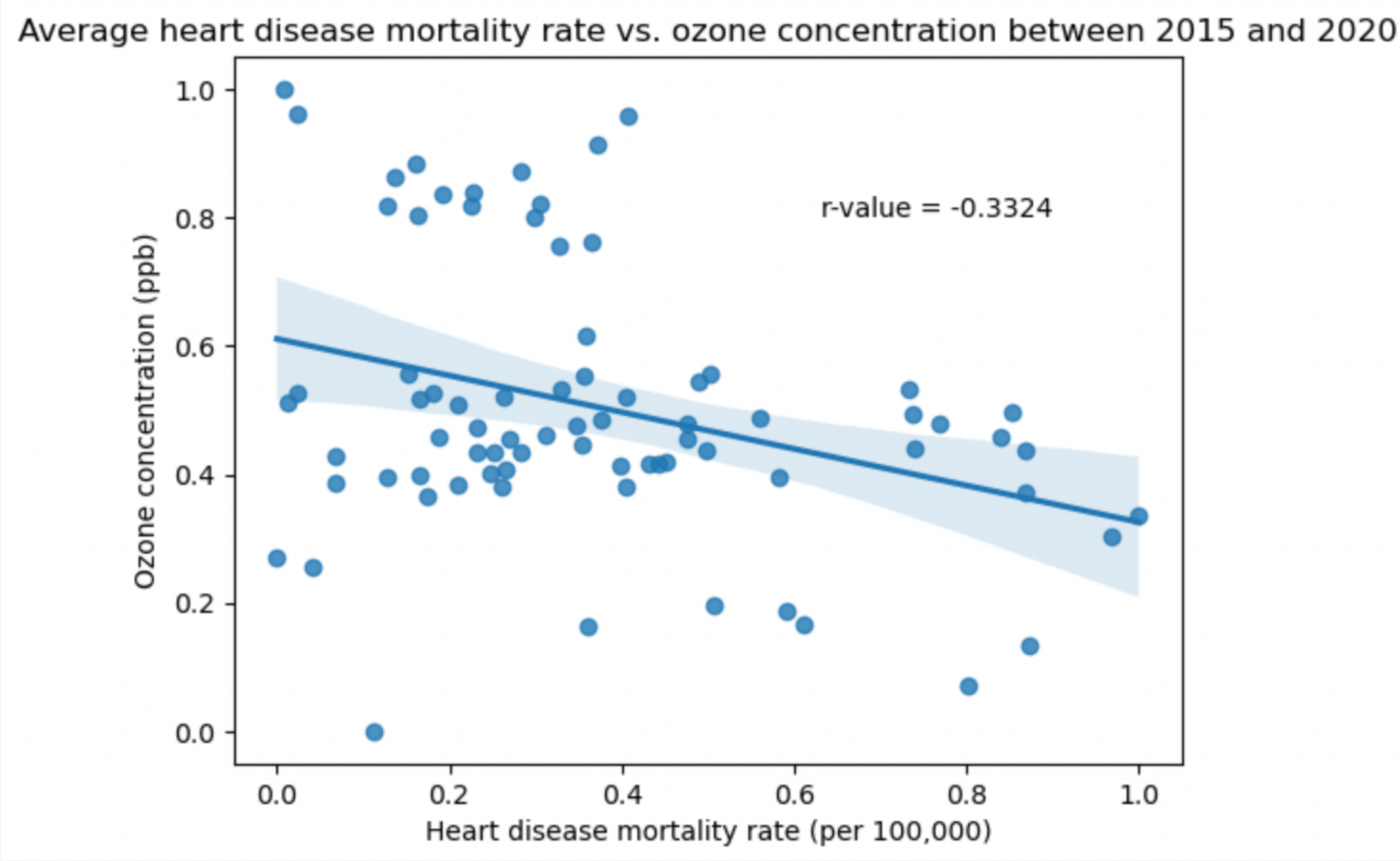

Average ozone concentration (ppb) between 2015-2017
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70


Average ozone concentration (ppb) in 2023
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70


Average ozone concentration (ppb) between 2018-2020
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70


Distribution of ambient ozone concentrations across the country
Ozone concentration/ppb
45
40
35
30
25
20
2015-2017
2018-2020
2023
Year

Linear Regression



Result and Conclusion

K-NN Classification model for prediction:

- Optimal K for recall: **5**
- f1 scores of 0.4 (class “high”) and 0.8 (class “low”) and accuracy score of 0.7 from train_test_split.
- The model predicted that in 2023, heart disease mortality risk decreases in many **Southwest** states - where it used to be the highest.
- The model predicted that overall, the number of states with high risk for heart disease above nation’s average has **decreased** in 2023.

Result and Conclusion

Correlation

- Although there is a decreasing trend in the number of states with high risk of heart disease mortality, ozone concentration maps show an **increasing** pattern over the years in all states.
- Linear regression plot and calculated r-value (**-0.3324**) shows a **negative correlation** between ambient ozone concentration and heart disease mortality rate.

Improvements

- **Complete dataset** with data for every state, consistent between all years.
- Include data from a **larger time frame** for linear regression.
- Incorporate **more features** to training set - e.g. concentration of other types of air particles that are also suspected to have an impact on heart disease - to increase accuracy of prediction model.
- Obtain data from other countries.

Thank you for listening!