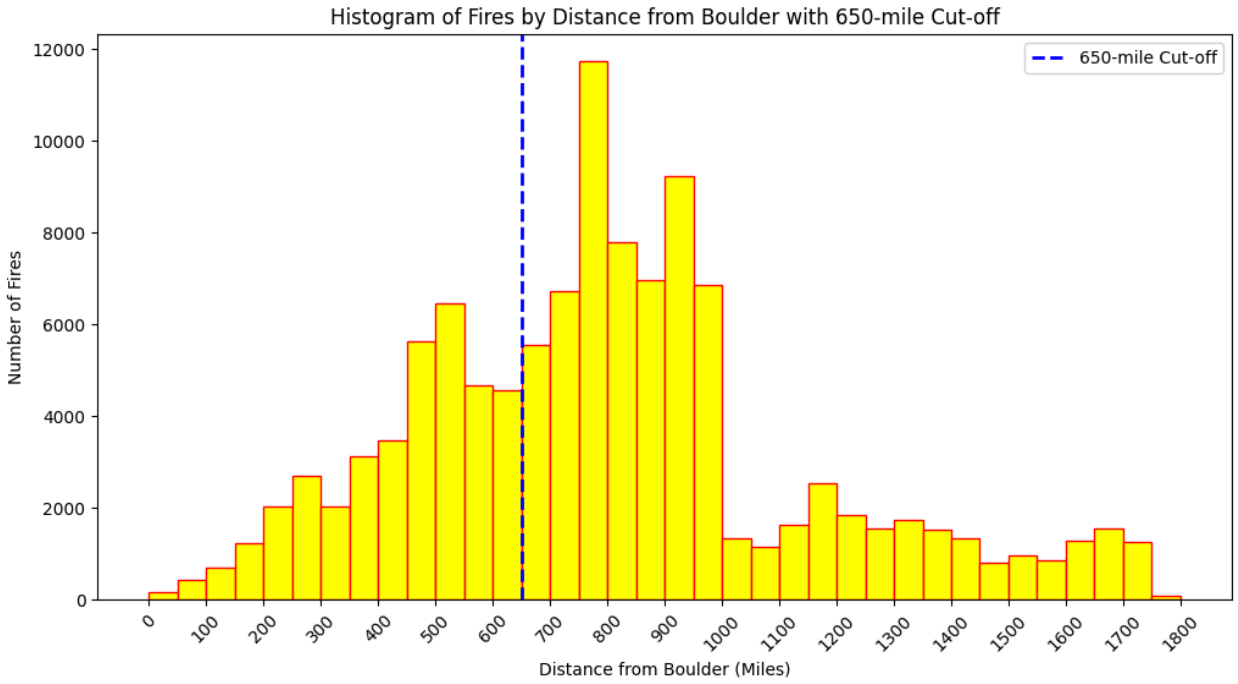
Radhika Sethi
rsethi3@uw.edu

Step 3: Analysis and Reflection Writeup

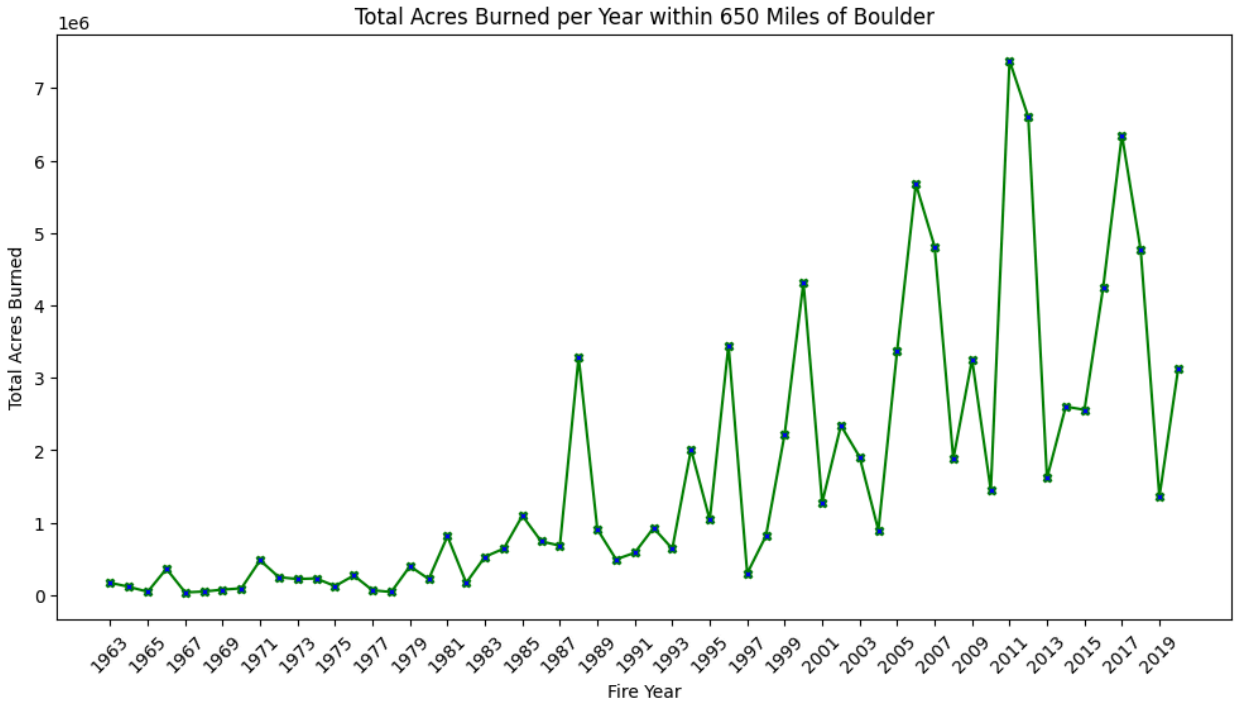
Visualization Explanations

1. Histogram of Fires by Distance from Boulder, CO



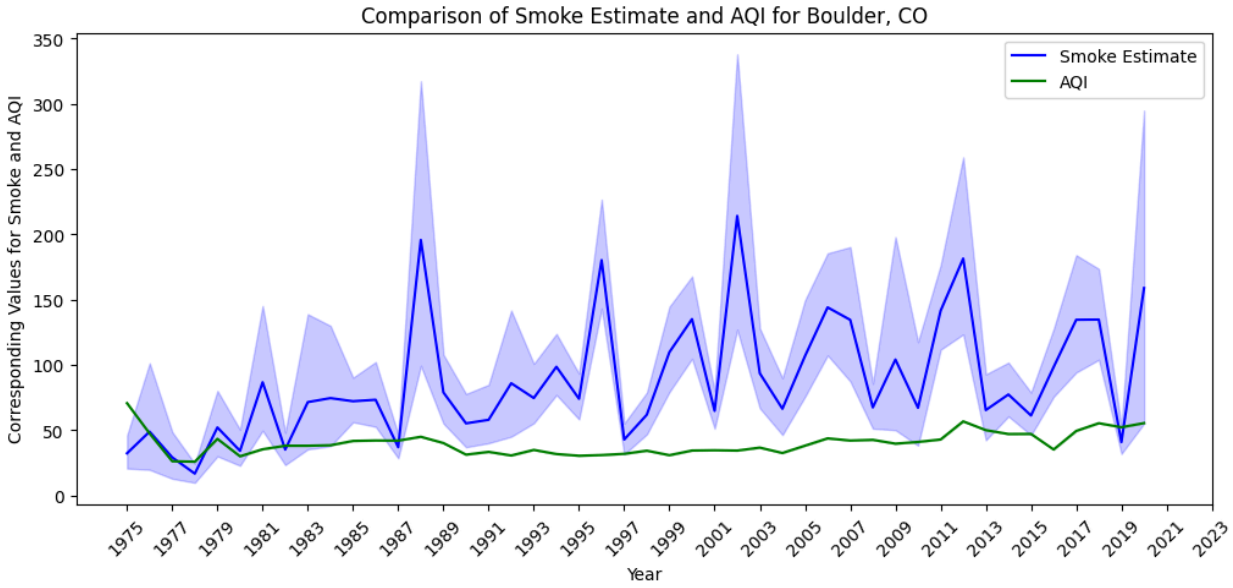
This histogram provides an overview of the distribution of wildfires based on their distance from Boulder, with each bar representing a 50-mile interval. The x-axis shows the distance from Boulder, extending up to 1800 miles, while the y-axis displays the number of fires within each interval. A blue dashed line marks a critical 650-mile cut-off, beyond which the influence of wildfires on Boulder's air quality is assumed to diminish significantly. Peaks around 500-600 miles and 900-1000 miles suggest clusters of frequent fire activity in those distances, possibly due to specific environmental or regional characteristics. This histogram sets the foundation for understanding how the proximity of fires affects potential smoke exposure in Boulder, giving context to later analyses on air quality impacts.

2. Total Acres Burned per Year within 650 Miles of Boulder



This line plot visualizes the total acres burned each year within the 650-mile radius of Boulder, which represents cumulative fire impact over time. The x-axis tracks years from 1963 to the present, and the y-axis shows the total acres burned, with some years experiencing notable spikes. These peaks likely correspond to larger, more intense fire seasons that contribute significantly to smoke production. The trend suggests a marked increase in fire activity from the 2000s onwards, possibly influenced by factors like climate change, shifting weather patterns, or changes in forest management practices. Observing this historical trend is crucial as it highlights the potential for increasing smoke exposure in Boulder over time, providing a background for the forecasting model.

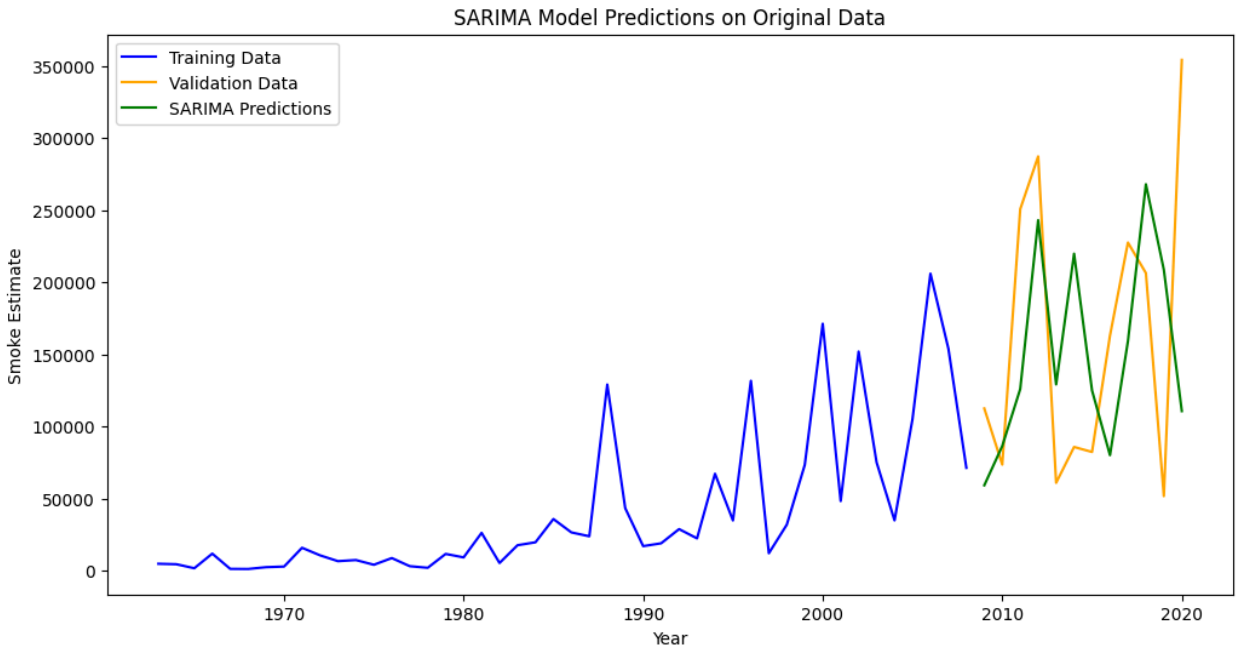
3. Comparison of Smoke Estimate and AQI for Boulder, CO



This line plot compares two key measures over time: Smoke Estimate and Air Quality Index (AQI) for Boulder, allowing us to assess the relationship between estimated smoke exposure and actual air quality conditions. The x-axis marks the years, while the y-axis shows the scaled values for Smoke Estimate (blue line) and AQI (green line). The Smoke Estimate is derived from wildfire data, accounting for fire size and distance from Boulder, while the AQI is an official metric from the EPA that reflects various air pollutants. By examining where these two metrics align or diverge, we can start to understand how effectively wildfire data predicts air quality and where other factors (such as urban pollution) may play a role. This visualization is essential to validate the effectiveness of using smoke estimates for forecasting air quality impacts in Boulder and will help refine the model further.

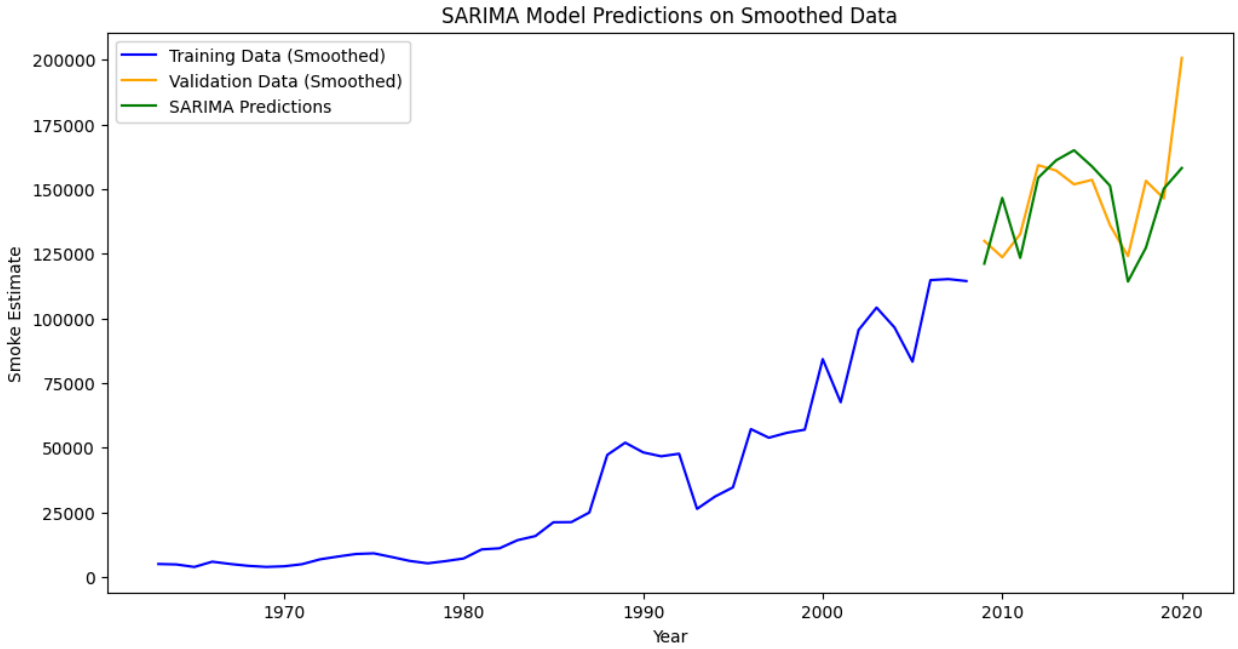
Forecasting Visualizations

4. SARIMA Model Predictions on Original Data



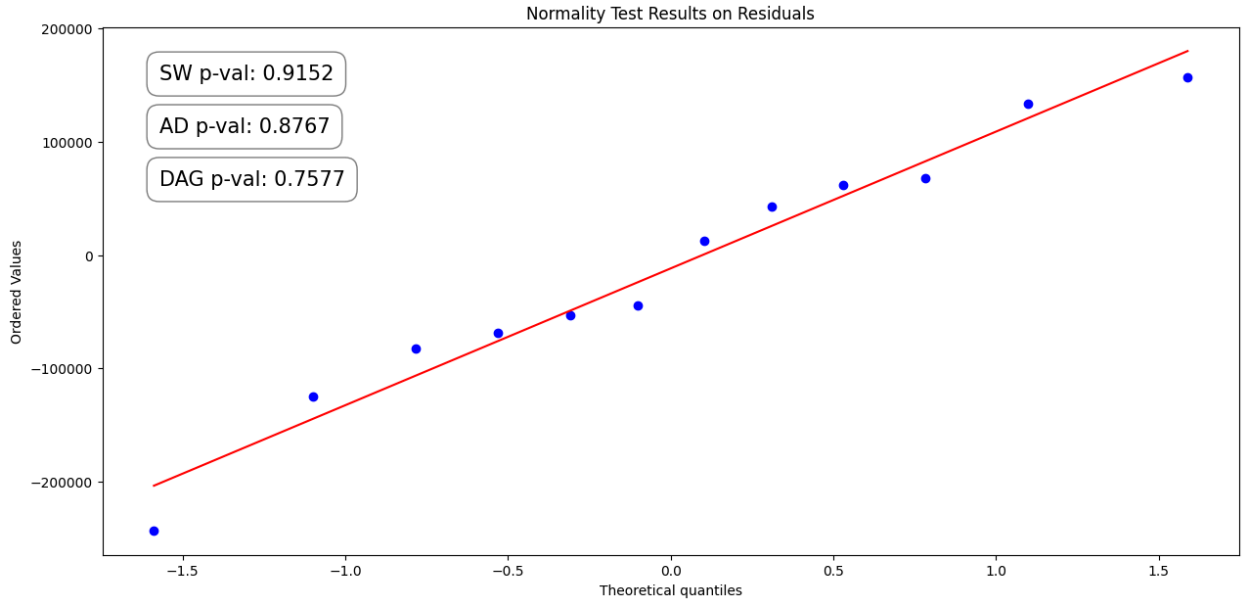
In this time series plot, we see the SARIMA model applied to the original smoke estimate data. The blue line represents training data used to fit the SARIMA model, while the orange line displays the validation data used to test its predictions. The green line represents the SARIMA model's forecasted smoke estimates, which can be directly compared against the validation data to assess accuracy. This plot is essential in evaluating how well the SARIMA model captures the seasonal patterns and trends in smoke estimates, especially without any data transformations. A close match between the predictions (green) and validation data (orange) indicates a good model fit, whereas significant deviations suggest the need for data smoothing or further model tuning.

5. SARIMA Model Predictions on Smoothed Data



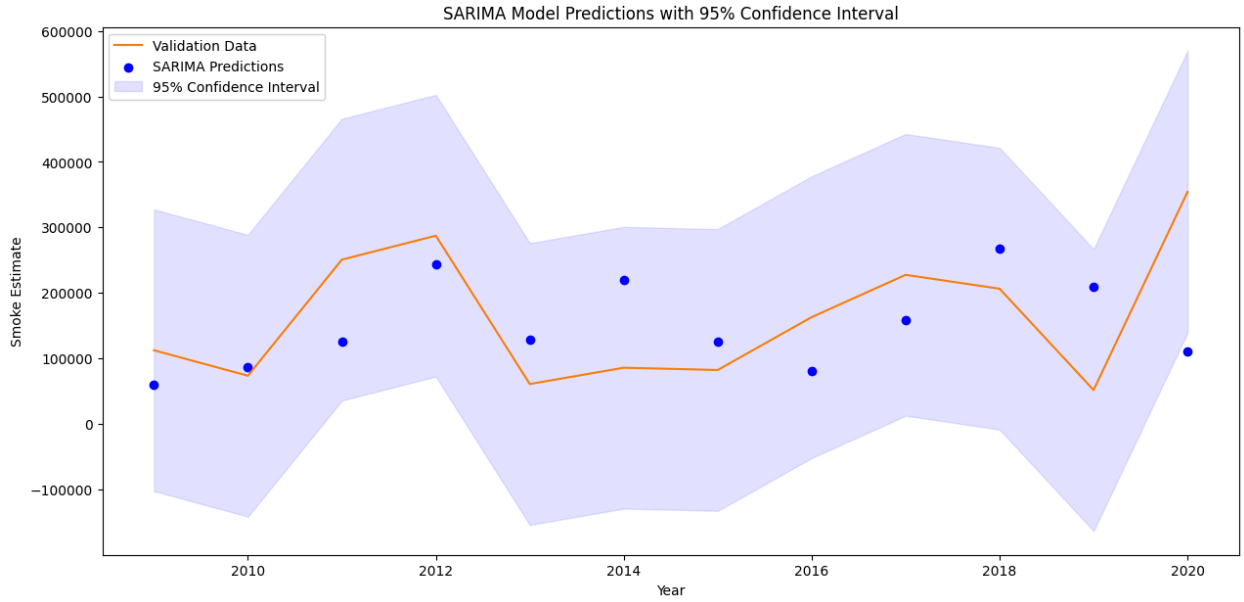
This plot applies the SARIMA model to smoothed smoke estimate data, which minimizes short-term noise and emphasizes long-term trends. Here, the blue line represents the smoothed training data, the orange line is the smoothed validation data, and the green line is the SARIMA model's predictions based on the smoothed input. Smoothing helps reduce volatility, allowing the model to better capture seasonal patterns or general trends that might be obscured by sharp fluctuations in the original data. Comparing this plot with the original SARIMA predictions reveals the impact of data preprocessing, showing how a cleaner dataset can lead to more stable, reliable forecasts. This is a key step in understanding how data preparation influences model outcomes, particularly in time series with high variability like smoke estimates.

6. Normality Test Results on Residuals



This Q-Q plot evaluates the residuals (errors) from the SARIMA model predictions to check if they follow a normal distribution, an indicator of a well-fitted model. Three normality tests—Shapiro-Wilk (SW), Anderson-Darling (AD), and D'Agostino's K-squared (DAG)—are used to statistically assess this normality, with each test returning a p-value. Higher p-values ($p > 0.05$) suggest that the residuals are not significantly different from a normal distribution, meaning the model errors are randomly distributed. The red line in the Q-Q plot represents a perfect normal distribution, while the blue points show the actual residuals. Residuals that closely follow the red line imply the model does not have a systematic bias, reinforcing confidence in the model's accuracy. This analysis is critical as it confirms the SARIMA model's residuals are random, validating the predictions for Boulder's smoke estimates.

7. SARIMA Model Predictions with 95% Confidence Interval



This plot presents the SARIMA model predictions for the validation period with a 95% confidence interval, allowing us to visualize prediction uncertainty. The orange line shows the actual validation data, blue dots represent the SARIMA predictions, and the shaded region displays the confidence interval. This interval is essential for evaluating the reliability of the model's predictions, offering an estimated range within which the true values likely fall. A narrower confidence interval suggests higher confidence in the predictions, while a wider interval indicates more uncertainty. For applications where predicting smoke impact on air quality is critical, this plot emphasizes the importance of factoring in uncertainty, helping users make informed decisions based on model predictions.

Reflection

This analysis deepened my understanding of how historical wildfire data can be used to estimate air quality impacts in Boulder. I learned the importance of preprocessing steps, such as data smoothing, in stabilizing volatile data and improving the accuracy of time series models like SARIMA. Working with smoke estimates and AQI data was eye-opening, particularly in seeing how environmental data can predict local air quality. The results highlighted the increasing intensity and frequency of wildfires over the past few decades and the corresponding rise in smoke exposure in Boulder, likely influenced by changing climate conditions and forest policies.

Key Learnings from Collaboration

Collaborating on this project was highly beneficial, as it allowed me to approach the analysis with new perspectives. For instance, a peer suggested adding confidence intervals to our SARIMA predictions, which helped me quantify the uncertainty in our forecasts. Another valuable suggestion was performing normality tests on residuals, which confirmed the robustness of the SARIMA model. Discussing these methods with classmates gave me practical insights into model validation techniques, which enhanced the credibility of my analysis. This collaborative process underscored the importance of sharing ideas, as feedback often leads to a stronger, more thorough approach.

Specific Attributions

Several techniques in this analysis were inspired by discussions with classmates, such as the implementation of confidence intervals for SARIMA predictions and normality testing of residuals. Additionally, using data smoothing before applying the SARIMA model was a method suggested in peer reviews, which contributed to a more reliable forecasting model. I would like to acknowledge the support of my classmates, whose insights helped shape the approach taken in this project and enriched the overall analysis.

Conclusion

This project demonstrates the potential of statistical modeling in environmental analysis, specifically using SARIMA to forecast smoke impacts on Boulder's air quality based on historical wildfire data. The findings underscore an increasing trend in wildfire intensity, likely contributing to higher smoke estimates in recent years. The combination of SARIMA forecasting, normality testing, and confidence intervals allowed me to create a well-rounded model with quantifiable uncertainty. This experience reinforced the importance of robust data processing and model validation in making meaningful environmental predictions, and I look forward to further exploring the intersection of data science and environmental studies.