

**Name:** Ayesha Zafar  
**Position:** Data Science  
**Duration:** Jun3 2025 – August 2025  
**Mentor:** Abdullah Farooqi & Muhammad Faizan  
**Organization:** 10pearls

**AQI Prediction System: Detailed Report**

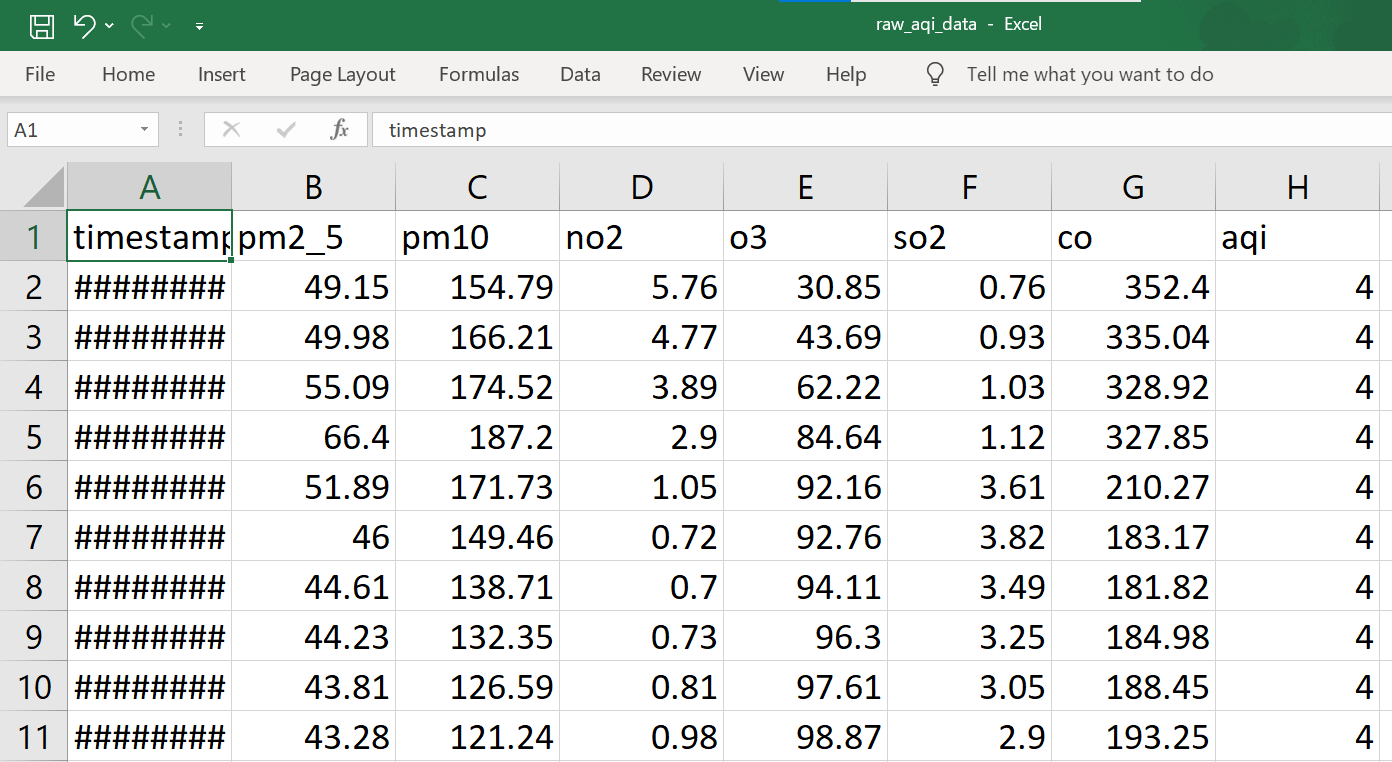
**Introduction**

The Air Quality Index (AQI) prediction system for Multan provides **short-term forecasts** of air quality by combining historical data, real-time API feeds, and machine learning models. The system is designed to be **serverless**, accessible via a **Streamlit Cloud dashboard**, and provides **3-day AQI forecasts** with visualizations and interpretable insights.

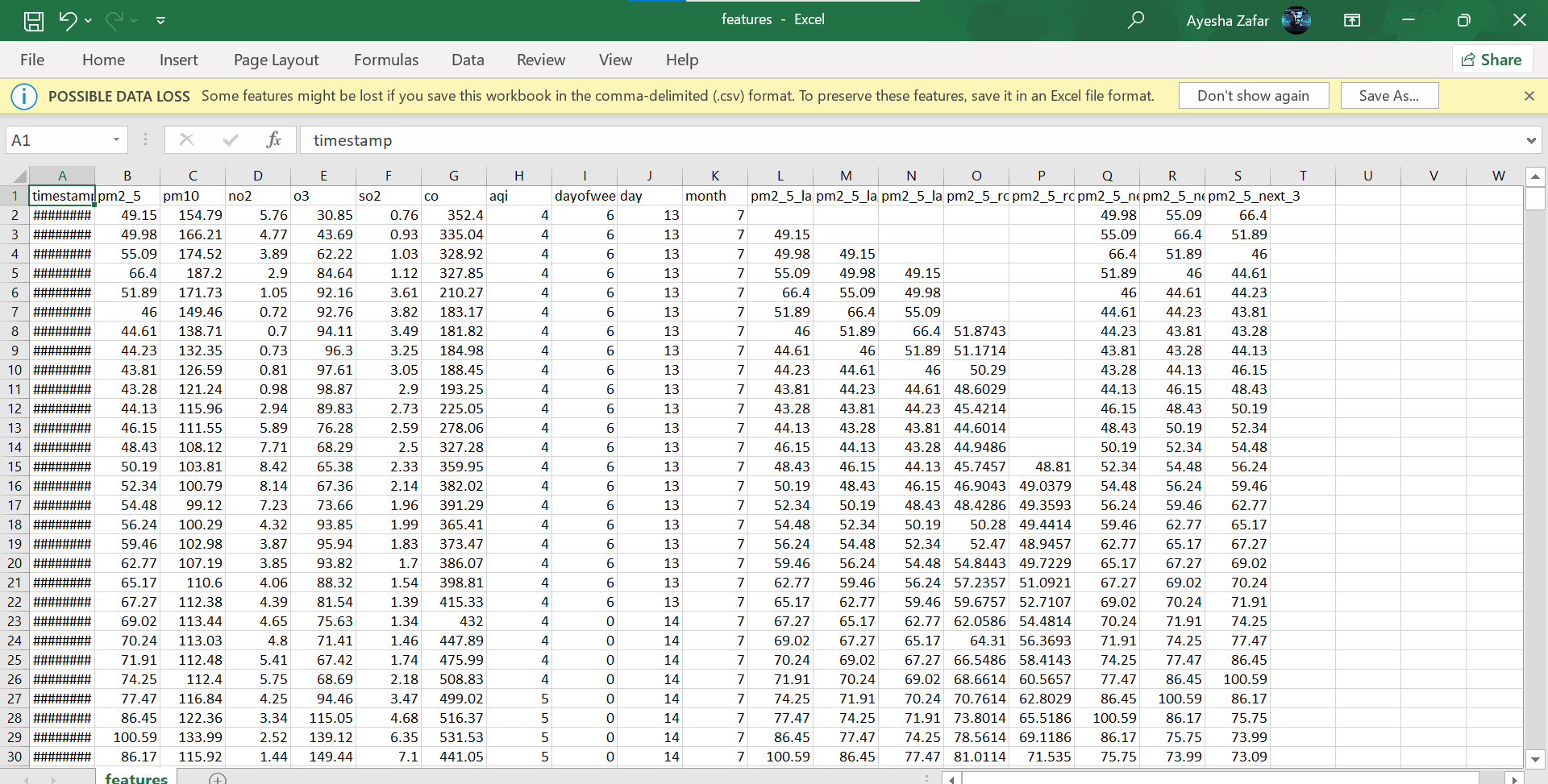
It aims to help citizens and stakeholders track air pollution trends and make informed decisions to minimize exposure to unhealthy air conditions.

**1) Data Acquisition**

* **Historical Data:** Fetched **past 30 days** of PM2.5 and weather data using **Pandas**. This provided a basis for generating lag features and understanding recent pollution trends.
* **Future Forecast Data:** Integrated **OpenWeather API** to collect predicted weather and pollution parameters for Multan. This ensures predictions account for upcoming conditions such as temperature, wind, and humidity.

By combining historical and forecasted data, the system can provide **accurate short-term predictions**. 

**2) Feature Engineering and Preprocessing**

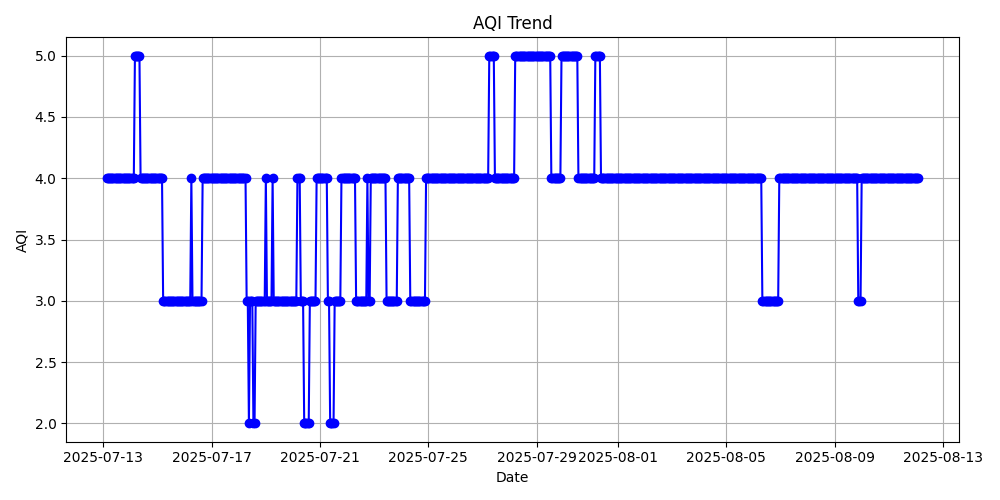
* **Lag Features:** Created lagged PM2.5 values (lag1, lag2, lag3) to capture short-term autocorrelations in pollution levels.
* **Rolling Features:** 7-day and 14-day rolling averages of PM2.5 to model trends and seasonal effects.
* **Date Features:** Added day, month, and day-of-week to capture weekly or monthly patterns.
* **Merged Dataset:** Combined historical PM2.5, weather data, and forecast features into a feature.csv file for model training.

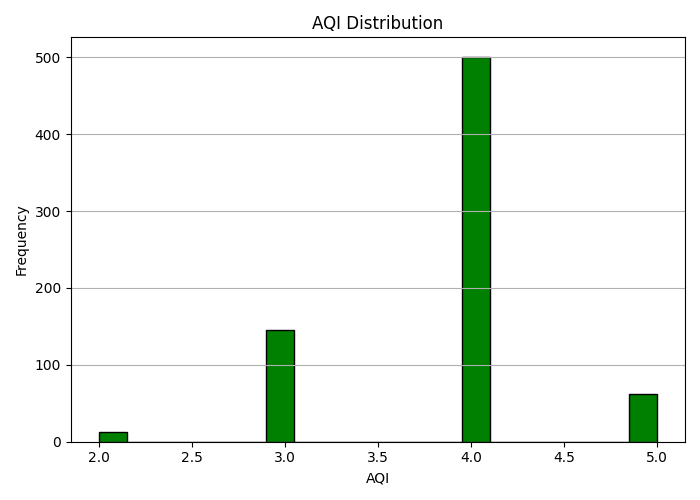
EDA was conducted to identify patterns, trends, and correlations:

* **Distribution Analysis:** Checked distribution of AQI values to identify typical pollution levels.
* **Trend Analysis:** Observed rising AQI during certain weather patterns (low wind speed + high temperature).
* **Correlation Heatmap:** Found strong correlation between PM2.5 and overall AQI, and moderate correlation with humidity and wind speed.
* **Seasonality Detection:** AQI levels were higher on certain days due to local environmental conditions.

**Insights from EDA:**

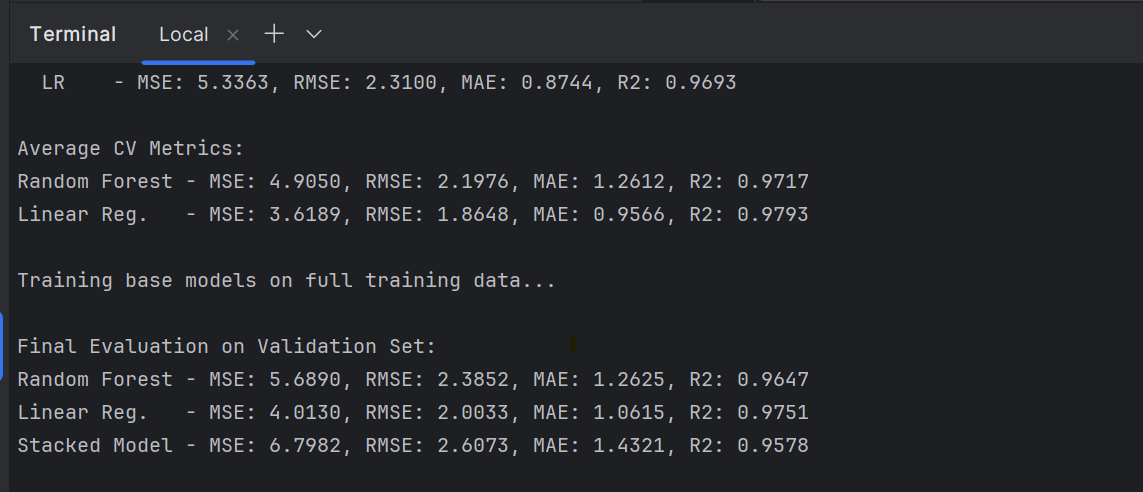
* PM2.5 is the most critical pollutant for AQI prediction in Multan.
* Humidity inversely correlates with AQI in certain weather patterns.
* Weather conditions such as low wind speed contribute to stagnant air and higher pollution levels.

****

****

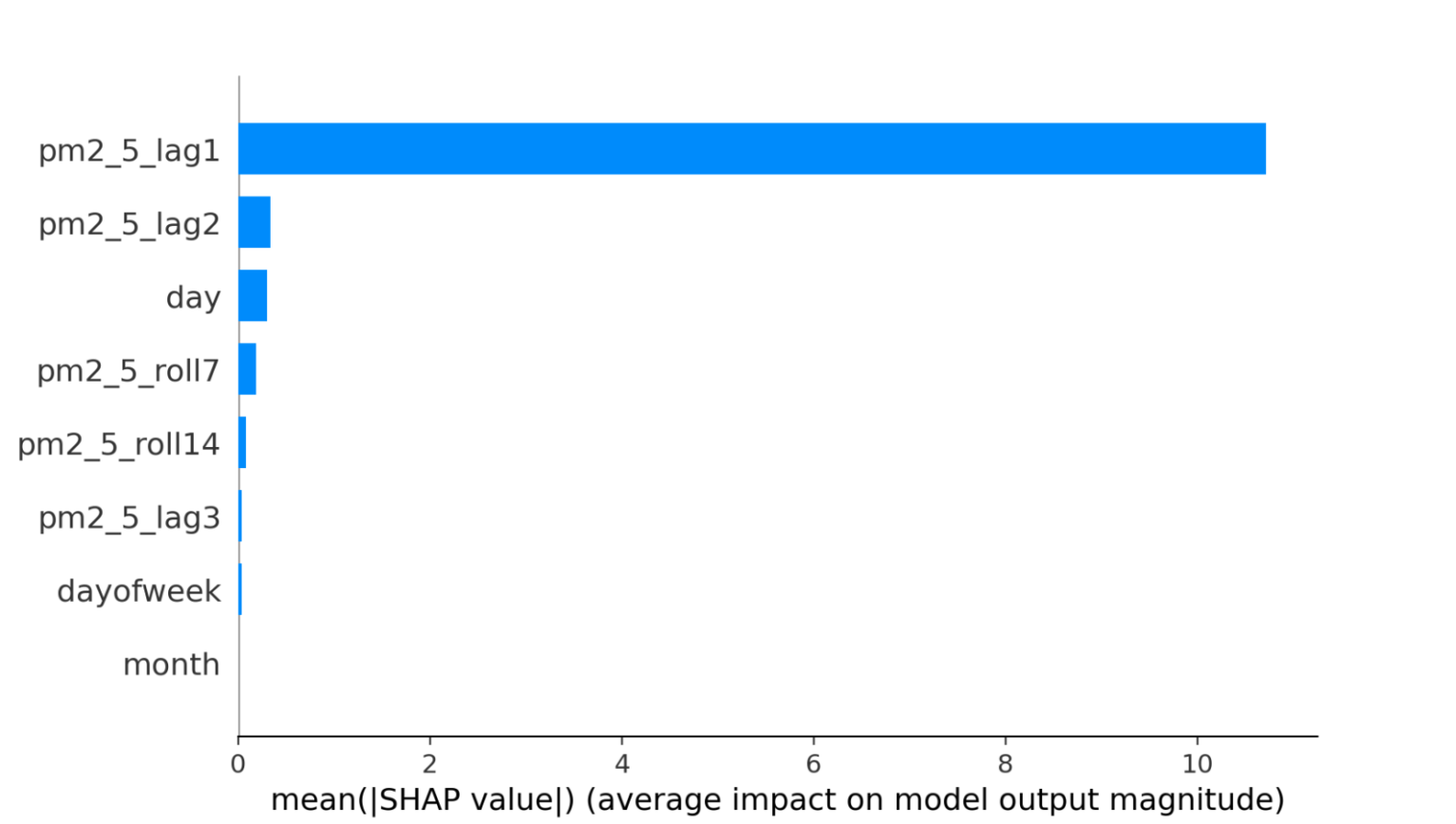
**3) Machine Learning Model Development**& SHAP Interpretability

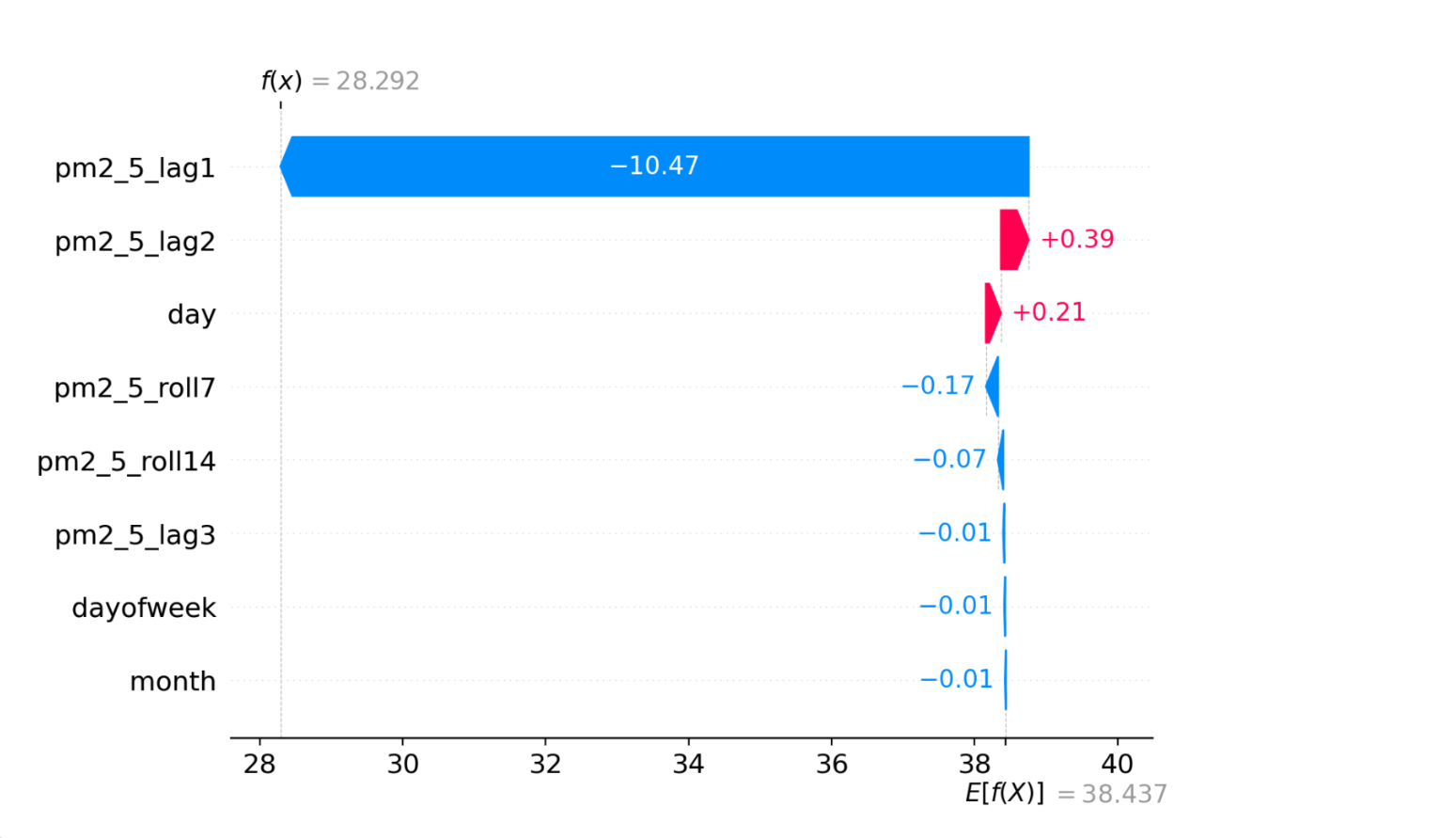
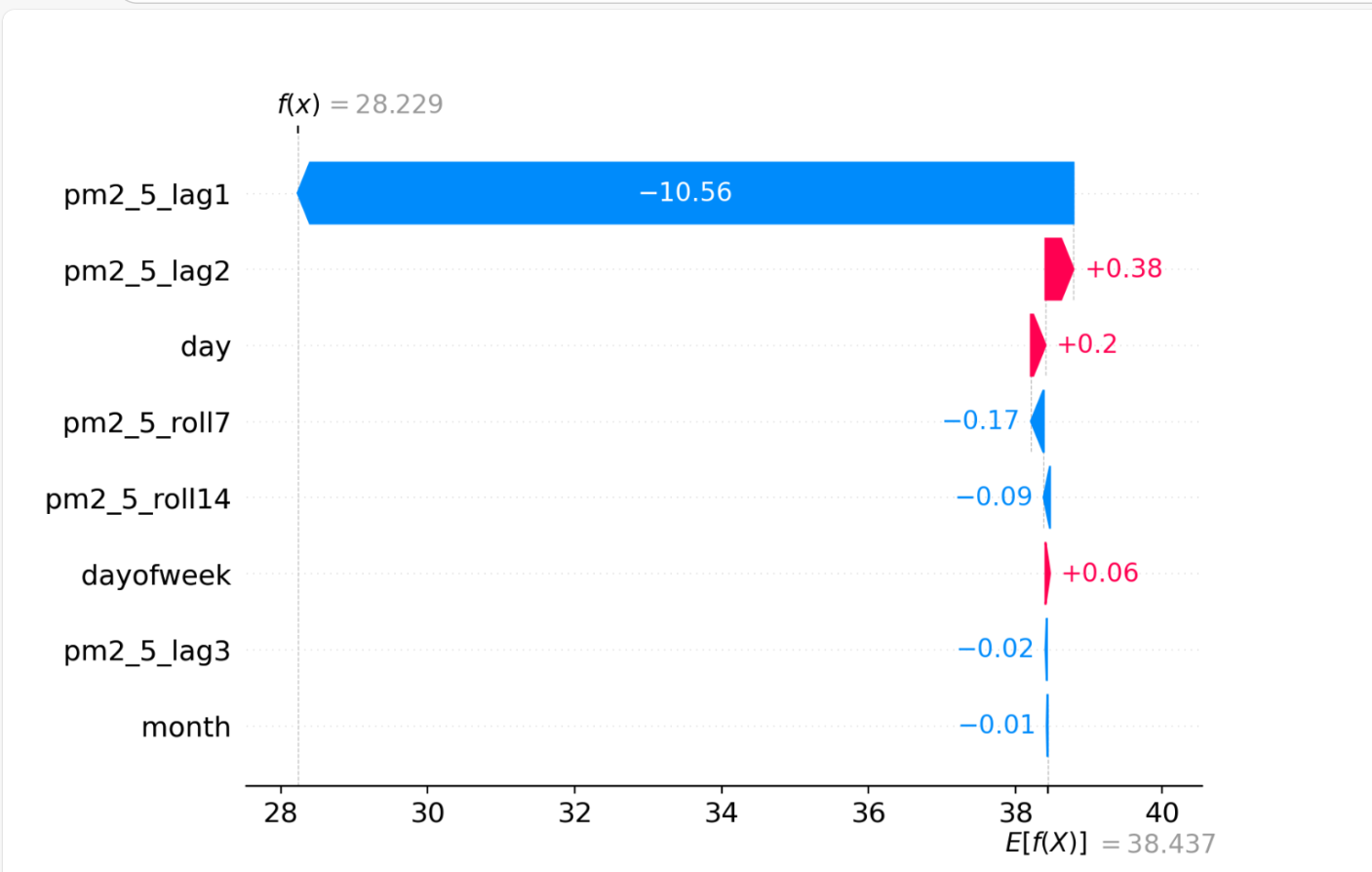
* **Models Used:**
  + **Random Forest Regressor** – robust for non-linear relationships.
  + **Linear Regression** – simple baseline model.
  + **Stacked Model** – ensemble combining Random Forest and Linear Regression to improve forecast accuracy.
* **Model Storage:** Models serialized to .pkl files for deployment and reuse in the Streamlit app.
* **Evaluation:** Validated on historical data using metrics such as **MAE** and **R²**, ensuring reliable short-term AQI predictions.

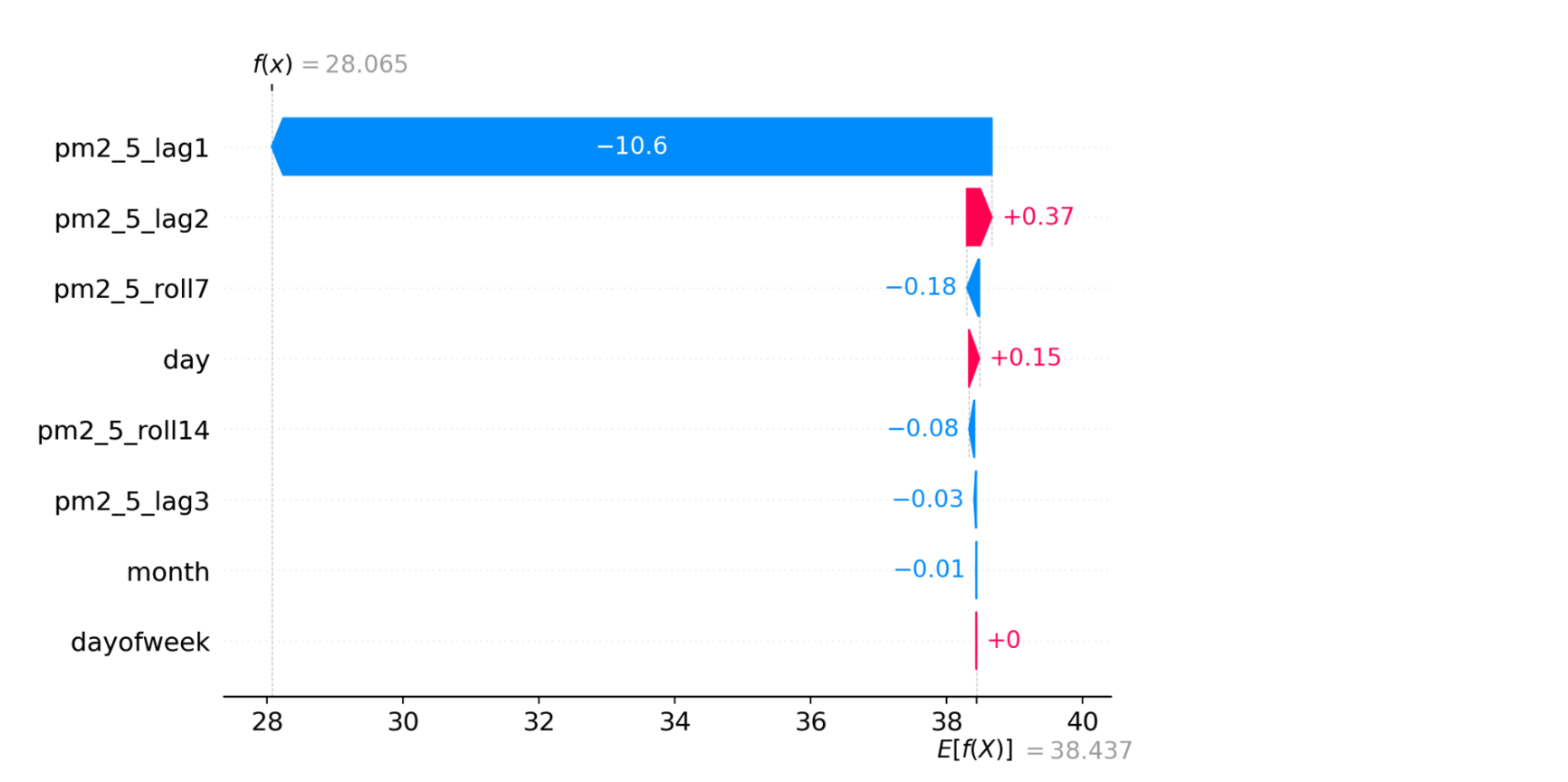


**SHAP (SHapley Additive exPlanations) Analysis:**

* SHAP was used to interpret the model predictions and understand **which features most influence AQI forecasts**.
* **Findings from SHAP:**
  + **PM2.5 levels** contributed the most to prediction changes.
  + **Temperature & wind speed** had significant influence on next-day AQI.
  + API-based forecasted weather features had stronger effects on 2-3 day predictions compared to historical lag features.







**4) 3-Day Forecasting**

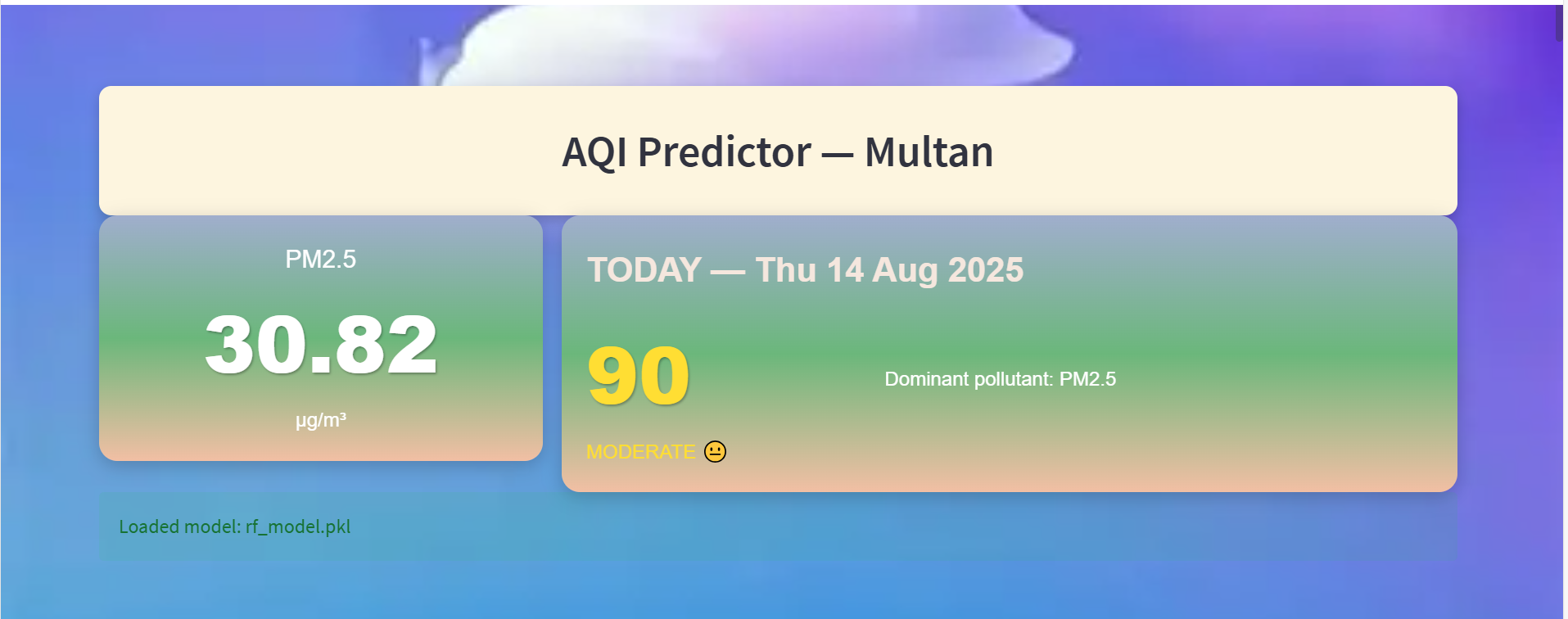
The system predicts AQI for the **next 3 days** using both historical PM2.5 trends and API-provided weather/pollution forecasts.

* **Prediction Process:**
  1. Extract latest PM2.5 and weather features.
  2. Generate lagged and rolling features for each future day.
  3. Apply the stacked model to forecast PM2.5, then convert to AQI using standardized formulas.
* **Outputs:** For each day, the dashboard shows:
  1. Forecasted PM2.5 concentration (µg/m³)
  2. AQI value
  3. Category (Good, Moderate, Unhealthy, etc.) with color-coded visualizations
  4. Emojis to quickly convey air quality severity
* **Trend Analysis:** Forecasts are plotted alongside historical data to show continuity and predicted fluctuations.

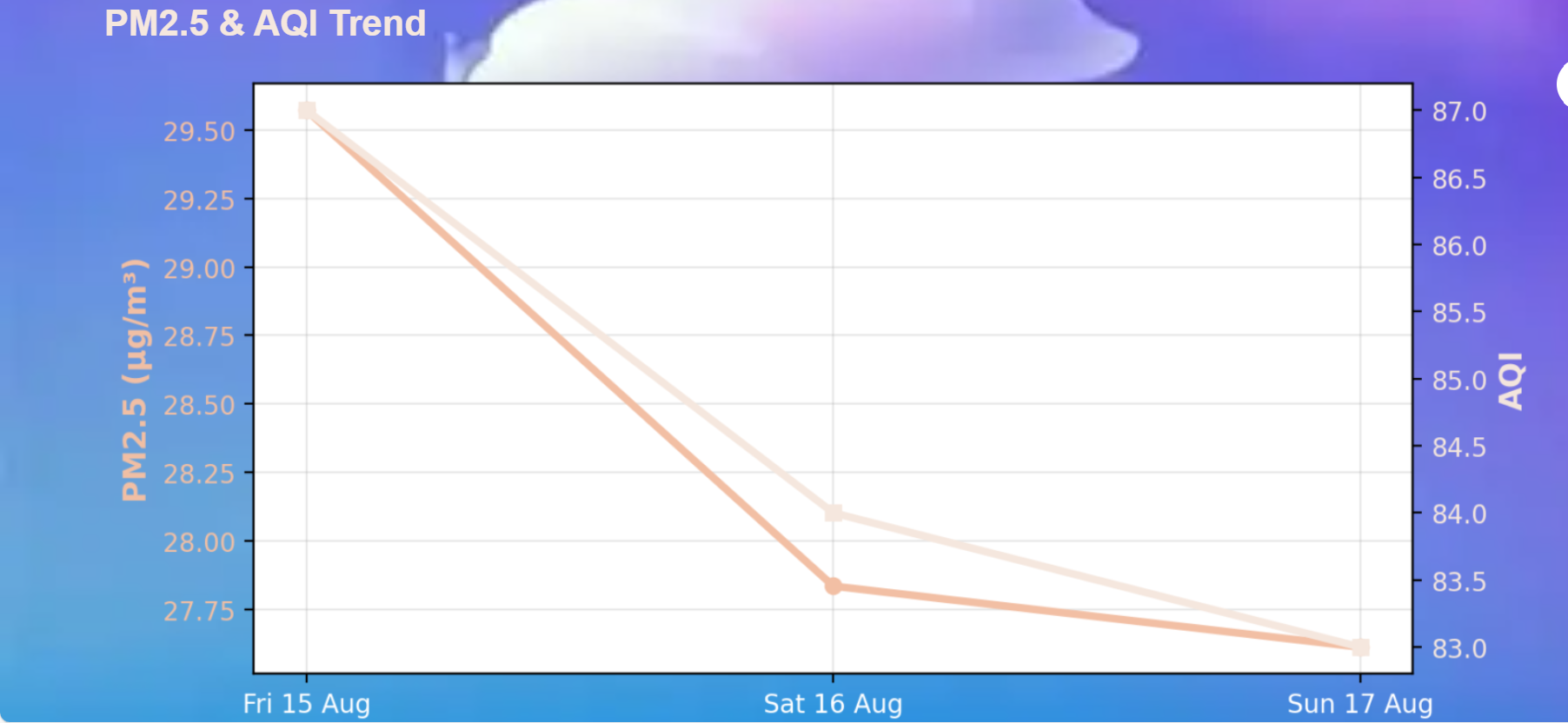
This enables users to **plan outdoor activities** and take precautions during high pollution periods.

**5) Streamlit Dashboard**

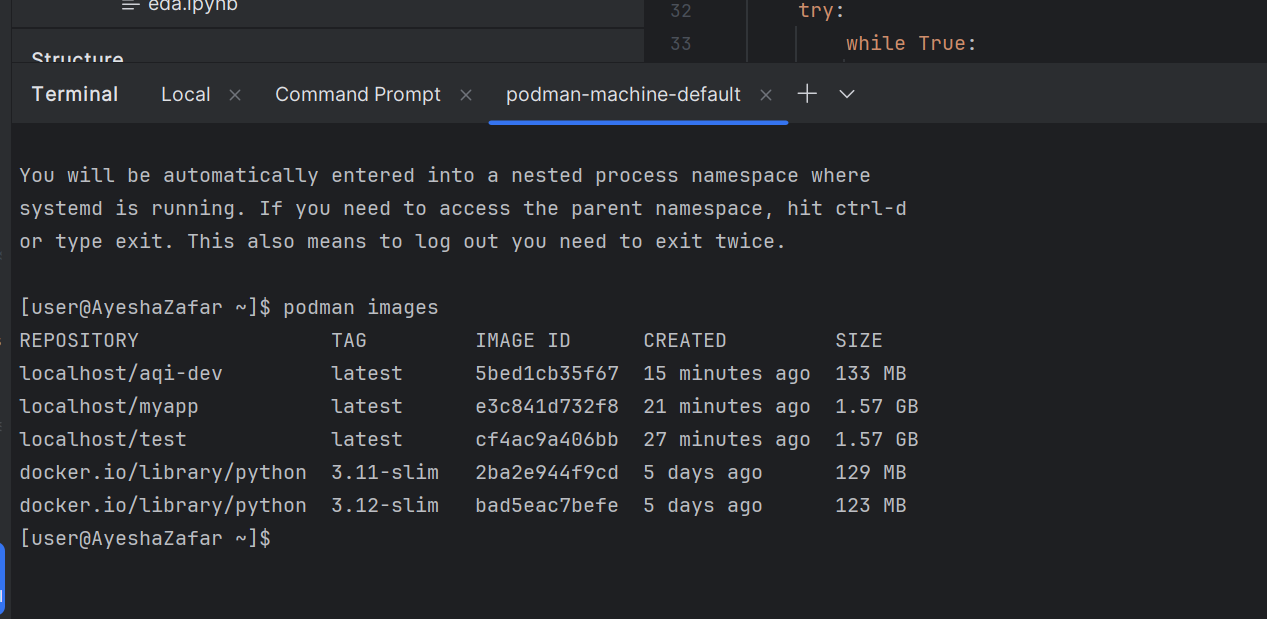
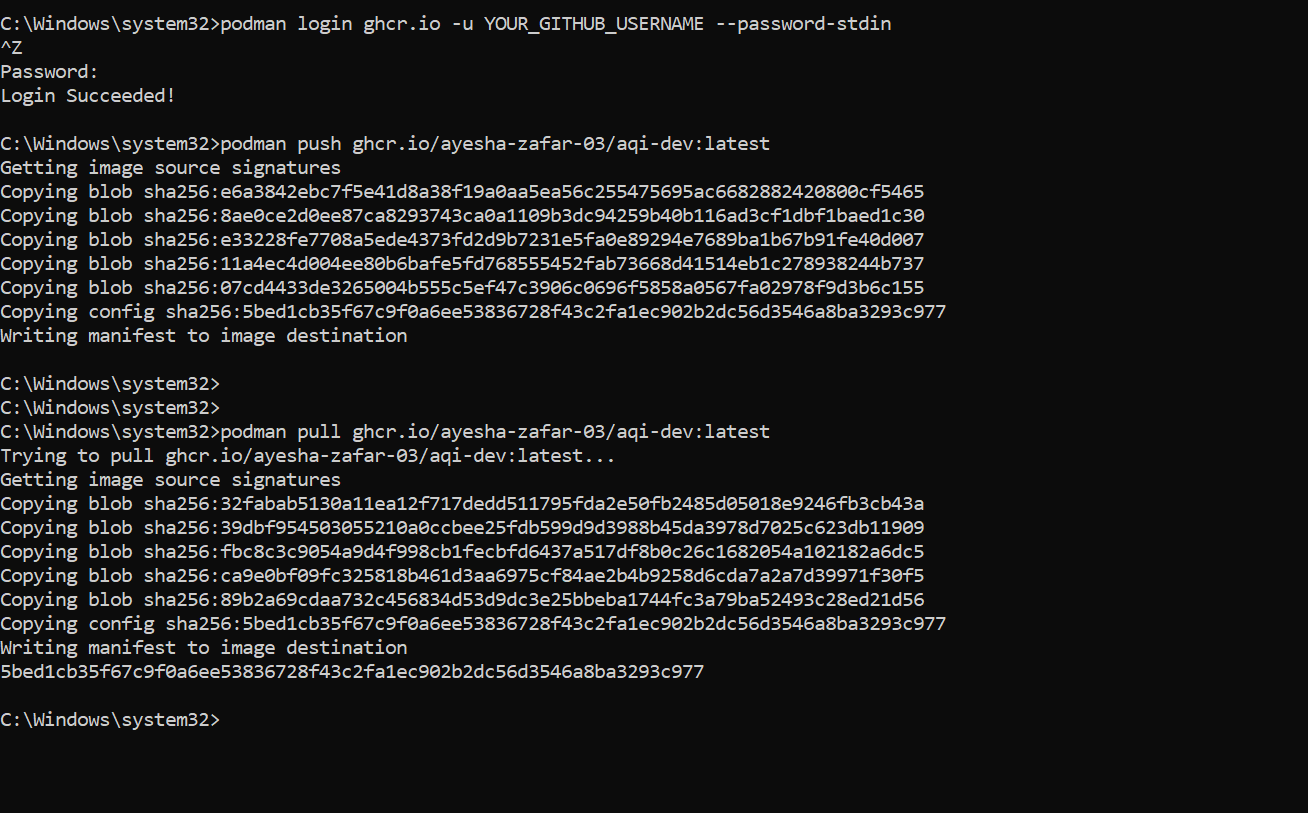
* **Interactive Interface:** Accessible via Streamlit Cloud.
* **Features:**
  + Current PM2.5 and AQI values
  + 3-day AQI forecast cards with color-coded categories
  + PM2.5 and AQI trend plots
  + SHAP global feature importance and per-day contribution plots for explainability
* **Design:** Custom CSS for a visually appealing, responsive, and user-friendly interface.







**6) Containerization & Serverless Deployment**

* **Containerization:** Used **Podman** to package the app, dependencies, and models.
* 
* **Container Storage:** Stored images on **GitHub Container Registry (GHCR)** to version and share the containerized app.
* 
* **Serverless Deployment:** Hosted on **Streamlit Cloud**, making it accessible online without server or infrastructure management.

**7) CI/CD Pipelines**

* **GitHub Actions:** Configured CI/CD to automate testing and deployment:
  + Code pushed to GitHub triggers automated build and tests.
  + Successful builds deploy updated models and app code to Streamlit Cloud.
* Ensures **continuous updates** without manual intervention and maintains consistency across deployments.

**8) Challenges Faced**

* Aligning historical data and API forecasts for consistent feature generation.
* Interpreting model predictions with SHAP to explain AQI drivers.
* Managing containerization and deployment for a serverless architecture.

**9) Future Work**

* Expand predictions to **other major cities** in Pakistan.
* Integrate additional environmental data (e.g., satellite imagery, industrial emissions) for higher prediction accuracy.

**Conclusion**

This AQI prediction system combines **historical trends**, **API forecasts**, **machine learning**, and **interpretability tools** to deliver reliable **3-day AQI forecasts** for Multan. The **serverless Streamlit deployment**, **containerization**, and **CI/CD pipelines** ensure accessibility, scalability, and continuous updates, making it a valuable tool for environmental monitoring and public health awareness.