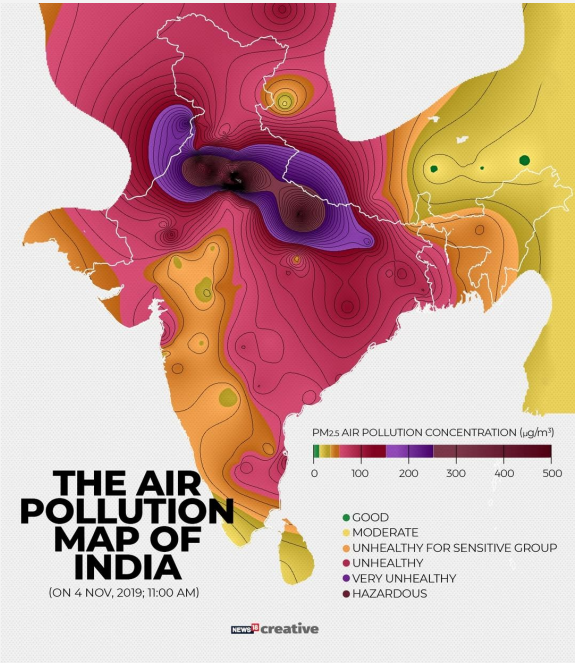
Data Visualization with Air Quality Data

Exploring India's Air Pollution Levels Over the Years



Aeroviz Group

Presented By: Reem Aboutaleb, Garam Lee, Jade Kas , Bobby Stitt , Mariam Zoair

NYU

Why study India's air quality?

World Health Organization

- Air pollution causes over **7M Deaths / year** (WHO).
- India is one of the most affected countries globally.

Project Goal

Analyze India's air pollution data to identify **local trends**,

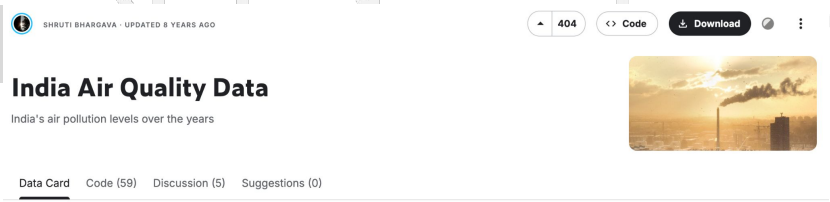
examine the correlation between **air quality and environmental policies**,

and explore **factors influencing pollution levels**.

(Data availability also varies by city and year, which may affect long-term trend accuracy.)

NYU

Data Set Overview



- **Source:** Kaggle – *India Air Quality Data (1987.01.01–2015.12.31)*
- **Data Size:** 435,742 rows × 13 columns
- **Collected By:** Central Pollution Control Board (CPCB), India
- **Variables:**
  - Includes pollutant concentrations (SO_2, NO_2, RSPM, SPM, $PM_{2.5}$), state, city, location type, and date
- **Coverage:** 26 states, 300+ cities

❖ **Dataset cleaned using Python (Pandas, NumPy, Seaborn, Matplotlib) –** missing values dropped or handled before visualization and statistical analysis.

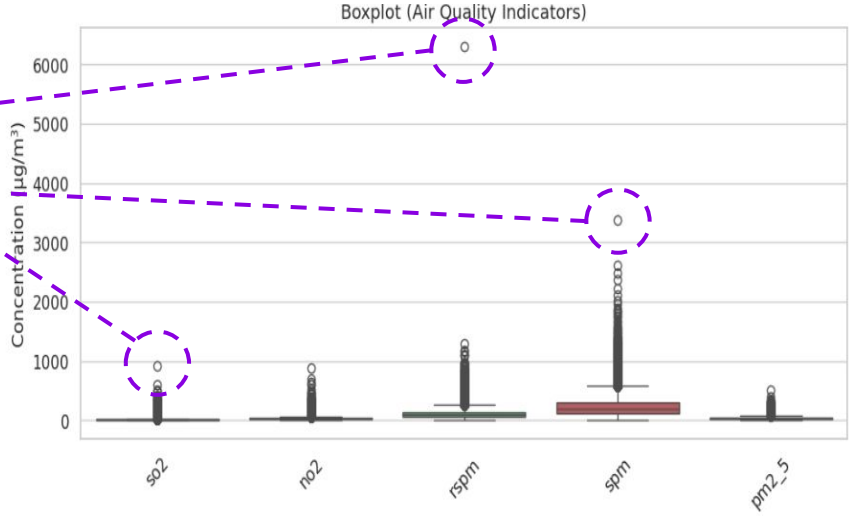
https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data/data

Descriptive Statistics: Pollutant Variables

	so2	no2	rspm	spm	pm2_5
Mean	10.83	25.81	108.83	220.78	40.79
Std Dev	11.18	18.50	74.87	151.40	30.83
Max	909.0	876.0	6307.03	3380.0	504.0

Extreme Outliers Found

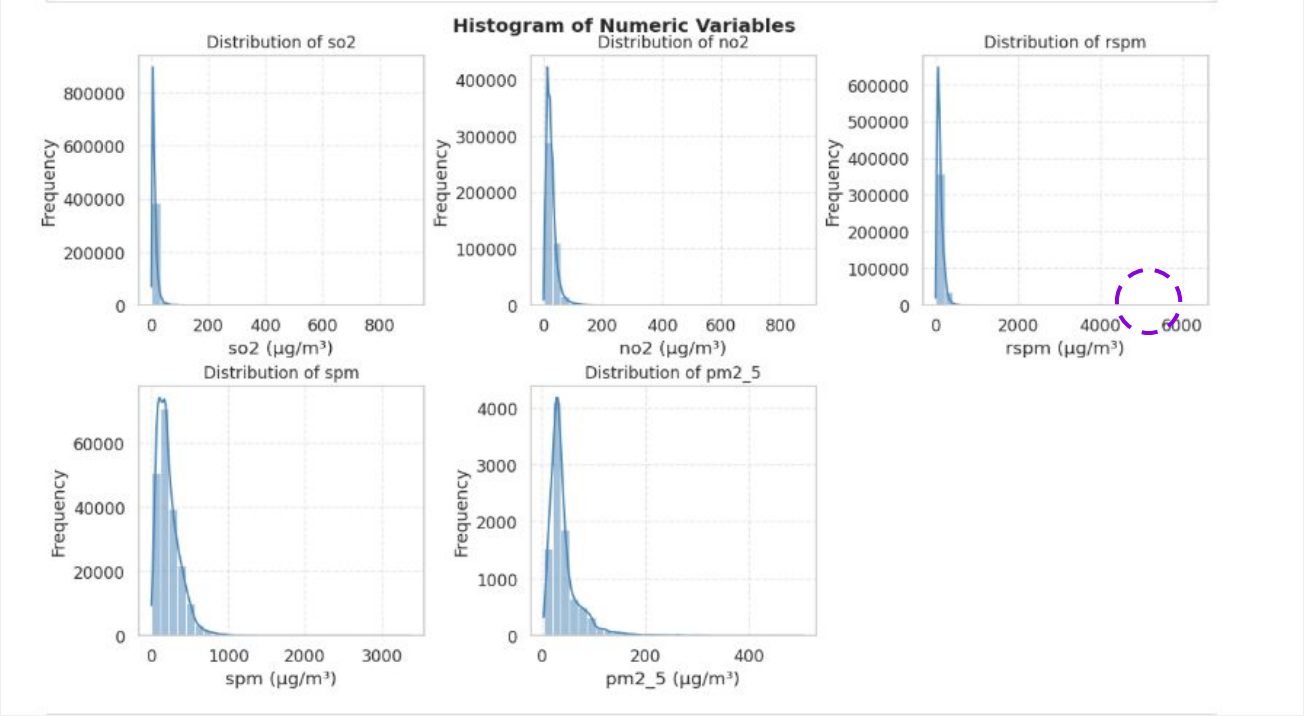
- SO2 Max: 909.0 (Tamil Nadu, 2011)
- RSPM Max: 6307.03 (Uttar Pradesh, 2010)
- SPM Max: 3380.0 (Rajasthan, 2001)



Caption : *RSPM and SPM show wider variability and multiple extreme outliers compared to SO_2 and NO_2.*

- We created boxplots for each pollutant (numerical variables) to detect variability and extreme values.
- RSPM and SPM show high variability and multiple outliers, while SO_2 and NO_2 are more consistent.
- This helps us identify pollutants with the **biggest fluctuations** — these might need closer monitoring or policy intervention.

Histogram: Distribution Visualization for Numerical Variables



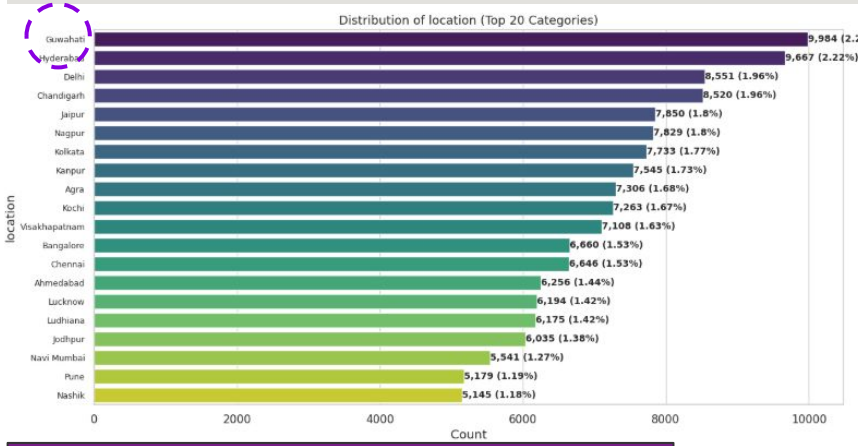
Why It Matters:

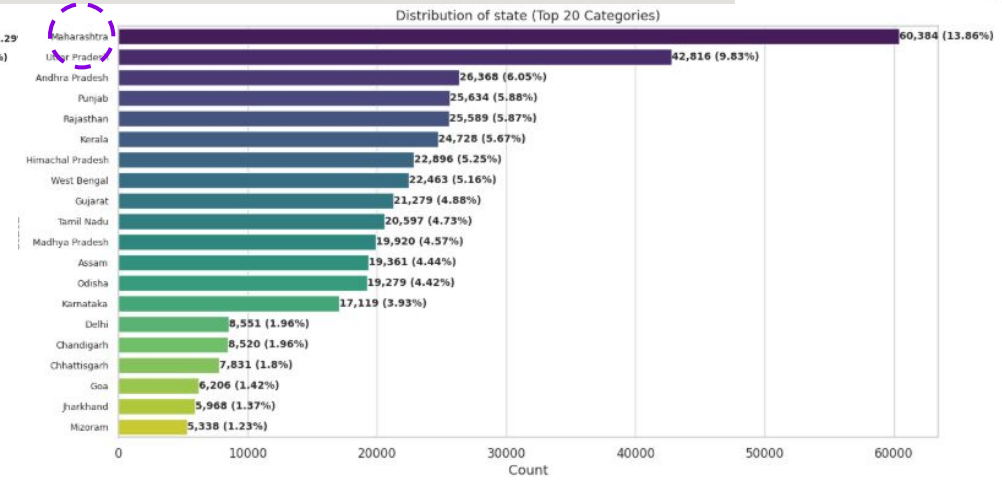
This **skewness** shows the presence of pollution spikes, which are critical events.
Indicating most days have low pollution, but occasional spikes drive poor air quality averages.
It also means we may **need to handle outliers carefully** before modeling or making policy-related conclusions.

- We plotted **histograms** using matplotlib for each pollutant to explore how their values are distributed and to detect skewness or extreme peaks.
- All pollutants are strongly **right-skewed,** meaning most measurements are low, but there are some very high spikes, especially for PM2.5 and RSPM.

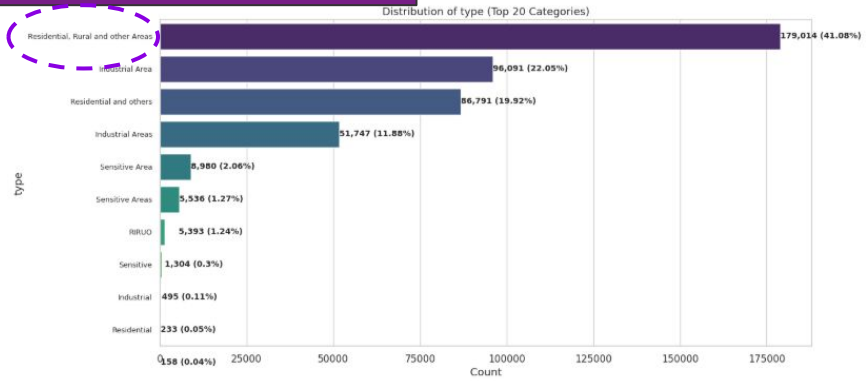
NYU

Bar Plots for Categorical Variables(count, ratio)





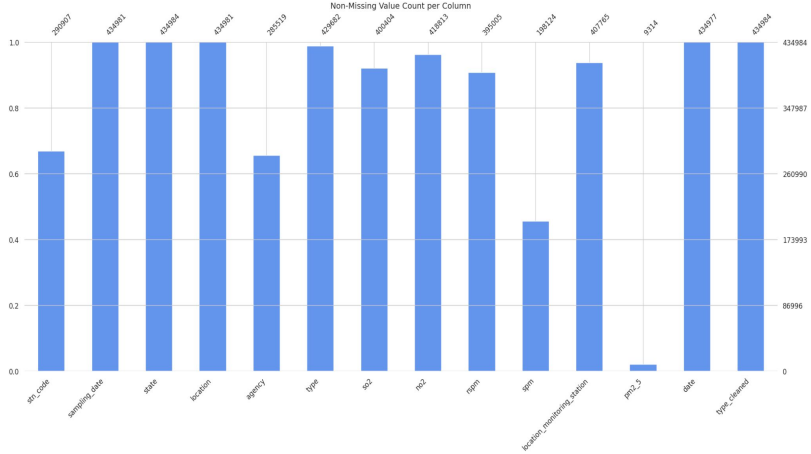
Uneven distribution can bias results and affect how representative the analysis is

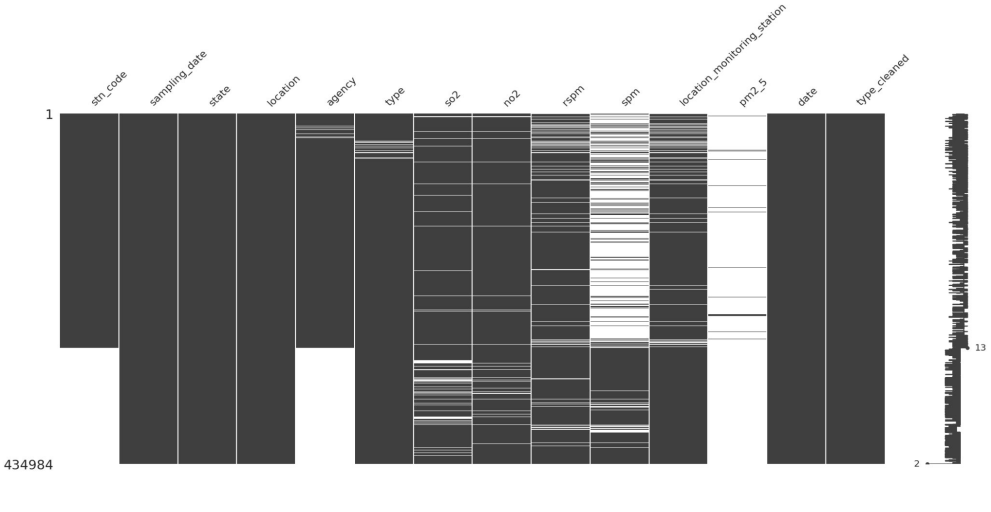


- We checked how data is **distributed by state, location, and site type.**
- Uttar Pradesh and Maharashtra have the most data.
- Residential areas dominate.
- Some states have very little data.

NYU

Missing Value Analysis





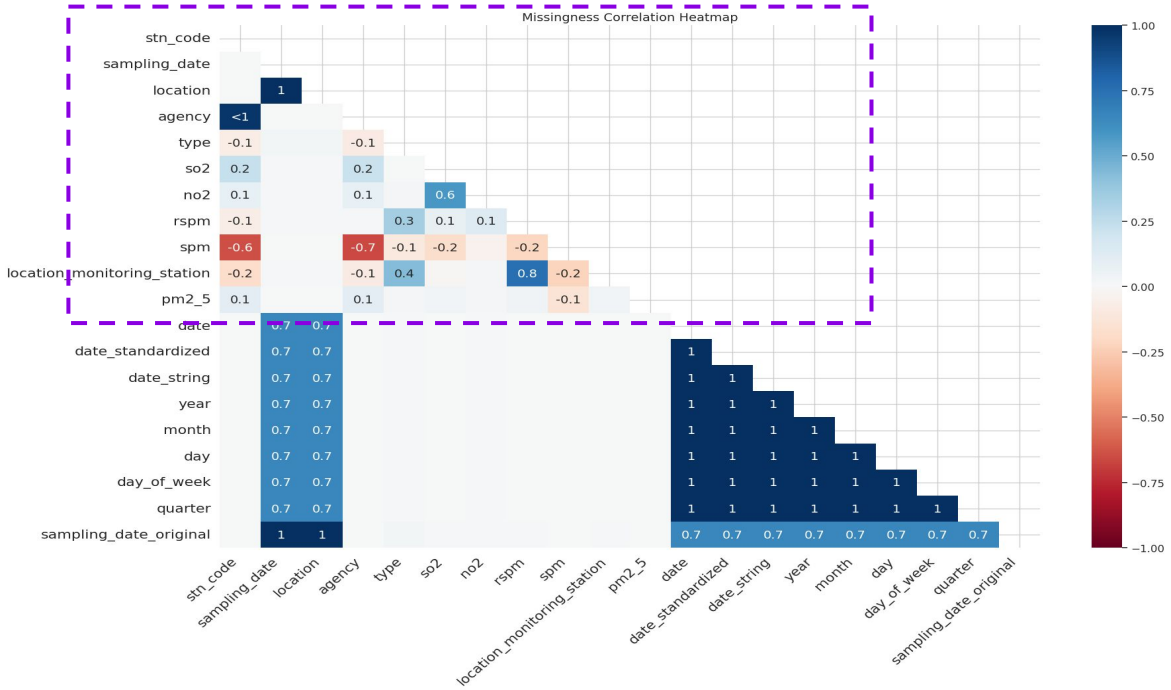
- We checked how much missing data exists in each column.
- Some pollutant columns have a lot of missing values, while other fields like location and date are mostly complete.

Why It Matters:

- This helps us identify which columns need **cleaning or imputation before analysis.**

NYU

Relations between Features

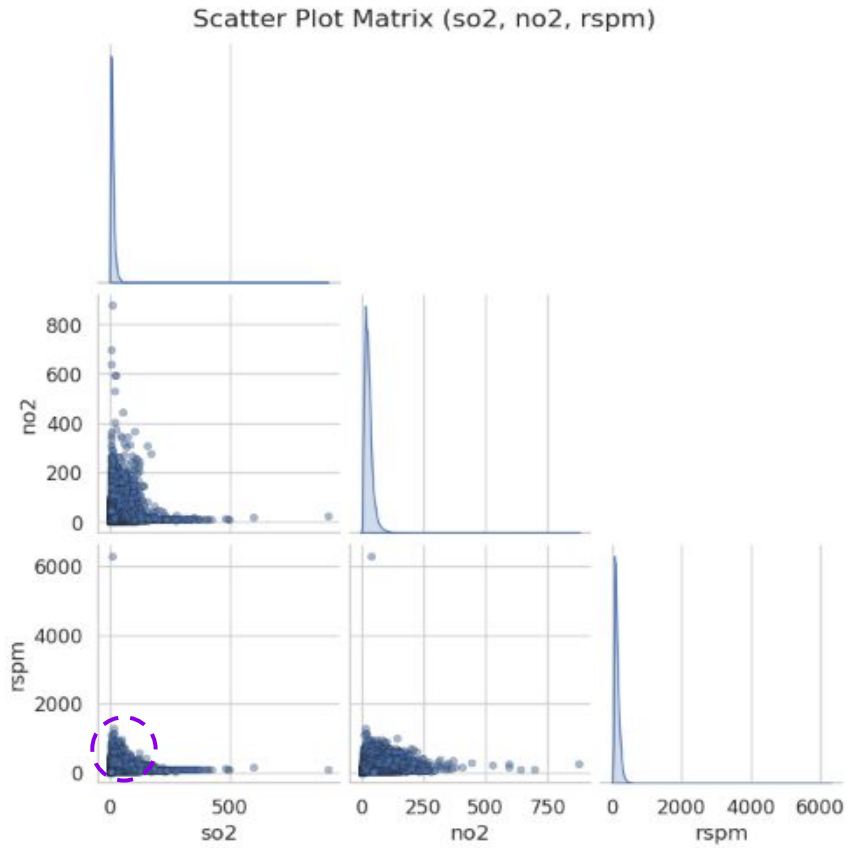


Why It Matters:

➔ Understanding where missing data occurs helps us decide how to clean and which columns to keep or drop.

➔ We created a heatmap to check how missing values are related across variables.
➔ Some pollutant columns have missing values that are correlated.
➔ Date columns have high correlations with each other, which is expected.

NYU



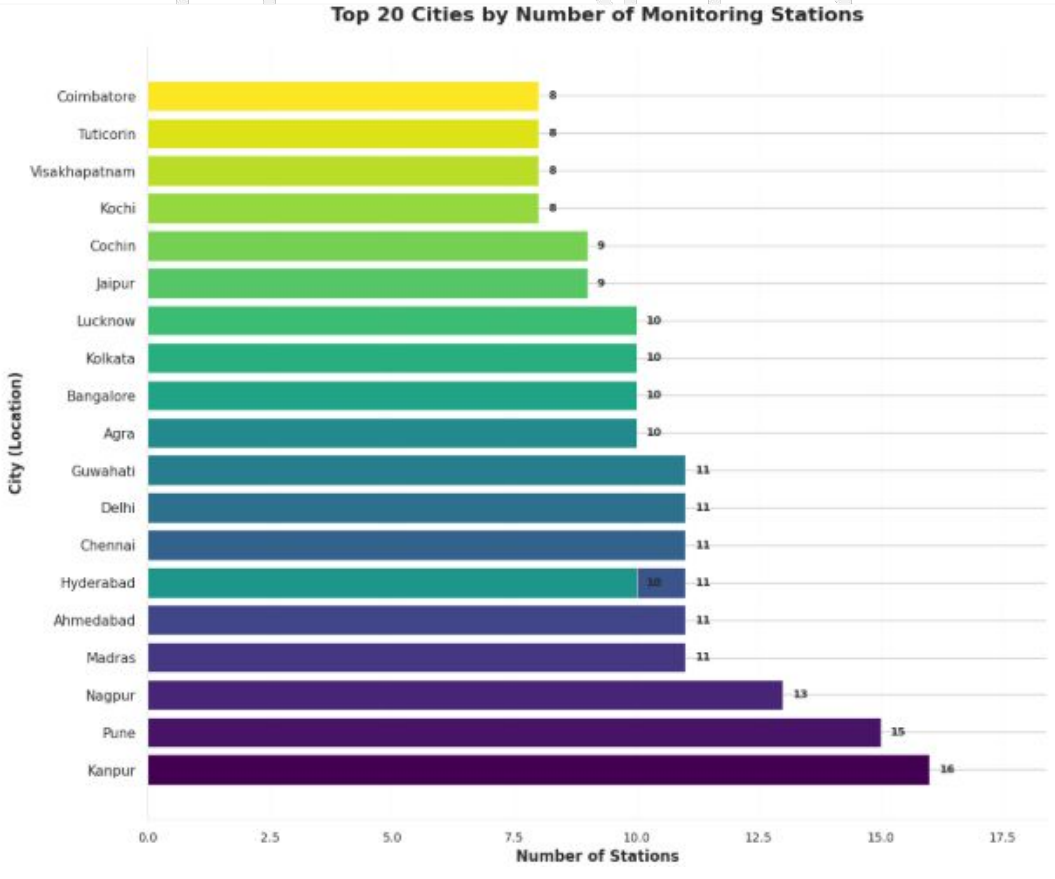
We created a scatter plot matrix to check relationships between pollutant variables.

- Most data points cluster at low values. There are some high-value spikes, especially for RSPM.
- The variables don't show a strong linear relationship.

Why It Matters:

➔ This helps us understand how pollutants behave together and shows that extreme values may need special handling before further analysis.

NYU



Number of monitoring stations per city:

state	location	num_stations
Uttar Pradesh	Kanpur	16
Maharashtra	Pune	15
Maharashtra	Nagpur	13
Tamil Nadu	Madras	11
Gujarat	Ahmedabad	11
Andhra Pradesh	Hyderabad	11
Tamil Nadu	Chennai	11
Delhi	Delhi	11
Assam	Guwahati	11
Uttar Pradesh	Agra	10
Telangana	Hyderabad	10
Karnataka	Bangalore	10
West Bengal	Kolkata	10
Uttar Pradesh	Lucknow	10
Rajasthan	Jaipur	9
Kerala	Cochin	9
Kerala	Kochi	8
Andhra Pradesh	Visakhapatnam	8
Tamil Nadu	Tuticorin	8
Tamil Nadu	Coimbatore	8

- We grouped the dataset by city and counted the number of monitoring stations per location.
- Cities like Coimbatore and Tuticorin have the fewest stations.

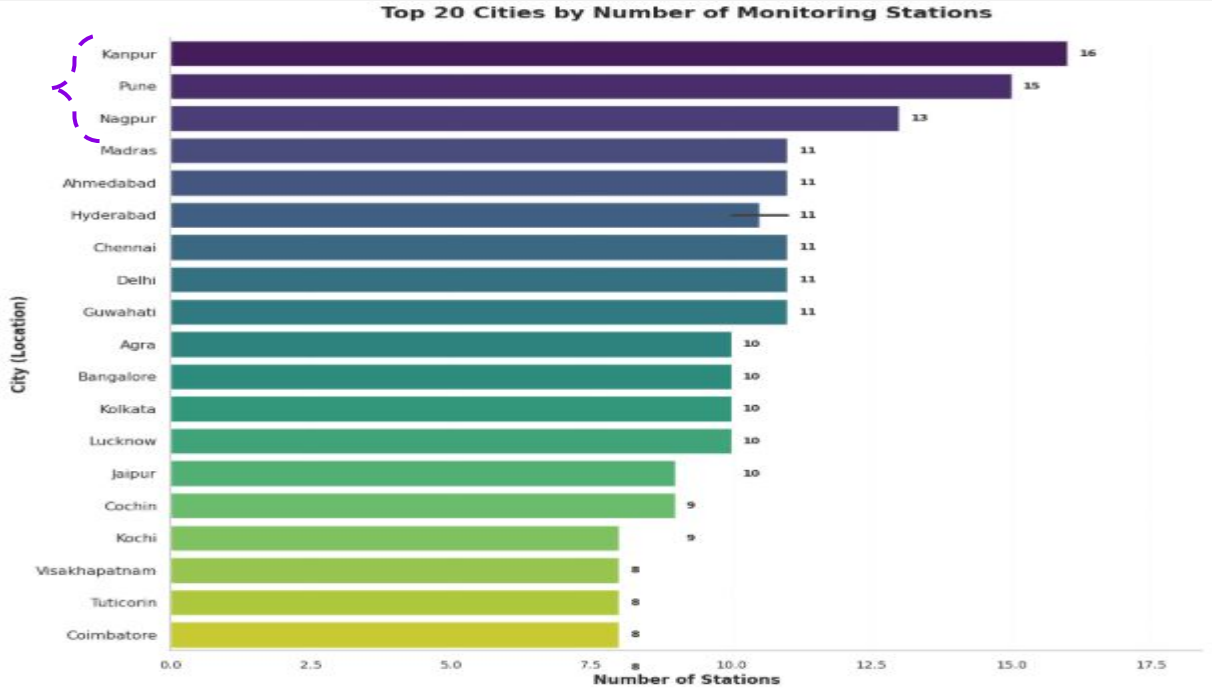
Why it matters:

➔ This shows that monitoring is unevenly distributed, which can affect how representative the dataset is.

NYU

Kanpur, Pune, and Nagpur have the highest number of stations.

Other cities have fewer monitoring sites.



➔ **We counted the number of monitoring stations in each city.**

Why It Matters:

- Data is concentrated in a few cities.
- This can affect how well our analysis reflects the entire country.

NYU

Conclusion

- India's air quality monitoring shows valuable coverage but uneven representation.
- Data analysis highlights where policies like NCAP should focus: North India, high-PM regions.

Future Analysis

Yearly trend visualization

Correlation analysis between pollutants

Predictive modeling (Linear Regression or Random Forest / Time-series based model)

NYU

Appendix

You can access the Notebook here

Notebook: https://colab.research.google.com/drive/13YxSaJ7e9HMLawu7NBQZYeTOoBRuhJZi?usp=sharing

You can access the Data here

DataSet: https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data/data

NYU