Something in the Air:

Wildfires and their impact on air quality

Caroline Kranefuss, Anna Li, Karina Mehta, Shaveen Saadee
Group 15 - December 12, 2025

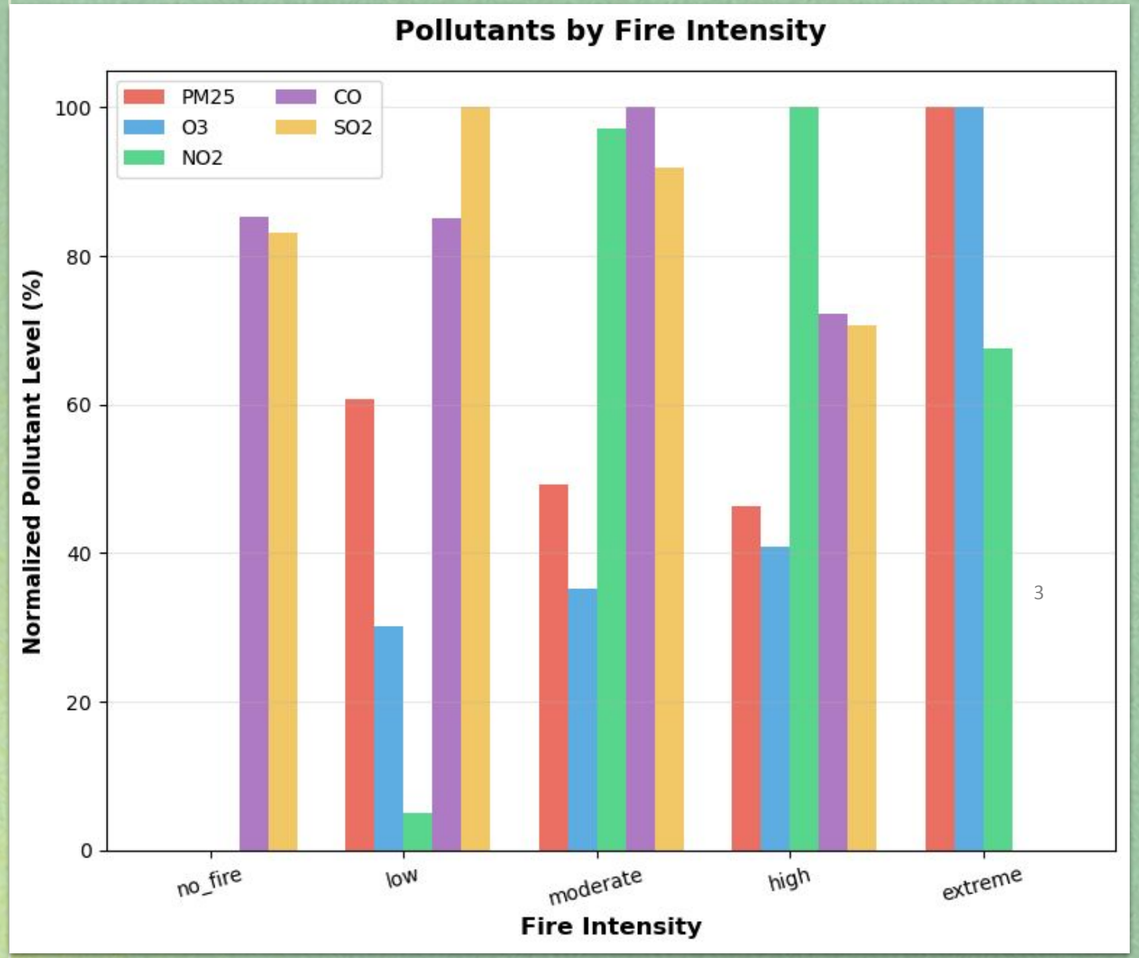
Data & Variables

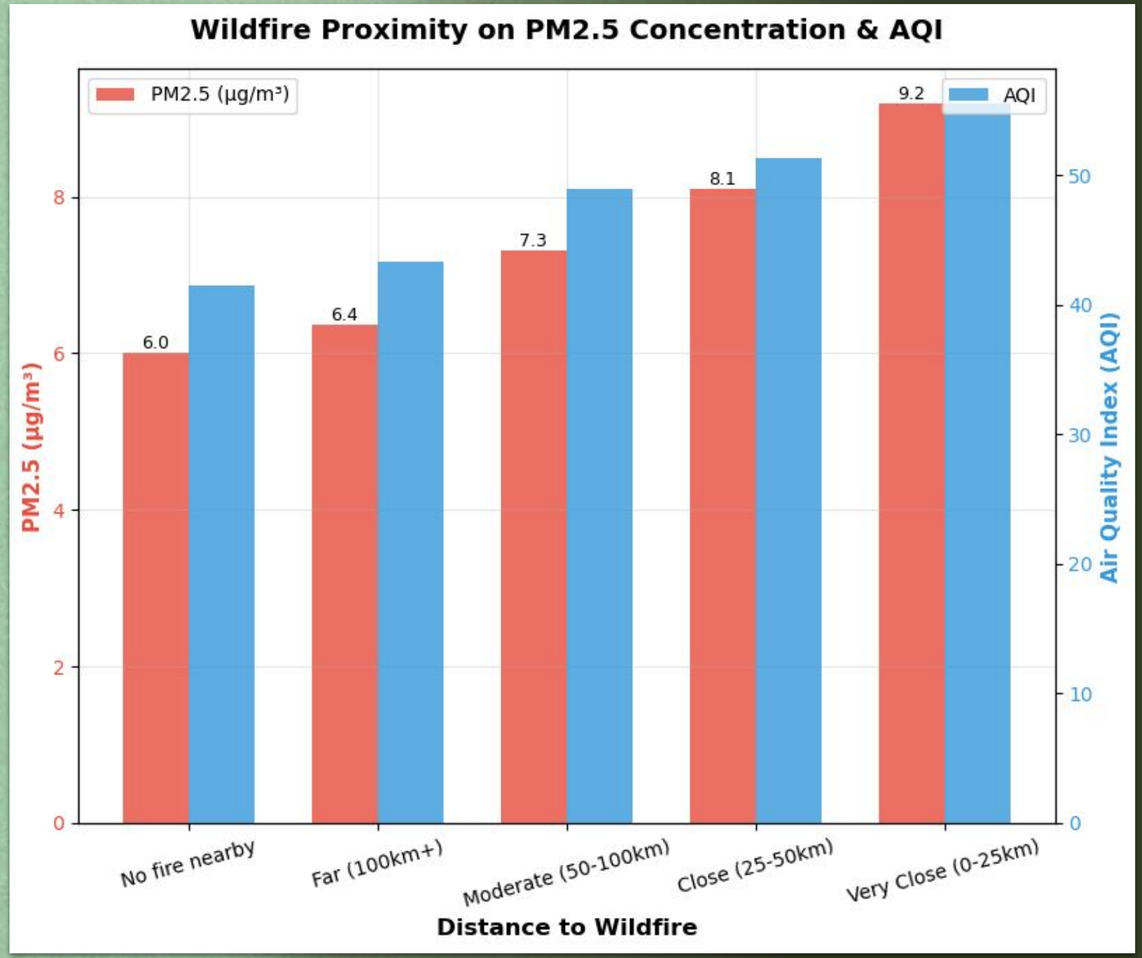
- **Sources:** NASA satellite fire detections, EPA air quality monitoring, and Open-Meteo API weather data
- **Dataset**: 20K observations from 75 monitoring sites across 40 states, each row corresponding to a daily observation

- ***Target variable:*** PM2.5 concentration (µg/m3), i.e., fine particulate matter

- ***Air quality measurements:*** CO, O3, NO2, SO2, AQI
- ***Fire features:*** distance to nearest fire, intensity, brightness, fire counts within 50-100 km radius
- ***Weather conditions:*** temperature, humidity, wind speed, precipitation, evapotranspiration, weather description
- ***Temporal factors:*** season, month, day of week, wildfire season

Exploratory Data Visualizations





Research Questions:

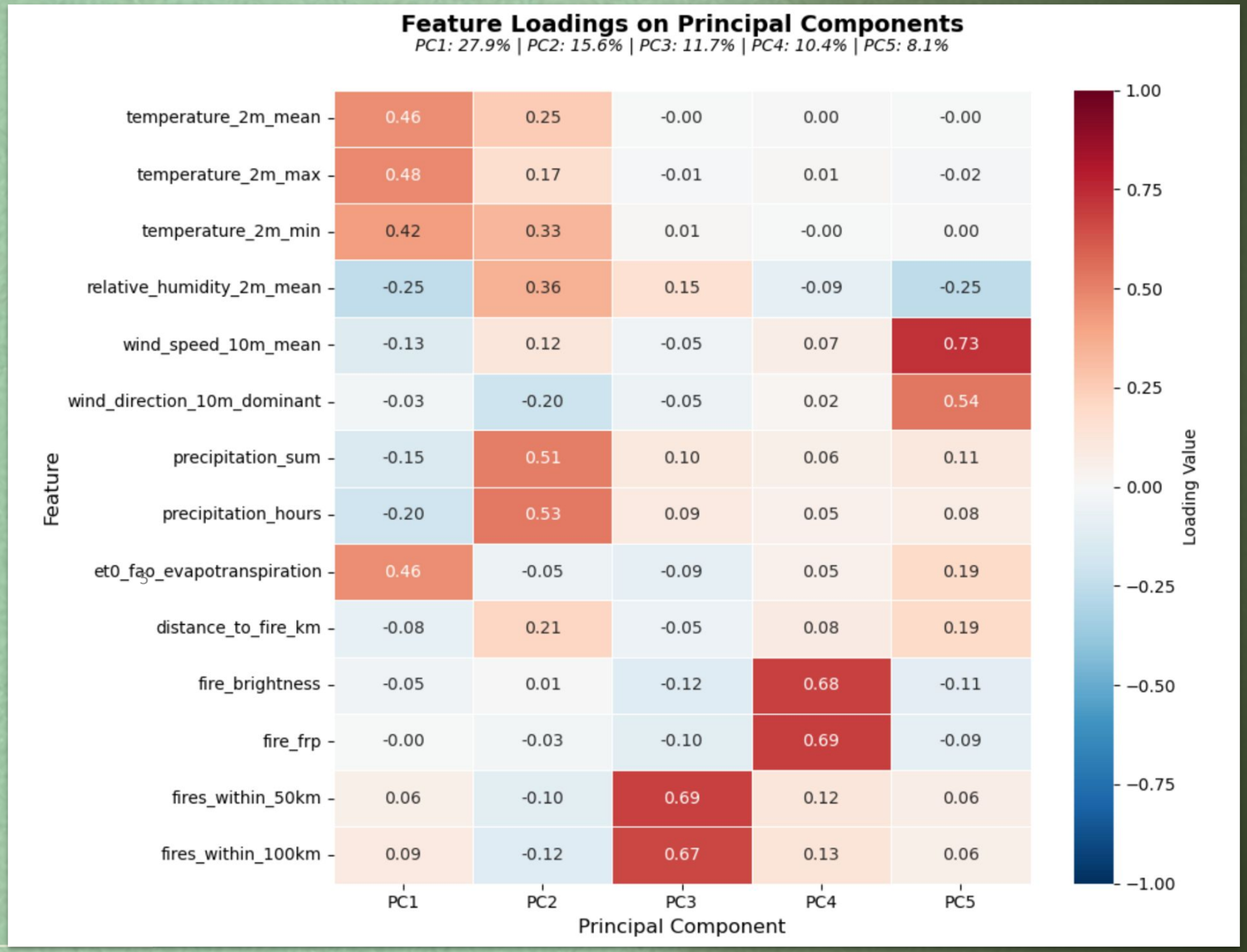
How does fire proximity affect PM2.5?

Do weather conditions moderate these fire impacts?

Principal Component Analysis

- First 5 PCAs capture **74% of variability**

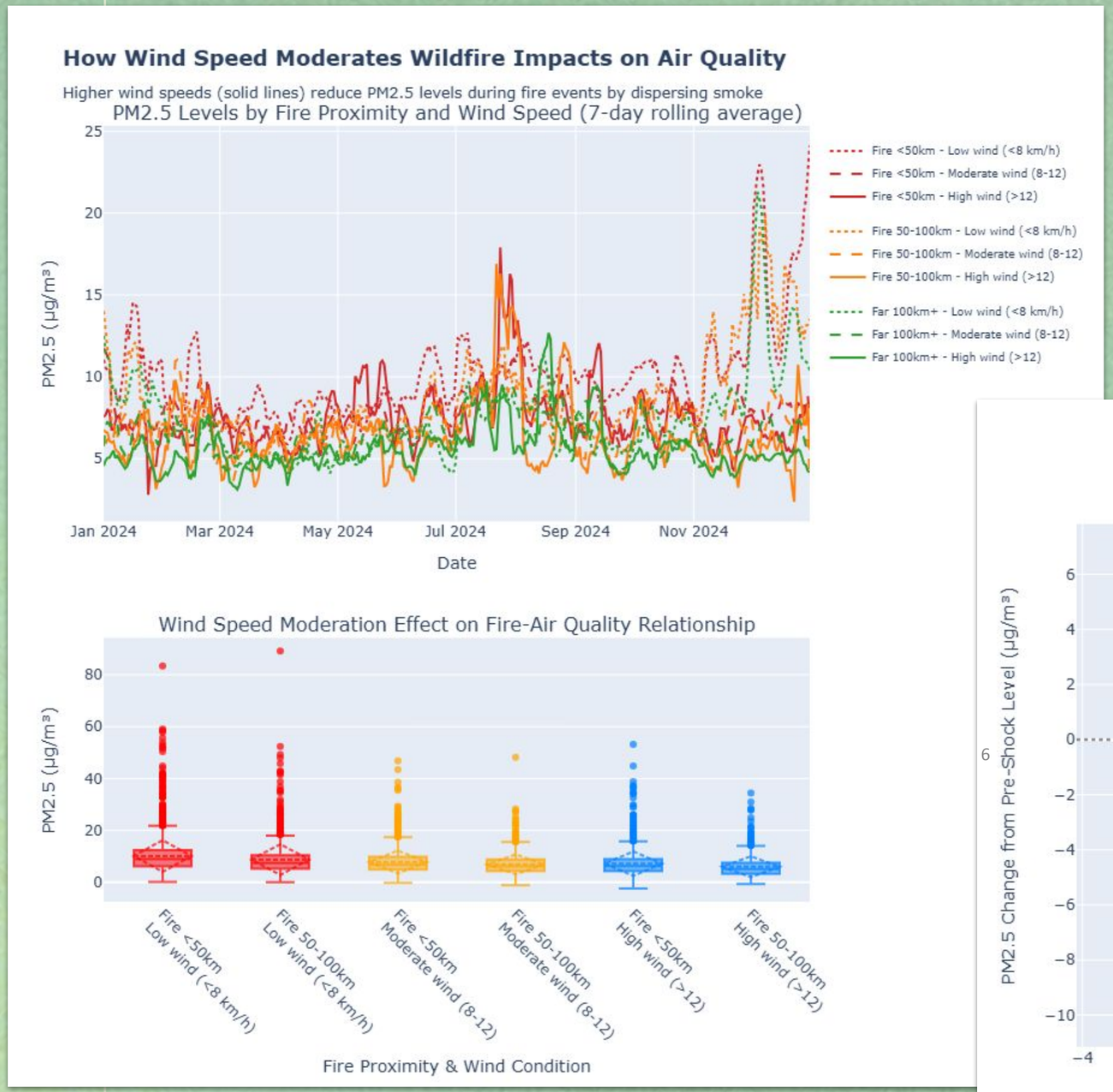
- **Natural groupings:**
  - PC1 ~ temperature
  - PC2 ~ rain
  - PC3 ~ fire proximity
  - PC4 ~ fire intensity
  - PC5 ~ wind



Time Series Analysis

High wind speeds 👍

Low humidity 👍



Conclusions for Stakeholders

- **Environment:** Encourage action by policymakers and individuals
- **Public health**: Issue air quality warnings and mind humid conditions

- **Prediction challenges despite multi-model approach**
- **However, successful capture of trends and conditions**
- **Actionable insights for protecting environmental and public health**

- **Model limitations:** Small dataset (despite bootstrapping), environmental factors can be extremely varied and complex, unpredictability of human-caused spikes in PM2.5, lack of data on days without fire, lack of understanding about healthy vs unhealthy PM2.5 values.

Thank you!

Questions?

MLR Model

streamlitApp

This analysis performs Multiple Linear Regression (OLS) with Lasso regularization for feature selection to predict standardized PM2.5 levels. **The results displayed below are pre-calculated.**

1. Baseline OLS Model Summary (Full Feature Set)

4a. OLS Model Summary (Train Data)

```
==============================================================================
Dep. Variable:                   PM25   R-squared:                       0.126
Model:                            OLS   Adj. R-squared:                  0.125
Method:                 Least Squares   F-statistic:                     166.1
Date:                Fri, 12 Dec 2025   Prob (F-statistic):               0.00
Time:                        01:59:58   Log-Likelihood:                -41390.
No. Observations:               13827   AIC:                         8.281e+04
Df Residuals:                   13814   BIC:                         8.290e+04
Df Model:                          12
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          7.8965      0.118     67.011      0.000       7.665       8.127
latitude                      -0.8261      0.044    -18.600      0.000      -0.913      -0.739
longitude                      0.0208      0.047      0.444      0.657      -0.071       0.113
relative_humidity_2m_mean      0.5926      0.050     11.754      0.000       0.494       0.691
wind_speed_10m_mean           -1.0019      0.044    -22.535      0.000      -1.089      -0.915
precipitation_sum             -0.4605      0.047     -9.843      0.000      -0.552      -0.369
fires_within_50km             -0.0828      0.049     -1.697      0.090      -0.178       0.013
fires_within_100km             0.3263      0.052      6.235      0.000       0.224       0.429
distance_to_fire_km           -0.5777      0.046    -12.553      0.000      -0.668      -0.487
fire_brightness               -0.1224      0.041     -2.961      0.003      -0.203      -0.041
dummy_cloudy                  -0.0198      0.131     -0.151      0.880      -0.276       0.236
dummy_rainy                   -0.9722      0.146     -6.658      0.000      -1.258      -0.686
dummy_snowy                   -1.5006      0.223     -6.739      0.000      -1.937      -1.064
==============================================================================
Omnibus:                    10303.467   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           351032.927
Skew:                           3.233   Prob(JB):                         0.00
Kurtosis:                      26.822   Cond. No.                         9.91
==============================================================================
```

Initial Flaw: The low R^2 (**0.126**) and highly insignificant coefficients for variables like `longitude` indicate poor model fit and multicollinearity, necessitating feature selection.

2. Final OLS Model Summary (Lasso-Selected Features)

4b. Final OLS Model Summary (Test Data)

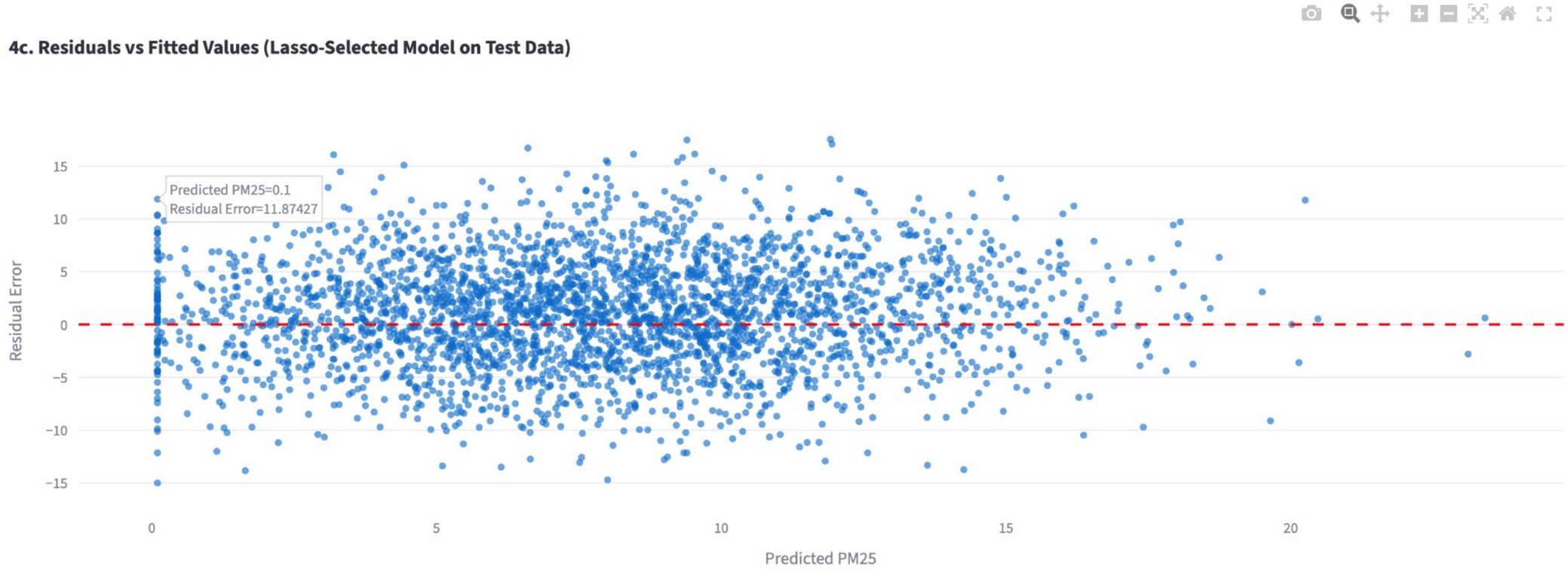
```
OLS Regression Results
================================================================================
Dep. Variable:                   PM25   R-squared (uncentered):                   0.611
Model:                            OLS   Adj. R-squared (uncentered):              0.610
Method:                 Least Squares   F-statistic:                              422.3
Date:                Fri, 12 Dec 2025   Prob (F-statistic):                        0.00
Time:                        02:00:30   Log-Likelihood:                         -9382.6
No. Observations:                2963   AIC:                                  1.879e+04
Df Residuals:                    2952   BIC:                                  1.885e+04
Df Model:                          11
Covariance Type:            nonrobust
================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
relative_humidity_2m_mean    -0.5166      0.121     -4.279      0.000      -0.753      -0.280
fires_within_100km            0.6705      0.137      4.880      0.000       0.401       0.940
latitude                     -0.7659      0.113     -6.767      0.000      -0.988      -0.544
wind_speed_10m_mean          -1.2154      0.108    -11.229      0.000      -1.428      -1.003
precipitation_sum            -0.5995      0.121     -4.938      0.000      -0.838      -0.361
fires_within_50km             1.5107      0.348      4.339      0.000       0.828       2.193
distance_to_fire_km          -0.5538      0.110     -5.013      0.000      -0.770      -0.337
fire_brightness              -0.2150      0.106     -2.020      0.043      -0.424      -0.006
dummy_cloudy                  7.6239      0.175     43.596      0.000       7.281       7.967
dummy_rainy                   7.5972      0.180     42.193      0.000       7.244       7.950
dummy_snowy                   7.6098      0.516     14.755      0.000       6.598       8.621
==============================================================================
Omnibus:                     2380.270   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           128787.976
Skew:                           3.386   Prob(JB):                         0.00
Kurtosis:                      34.580   Cond. No.                         6.39
==============================================================================
```

streaml

Improvement: After Lasso Regularization removed non-contributing features (like longitude), the R-squared (**0.611**) is **significantly higher**, and all remaining predictors are now statistically significant ($p < 0.05$). This indicates a much more robust and explanatory model.

odel Residual Examination

4c. Residuals vs Fitted Plot (Lasso-Selected Model)



Model Robustness: The residual plot shows a concentration of positive residuals at higher predicted PM2.5 values (heteroscedasticity). This suggests the model is **less accurate at predicting extreme high PM2.5 events** but is generally robust for normal air quality predictions.

Conclusion: Now all our predictors are statistically significant and our r-squared and adjusted r-squared values are much higher. Over 60% of the variation in the data are explained by our predictors. The residual plot for the model fitted to the tested model is different from the one previously seen; this is because of the lasso regularization that was performed. Now residuals tend positive, but most are close to zero. Their patterns do not change as the fitted values change. This is indicative of a good model. An RMSE of 5.741 on a scale of 87.4 means that our predictions are off by about 6.5% of the range on average. This means that our predictions are usually reasonable.

Logit Model

5. Logistic Regression: High PM2.5 Risk Classification

This model classifies air quality as **'High Risk'** (PM2.5 **above** the median) or **'Low Risk'** (PM2.5 **at or below** the median) based on scaled features.

1. Key Metrics and Confusion Matrix

5a. Key Metrics

Metric	Value
Accuracy	0.6652
Log Loss	0.6143
Threshold	6.25 (PM25)

The overall accuracy of **~67%** is acceptable. However, the high **Log Loss of 0.6143** indicates a flaw: the model's predicted probabilities are unreliable, meaning it is often highly confident in predictions that turn out to be incorrect.

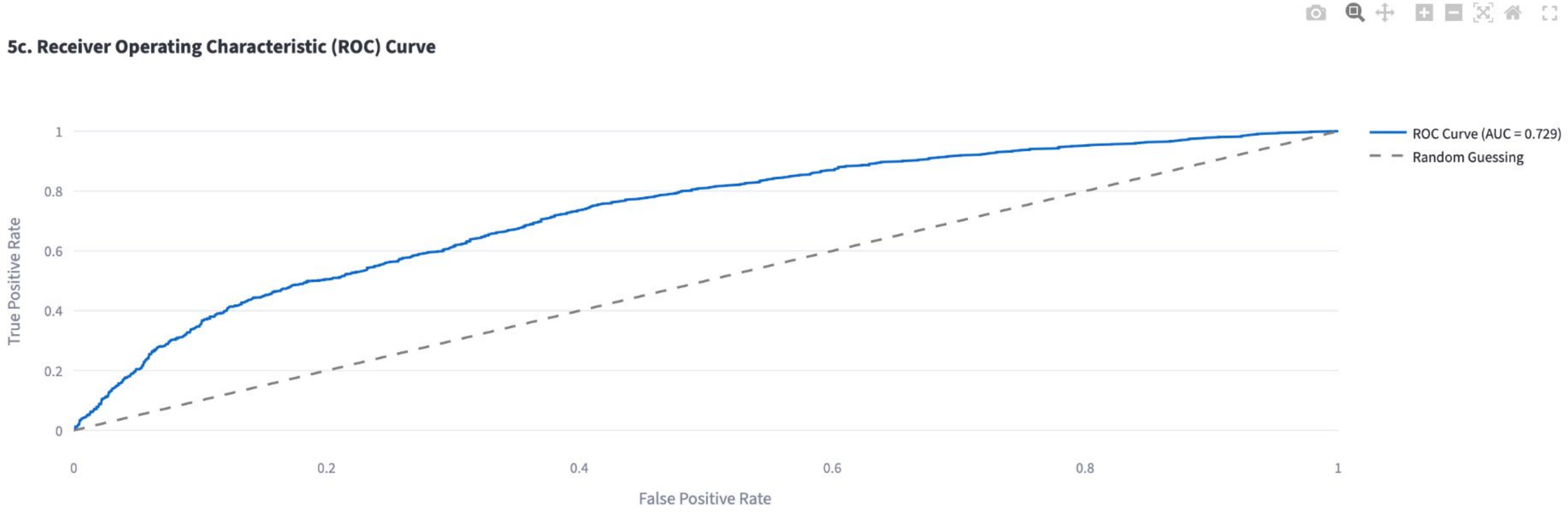
5b. Confusion Matrix

	Predicted Low PM25 (0)	Predicted High PM25 (1)
Actual Low PM25 (0)	923	516
Actual High PM25 (1)	476	1048

The matrix shows roughly twice as many true positives and true negatives as false positives and false negatives, indicating a reasonable balance in classification errors.

2. ROC Curve and Discrimination

5c. Receiver Operating Characteristic (ROC) Curve



The **Area Under the Curve (AUC)** is **0.729**. This indicates that the model has **good discrimination** between high and low PM2.5 risk, with a 72% chance of correctly distinguishing between the classes.

The ROC curve shows the trade-off between the True Positive Rate and the False Positive Rate. A curve that rises slowly (as seen here) indicates that discrimination ability is only slightly better than random guessing in some regions of probability.

KNN Model

streamlitApp

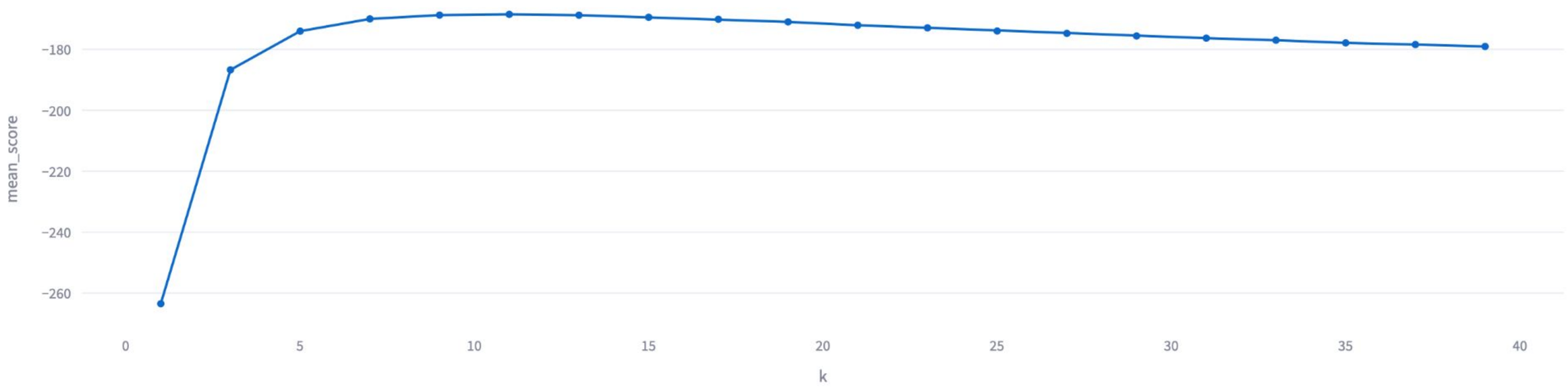
K-Nearest Neighbors Regression: Predicting AQI

KNN regression is employed to predict AQI (PM2.5) by using the fire, weather, and location datasets. KNN operates by measuring environmental similarity and averaging the AQI values of the K closest neighbors.

1. Model Tuning

2a. Cross-Validation for Optimal K

2a. Cross-Validated Negative MSE vs. K (Best K = 11)



Optimal k is **11**, chosen to minimize prediction error (Negative MSE). This means each prediction uses the **11 closest points**.

Manage app

2. Key Metrics

2b. Key Metrics on Test Set

streamlitApp

Metric	Value	Interpretation
Best K	11	Optimal number of neighbors used in the final model.
RMSE	12.574 (AQI units)	Average prediction error in AQI units.
R^2 Score	0.507	Percentage of AQI variance explained by the model.

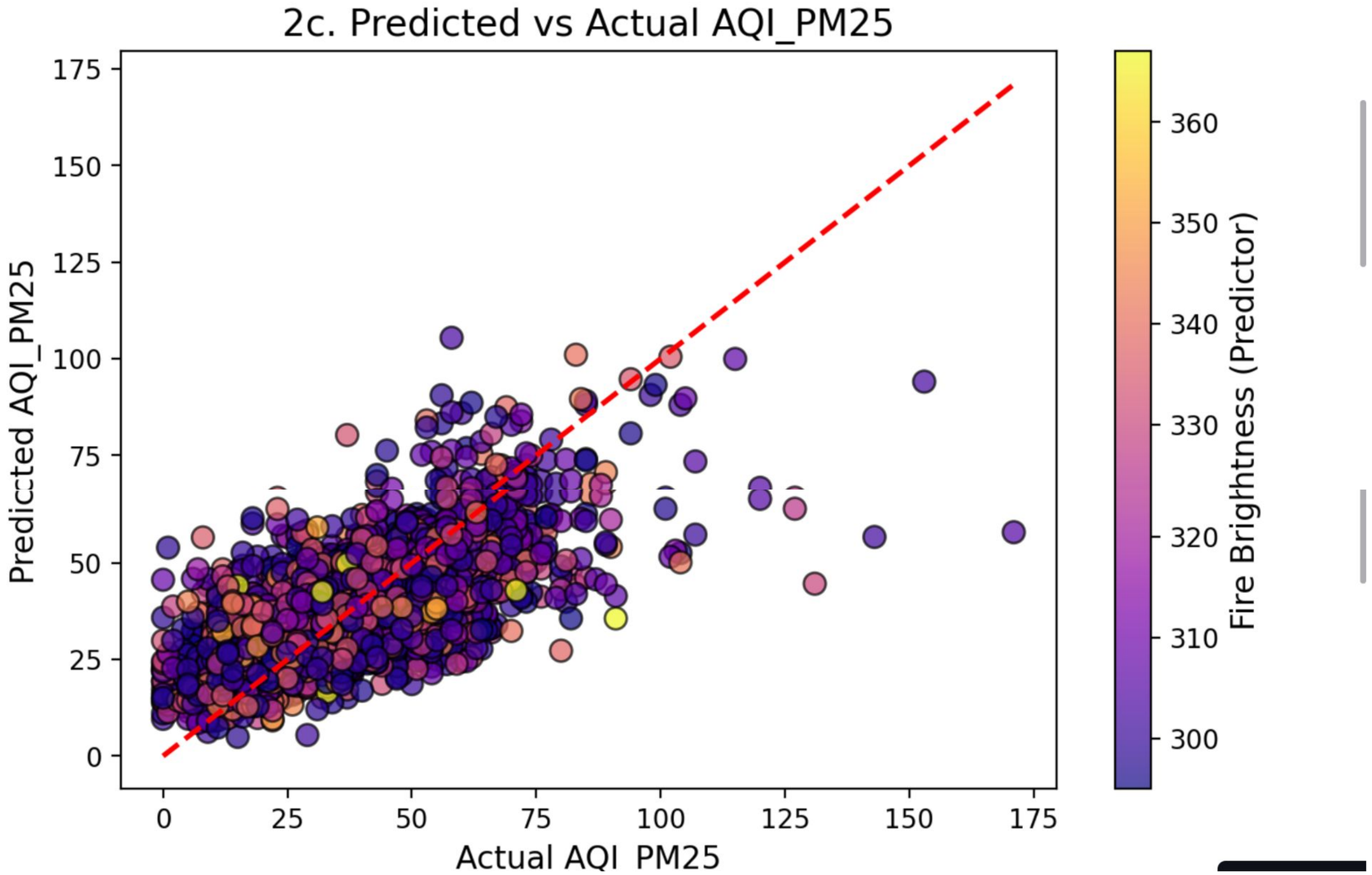
The R^2 score of **~51%** means the model explains approximately 51% of the variation in AQI, suggesting that other factors (e.g., traffic, industrial emissions) contribute significantly to the remaining 53% of unexplained variation.

The cross-validation RMSE of ± 12.98 **AQI units** means the model's predictions are, on average, off by about 13 AQI units.

3. Predicted vs. Actual Performance & Diagnostics

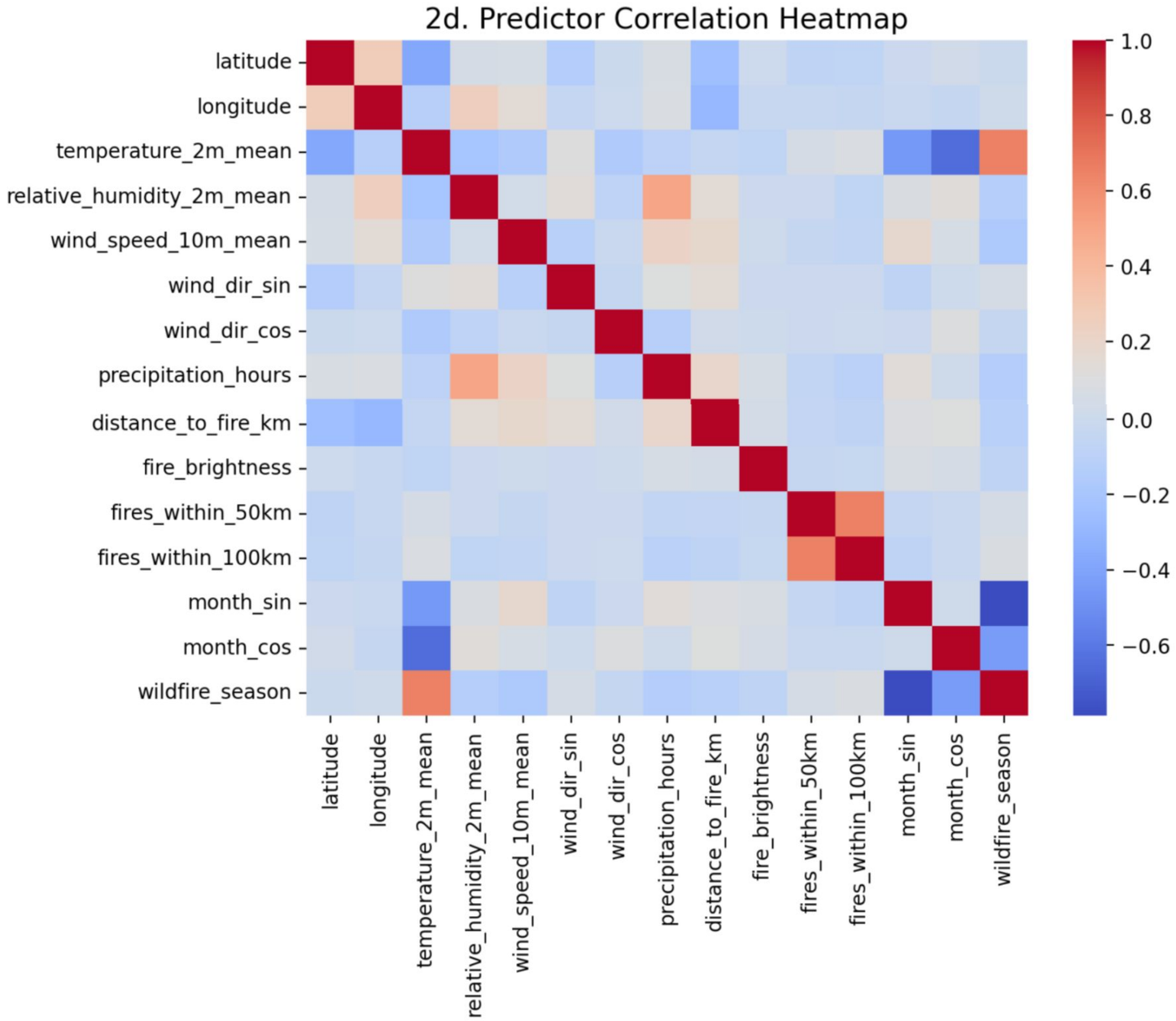
streamlitApp

2c. Predicted vs Actual AQI_PM25



Underprediction at Higher Values: As actual AQI increases beyond approximately 75, the predicted values start falling below the perfect prediction line. This indicates the model **underestimates high AQI events**, which is a common effect of KNN's smoothing (averaging) nature on extreme outliers.

2d. Predictor Correlation Heatmap



Multicollinearity, while not breaking the KNN model, was minimized by removing redundant variables like 'Fire FRP' (strongly correlated with 'fire brightness') to increase computational efficiency.

K-Means Clustering Model

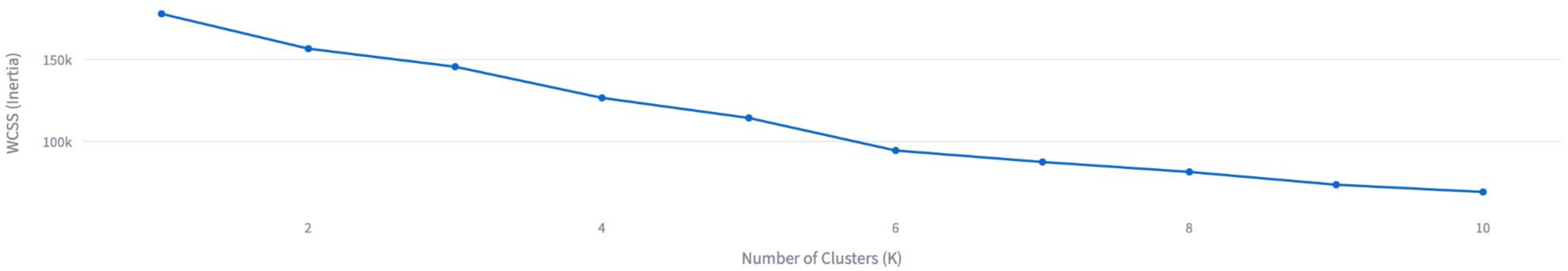
1. K-Means Clustering: Unsupervised Grouping

K-Means clustering is utilized to discover natural, unlabelled groupings, or environmental regimes, in the fire and weather data.

1. Optimal Cluster Selection

1a. Elbow Method (WCSS)

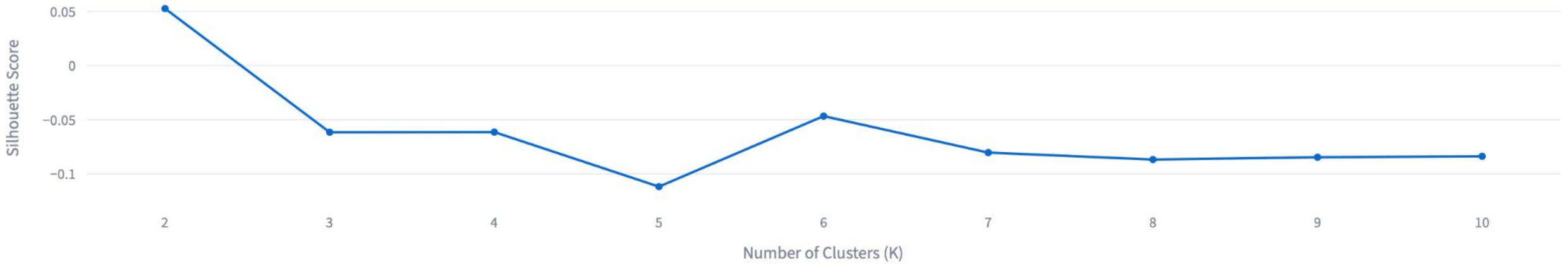
1a. Elbow Method: WCSS vs. K



The **steep decrease in WCSS from K=1 to K=3** (the 'elbow') suggests that $K = 2$ to $K = 5$ clusters potentially capture most of the patterns in the data.

1b. Silhouette Score

1b. Silhouette Scores for different K



The highest Silhouette Score, a very high **0.88** for $K = 2$, indicates clusters that are theoretically well-separated. Scores for $K \geq 3$ are less than 0.25, suggesting weak or overlapping structures.

2. Cluster Characteristics (K=2)

2a. Scaled Feature Centers

Cluster	latitude	longitude	temperature_2m_mean	wind_speed_10m_mean	precipitation_sum	fires_within_50km	fires_within_100km	distance_to_fire_km	fire_brightness
0	-0.6421	-0.4944	0.5787	-0.3641	-0.2076	0.0824	0.1305	-0.0139	-0.0538
1	0.5778	0.4449	-0.5207	0.3276	0.1868	-0.0742	-0.1174	0.0125	0.0484

Values are standardized (mean=0, std=1). Positive values indicate above-average feature levels for that cluster.

- **Cluster 1** (The high-PM2.5 cluster) is associated with **higher temperature** and **more nearby fires**.
- **Cluster 0** (The low-PM2.5 cluster) is associated with **higher wind speed**, **greater precipitation**, and **longer distances from fires**.

2b. Average PM25 per Cluster

cluster	PM25_Mean
0	8.7079
1	6.2707

Cluster 1 (PM25: **6.27**) captures **extreme pollution events** (only 2 points), while **Cluster 0** (PM25: **8.71**) captures **normal air quality** (almost the entire dataset).

Overall Conclusion: The clustering is **highly unbalanced**. The K-Means model is largely grouping the majority of the data together while isolating extreme outliers, indicating the model is not finding meaningful, balanced subgroups. This suggests that most of the data are within average environmental conditions, and the model is unable to find meaningful patterns beyond extreme cases.