# Technical Deliverable

University of Virginia

STAT 4996 Capstone

Team M8
Bella Binder (imb6bwd)
Karmina Buensuceso (kob4zh)
Shamir Centeno Padilla (rns7bw)
Jiayi Niu (nwb4td)

December 8, 2023

# 1. Motivation

During the summer of 2023, smoke from wildfires in Canada negatively impacted the air quality across the United States. The air quality was visibly "very unhealthy," which was especially unusual for areas along the East Coast. Because we found this period of time so striking, we decided to conduct research on wildfires, air quality, and air quality maintenance strategies. Over the past four decades, wildfires have become a bigger problem. Across the U.S., wildfires have burned larger areas, accumulated higher costs, and caused more deaths (MacCarthy et al., 2023). Furthermore, the smoke from wildfires has been tied to increased air pollutants and decreased air quality. The causes for wildfires range from climate change to shifting populations (MacCarthy et al., 2023); therefore, we wanted to look into these factors' effects on air quality as well. Our objective was to investigate how different environmental factors contribute to air quality, and provide potential solutions to this ongoing problem.

# 2. Research Question

How do environmental factors (e.g., wildfires and their pollutants, biodiversity, weather, and population) influence air quality? Which factor is most influential on the number of days within a year with a good air quality index (AQI)? What action plans can be implemented to maintain a good AQI?

# 3. Data

*(Refer to References to view the sources of our data.)*

## 3a. Data Description

| Variable Type | Description |
|---|---|
| **Response Variables** | |
| Ratio of Good AQI Days | The percentage of days within a year measuring good AQI levels |
| Ratio of Bad AQI Days | The percentage of days within a year measuring unhealthy, very unhealthy, and hazardous AQI levels |
| **Explanatory Variables** | |
| Number of Plant and Animal Species | The number of plant and animal species |
| Number of Wildfires | The number of wildfires which took place |
| Acres Burned | The area burned in acres |
| Average Tree Cover Loss | The average number of Hectares of trees lost |
| Average Carbon Emissions | The gross carbon emissions in Mg (megagrams) CO2e (carbon dioxide equivalents) |
| Total Precipitation | The total precipitation in inches |
| Average Temperature | The annual average temperature in Fahrenheit |
| Population | The total population |
| Region | The region where the state resides (1 = Northeast, 2 = Midwest, 3 = South, 4 = West) |

Table 1: Description of Response and Explanatory Variables

## 3b. Data Sources

The response variables involving AQI days are both from the United States Environmental Protection Agency. For explanatory variables, the number of plant and animal species is from NatureServe. The number of wildfires which took place as well as the area burned in acres are from the Insurance Information Institute and Statista. The average number of Hectares of trees lost and the gross carbon emissions in Mg CO2e are from Global Forest Watch. The total precipitation in inches and the annual average temperature in

Fahrenheit are both from the National Centers for Environmental Information. Finally, the total population as well as the region where the state resides are both from the United States Census Bureau.

### 3c. Data Cleaning

For our final data set, we used a variety of statistical approaches to transform the original data into variables relevant to our research project.

The first variables we cleaned were our response variables (ratio of good days and ratio of bad days within 2021 or 2022). The Air Quality Index Report from the EPA included the number of days in which the EPA collected AQI information, and from those days, detailed when the AQI was in good, moderate, or unhealthy conditions by state. The manner by which we obtained our response variable was by dividing the number of good or bad days by the number of days an AQI was collected by the EPA. When we obtained that decimal, we then multiplied the values by one hundred to obtain the ratio of good or bad AQI days within the years for each state.

Another variable we used within our data contents was the average tree cover loss by hectares. The observations of tree cover loss are separated by percent canopy (0%, 10%, 15%, 20%, 25%, 30%, 50%, 75%). Because of this, for each state, we took the average across the thresholds to obtain the average tree cover loss and get a better understanding of whether the amount of tree density loss affects the ratio of good AQI days within the United States.

## 4. Evolution of Project

We changed the question of the student who nominated this project as a capstone topic. The question was: "How does tree density in selected areas affect AQI post natural disasters? Does plant species diversity have a positive effect on AQI, especially when considering recovery speed of an area post natural disaster?" Our evolved question is: How do environmental factors (e.g., wildfires and their pollutants, biodiversity, weather, and population) influence air quality? Which factor is most influential on the number of days within a year with a good air quality index (AQI)? What action plans can be implemented to maintain a good AQI?

We changed our response variable from average annual AQI to the ratio of days in which AQI was considered healthy or unhealthy, very unhealthy, and hazardous. We felt that the ratio was a better representation of a state's AQI because it had more variation. Additionally, in the United States, average annual AQI should be healthy, so not much information can be deduced from that.

We deleted some variables, specifically "tree cover percentage in 2017" and "total number of plant species in 2022" to perfectly match our 2021 and 2022 data. In addition, we also included some new variables after our initial progress presentation to deduce more information about whether those new environmental factors have a significant effect on the percentage of good AQI levels within a year.Likewise, we deleted the data for Alaska, Hawaii, and the District of Columbia due to the lack of valid data source for certain variables we attempted to investigate.

# 5. Analysis

## 5a. Exploratory Data Analysis

### Correlation Coefficients

```
##                                   Correlation Coefficient
## Ratio of Good AQI Days                          1.0000000
## Ratio of Bad AQI Days                          -0.5411318
## Number of Plant and Animal Species             -0.5615354
## Number of Wildfires                            -0.3204044
## Acres Burned                                   -0.4468585
## Average Tree Cover Loss                        -0.2418662
## Average Carbon Emissions                       -0.1741774
## Total Precipitation                             0.4590496
## Average Temperature                            -0.2567205
## Population                                     -0.3030448
## Region                                         -0.4550431
```

```
##                                   Correlation Coefficient
## Ratio of Good AQI Days                       -0.541131758
## Ratio of Bad AQI Days                         1.000000000
## Number of Plant and Animal Species            0.517944763
## Number of Wildfires                           0.319684679
## Acres Burned                                  0.705920617
## Average Tree Cover Loss                       0.486985169
## Average Carbon Emissions                      0.458492469
## Total Precipitation                          -0.468317693
## Average Temperature                          -0.005440308
## Population                                    0.230358803
## Region                                        0.410960273
```

We generated correlation coefficients to determine the linear relationship between each of our explanatory variables and our response variables: either Ratio of Good AQI Days or Ratio of Bad AQI Days. Although none of our correlations were particularly strong, we found a few that were moderate.

The Ratio of Good AQI Days had moderate negative linear correlations with Number of Plant and Animal Species and Acres Burned. The Ratio of Good AQI Days also had a moderate positive linear correlation with Total Precipitation.

The Ratio of Bad AQI Days had a moderate negative linear correlation with Total Precipitation. The Ratio of Bad AQI Days also had moderate positive linear correlations with Number of Plant and Animal Species, Acres Burned, Average Tree Cover Loss, and Average Carbon Emissions.

### Scatterplots

We also generated scatterplots to visualize the linear relationship between each of our explanatory variables and our response variables: either Ratio of Good AQI Days or Ratio of Bad AQI Days. The scatterplots confirmed the observations that we made from our correlation coefficients.

According to our Exploratory Data Analysis, the Ratio of Bad AQI Days seemed to be the better choice for our response variable because it showed stronger correlations with our explanatory variables. However, more diagnostic tests and assumption checks are needed to decide if Ratio of Bad AQI Days is an appropriate response variable for our project.

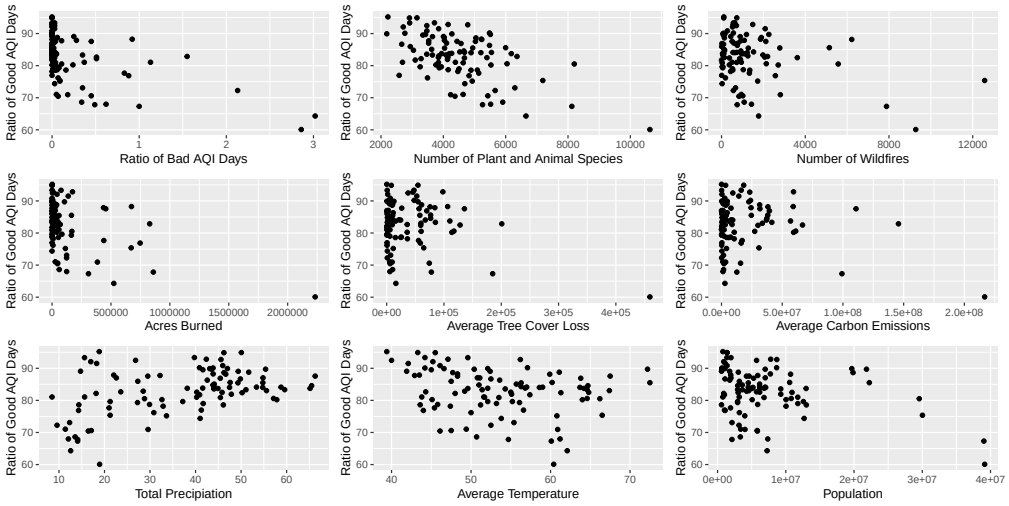


Figure 1: Scatterplot Matrix of Ratio of Good Days as Response

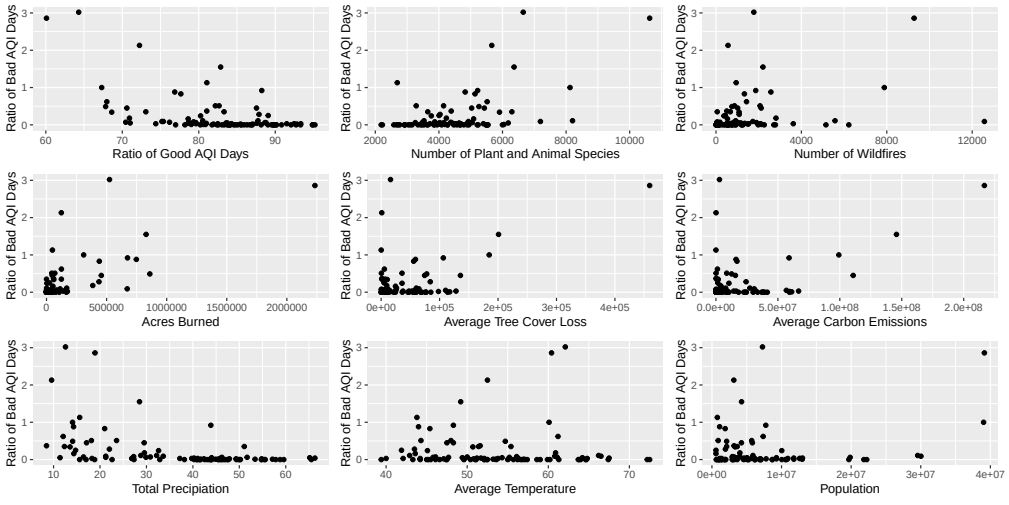


Figure 2: Scatterplot Matrix of Ratio of Bad Days as Response

## Boxplots

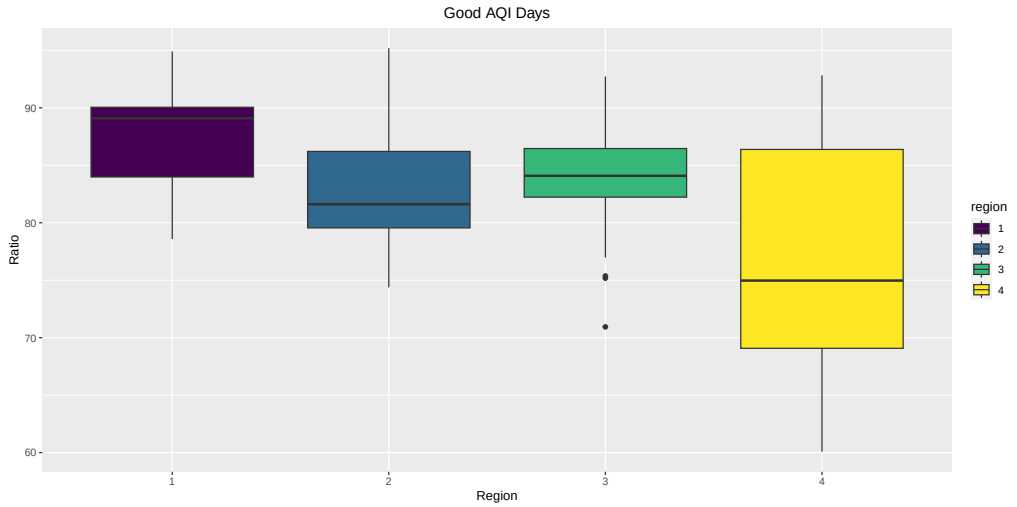


Figure 3: Boxplots of Ratio of Good Days vs Different Regions

Based on the boxplots of "Good AQI Days," the West had the largest range of ratios of good AQI days. The South had outliers displaying that that region had several quite low ratios of good AQI days. In addition to this, the Northeast had the highest median. Finally, the Midwest and the Northeast had almost the same maximum amount of good AQI days.

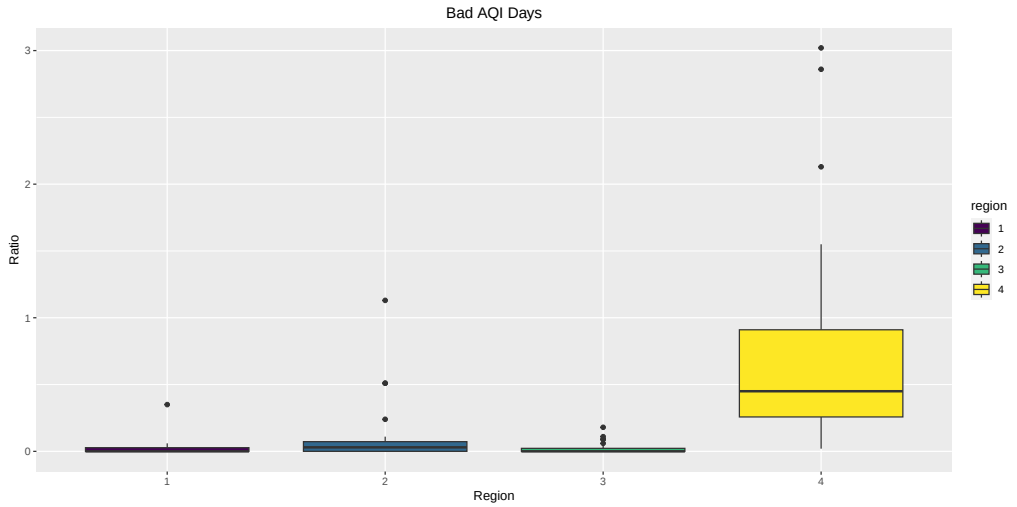


Figure 4: Boxplots of Ratio of Bad Days vs Different Regions

Based on the boxplots of "Bad AQI Days," the West had the largest range of ratios of bad AQI days. Additionally, the West had the highest median. All of the regions had outliers displaying that they had several high ratios of bad AQI days. However, the Midwest and West had the highest outliers of ratios of bad AQI days.

## 5b. Methods

Our analysis approach consisted of building a multiple linear regression model, where we gathered insight about which environmental factors have a significant effect on the percentage of good AQI levels in a year for each state in the United States. Since we combined the data for 2021 and 2022, we focused on building one model. As previously mentioned, we had two candidates for our response variable, which were "Ratio of Good AQI Days" and "Ratio of Bad AQI Days". The EDA shows that "Ratio of Bad AQI Days" is deemed to be a better option. We will then check the assumptions underlying linear regression to see if this decision is valid.

We first built a multiple linear regression model fitting the number of species, number of wildfires, acres burned, average tree loss, average carbon emissions, total precipitation, average temperature, population and region for the ratio of good AQI days.

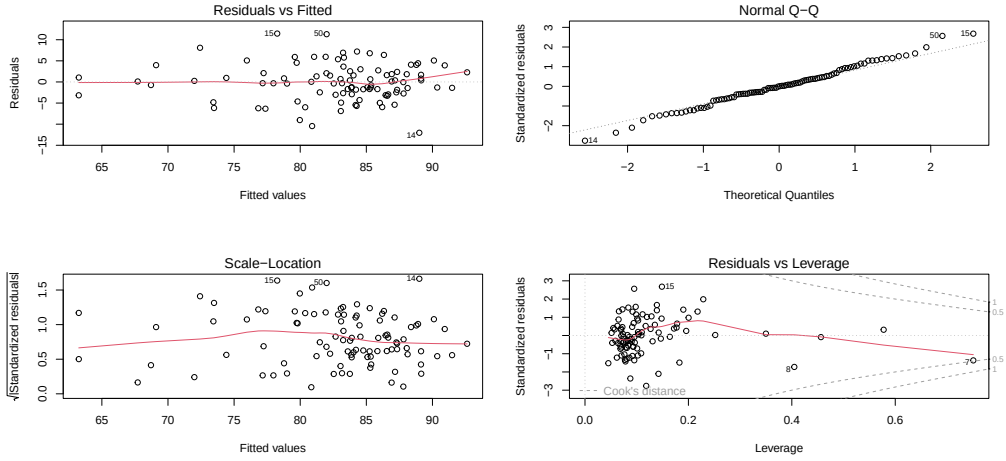


Figure 5: Diagnostic Plots of Ratio of Good Days as Response

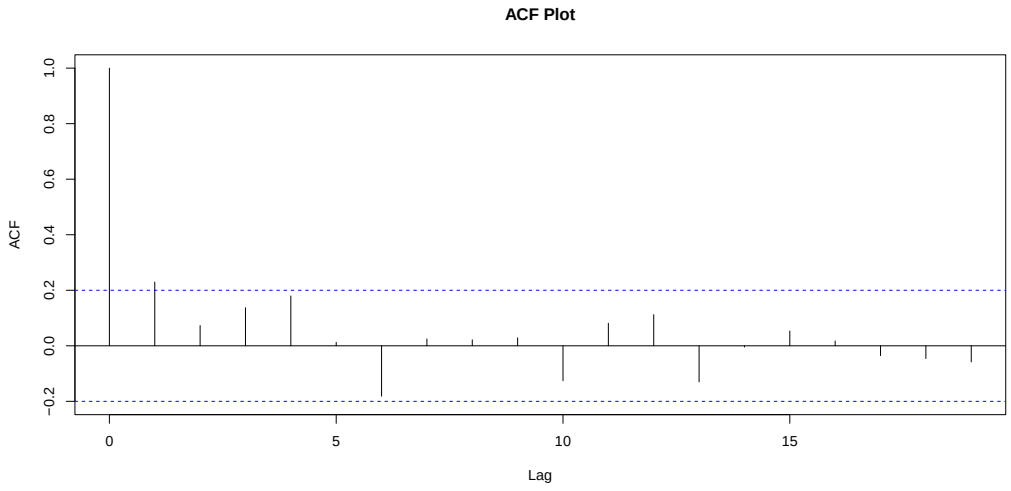


Figure 6: ACF Plot of Ratio of Good Days as Response

Based on the diagnostic plots, we were able to check the assumptions of the model and transform the variables if needed. The data points were approximately evenly scattered on both sides of the regression line, as we moved from left to right. We see this in the residuals vs fitted graph, the red line is approximately linear, indicating that the residuals have mean 0. Assumption of constant variance is met as the vertical spread of the data is constant, indicating constant variance. Assumption of independence is met as there is no significant correlation between variables on the ACF Graph for lag after 0. Assumption of normality is met as the data points approximately fall onto the qq line.

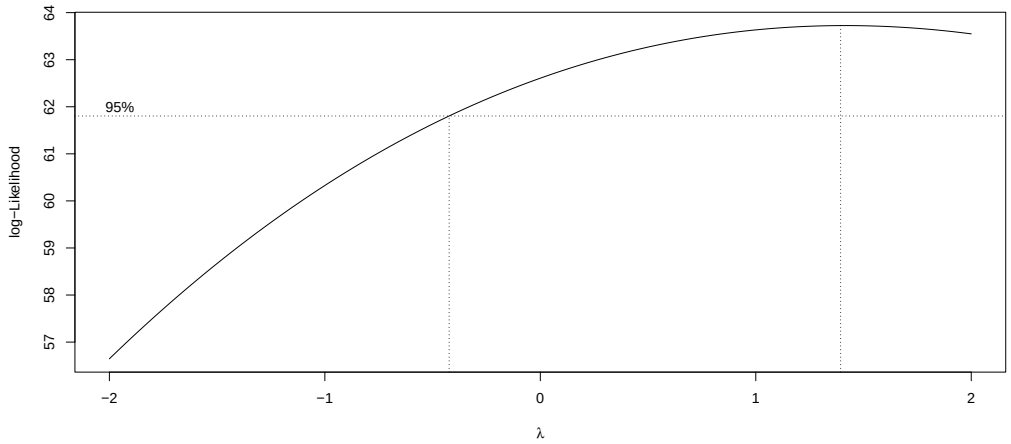


Figure 7: Box-Cox Log-Likelihood Plot Ratio of Good Days as Response

The Box-Cox is used to stabilize variance and make the data more closely follow a normal distribution, thereby satisfying the assumptions of linear regression models. The plot shows that 1 is inside the 95% CI for lambda value. It indicates that no transformation is needed, so the data already meets the assumptions of constant variance.

We then build a multiple linear regression model with the ratio of bad days as the response variable.

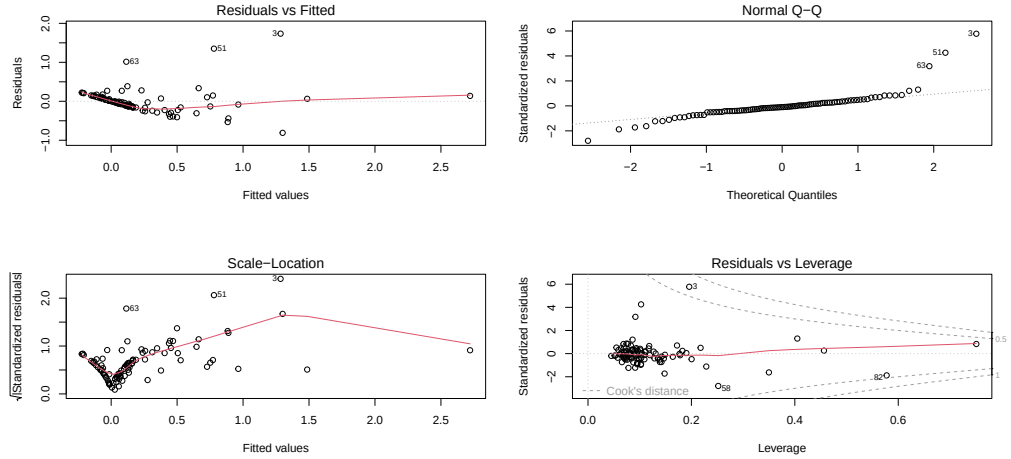


Figure 8: Diagnostic Plots of Bad Days as Response

In our analysis, we initially considered using "Ratio of Bad AQI Days" as the response variable for our linear regression model. However, diagnostic checks revealed that this choice violated key assumptions of linear regression, including linearity, constant variance, and normality of residuals. Given these violations, we eventually decided to switch our focus to "Ratio of Good AQI Days" as the response variable. Notably, the Box-Cox transformation indicated that lambda = 1 falls within the 95% Confidence Interval for 'Good Days Ratio,' suggesting that this variable likely already satisfies the assumptions of a linear regression model. This made "Good Days Ratio" a more suitable and statistically robust choice for our analysis, allowing for a simpler and more interpretable model. Therefore, we proceeded with "Good Days Ratio" as the response variable for subsequent modeling and evaluation.

For the "Ratio of Good Days" model, we created a VIF plot to see the effect and multicollinearity and if we would need to deal with it.

```
##     num_species   num_wildfires    acres_burned   avg_tree_loss avg_c_emissions
##        4.935435        3.502931        3.779817       22.101531       16.676190
##    total_precip        avg_temp      population         region2         region3
##        4.667115        3.535261        3.573182        2.688655        4.271246
##         region4
##        5.805900
```

A VIF value greater than 5-10 is generally considered indicative of high multicollinearity. "Average tree cover loss" and "average carbon emissions" have VIF values significantly above the common threshold of 5-10, indicating high multicollinearity. All other variables have VIF values below 5, suggesting moderate to low multicollinearity.

Multicollinearity can distort the estimates and reduce the interpretability of the model, making it challenging to discern the individual impact of each predictor. To address this, we opted for Ridge and Lasso regression techniques, both of which are regularization methods designed to handle multicollinearity effectively and reduce variance. Ridge regression adds a penalty term to the least squares loss function, effectively shrinking the coefficients towards zero but not eliminating any variable within our model. This method is particularly useful when we have reason to believe that all the predictors should be included in the model for their substantive importance. Lasso (Least Absolute Shrinkage and Selection Operator) regression shrinks the coefficients towards zero and also sets some coefficients to zero, effectively performing variable selection. This is advantageous when we suspect that the model may be overfitted with unnecessary variables. Both Ridge and Lasso regression methods offer a balance between bias and variance, thereby improving the model's performance when multicollinearity is present. Therefore, we decide to resume our analysis through multiple linear regression, ridge regression, and lasso regression.

In our data, we have collected observations from both 2021 and 2022, which we used to train and test our model. In our preliminary EDA, we noticed that there were some major similarities between the data we gathered from 2021 and 2022. However, there were also some discrepancies between variables such as total annual precipitation which resulted in being highly significant with 2021 data but less significant when using 2022 data. Because of this challenge, we implemented a k-fold cross validation procedure to refit the model (Linear Regression, Ridge Regression, Lasso Regression) with various training and testing sets formed from the entire data, in order to obtain an accurate estimate of the fitted model. We found the optimal regularization parameter for ridge and lasso regression. In addition, since the observations were randomly divided into k groups (set a seed), we assessed which model is the most optimum when comparing metrics such as MSE and reduced variance for these three models. With this approach, we were certain that the cross validation technique will counteract the discrepancies between the 2021 and 2022 data and also avoid overfitting to determine a more accurate estimate of model prediction performance.

## 5c. Model Building Process

In response to the presence of multicollinearity within our model, we decided that the first statistical learning methods we wanted to initially fit and analyze were both ridge and lasso regression. These two shrinkage methods, although very similar to one another, offer an opportunity to retain all of the predictor variables within the model without any reduction of variable interpretation (11 predictor variables). This is advantageous, since we want to initially observe if we can mitigate the variance by introducing a penalty term, and also consider whether there are any alarming drawbacks such as missed coefficient interpretation and increase in bias.

To commence our analysis, we started by splitting the population data set into a training and testing set, adopting a 50/50 split and ensuring randomization for both sets to maintain their representativeness. Next, we followed up by carrying out a 10-fold cross validation procedure. This process served a dual purpose: determining the optimal tuning parameters and eliminating discrepancies between the observations by allowing 10-fold groups to be fitted as test data. Below, we can observe the best tuning parameter for both shrinkage methods by plotting them against the test error.

**Ridge Regression**

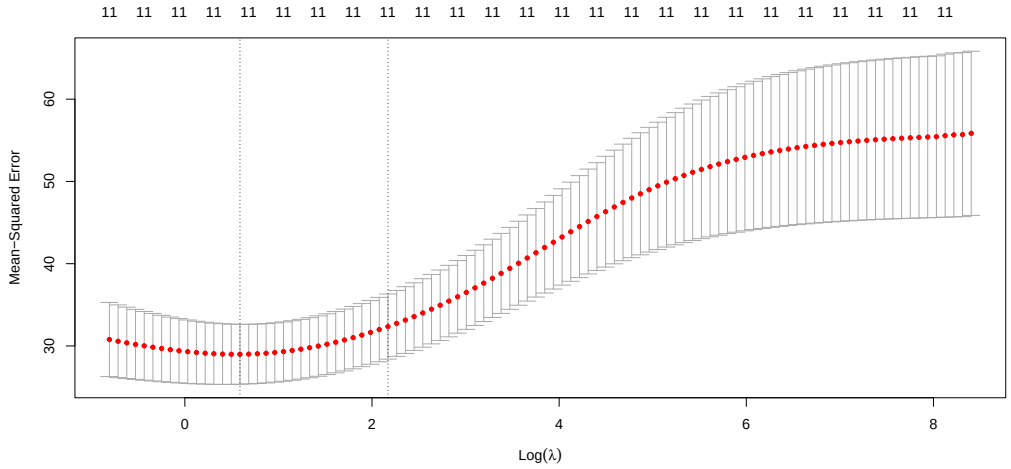


Figure 9: Mean-Squared Error vs Log of Ridge Regression

**Lasso Regression**

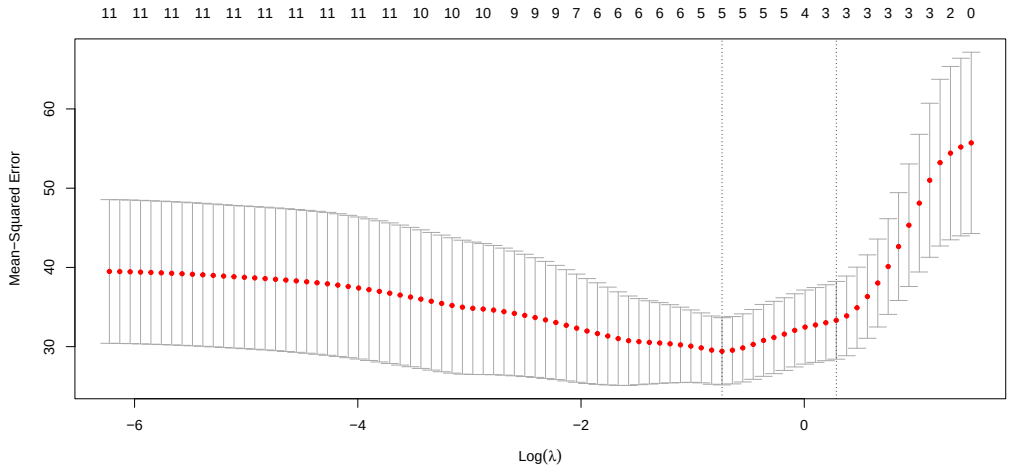


Figure 10: Mean-Squared Error vs Log of Lasso Regression

Once we found the optimum tuning parameter for both shrinkage methods, we rift both of the models, and interpreted each of the coefficients from both models. Due to the nature of Lasso Regression, the learning method led to the regularization of five predictor coefficients, effectively setting them to zero. This characteristic of Lasso allows for a simple model with only a subset of most influential predictors.

Conversely, In Ridge Regression, the majority of the coefficients shrinked to be near zero, when compared to the coefficients of the OLS model. This shrinkage method showcased the ability to mitigate the multicollinearity-induced coefficients that were observed in OLS.

In assessing the predictive performance among all the models, we aimed to determine accuracy by analyzing the test Mean Squared Error (MSE) of all three models. This is crucial within our model building process, since we want to ascertain which statistical learning method is most ideal to answer our research question.

The ridge regression model resulted in having the best predictive performance among the models tested. This can be concluded from the lower test MSE that was retrieved from the ridge regression model. The OLS regression model, which is equivalent to ridge regression when tuning parameter is set to zero, resulted in having a test MSE of 31.07365 with the same cross validation procedure. This test MSE was the highest

| Model | TestMSE |
|---|---|
| Ridge (10 k-fold) | 29.17949 |
| Lasso (10 k-fold) | 30.34487 |
| OLS Regression (10 k-fold) | 31.07365 |

Table 2: Comparison of Model Test MSE

among the three models, suggesting that without any regularization, the OLS model may be overfitting the training data and thus performing worse on the test data set. However, according to the test MSE results from the shrinkage methods, it can be argued that the reduction in test error is relatively insignificant. In addition, we reduced the variance at the expense of introducing bias, thus leading to a less accurate and valid interpretation for the coefficients. Because of this, we decided to proceed with the OLS for better interpretability, rather than predictive performance, and operate the stepwise regression, with the goal of further improving the test error.

**The Estimated Ordinary Least Regression Equation**

$\hat{y} = 105.7853 - 0.00244 * NumberofPlantandAnimalSpecies + 0.0005 * NumberofWildfires - 0.0000009 * AcresBurned + 0.000007 * AverageTreeCoverLoss - 0.0000001 * AverageCarbonEmissions + 0.24 * TotalPrecipitation - 0.43 * AverageTemperature - 0.0000001 * Population - 2.1 * MidwestRegion + 2.5 * SouthRegion + 1.1 * WestRegion$

We chose to use stepwise regression due to the need to refine the model developed through Ordinary Least Squares (OLS) regression. While OLS regression is beneficial for its simplicity and provision of unbiased estimates, it inherently includes all available predictors. This inclusion can lead to a model that is overly complex, potentially suffering from issues like overfitting or multicollinearity, where the large number of variables obscures the true relationships in the data. Stepwise regression's process of adding and removing variables allows for the construction of a model that retains only those variables that have a meaningful impact on the dependent variable.

To accomplish this, we conducted a both forward and backward stepwise regression and analyzed certain important metrics such as Residual Sum of Squares (RSS), which represents the sum of the squares of the residuals, indicating the unexplained variance by the model and the Akaike Information Criterion (AIC), which represents the balances between complexity and fit of the predictor variables.

```
## Step:  AIC=303.32
## good_days_ratio ~ num_species + total_precip + avg_temp + num_wildfires +
##     region + avg_c_emissions + acres_burned
##
##                   Df Sum of Sq    RSS    AIC
## <none>                          1836.4 303.32
## - num_wildfires    1     39.83 1876.2 303.38
## + avg_tree_loss    1     15.49 1820.9 304.50
## + population       1     11.39 1825.0 304.72
## - region           3    149.12 1985.5 304.81
## - acres_burned     1    109.84 1946.2 306.89
## - total_precip     1    119.66 1956.0 307.38
## - avg_c_emissions  1    154.18 1990.6 309.05
## - num_species      1    280.67 2117.1 314.97
## - avg_temp         1    453.13 2289.5 322.49
```

```
##
## Call:
## lm(formula = good_days_ratio ~ num_species + total_precip + avg_temp +
##     num_wildfires + region + avg_c_emissions + acres_burned,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5032  -2.2920  -0.2264   2.5758  12.1033
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.142e+02  5.271e+00  21.670  < 2e-16 ***
## num_species     -2.445e-03  6.744e-04  -3.626 0.000488 ***
## total_precip     1.578e-01  6.667e-02   2.367 0.020164 *
## avg_temp        -5.193e-01  1.127e-01  -4.607 1.41e-05 ***
## num_wildfires    5.542e-04  4.057e-04   1.366 0.175576
## region2         -1.533e+00  1.779e+00  -0.862 0.391356
## region3          3.430e+00  1.838e+00   1.866 0.065398 .
## region4         -2.161e+00  2.572e+00  -0.840 0.403192
## avg_c_emissions  7.244e-08  2.696e-08   2.687 0.008652 **
## acres_burned    -6.509e-06  2.870e-06  -2.268 0.025830 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.621 on 86 degrees of freedom
## Multiple R-squared:  0.6317, Adjusted R-squared:  0.5932
## F-statistic: 16.39 on 9 and 86 DF,  p-value: 2.484e-15
```

Following the stepwise regression, our comprehensive analysis led to the conclusion that 9 out of the original 11 predictor variables were deemed significant, due to their AIC value, and were retained in the model. These variables included:

1) Number of Wildfires
2) Region (dummy variables with Midwest, South and West)
3) Acres Burned
4) Total Precipitation
5) Average Carbon Emissions
6) Number of Plant and Animal Species
7) Average Temperature

In our analysis, a discrepancy was noted regarding the region variable in the stepwise regression compared to the final model summary. Such discrepancies can occur due to the nature of stepwise regression, which selects variables based on their impact on the AIC value. The region variable's individual categories may not significantly change the AIC during the stepwise process but appear significant in the final model. This does not compromise the model's validity; rather, it underscores the complexity of model selection methods. In our report, we focus on the implications of the final model, which incorporates region2, region3, and region4, indicating their significant contribution to predicting the good AQI days ratio.

The model's performance was further validated through key metrics. The Residual Sum of Squares (RSS) yielded a value of 1836.4, indicating the unexplained variance by the model. Additionally, the Akaike Information Criterion (AIC) was calculated to be 303.32. These results collectively suggest a commendable balance between model complexity and fit, reinforcing the robustness and effectiveness of the final model in capturing essential relationships within the data.

**The Final Regression Equation**

$$\begin{aligned}\hat{y} = 114.2 &- 0.0024 \times \text{Number of Plant and Animal Species} \\ &+ 0.158 \times \text{Total Precipitation} \\ &- 0.519 \times \text{Average Temperature} \\ &+ 0.000554 \times \text{Number of Wildfires} \\ &- 1.533 \times \text{Midwest Region} \\ &+ 3.430 \times \text{South Region} \\ &- 2.161 \times \text{West Region} \\ &- 0.00000007 \times \text{Average Carbon Emissions} \\ &- 0.000006 \times \text{Acres Burned} \end{aligned} \tag{1}$$

# 6. Results and Interpretations

**Observations from the Stepwise Process**

In the stepwise regression process, several key observations have helped us understand the significance of various variables in explaining the variation in the "Ratio of Good AQI Days." First and foremost, the Akaike Information Criterion (AIC) established that the model kept variables that indicated statistical significance or contributed to an improved model fit. Furthermore, the number of wildfires was found as a variable whose removal resulted in a modest rise in the AIC. This implies that, while its contribution to the model is minor, it is however relevant and so has remained in the model. Certain factors, such as average tree cover loss and population, were, on the other hand, removed from the final model. This was based on the understanding that their inclusion would have led to an increase in the AIC, indicating that they did not contribute significantly to explaining the "Ratio of Good AQI Days." Furthermore, the impact of removing variables like average temperature, number of species, total precipitation, average temperature, acres burned, and average carbon emissions were substantial, as evidenced by the significant increase in the AIC.

These variables also demonstrated significant p-values. The total precipitation is significant with a p-value substantially below 0.05, suggesting a potent positive influence on air quality indicators. In contrast, the negative correlation of number of plant and animal species and the positive correlation of average carbon emissions with good AQI days, significant at the 5% and 1% levels respectively, indicating that less biodiversity and more carbon emissions–counterintuitively–are associated with more good AQI days within the context of this model. Average temperature presents a negative association, also highly significant, suggesting that higher temperatures may adversely affect the ratio of good AQI days. Meanwhile, acres burned is significant at the 5% level, reinforcing the notion that larger areas affected by fires could deteriorate air quality. The model exhibits a high level of explanatory power with an Adjusted R-squared value of 0.5932, indicating that a significant portion of the variance in the Ratio of Good AQI Days is captured by the variables included. Overall, the F-statistic's p-value approaches zero, further affirming the overall significance of the model.

## 6a. Model Comparison Between OLS and Stepwise Regression

In the analysis, two distinct modeling approaches were employed to gain insights into the factors influencing the "Ratio of Good AQI Days." Initially, an Ordinary Least Squares (OLS) model was constructed and it included all available predictors we have through a 10-fold cross validation procedure. However, as we mentioned above, this inclusive approach carries the risk of potential overfitting or multicollinearity issues, where the model might become too complex or sensitive to noise in the data. To address the concerns, we then utilized stepwise regression. The stepwise regression would strike a balance between model complexity and predictive power, ensuring that only the most relevant variables were retained. Both modeling approaches yielded consistent findings, identifying several variables, such as number of plant and animal species, total precipitation, and average temperature, as significant predictors of influencing air quality. Notably, stepwise

regression shed light on the role of number of wildfires, suggesting that it may have a marginal impact on the model. This observation was less obvious in the OLS model, which by default retained all predictors.

| Model | TestMSE |
|---|---|
| Reduced OLS Regression (10 k-fold) | 25.74997 |
| Ridge (10 k-fold) | 29.17949 |
| Lasso (10 k-fold) | 30.34487 |
| OLS Regression (10 k-fold) | 31.07365 |

Table 3: Comparison of Model Test MSE

Comparing the reduced model from the complete model and both shrinkage methods, the reduced model resulted in having a significant lower test MSE compared to all three other models. This suggests that the reduced model obtained from Stepwise regression is the most efficient in predicting the ratio of good AQI days and also allows the most room for interpretation, given that the relationship between the environmental factors and the response variable are much more suitable. Thus, the stepwise approach provided a more comprehensive understanding of the variable contributions to air quality prediction.

## 6b. Coefficients Interpretation

*(Bolded variables are labeled as significant as having a p-value less than 0.05.)*

1) Positive Coefficients:

- Number of Wildfires, surprising as we would expect more wildfires to degrade air quality
- **Average Carbon Emissions**, surprising as we would expect more carbon emissions to degrade air quality
- **Total Precipitation**, expected as rain can clear airborne pollutants
- **South Region**, expected as it is the western part of the US and it incorporates California which is the state that has the most number of wildfires

2) Negative Coefficients:

- **Number of Plant and Animal Species**, surprising as we would expect more biodiversity to increase air quality
- **Acres Burned**, expected as larger wildfires lead to fewer good AQI days
- **Average Temperature**, expected as hotter days lead to possibility of wildfires, thus decreasing AQI days
- Population, expected as higher pollution with larger populations
- South Region and West Region reflect regional differences

## 6c. Suprises

- The signs of the coefficients for the models are somehow different from our correlation matrix. We think the reason might be that the correlation matrix only shows bivariate relationships, not accounting for the influence of other variables.
- The positive association of number of wildfires, average carbon emissions with good AQI days in the model is surprising and counterintuitive. These findings may suggest that the models are capturing complex interactions or that there are missing variables that could explain these relationships better.
- The negative association of number of plant and animal species with good AQI days in the model is also unexpected.

# 7. Conclusion

In conclusion, the analysis of environmental factors, including wildfires and their pollutants, biodiversity, weather, and population, has shed light on their intricate influence on air quality. Our study revealed that several variables significantly impact the number of days within a year with a good air quality index (AQI). **Notably, number of species, total precipitation, average temperature, acres burned, and average carbon emissions were identified as key factors affecting air quality**. This determination is based on the fact that all of these variables exhibit p-values below the commonly accepted significance threshold of 0.05. The unexpected positive associations of variables like number of wildfires and average carbon emissions with good AQI days suggest complex interactions at play, which could be due to the influence of other unaccounted-for variables or intricate relationships within the dataset.

As we answered our research question, it is pivotal to delve into some practical implications regarding air quality control. Based on the insights gained from our analysis, we come up with some recommendations to help guide efforts in improving and maintaining good air quality. Given the importance of the number of species, which is biodiversity, as highlighted in our study, it would be better to encourage the development of green spaces, such as parks, community gardens, and green corridors in urban areas. Implementing conservation programs for protecting and restoring habitats should also be considered. Also given the statistically significant positive impact of total precipitation on air quality, efforts to preserve and enhance precipitation patterns should be considered. This might involve strategies to mitigate drought conditions and ensure that regions receive adequate rainfall to help clear airborne pollutants. In addition, in regions susceptible to wildfires, addressing rising temperatures is important. Action plans should include measures to prevent wildfires, such as controlled burns and firebreaks, or like urban planning strategies that enhance heat resilience. Furthermore, the extent of land affected by wildfires, represented by acres burned, also plays a crucial role in air quality. Effective wildfire prevention and management strategies like increasing firefighting capabilities, implementing controlled burns, and educating the public about fire safety can be considered as well. Finally, the surprising finding that average carbon emissions have a statistically significant positive impact on air quality should be explored further. However, this does not necessarily mean more pollution leads to better air quality, as we mentioned before, the combined effects or the intricate interplay is not captured in the regression model. We should still consider reducing carbon emissions should remain a priority in environmental policies.

Another important insight is that we have the South Region being slightly significant. The positive coefficient associated with this region indicates that it tends to have more good air quality days compared to other regions. It's essential to consider the unique environmental and climatic factors of the South, such as humidity and temperature. Practical implications for air quality control apply to all regions, but it's important to acknowledge that the parameters and challenges may vary from one region to another.

In sum, our research contributes to the broader goal of environmental health and public health by fostering clean and healthy air for all. By paying attention to the significant factors identified in our analysis, we hope that people can work towards a future with improved air quality, benefiting society as a whole.

# 8. Considerations and Limitations

When looking at our results and interpretations, it is important to also consider their limitations.

In our project, we can identify correlations, but we cannot definitely conclude any causation. There are two reasons for this. First, the environment has a lot going on at once (e.g., animal populations, sunlight, mining), so the factors that we research can be confounded by those that we do not. For this same reason, there are also possibilities of bidirectional causation. For example, AQI may affect the number of plant and animal species and temperature instead of the other way around.

Additionally, our data is extremely limited. Because our comprehensive data set was dependent on pre-published, public data, it was difficult to find exactly what we were looking for. For example, many of the factors that we were interested in were not measured recently (e.g., tree density is only measured every five years), and it was difficult to align annual data as factors were available for different years (e.g., number of

plant species in 2022, tree density in 2017). Also, most publicly available environmental data is measured at a country-level as opposed to a state-level, which is how we chose to index our data set. Considering these reasons, we only have data from 2021 and 2022, so our model may not be as accurate as we had hoped. Many of our factors can fluctuate from year-to-year, so data from a longer period of time would help create a more accurate model.

It may be because of this that we found tree cover loss to be insignificant and number of plant and animal species to be negatively correlated with the annual ratio of days with good AQI despite past research saying otherwise. According to a 2015 publication from the Convention on Biological Diversity and the World Health Organization, plant cover and diversity are important for improved air quality. Increased air temperature leads to increased air pollution, but different trees and vegetation cool the air through evaporation and shade. Furthermore, most plants clean the air by removing air pollutants through their leaves and surface. "[I]n areas with 100% tree cover (i.e. contiguous forest stands),... short term improvements in air quality (1 hour) [is] as high as 16% for O3 and SO2, 13% for particulate matter, 8% for NO2, and 0.05% for CO" (Nowak et al., 2015).

# 9. Future Work

We have four schemes for future work. Our first scheme involves pollutants' details. One of our quantitative, explanatory variables is "Average Carbon Emissions". If we go into detail about the amount of pollutants in the air besides carbon dioxide, we can see multiple pollutants contributing to AQI. According to the National Weather Service, "EPA calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide" (National Weather Service).

Our second scheme involves environment types. One of our qualitative, explanatory variables is "Region". We think we should have a qualitative, explanatory variable that is for the environment type of each state. Examples of an environment type for states would be "Desert" or "Forest". If we go into detail about environment types of states, we can see how the environments contribute to the amount of wildfires.

Our third scheme involves water coverage. One of our quantitative, explanatory variables is "Total Precipitation". We could use the United States Geological Survey's table on states' water area. The table involves square miles of water and water coverage percentage. If we go into detail about the amount of water coverage by state, we can see how the water coverage contributes to the amount of wildfires.

Our final scheme involves years. We could expand on the years of the data, thus the model will be accurate.

# 10. References


1. Burgueño Salas, E. (2023, August 7). Number of wildfires in the United States in 2022, by state. Statista.https://www.statista.com/statistics/1269724/number-of-us-wildfires-by-state/
2. Burgueño Salas, E. (2023, August 15). Acres burned by wildfires in the United States in 2022, by state. Statista. https://www.statista.com/statistics/217072/number-of-fires-and-acres-burned-due-to-us-wildfires/
3. Global Forest Watch. (n.d.). United States deforestation rates and statistics. Retrieved September 12, 2023, from https://www.globalforestwatch.org/dashboards/country/USA/?category=forest-change&map=eyJjYW5Cb3VuZCI6dHJ1ZX0%3D
4. Insurance Information Institute. (n.d.). Archived tables. Retrieved September 10, 2023, from https://www.iii.org/table-archive/23284
5. MacCarthy, J., Richter, J., Tyukavina, S., Weisse, M., & Harris, N. (2023, August 29). The latest data confirms: Forest fires are getting worse. World Resources Institute. https://www.wri.org/insights/global-trends-forest-fires#:~:text= Both%20the%20annual%20cost%20and,over%20the%20past%20four%20decades.
6. National Centers for Environmental Information. (n.d.). Climate at a glance statewide mapping. Retrieved September 19, 2023, from https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/mapping/110/pcp/202112/12/value
7. National Weather Service. (n.d.). Air quality index. Retrieved November 25, 2023, from https://www.weather.gov/safety/airquality-aqindex#:~:text=EPA%20calculates%20the%20AQI%20for,sulfur%20dioxide%2C%20and%20nitrogen%20dioxide
8. NatureServe. (2021). Species at risk [map]. Retrieved October 19, 2023, from https://www.natureserve.org/sites/default/files/NatureServe_AnnualReport_2021_map.pdf
9. NatureServe. (2023). Biodiversity in focus: United States edition. Retrieved October 19, 2023, from https://www.natureserve.org/bif
10. Nowak, D. J., Jovan, S., Branquinho, C., Augusto, S., Ribeiro, M. C., & Kretsch, C. E. (2015, June 3). Biodiversity, air quality and human health. In Convention on Biological Diversity & World Health Organization (Eds.), Connecting global priorities: biodiversity and human health: a state of knowledge review (pp. 63-74). https://www.who.int/publications/i/item/9789241508537
11. United States Census Bureau. (n.d.). Census regions and divisions of the United States [map]. Retrieved September 19, 2023, from https://www2.census.gov/geo/pdfs/maps-data/maps/ reference/us_regdiv.pdf
12. United States Census Bureau. (n.d.). State population totals: 2020-2022. Retrieved September 21, 2023, from https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html
13. United States Environmental Protection Agency. (n.d.). Air quality index report. Retrieved September 5, 2023, from https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report