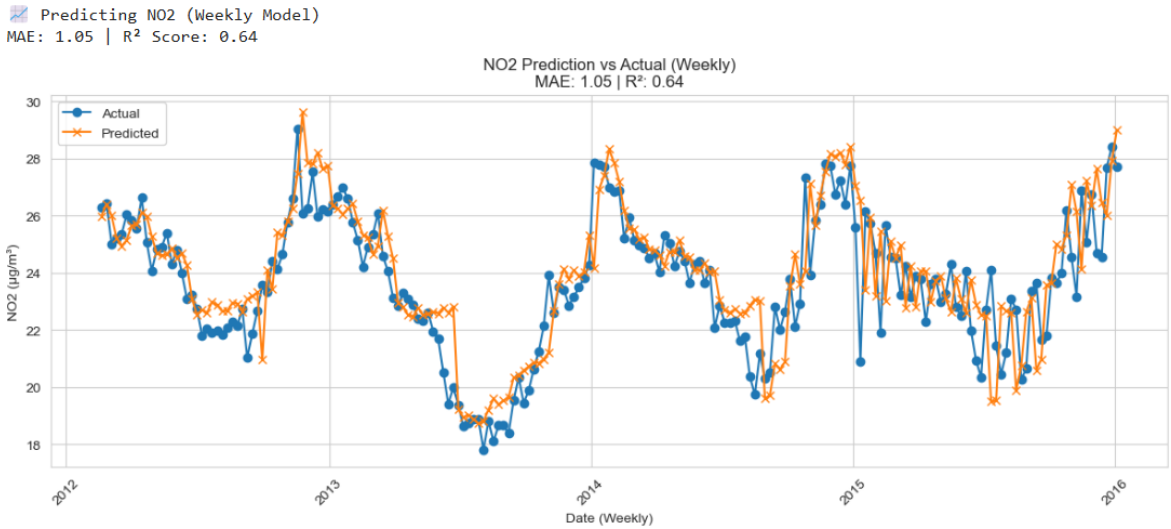
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”
Topic: A quick run on all 1-4 parts. Model: random forest

1. Dataset Exploration

2. Temporal Analysis

3. Regional Trends

4. Predictive Modeling



Extra: maybe web service for data science?

Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

1. Dataset Exploration

- Dataset: India Air Quality Data (Kaggle)
- Objective: Understand air pollution trends and explore correlation with environmental policies.
- Data Info:
    - Source: Historical Daily Ambient Air Quality Data
    - Key Features: sampling _date, state, city, pollutants like SO2, NO2, PM2.5, etc.
    - Preprocessing Steps:
        - Converted sampling_date to datetime
        - Extracted year, month, and week
        - Checked for missing dates and data integrity

Initial Findings:

- Coverage spans multiple years and regions
- Varying availability of pollutants across observations

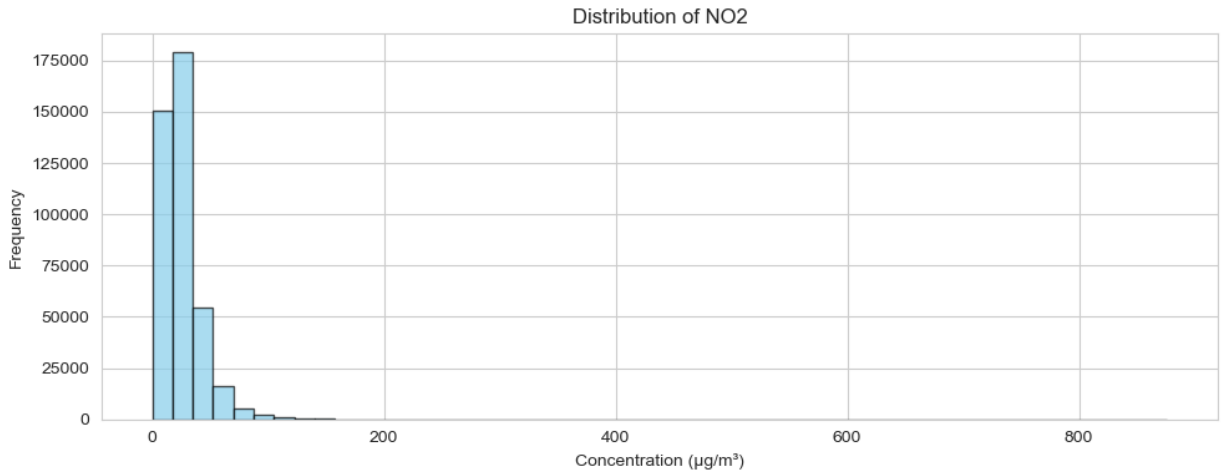
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

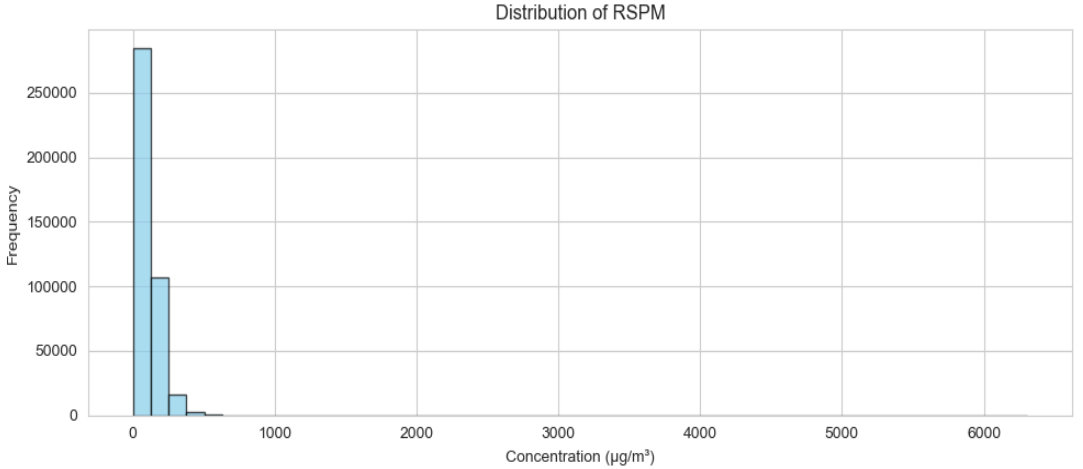
1. Dataset Exploration

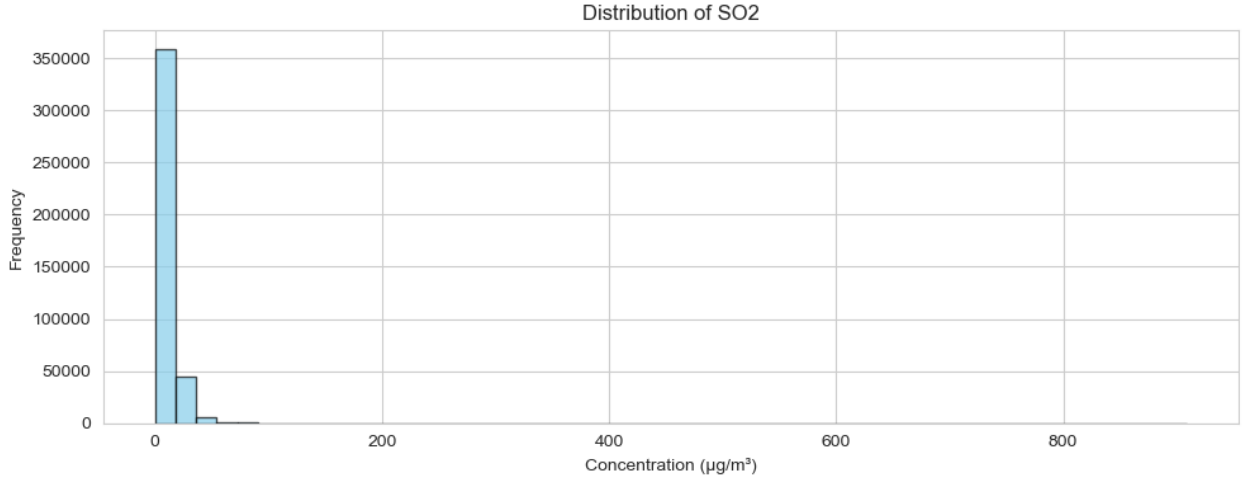
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                      Non-Null Count   Dtype  
---  ------                      --------------   -----  
 0   stn_code                    291665 non-null  object 
 1   sampling_date               435739 non-null  object 
 2   state                       435742 non-null  object 
 3   location                    435739 non-null  object 
 4   agency                      286261 non-null  object 
 5   type                        430349 non-null  object 
 6   so2                         401096 non-null  float64
 7   no2                         419509 non-null  float64
 8   rspm                        395520 non-null  float64
 9   spm                         198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object 
 11  pm2_5                       9314 non-null    float64
 12  date                        435735 non-null  object 
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

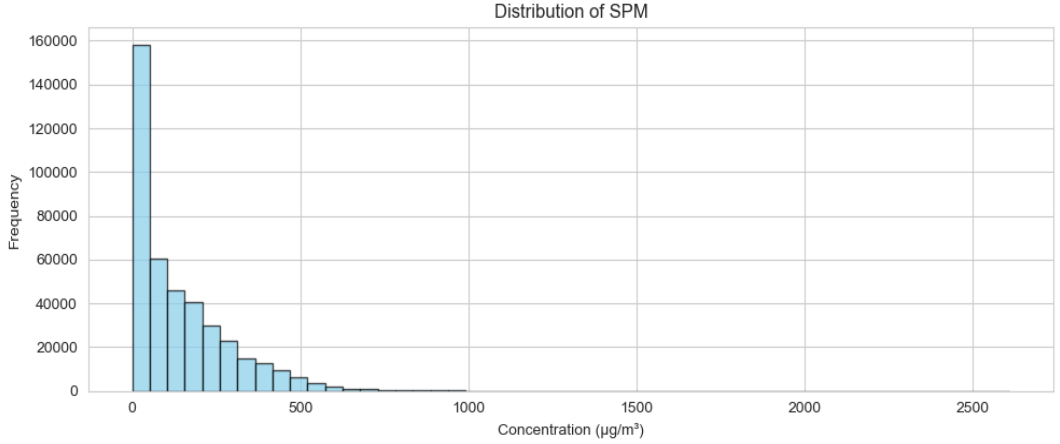
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

Pollutant variables’ distributions









Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

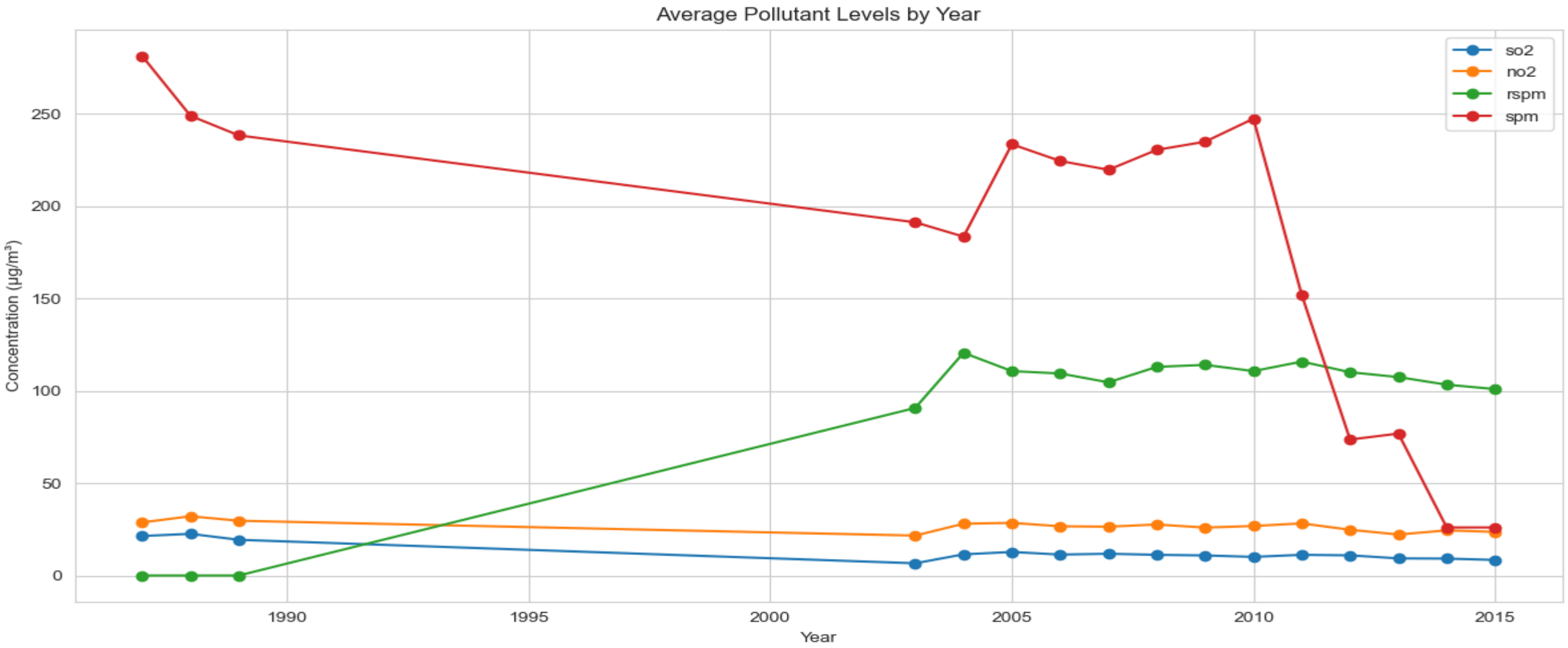
2. Temporal Analysis

- Goal: Discover trends over time and detect seasonal effects
- Techniques Used:
    - Time feature extraction (year, month, week)
    - Grouped statistics and line plots (notebook shows monthly/weekly patterns)

Findings:
- Visible long-term trends in pollutants
- Potential seasonality (e.g., spikes in winter for PM2.5)

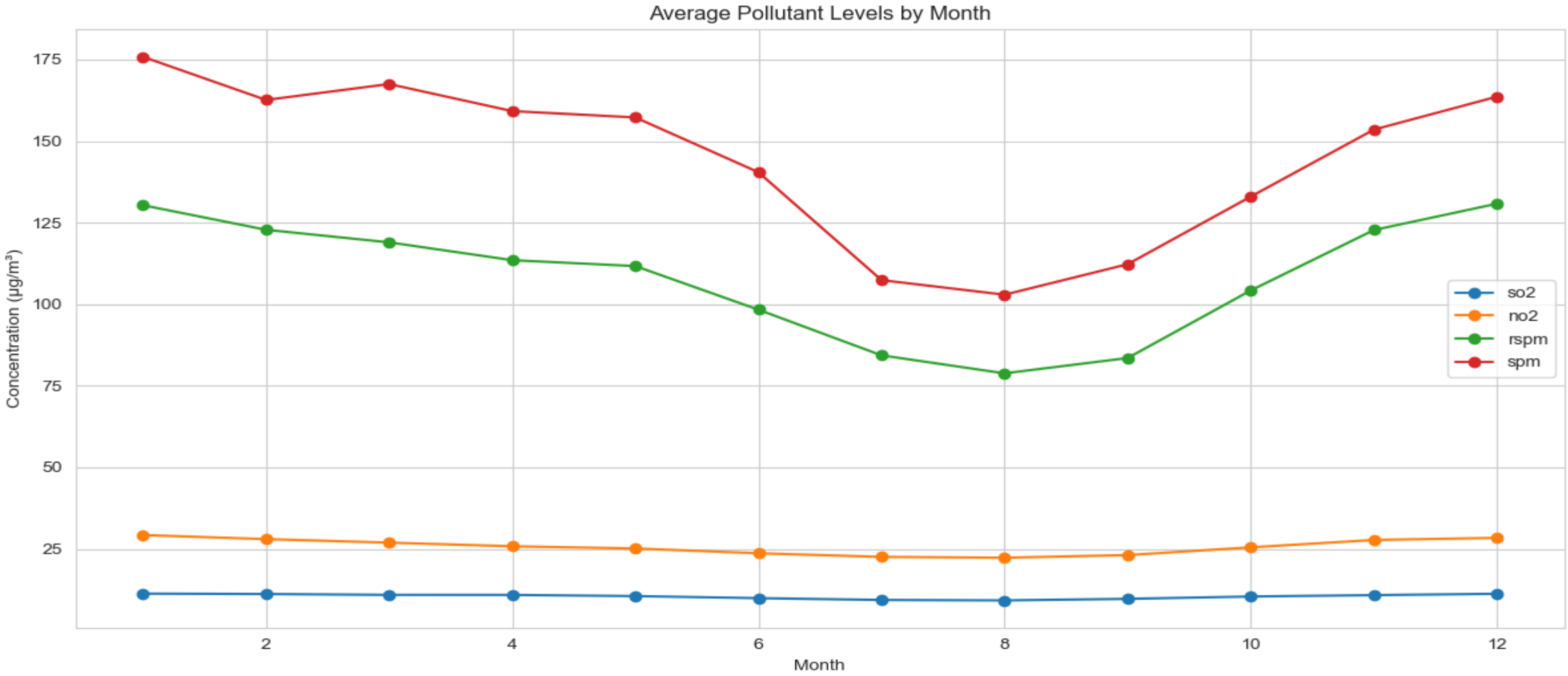
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

2. Temporal Analysis



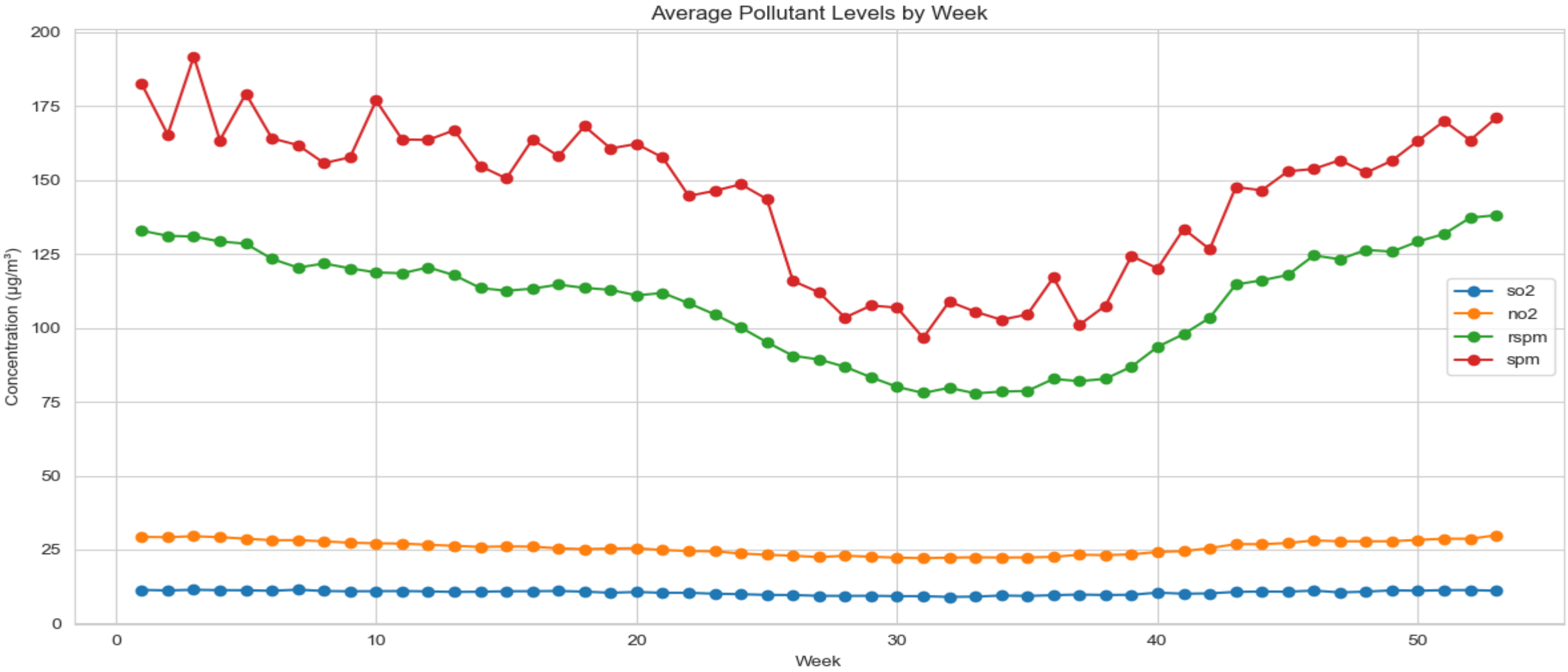
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

2. Temporal Analysis



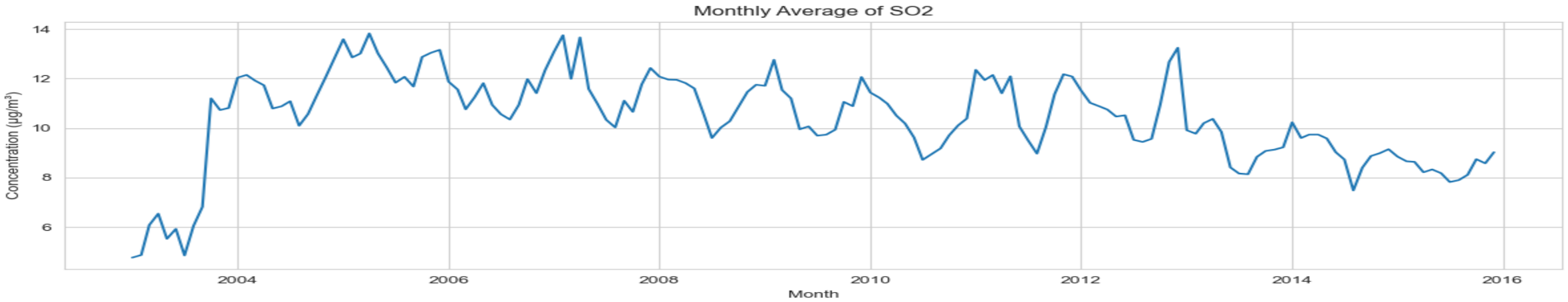
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

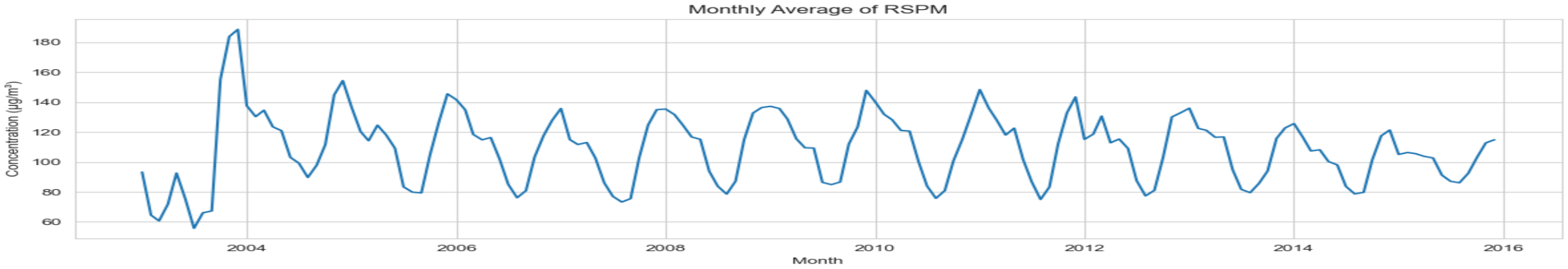
2. Temporal Analysis



Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

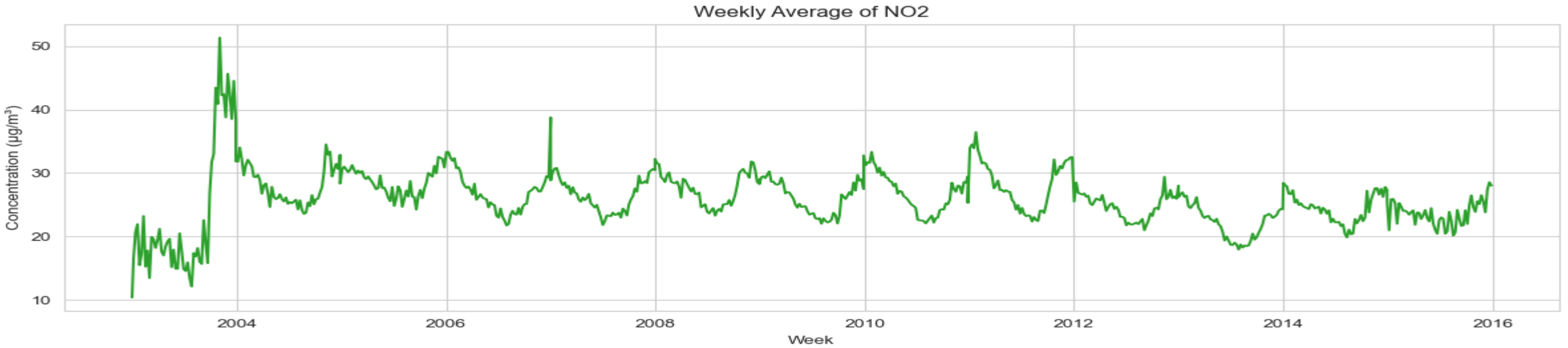
2. Temporal Analysis





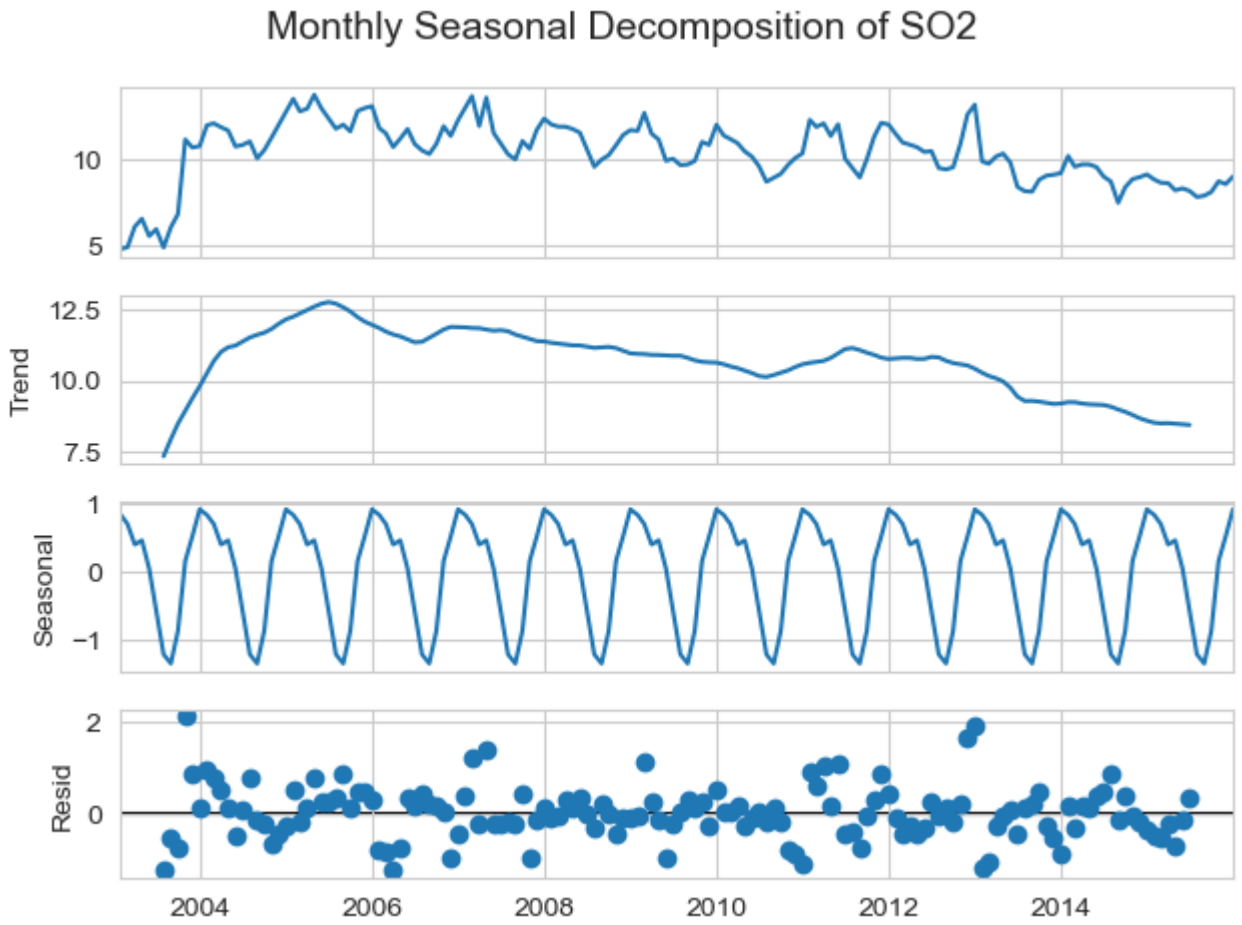
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

2. Temporal Analysis



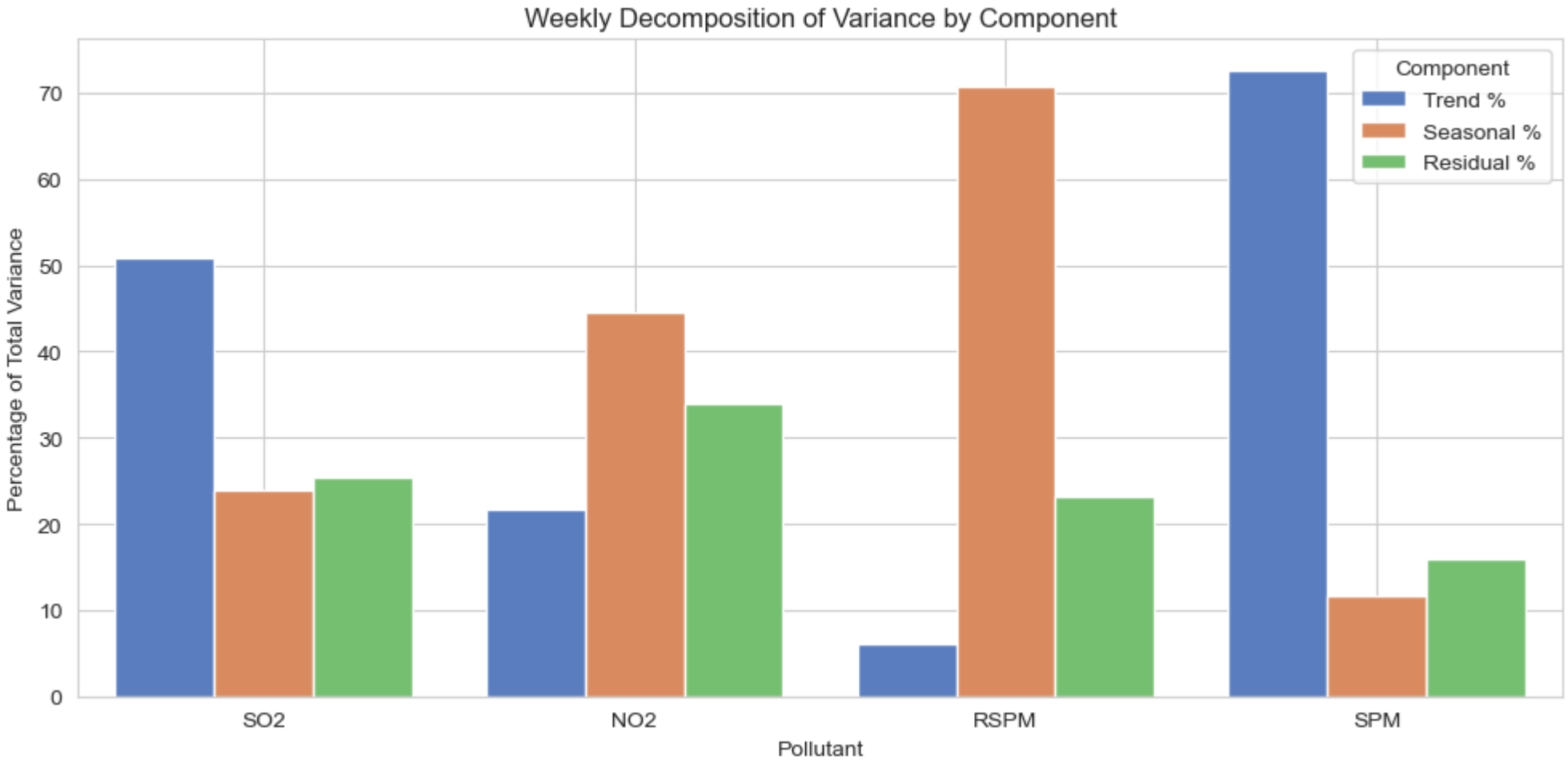
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

2. Temporal Analysis: component analysis



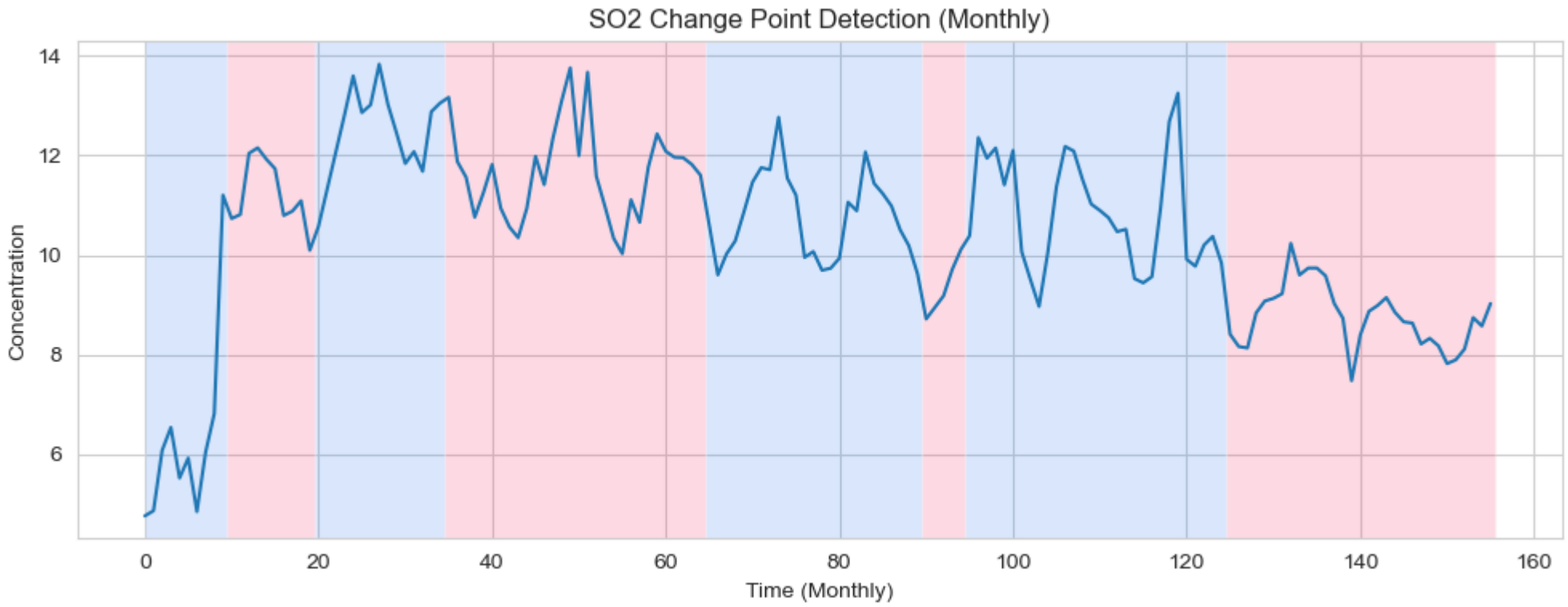
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

2. Temporal Analysis: component analysis



Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

2. Temporal Analysis: regime change analysis



Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

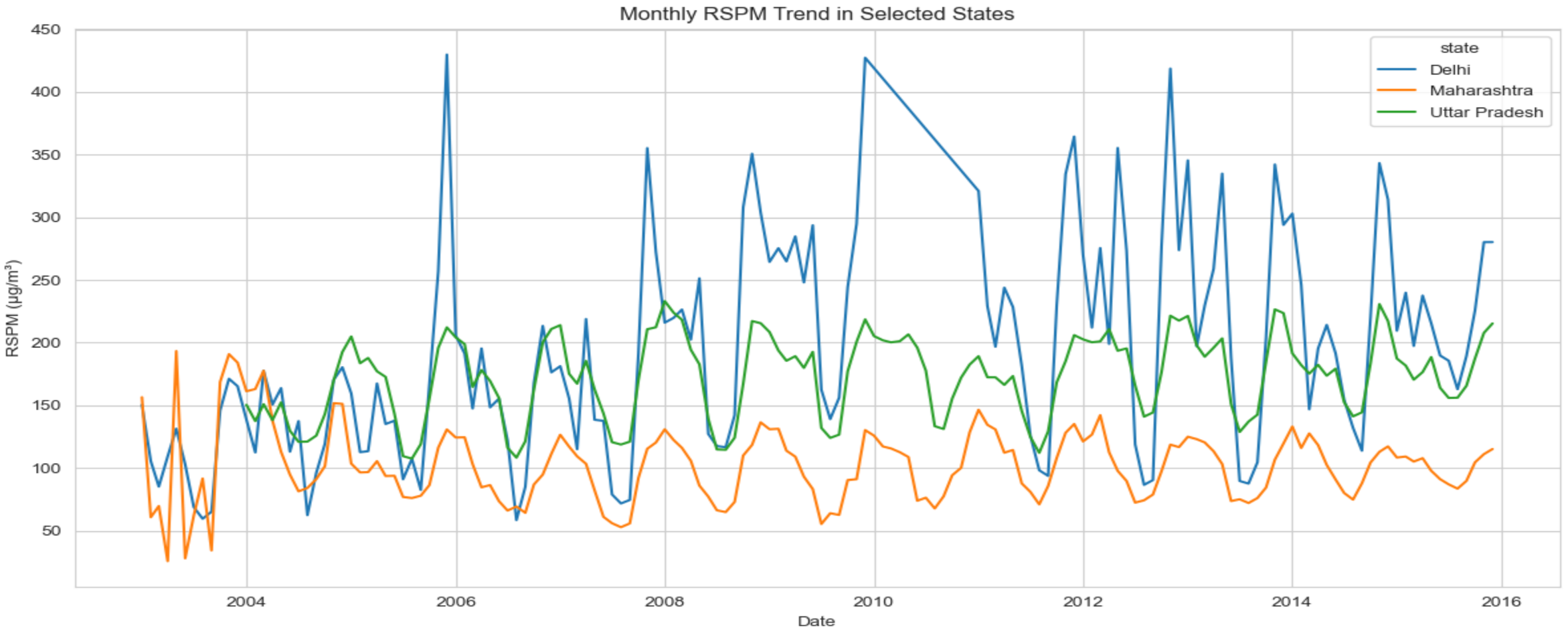
3. Regional Analysis

- Objective: Understand spatial variation in air pollution
- Methods:
  - Grouped data by state and city
  - Used summary statistics and visualizations (e.g., box plots, bar charts)

- Findings:
  - Certain cities/states consistently show higher pollution
  - Highlights of regional disparities and potential policy impact areas

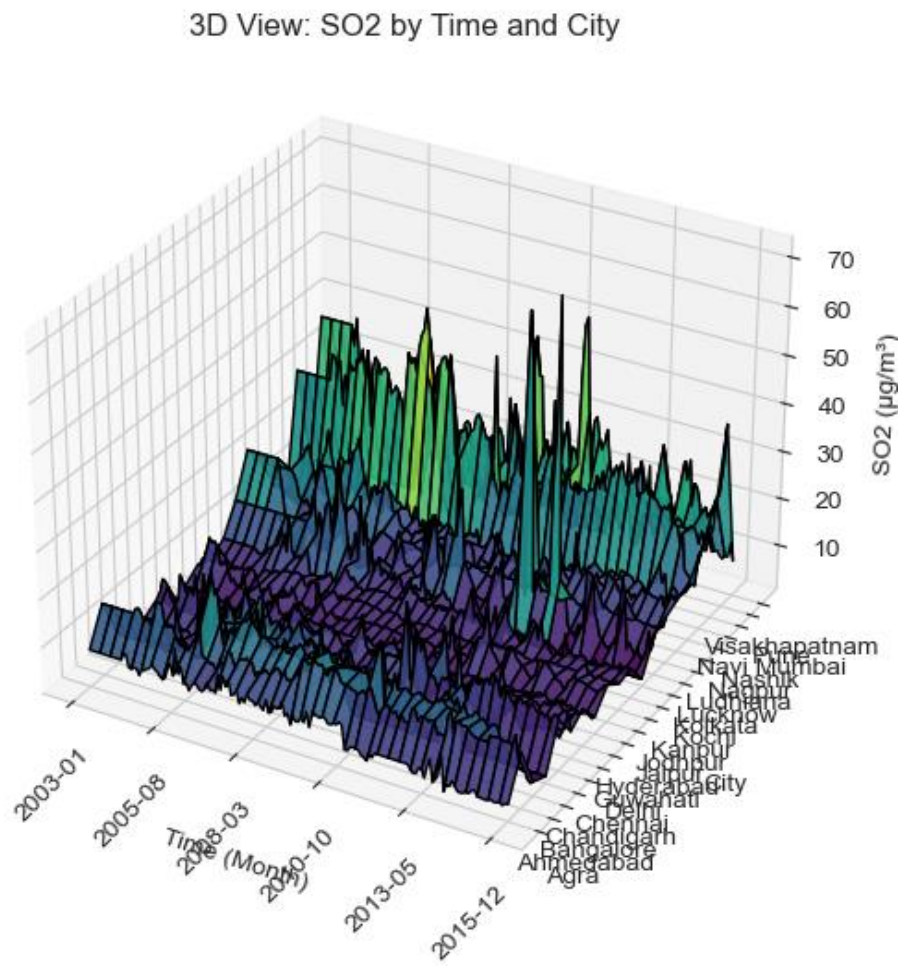
Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

3. Regional Analysis



Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

3. Temporal + Regional mixed effects analysis



Anova test for Mixed Effects (temporal + regional)

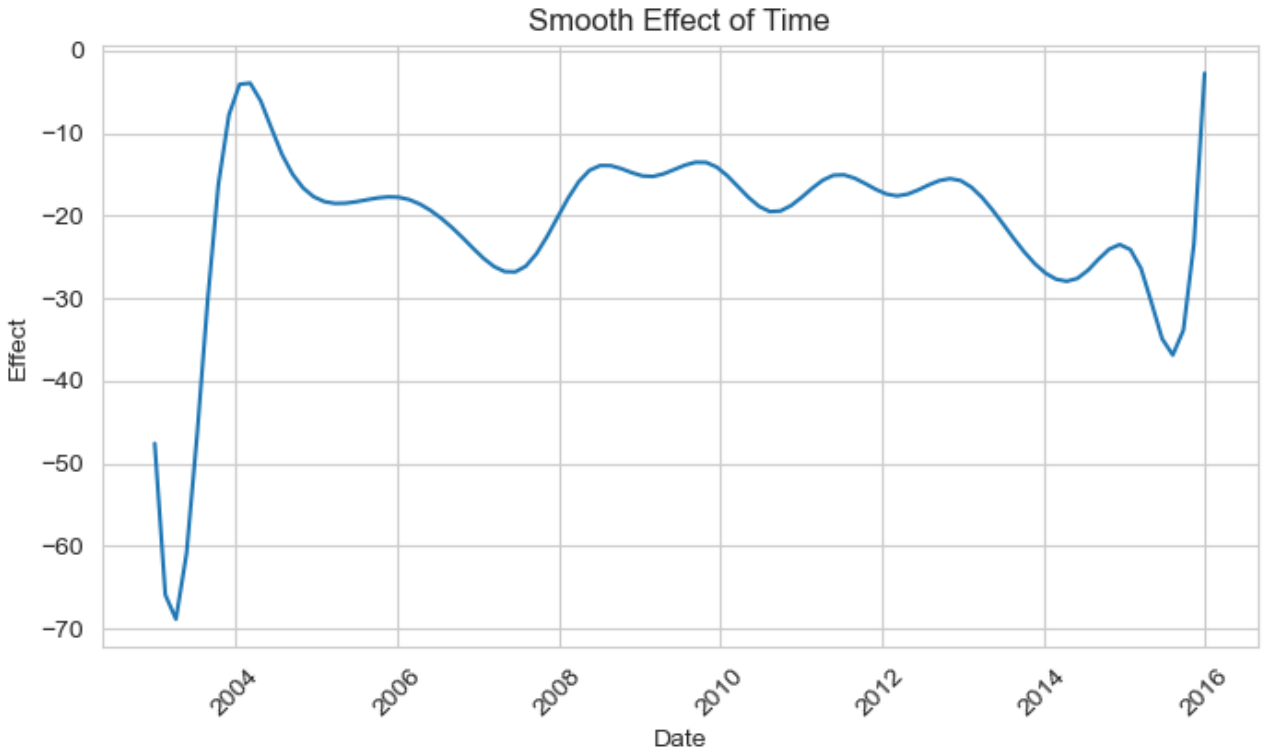
Mixed Linear Model Regression Results

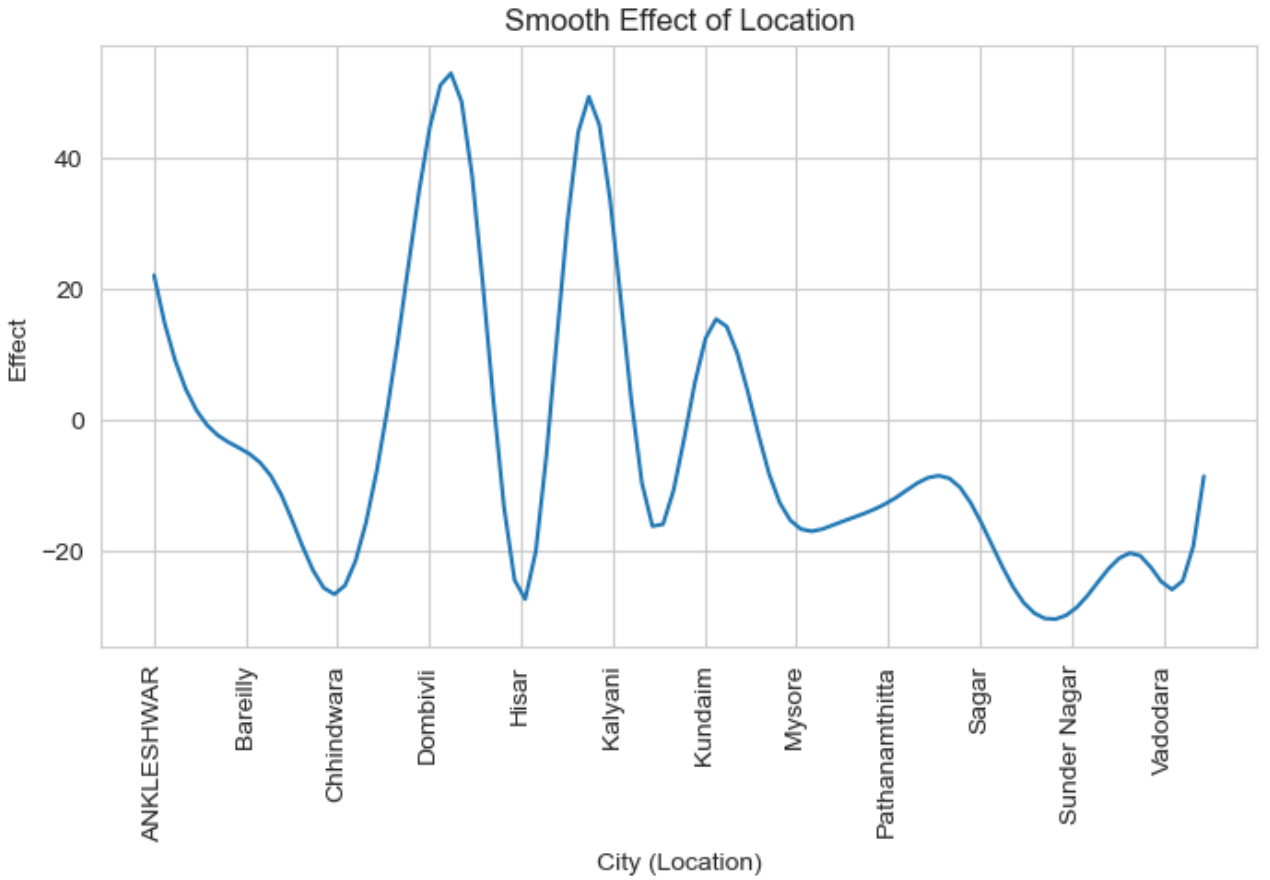
Model:	MixedLM	Dependent Variable:	rspm
No. Observations:	410803	Method:	REML
No. Groups:	287	Scale:	3623.2091
Min. group size:	1	Log-Likelihood:	-2267032.3882
Max. group size:	9739	Converged:	Yes
Mean group size:	1431.4		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	104.214	2.546	40.935	0.000	99.224	109.204
week	-14.335	0.326	-44.028	0.000	-14.973	-13.697
Group Var	1834.963	2.554				

Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

3. Temporal - Regional Analysis





Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

4. Predictive Modeling:

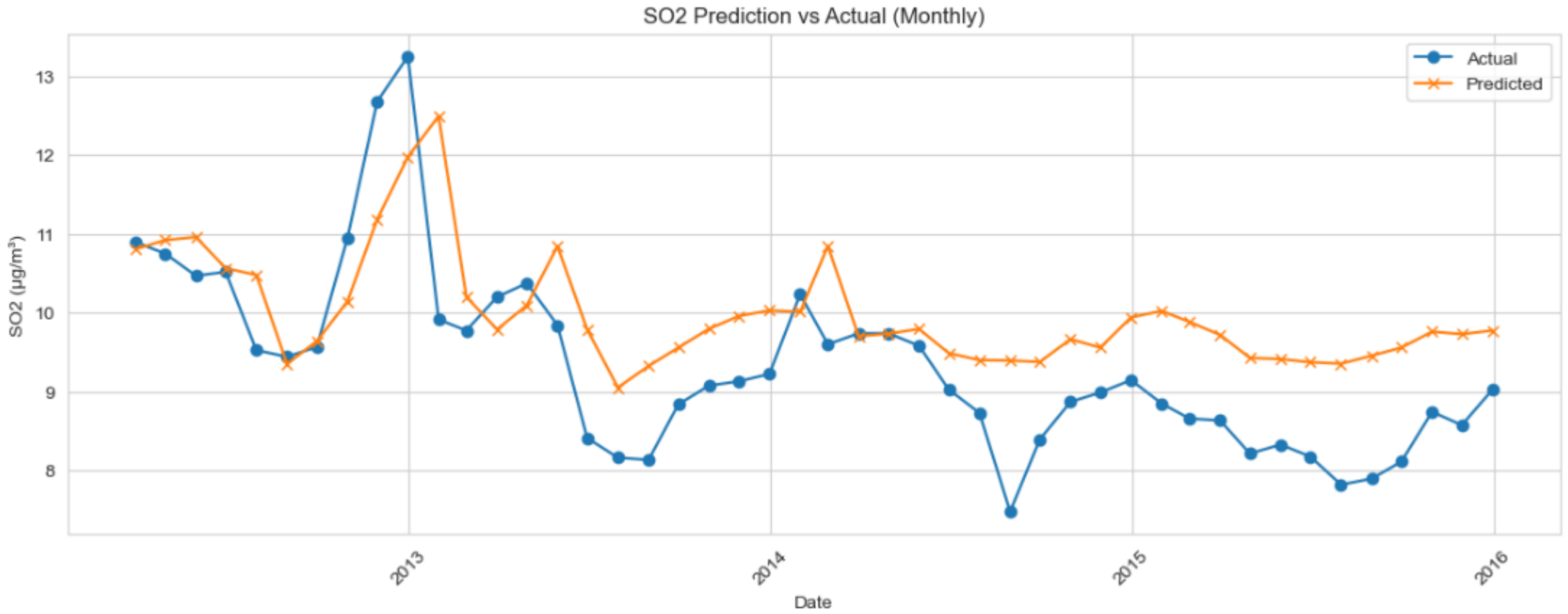
STEP-BY-STEP PLAN:
1. Prepare time-based features (month, year, week, lagged values)
2. Train/Test Split (time-aware)
3. Fit models (e.g., Random Forest, Linear Regression, XGBoost)
4. Evaluate performance (MAE, RMSE, R^2)
5. Visualize predictions

For this testing, we used:
```
model = RandomForestRegressor(n_estimators=100, random_state=42)
```

Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

4. Predictive Modeling:

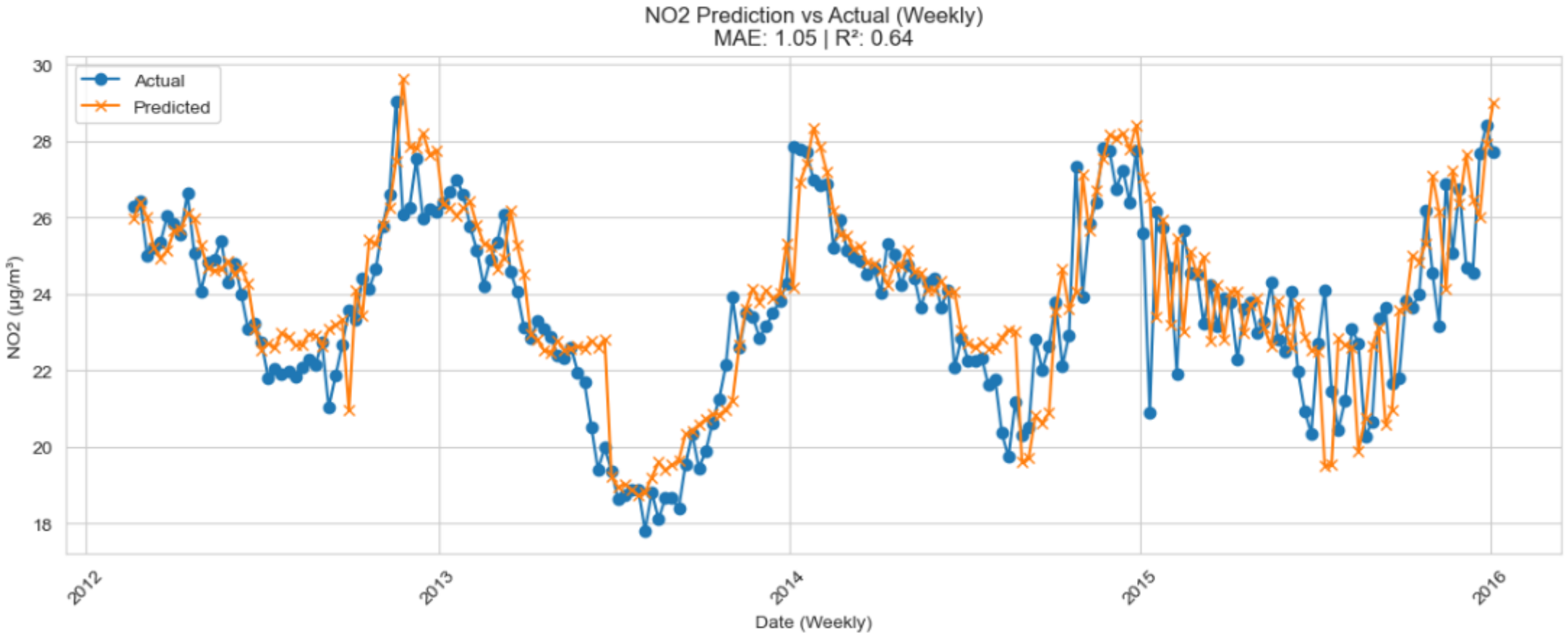
📈 Predicting SO2 (Monthly Model)
MAE: 0.85 | R^2 Score: 0.24
<Figure size 1000x400 with 0 Axes>



Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

4. Predictive Modeling:

Predicting NO2 (Weekly Model)
MAE: 1.05 | R^2 Score: 0.64



Title: "Data Visualization with Air Quality Data"
Dataset: Air Quality in India
Team: “Data Scientist”

Extra: maybe web service for data science?

Data science + software engineering