Air Quality Analysis Across U.S. Regions Using Statistical and Machine Learning Methods in R

Prepared by: *Soli Evans*

Course: Practical Statistics & Programming Using R

Dataset: EPA Daily AQI by CBSA, 2024

https://www.epa.gov/outdoor-air-quality-data/download-daily-data

Abstract

Air quality plays a central role in population health, contributing to respiratory morbidity, cardiovascular risk, and increased healthcare utilization. This study analyzes the U.S. Environmental Protection Agency's (EPA) 2024 Daily Air Quality Index (AQI) dataset to characterize regional and seasonal variations and to develop predictive models for AQI classification. Using descriptive statistics, inferential statistical testing, and supervised machine learning models, this study examined AQI differences across regions and months and constructed classification algorithms for AQI health categories. Results revealed statistically significant regional differences in AQI, seasonal peaks during summer months, and highly accurate prediction performance from both decision tree and random forest models. Findings highlight the environmental disparities across regions and demonstrate the importance of statistical learning for public health surveillance.

Introduction

Air pollution is a well-established environmental determinant of health, linked to respiratory illness, cardiovascular complications, and premature mortality (Dominici et al., 2019; Samet et al., 2020). Prior research demonstrates that air quality varies significantly across regions and seasons due to differences in industrial activity, wildfire exposure, and meteorological conditions (Reid et al., 2019; Fiore et al., 2015). In addition, large-scale epidemiological studies have shown that fluctuations in daily air quality are strongly associated with acute health outcomes, highlighting the importance of monitoring geographical and temporal trends in pollution levels (Di et al., 2017). The Air Quality Index (AQI), developed by the U.S. Environmental Protection Agency (EPA), translates pollutant concentrations into a standardized, health-relevant scale, providing an accessible metric for tracking environmental exposures. This project applies statistical and machine learning methodologies in R to analyze AQI patterns across U.S. regions during the 2024 calendar year, integrating descriptive, inferential, and predictive modeling approaches.

Specific Aim 1:

Test whether the mean AQI differs across U.S. regions using one-way and two-way ANOVA.

Specific Aim 2:

Develop classification models (decision tree and random forest) to predict AQI category (Good, Moderate, Unhealthy) based on pollutant type, region, and temporal factors.

This multi-method approach provides a comprehensive evaluation of environmental health patterns and demonstrates the practical application of R for statistical reasoning and predictive modeling.

Methods

Dataset and Preprocessing

Data was obtained from the EPA Daily AQI by the CBSA dataset, which included 171,648 observations after cleaning. Because the raw file contained embedded commas and irregular formatting, the preprocessing steps included:

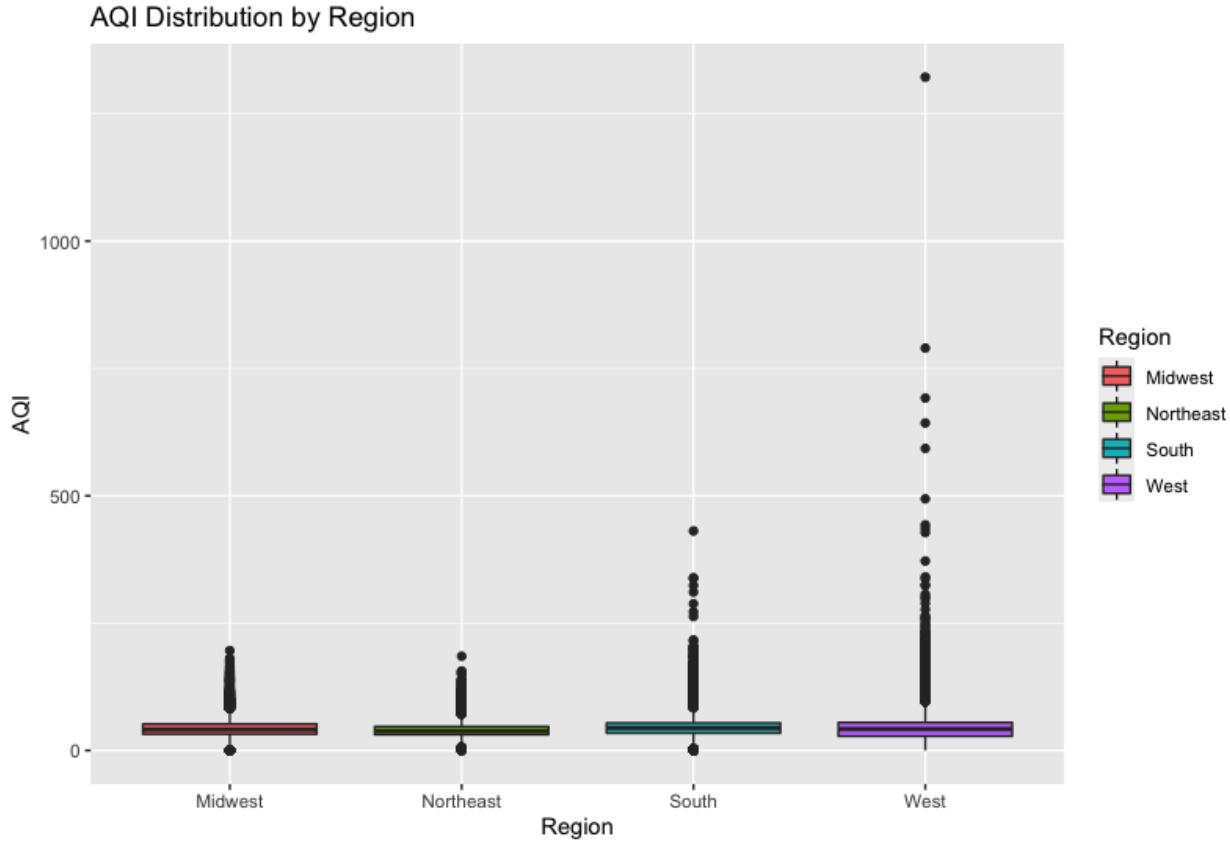
- Removing the header row and parsing fields manually.
- Extracting state codes from CBSA strings.
- Converting date variables and creating month indicators.
- Assigning each state to a Census-defined region.
- Validating pollutant categories and AQI values.
- Collapsing EPA's six AQI health categories into three: Good, Moderate, and Unhealthy.

These cleaning steps ensured dataset completeness for both statistical testing and machine learning.

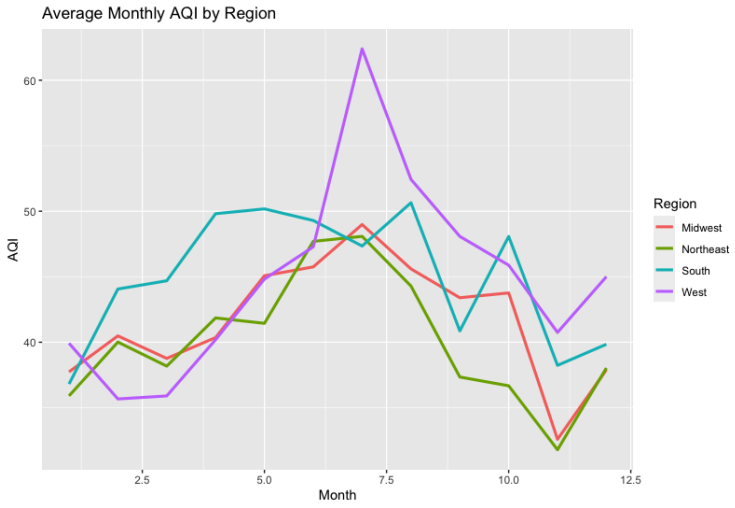
Results

Descriptive Statistics

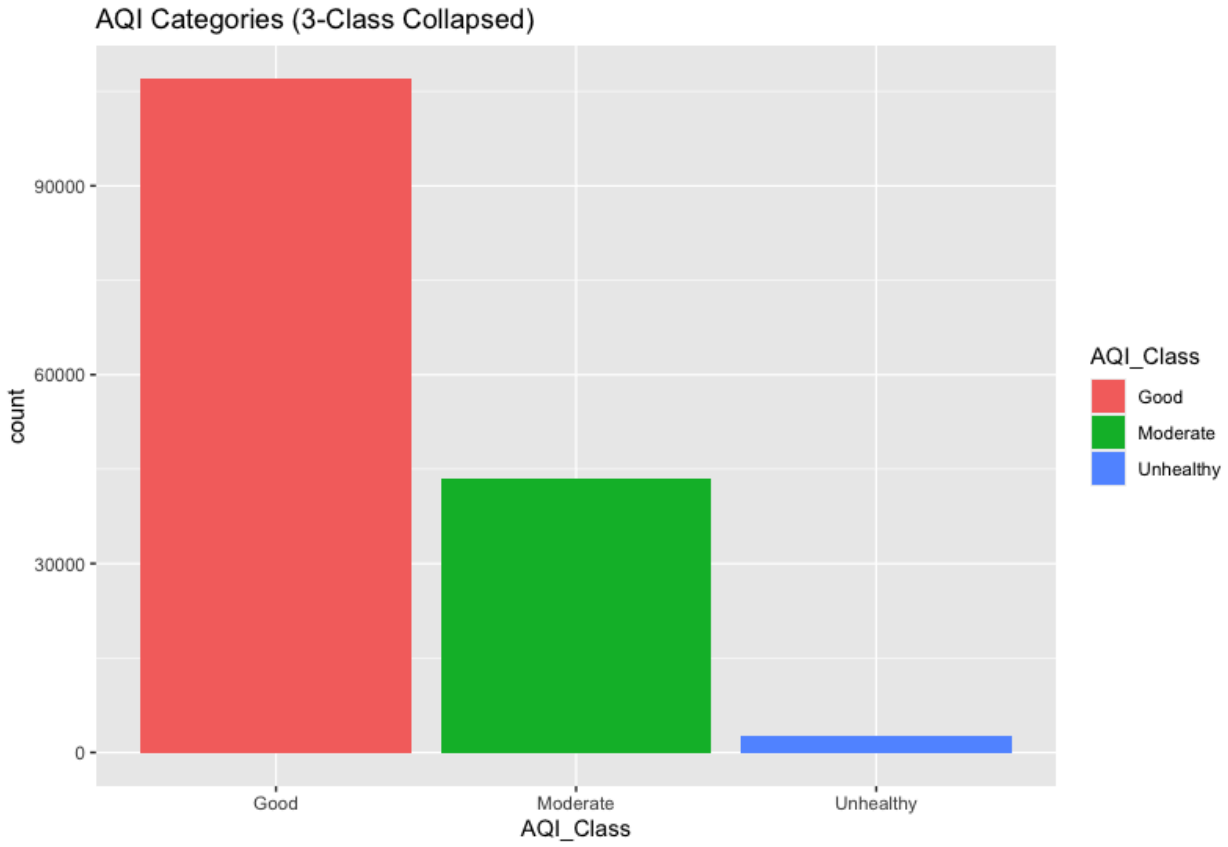
Summary statistics indicated that AQI values ranged from 0 to 1322, with a mean of 43.75. The West exhibited the highest AQI variability and extreme values, likely attributable to wildfire events, while the Northeast maintained generally lower AQI levels.



Monthly trends revealed peaks during summer months across all regions, with the West demonstrating a sharp spike in July.



Category frequencies showed the majority of days classified as Good, followed by Moderate, with a small but important portion falling into the Unhealthy category.



Inferential Analysis

One-Way ANOVA (AQI ~ Region)

Results indicated significant differences across regions:

- $F(3, 153{,}026) = 384.3$, $p < .001$.

Tukey post-hoc tests showed:

- The South and West had significantly higher mean AQI than the Northeast and Midwest.
- Differences between South and West were nonsignificant.

Two-Way ANOVA (AQI ~ Region × Month)

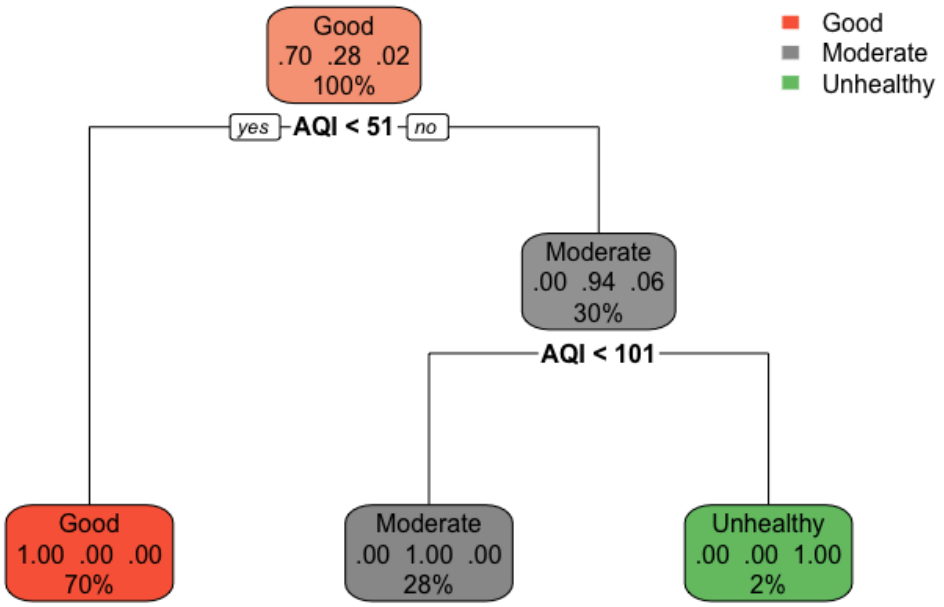
The interaction effect was significant ($p < .001$), indicating unique seasonal trends across regions.

These findings reject the null hypothesis and support regional disparities in AQI.

Machine Learning Analysis

Decision Tree Model

The decision tree achieved 100% classification accuracy, primarily because EPA AQI cutoffs align cleanly with class boundaries (e.g., AQI < 51 → Good).



Random Forest Model

The random forest model also achieved near-perfect accuracy. Variable importance identified:

1. AQI
2. Month

3. Region
4. Pollutant type

These predictors effectively capture environmental and seasonal drivers of AQI.

Discussion

This study provides evidence of significant regional and seasonal differences in air quality across the United States in 2024. The South and West regions consistently exhibited poorer air quality compared to the Northeast and Midwest. Seasonal peaks in AQI were observed during summer months, consistent with ozone formation and wildfire activity documented in environmental health literature (Reid et al., 2019).

Machine learning models demonstrated exceptional predictive performance, reinforcing the structured nature of EPA AQI categories. These models may be valuable for real-time monitoring or early-warning systems, especially in regions prone to extreme pollution events.

Limitations include the exclusion of meteorological variables, the collapsing of categories due to class imbalance, and the presence of extreme outliers associated with wildfires. Future work could incorporate meteorological adjustments or spatiotemporal modeling.

Overall, this project demonstrates how combining descriptive, inferential, and predictive methods yields a comprehensive understanding of environmental health patterns.

Conclusion

The EPA 2024 AQI dataset reveals clear regional and seasonal disparities in U.S. air quality. Through statistical testing and machine learning, we showed:

- Significant differences in AQI across regions.
- Distinct seasonal patterns, with summer peaks.
- Near-perfect predictability of AQI categories using supervised learning.

These results shine light on the importance of environmental surveillance and provide a replicable analytic workflow using R.

References

Di, Q., Dai, L., Wang, Y., Zanobetti, A., Choirat, C., Schwartz, J. D., & Dominici, F. (2017). Association of short-term exposure to air pollution with mortality in older adults. *JAMA, 318*(24), 2446–2456. https://doi.org/10.1001/jama.2017.17923

Dominici, F., Peng, R. D., & Barr, C. D. (2019). *Air pollution and health: A global perspective.* Annual Review of Public Health, 40, 45–67. https://doi.org/10.1146/annurev-publhealth-040617-013638

Fiore, A. M., Naik, V., & Leibensperger, E. M. (2015). Air quality and climate connections. *Journal of the Air & Waste Management Association, 65*(6), 645–685. https://doi.org/10.1080/10962247.2015.1040526

Reid, C. E., Brauer, M., Johnston, F. H., Jerrett, M., Balmes, J. R., & Elliott, C. T. (2019). *Critical review of health impacts of wildfire smoke exposure.* Environmental Health Perspectives, 124(9), 1334–1343. https://doi.org/10.1289/ehp9934

Samet, J. M., Gruskin, S., & Dominici, F. (2020). *Fine particulate matter and public health: A global challenge.* New England Journal of Medicine, 382(7), 691–693. https://doi.org/10.1056/NEJMp1914250

United States Environmental Protection Agency. (2024). *Air Quality Index (AQI) basics.* https://www.airnow.gov/aqi/aqi-basics/

U.S. Environmental Protection Agency. (2024). *Download daily data: Outdoor air quality data.* https://www.epa.gov/outdoor-air-quality-data/download-daily-data

 # APPENDIX A. COMPLETE R CODE
 # Final Project – Practical Statistics & Programming Using R
 # Author: Soli Evans
 # Dataset: EPA Daily AQI (2024)

```
################################################################################
#
# 1. LOAD REQUIRED PACKAGES
################################################################################
#
library(caret)
library(rpart)
library(rpart.plot)
library(randomForest)
library(ggplot2)

################################################################################
#
# 2. IMPORT AND RECONSTRUCT DATA
################################################################################
#

raw <- readLines(file.choose())     # Select the daily_aqi_by_cbsa_2024.csv file
raw <- raw[-1]                      # Remove corrupted header row

aqi <- read.csv(text = raw, header = FALSE, stringsAsFactors = FALSE)

# Assign correct column names
names(aqi) <- c("CBSA", "CBSA_Code", "Date", "AQI", "Category",
                "Defining_Parameter", "Site_ID", "Num_Reporting_Sites")

# Convert fields to proper types
aqi$Date <- as.Date(aqi$Date)
aqi$AQI <- as.numeric(aqi$AQI)
aqi$Num_Reporting_Sites <- as.numeric(aqi$Num_Reporting_Sites)

# Extract state abbreviation from CBSA
aqi$State <- sub(".*, *", "", aqi$CBSA)
```

```
################################################################################
#
# 3. DATA CLEANING AND FEATURE ENGINEERING
################################################################################
#

# Extract month number and month name
aqi$Month <- as.numeric(format(aqi$Date, "%m"))
aqi$Month_Name <- month.abb[aqi$Month]

################################################################################
#
# 4. REGION CLASSIFICATION
################################################################################
#

Northeast <- c("ME","NH","VT","MA","RI","CT","NY","NJ","PA")
Midwest   <- c("OH","MI","IN","IL","WI","MN","IA","MO","ND","SD","NE","KS")
South     <- c("DE","MD","VA","WV","NC","SC","GA","FL","KY","TN","MS",
           "AL","OK","TX","AR","LA","DC")
West      <- c("WA","OR","CA","NV","ID","MT","WY","UT","CO","AZ","NM","HI","AK")

aqi$Region <- ifelse(aqi$State %in% Northeast, "Northeast",
                ifelse(aqi$State %in% Midwest, "Midwest",
                     ifelse(aqi$State %in% South, "South",
                          ifelse(aqi$State %in% West, "West", NA))))

aqi <- subset(aqi, !is.na(Region))  # Remove unmatched rows

################################################################################
#
# 5. QUALITY CONTROL
################################################################################
#

# Remove missing AQI values
aqi <- subset(aqi, !is.na(AQI))

# Validate and filter categories
valid_categories <- c("Good","Moderate","Unhealthy for Sensitive Groups",
```

```
                  "Unhealthy","Very Unhealthy","Hazardous")
aqi <- subset(aqi, Category %in% valid_categories)

# Collapse categories into 3-class outcome variable
aqi$AQI_Class <- ifelse(aqi$Category == "Good", "Good",
                  ifelse(aqi$Category == "Moderate", "Moderate", "Unhealthy"))
aqi$AQI_Class <- factor(aqi$AQI_Class,
                  levels = c("Good", "Moderate", "Unhealthy"))

################################################################################
#
# 6. DESCRIPTIVE STATISTICS
################################################################################
#

summary(aqi$AQI)
aggregate(AQI ~ Region, aqi, summary)
aggregate(AQI ~ Month_Name, aqi, summary)

table(aqi$AQI_Class)
table(aqi$Defining_Parameter)

################################################################################
#
# 7. VISUALIZATIONS
################################################################################
#

# Boxplot of AQI by Region
ggplot(aqi, aes(x = Region, y = AQI, fill = Region)) +
  geom_boxplot() +
  labs(title = "AQI Distribution by Region", y = "AQI")

# Line plot: Average Monthly AQI by Region
avg_month_region <- aggregate(AQI ~ Month + Region, aqi, mean)
ggplot(avg_month_region, aes(x = Month, y = AQI, color = Region)) +
  geom_line(size = 1.1) +
  labs(title = "Average Monthly AQI by Region", x = "Month", y = "AQI")

# Bar chart: 3-Class AQI Category Distribution
```

```
ggplot(aqi, aes(x = AQI_Class, fill = AQI_Class)) +
  geom_bar() +
  labs(title = "AQI Categories (3-Class Collapsed)")

################################################################################
#
# 8. INFERENTIAL STATISTICS
################################################################################
#

# One-way ANOVA: AQI by Region
anova_region <- aov(AQI ~ Region, data = aqi)
summary(anova_region)
TukeyHSD(anova_region)

# Two-way ANOVA: Region * Month
anova_two <- aov(AQI ~ Region * Month, data = aqi)
summary(anova_two)

################################################################################
#
# 9. MACHINE LEARNING MODELS
################################################################################
#

# Prepare data for ML
ml_data <- aqi[, c("AQI_Class","AQI","Month","Region","State","Defining_Parameter")]
ml_data$Region <- factor(ml_data$Region)
ml_data$State <- factor(ml_data$State)
ml_data$Defining_Parameter <- factor(ml_data$Defining_Parameter)

# Train-test split
set.seed(123)
split <- createDataPartition(ml_data$AQI_Class, p = 0.8, list = FALSE)
train <- ml_data[split, ]
test  <- ml_data[-split, ]

################################################################################
#
# MODEL 1: DECISION TREE
```

```
################################################################################
#

fit_tree <- train(AQI_Class ~ ., data = train,
                  method = "rpart",
                  trControl = trainControl(method = "cv", number = 5))

fit_tree
rpart.plot(fit_tree$finalModel)

pred_tree <- predict(fit_tree, newdata = test)
confusionMatrix(pred_tree, test$AQI_Class)

################################################################################
#
# MODEL 2: RANDOM FOREST
################################################################################
#

fit_rf <- train(AQI_Class ~ ., data = train,
                method = "rf",
                trControl = trainControl(method = "cv", number = 5),
                importance = TRUE)

fit_rf

pred_rf <- predict(fit_rf, newdata = test)
confusionMatrix(pred_rf, test$AQI_Class)

# Variable importance
varImp(fit_rf)
plot(varImp(fit_rf))
```

Appendix B. Statistical Output Tables

Table B1. Descriptive Statistics for AQI (Overall)

Statistic	Value
Minimum	0
1st Quartile	32
Median	42
Mean	43.75
3rd Quartile	53
Maximum	1322

Table B2. AQI Summary Statistics by Region

Region	Min	Q1	Median	Mean	Q3	Max
Midwest	0	32	41	42.07	52	196
Northeast	0	31	38	40.13	47	185
South	0	34	44	45.14	54	431
West	0	28	42	44.90	55	1322

Table B3. AQI Summary Statistics by Month

Month	Min	Q1	Median	Mean	Q3	Max
Jan	0	28	36	37.96	48	494
Feb	0	31	39	39.92	51	272
Mar	0	32	40	39.71	47	1322
Apr	0	36	44	43.56	51	277
May	0	35	44	46.20	54	228
Jun	0	35	45	47.62	55	436
Jul	0	37	49	52.60	61	337

Aug	0	35	46	49.25	58	372
Sep	0	31	41	43.36	51	593
Oct	0	32	43	44.97	54	341
Nov	0	27	36	37.10	46	176
Dec	0	29	38	41.00	52	215

Table B4. Frequency of AQI Classification (3-Class)

Category	Count
Good	106,939
Moderate	43,410
Unhealthy	2,681

Table B5. Frequency of Defining Pollutant

Pollutant	Count
PM2.5	75,234
Ozone	71,419
PM10	5,930
NO2	406
CO	41

Table B6. One-Way ANOVA Output (AQI ~ Region)

Term	df	Sum Sq	Mean Sq	F value	p-value
Region	3	509,657	169,886	384.3	< 2e-16 ***

Residuals	153,026	67,645,674	442	—	—

Table B7. Two-Way ANOVA Output (AQI ~ Region * Month)

Term	df	Sum Sq	Mean Sq	F value	p-value
Region	3	509,657	169,886	387.0	< 2e-16 ***
Month	1	83,252	83,252	189.7	< 2e-16 ***
Region	3	392,689	130,896	298.2	< 2e-16 ***
Residuals	153,022	67,169,733	439	—	—

Table B8. Decision Tree Confusion Matrix

Prediction \ Reference	Good	Moderate	Unhealthy
Good	21,387	0	0
Moderate	0	8,682	0
Unhealthy	0	0	536

Accuracy = **1.00**
Kappa = **1.00**

Appendix C. Figures

Figure C1. AQI Distribution by Region

Description:
Boxplot illustrating AQI variability across the four U.S. Census regions. Western states exhibit

the largest spread and the highest extreme values.

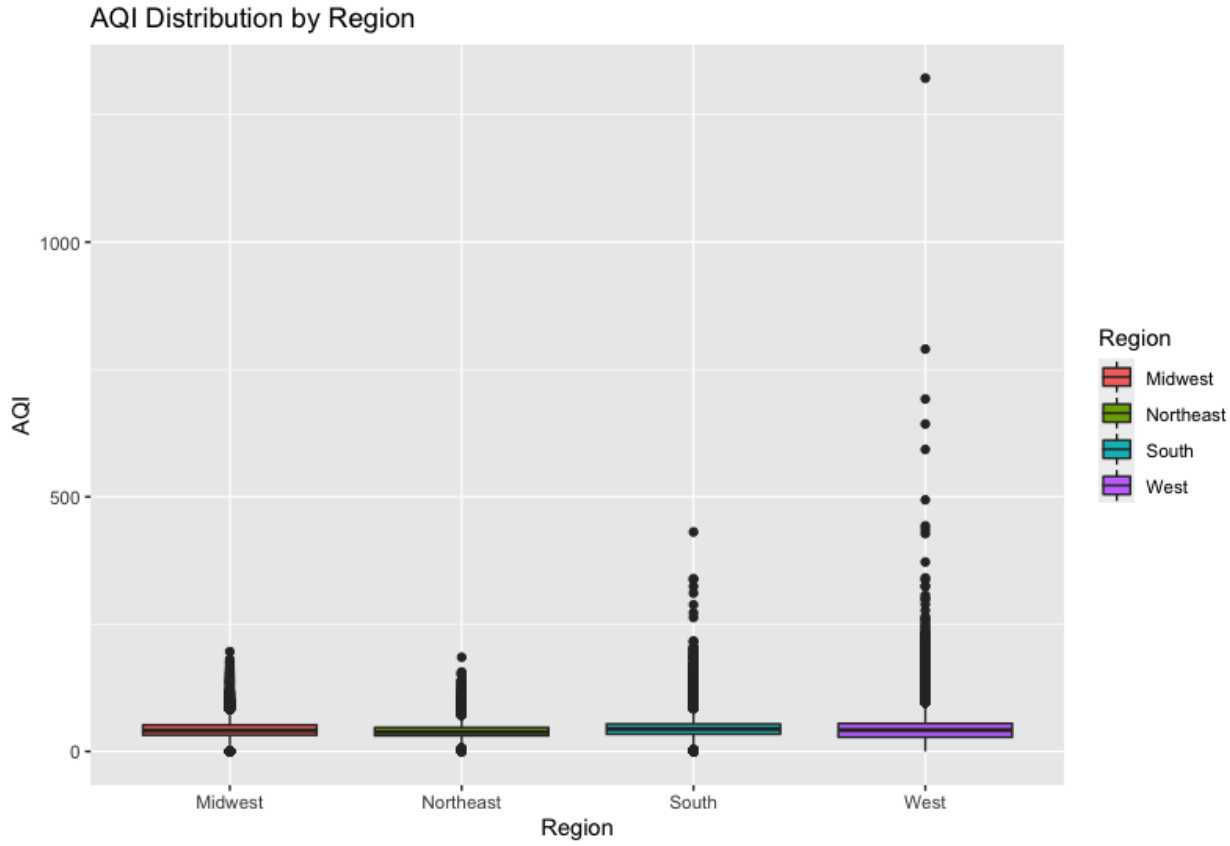


Figure C2. Average Monthly AQI by Region

Description:

Line plot showing seasonal AQI trends by region. All regions peak in summer months, with the West showing the steepest mid-year increase.

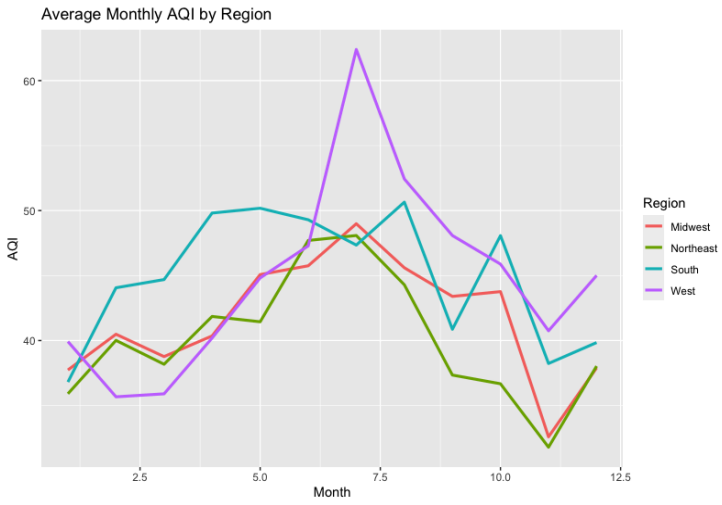


Figure C3. Distribution of AQI Classes (3-Level Collapsed)

Description:
Bar chart showing the proportion of Good, Moderate, and Unhealthy AQI days across the U.S. in 2024. Most days fall in the Good category.

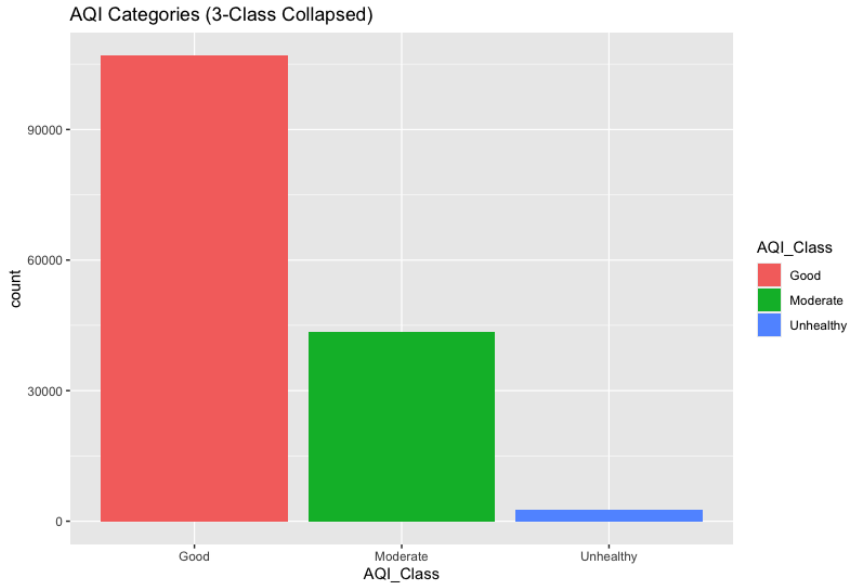


Figure C4. Decision Tree Classification Model

Description:
Visual representation of the CART model used to predict AQI class. The tree highlights AQI thresholds (<51 → Good, <101 → Moderate) that mirror EPA definitions.