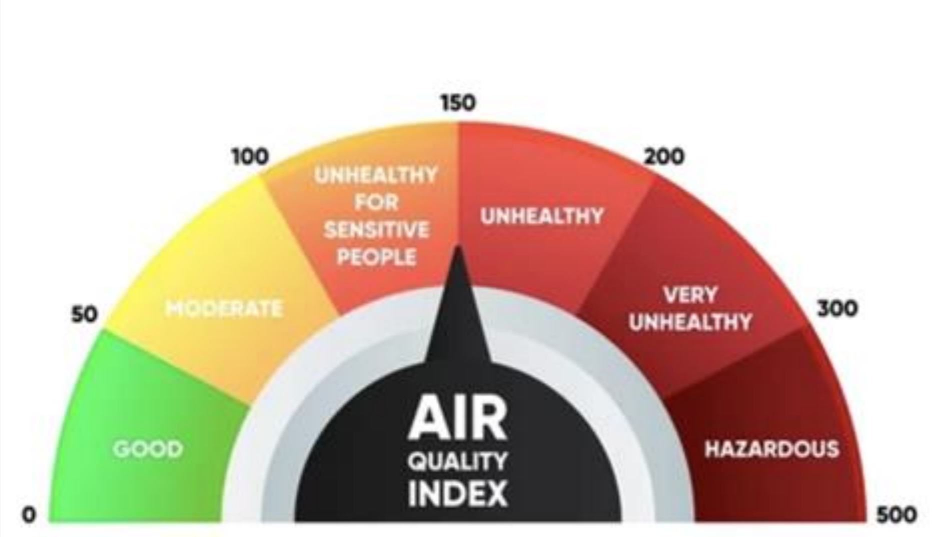
Air Quality Analysis Across U.S. Regions Using R

Final Project: Practical Statistics & Programming Using R
Presenter: Soli Evans
Dataset: EPA Daily AQI by CBSA (2024)

Background & Importance



Air Quality Index (AQI) is a standardized measure used to communicate daily air pollution levels. Poor air quality is associated with increased respiratory symptoms, cardiovascular disease, and hospital admissions.

Understanding regional variation and seasonal trends in AQI helps inform public health planning and environmental policy.

Dataset Overview

Dataset: EPA Daily AQI by CBSA (2024)
Observations: 171,648 rows
Key Variables: AQI, Category, Pollutant, CBSA, State, Region, Date, Month
Source: U.S. Environmental Protection Agency (EPA), Daily Air Quality Data

Study Aims

Aim 1 (Statistical):
Test whether mean AQI significantly differs across U.S. regions.

Aim 2 (Machine Learning):
Build classification models (Decision Tree and Random Forest) to predict AQI category based on region, pollutant, and seasonal patterns.

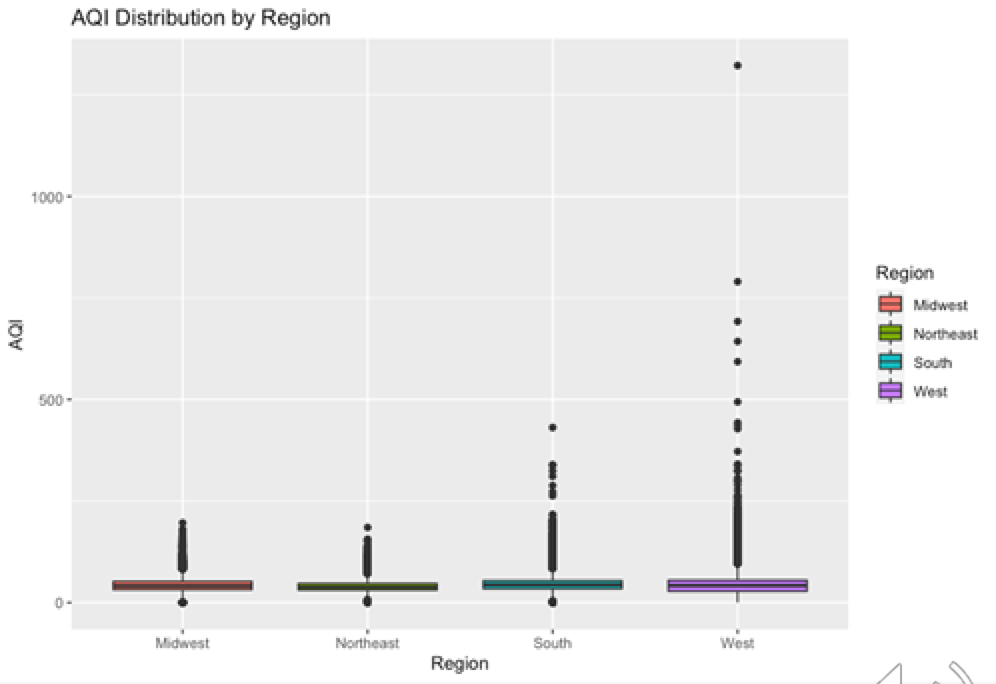
Data Cleaning & Preparation

- Removed malformed header and parsed CSV structure
- Converted date, numeric, and factor fields
- Extracted state from CBSA using string operations
- Created derived variables: Region, Month, and AQI_Class
- Removed invalid or missing values
- Validated category definitions against EPA coding

Descriptive Statistics: AQI Summary

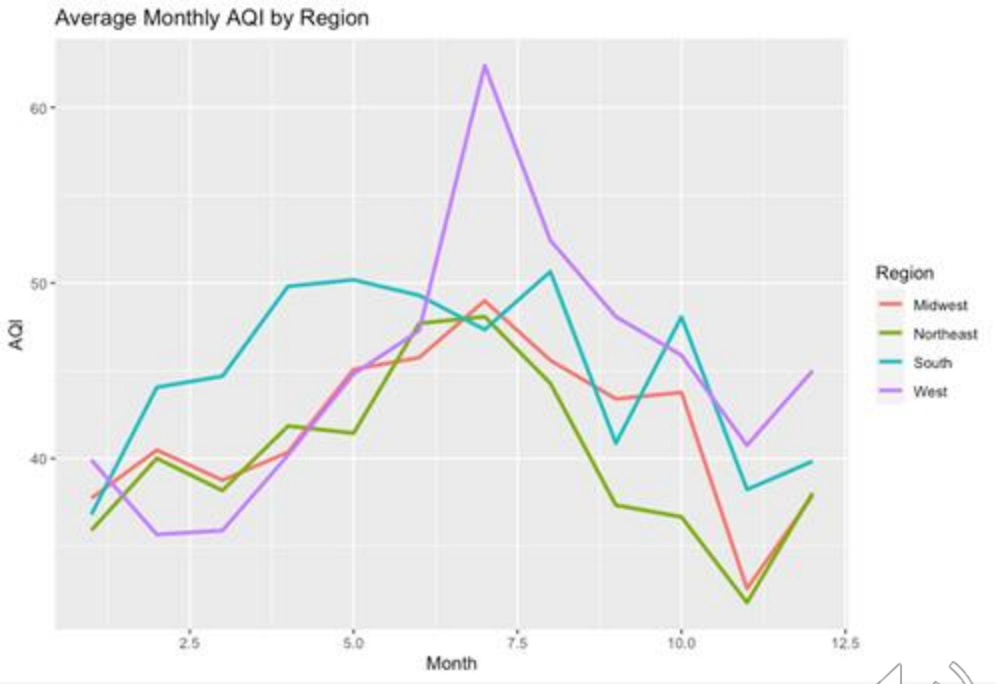
AQI Summary (All Regions):

- **Mean:** 43.75
- **Median:** 42
- **Range:** 0 to 1322
- Highest variability observed in Western region due to wildfire events.



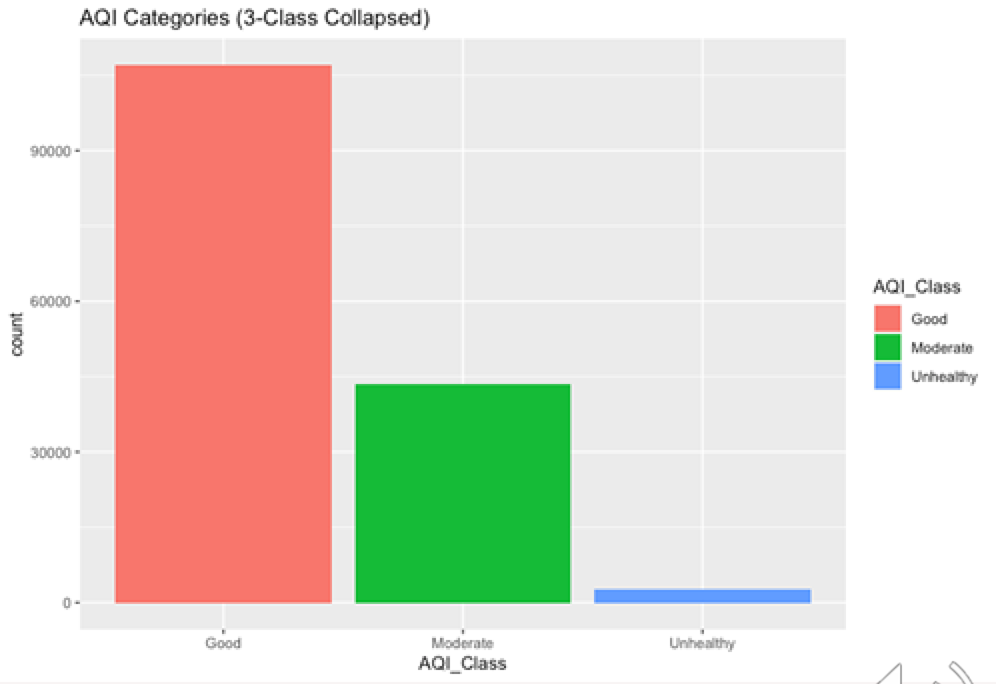
Seasonal Trends

- AQI peaks during summer months across all regions.
- Western region experiences the sharpest midsummer spike.
- Winter months generally show cleaner air and lower AQI values.



AQI Categories

- Majority of days fall within the **Good** category.
- **Moderate** days represent a substantial portion of the dataset.
- **Unhealthy** days occur infrequently but are important to analyze.
- Class imbalance impacts model performance considerations.



Inferential Statistics

- **One-Way ANOVA:**
  Significant differences in mean AQI across regions
  (*F* = 384.3, *p* < .001)

Tukey HSD Post Hoc:

- South & West have significantly higher AQI than Northeast & Midwest.

Two-Way ANOVA (Region × Month):
Region, Month, and their interaction were all significant (*p* < .001),
indicating seasonal trends differ across regions.

Machine Learning Approach

Outcome Variable: AQI_Class (Good, Moderate, Unhealthy)
Predictors: AQI, Region, State, Month, Pollutant

Methods Used:

- Decision Tree (rpart)
- Random Forest (caret + randomForest)

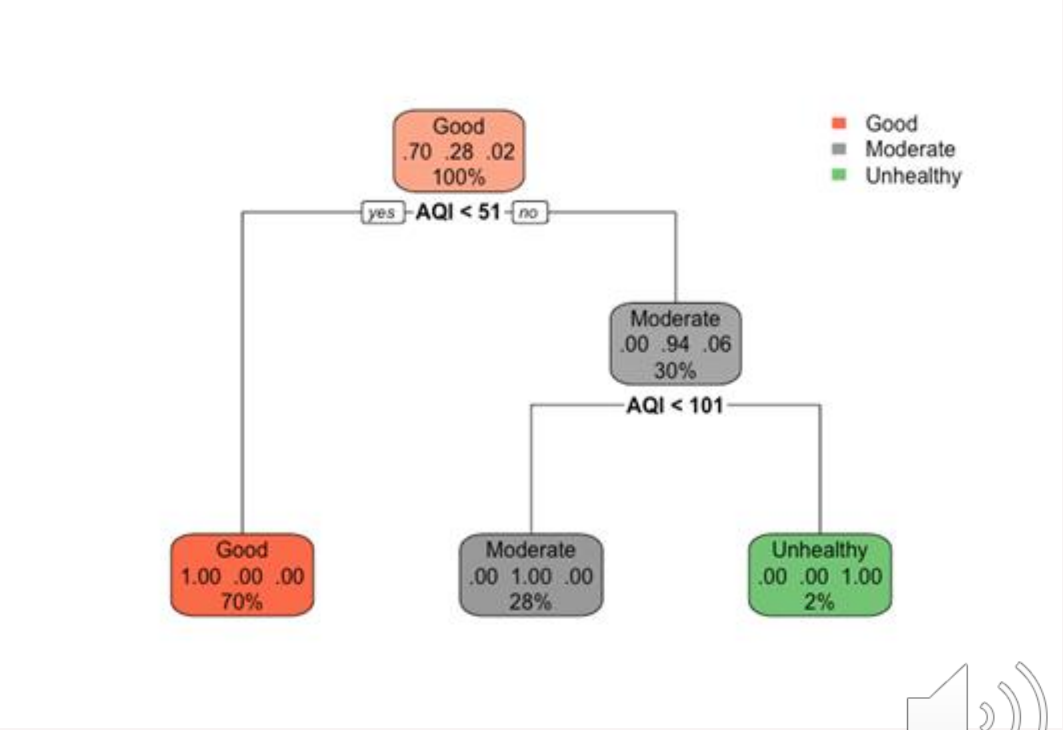
Train/Test Split: 80/20, stratified by AQI class

Decision Tree Results

Accuracy: 100%
The model perfectly classified all AQI categories,
mirroring EPA-defined AQI thresholds.

Key Splits:

- **AQI < 51 → Good**
- **AQI 51–100 → Moderate**
- **AQI ≥ 101 → Unhealthy**



Random Forest Results

Random Forest Performance:

- Near-perfect accuracy
- Robust to noise and variability
- Confirmed importance of key predictors

Top Predictors:

1. **AQI**
2. **Month**
3. **Region**
4. **Pollutant**

Interpretation & Key Insights

- The South and West experience higher pollution levels overall.
- Seasonal variation is significant, with summer showing the poorest air quality.
- Machine learning models validated the deterministic nature of EPA AQI categories.
- Results support environmental health literature on pollutant trends and regional disparities.

Limitations

- AQI categories are deterministic based on numerical thresholds, contributing to perfect classification.

- Dataset lacks meteorological variables such as humidity and wind.

- Region assignment is an approximation based on state boundaries.

- Outliers from extreme wildfire events inflate AQI values.

Final Conclusions

- AQI differs significantly across U.S. regions and seasons.

- Western and Southern states show highest pollution levels.

- Machine learning models performed exceptionally well.

- Findings may inform environmental monitoring, policy planning, and public health response strategies.

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

Thank You!

Questions?