Conclusions Drawn

1. Overall Observation from Data (EDA)

- PM2.5 and NO2 levels spike near wildfires, with **PM2.5 highest when fires are 1–10 km away**.
- NO2 remains high regardless of wildfires → indicates **traffic/industrial contribution**.
- Other pollutants (CO, O3, SO2) are minimally affected.
- Seasonal patterns: **summer > winter > fall > spring** in PM2.5.
- Geographic hotspots: **California and the Western US**, with some high-priority sites identified (Los Angeles-North Main Street, Rubidoux, North Birmingham).
- **Data limitations:** Granular data makes visual interpretation challenging → further aggregation may improve insights.

Priority Monitoring

- Sites ranked by a priority score (PM2.5 × nearby fires) highlight areas most affected:
    - **Top sites:** Los Angeles-North Main Street, Rubidoux, North Birmingham.
- **Recommendation:** Allocate additional air quality monitoring resources to these high-priority areas.

2. Model-Based Conclusions

a. Multiple Linear Regression (OLS)

- Lasso-regularized regression shows wildfire proximity and intensity drive PM2.5, explaining over **60% of the variation** with an RMSE of ~6.5% of the range.
- Residuals are small and evenly distributed, indicating the model reliably captures trends.
- These results demonstrate that **wildfires have a meaningful and measurable effect on air quality**, though extreme PM2.5 spikes remain harder to predict.

b. Ridge & Lasso

- Both regularization methods improved predictive accuracy over standard OLS, with Lasso slightly outperforming Ridge (lower RMSE).
- **Conclusion:** Regularization helps stabilize coefficients and better captures the effects of wildfire proximity and intensity, making predictions more reliable, though extreme PM2.5 events remain challenging.

c. Logistic Regression

- Classifies whether PM2.5 exceeds a threshold: **accuracy ~67%, AUC 0.72**.

- **Conclusion:** Proximity to fires and certain weather conditions can **reasonably predict high-PM2.5 events**, but confidence in probabilities is limited.

d. K-Nearest Neighbors (KNN)

- Useful for capturing local or non-linear effects for continuous PM2.5: **RMSE ~13, R^2 ~0.50**.
- Performs well for typical air quality but **underpredicts extreme pollution events**.
- **Conclusion:** Wildfire, weather, and location together **explain ~50% of PM2.5 variation**, making KNN useful for anticipating day-to-day pollution patterns.

e. K-Means Clustering

- Most data in one cluster; extreme pollution events isolated in a second, tiny cluster.
- **Conclusion:** Most days experience average conditions; extreme events are rare but identifiable.

.

Implications for Stakeholders

- **Air quality alerts:** Use the model predictions to issue warnings during nearby wildfires, especially in high-priority regions.

- **Resource allocation:** Deploy additional sensors and public health measures in high-risk areas and during summer/winter peaks.

- **Policy planning:** Use model insights to allocate resources efficiently, especially for wildfire-related pollution spikes.

By combining EDA with multiple predictive models, we find that PM2.5 is strongly influenced by wildfire proximity, intensity, and seasonality. Lasso-regularized linear regression now provides the best overall predictive accuracy (R^2 ~0.61, RMSE ~5.7), capturing day-to-day trends, while KNN remains useful for identifying local patterns and non-linear effects. Extreme pollution events remain challenging to predict but can be flagged for priority monitoring and public health interventions.

Research Question

1. How does fire proximity affect PM2.5 levels?

PM2.5 levels rise noticeably as wildfires get closer, peaking when fires are within 1–10 km of monitoring sites. This nonlinear relationship shows that proximity is a key driver of particulate pollution, and KNN predictions confirm that nearby fires significantly elevate PM2.5 on any given day.

2. Which pollutants are most impacted by wildfires?

PM2.5 is the most affected pollutant, showing clear increases near fires. NO2 also rises slightly with nearby fires but remains high due to traffic and industrial sources. CO, O3, and SO2 show minimal change, indicating that wildfires primarily influence fine particulate matter and, to a lesser extent, NO2.

3. Do weather conditions moderate fire impacts on air quality?

annajli1298@gmail.comfor you to fill out :)

4. Can we cluster regions by air quality response patterns?

We attempted to cluster regions based on air quality, wildfire activity, and weather patterns using K-Means. The results showed that most of the data fell into a single large cluster representing average conditions, while extreme pollution events formed a very small, separate cluster. This indicates that while extreme events can be identified, the majority of regions experience similar air quality patterns, making it difficult to define meaningful subgroups for typical conditions. Essentially, clustering is more useful for detecting outliers than for grouping regions by standard air quality responses.