Air Quality Predictive Dashboard: Analysis and Forecasting

A Data Science Project on Environmental Monitoring, Predictive Modeling of PM2.5 Concentration, and Interactive Dashboard Visualization.

> Good morning. I'm presenting my final project: an Air Quality Predictive Dashboard. The goal was to go from raw, hourly air quality data to an integrated system capable of insightful analysis and accurate short-term pollutant forecasting.

The Challenge: Predicting Fine Particulate Matter (PM2.5)

Project Goal

Develop robust models to analyze and forecast hourly **PM2.5 concentration**, a critical measure of air pollution that significantly impacts public health.

Accurate Forecasting

Predict upcoming high-pollution events with high confidence.

Interactive Visualization

Deliver actionable insights via a user-friendly Streamlit dashboard.

Data Context: Realistic, hourly air quality readings from a monitoring station (e.g., Delhi), including PM2.5, PM10, NO2, Temperature, and Wind Speed.

M1: Data Understanding & Preprocessing

Data Cleaning

Handled missing values using a **Forward-Fill** approach to maintain the time-series integrity.

Feature Engineering

Converted timestamps, set index, and ensured data was ready for time-series analysis.

Distribution Analysis

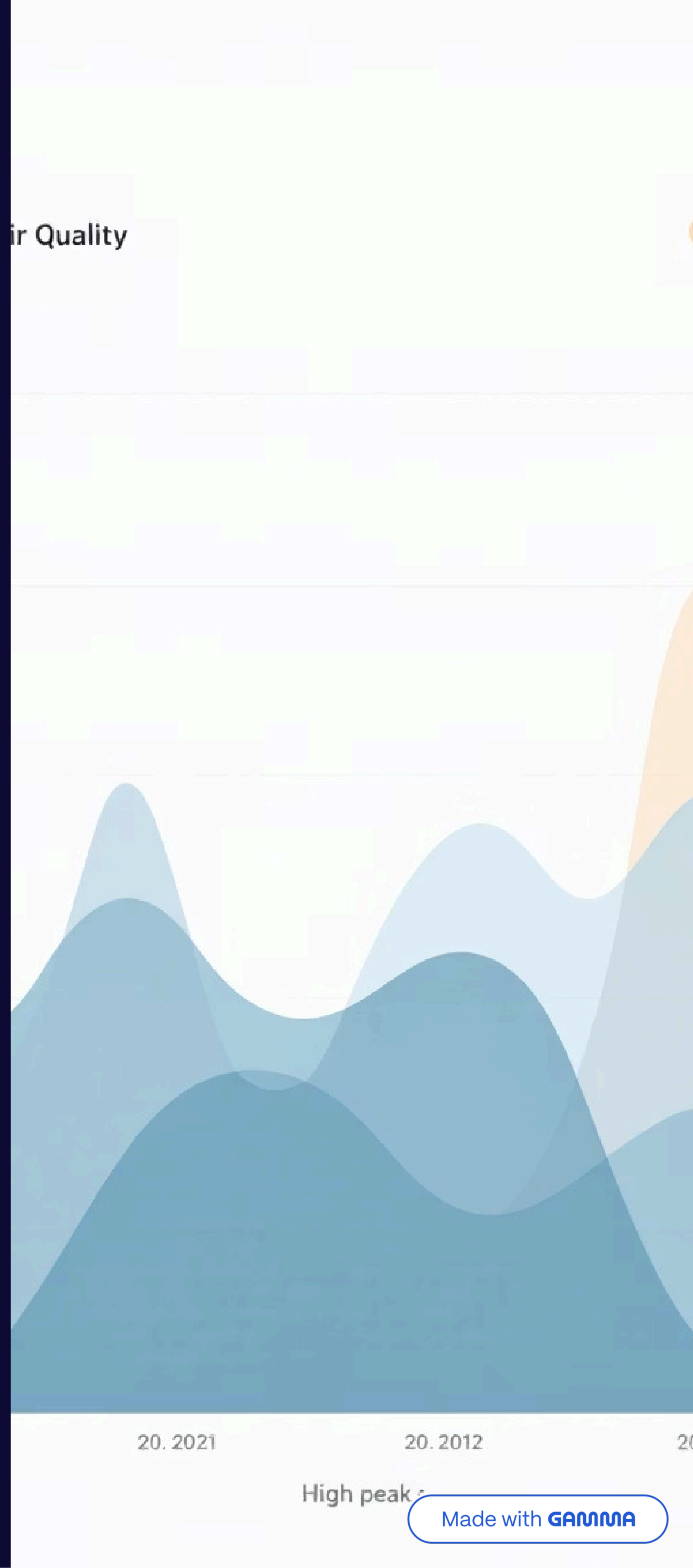
Confirmed PM2.5 is significantly **right-skewed**, emphasizing the need to model extreme values.

Key Findings: Temporal Patterns

Exploratory Data Analysis (EDA) revealed clear **diurnal (hourly)** and **seasonal** patterns in pollutants, suggesting strong time-series components.

Key Findings: High Correlation

Identified a very strong **linear relationship** between PM10 and PM2.5. PM10 will serve as a primary explanatory variable for Model 1.



M2: Model 1 - Predicting PM2.5 from PM10 (Causal Model)

The high correlation discovered in M1 allowed us to build a simple, highly effective causal model where PM10 acts as the sole feature to predict PM2.5 concentration.

- **Objective:** Use PM10 as the sole feature to predict PM2.5 concentration.
- **Methodology:** Simple **Linear Regression**, trained on a 70% train / 30% test split.

·0.95

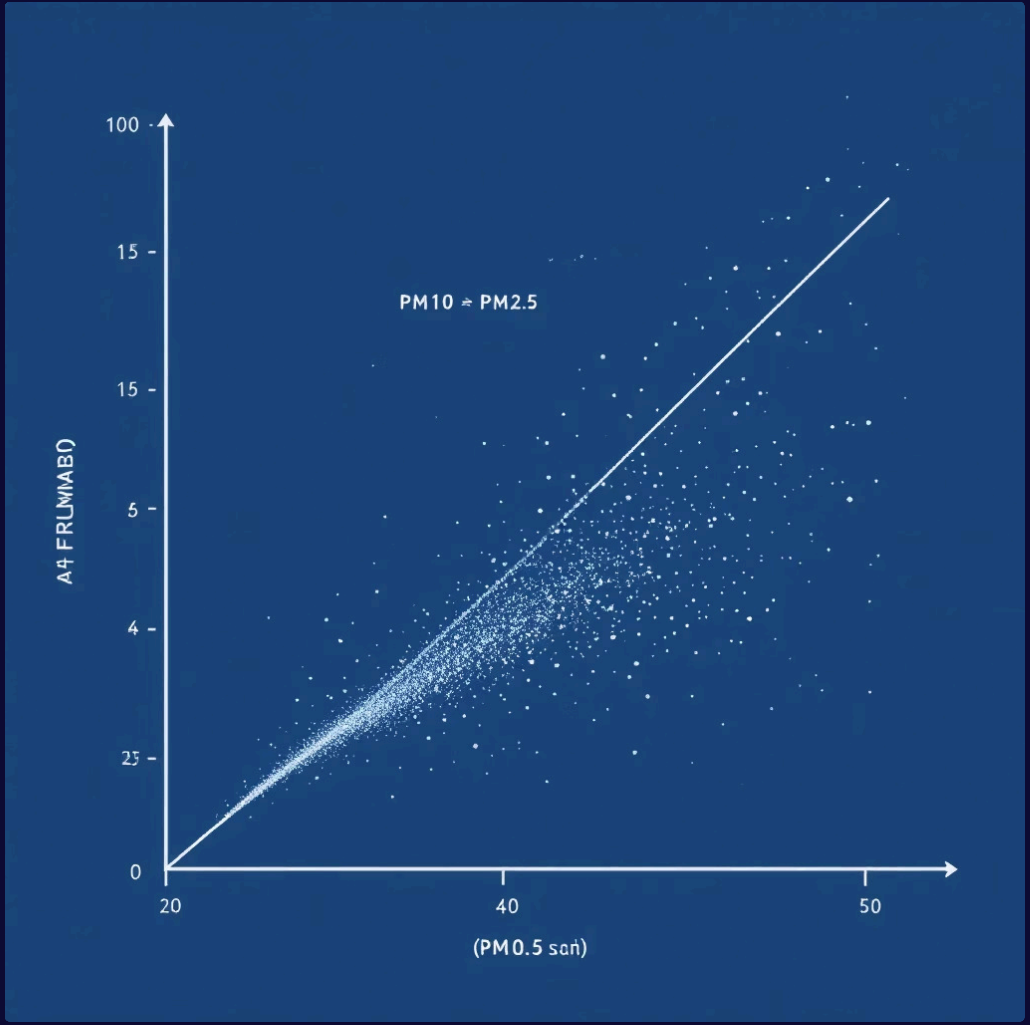
R-squared ($R˙2$)

Strong model fit, indicating PM10 explains 95% of the variance in PM2.5.

·8.5

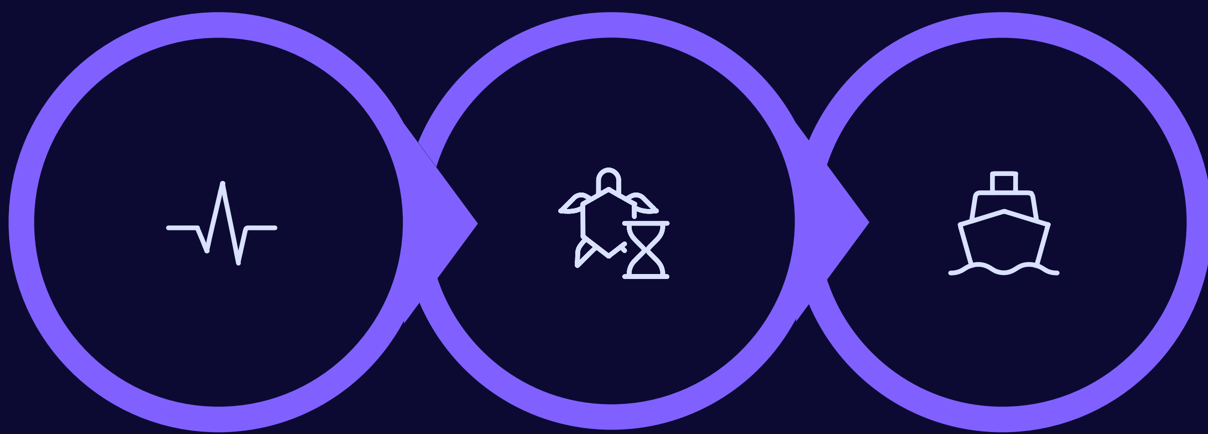
Mean Absolute Error (MAE)

Prediction error is, on average, low (in $\mu g/m^3$ on the test set).



M3: Model 2 - 24-Hour PM2.5 Forecast (Time Series)

For true time-series prediction, we implemented a Lag-Feature Linear Regression model, leveraging the principle that a variable's immediate past is the strongest predictor of its near future.



Raw Time Series

Lag Features

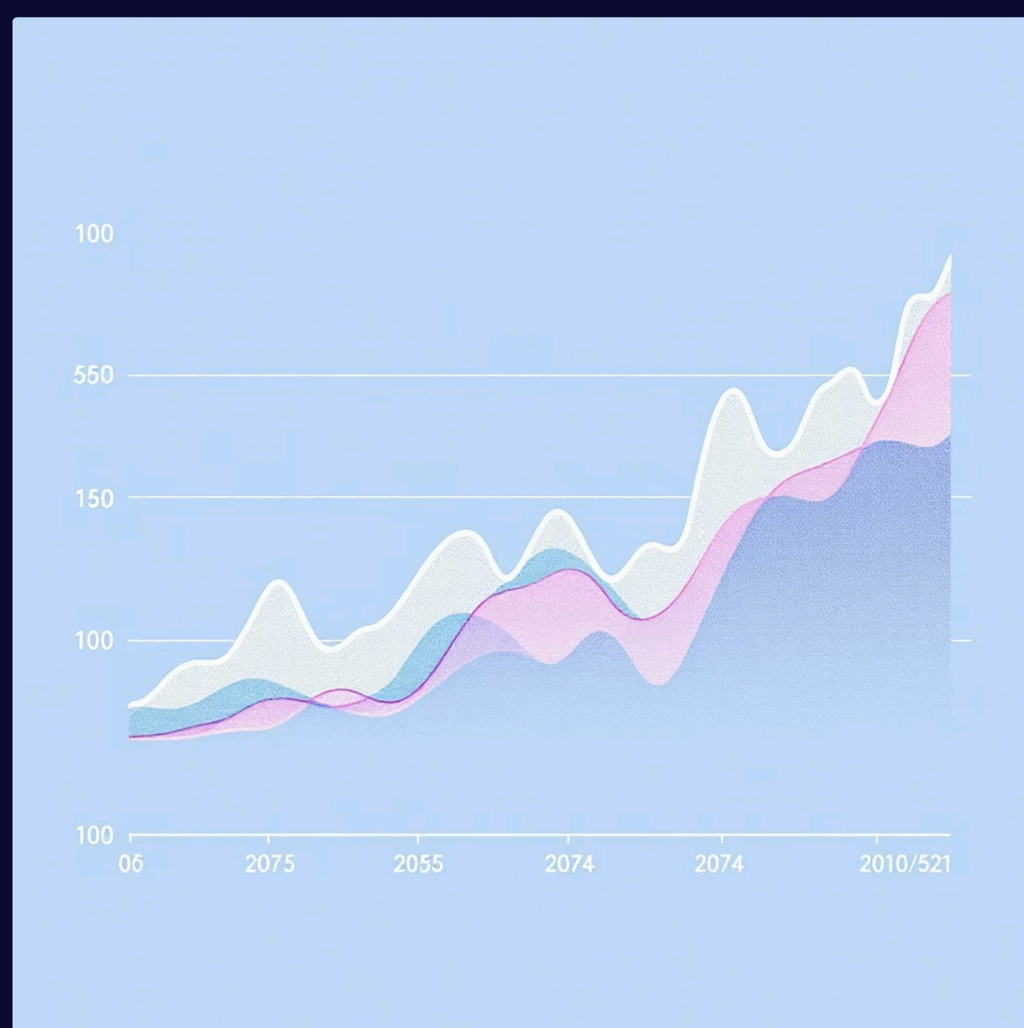
Train Linear Model

Methodology Details

- Created **Lag Features** (e.g., PM2.5 at $t-1$, $t-2$, $t-3$ hours) to capture autocorrelation.
- Trained a new Linear Regression model utilizing these historical values to forecast the value at time t.

Key Result

Model Coefficients showed high values for the most recent lags (especially **t-1**), confirming that the air quality one hour ago is the single strongest predictor. This enables robust hour-by-hour forecasting.



M4: The Interactive Predictive Dashboard

The analysis and models were integrated into a single, user-friendly **Streamlit** web application, making insights and predictions immediately accessible to end-users like environmental agencies.

Tech Stack

Built using Python, Pandas for data processing, Scikit-learn for modeling, and Streamlit for rapid dashboard development.

Data Overview

Visual display of statistical summaries and pollutant distributions (outputs from M1).

Causal Model Results

Interactive visual validation of the high-performing PM10 → PM2.5 prediction model (M2 outputs).

24-Hour Forecast

Visualization of actual vs. predicted time series data using the Lag-Feature model (M3 outputs).

Summary and Next Steps

Project Success

Successfully identified key correlations, developed two high-performing predictive models (Causal & Time Series), and integrated findings into a professional dashboard.

Non-Linear Models

Explore advanced non-linear techniques, such as Random Forest or LSTM Neural Networks, for potentially improved accuracy in capturing complex dynamics.

External Data Integration

Incorporate real-time external features, like traffic data, local industrial activity, or meteorological forecasts, for contextual prediction.

Forecast Expansion

Expand the current 24-hour forecast horizon to 48 or 72 hours, increasing the tool's utility for proactive policy-making.

This project demonstrated a full data science workflow, resulting in effective models for air quality prediction. The future work focuses on evolving the system into a truly operational and robust environmental forecasting tool.