📊 OVERALL DATASET DESIGN (City-Specific, Daily Time-Series)

You will mainly maintain **4 raw datasets** and **1 engineered ML dataset**.

🗂 DATASET 1 — Air & Water Sensor Data (`sensors_daily.csv`)

Column Name	Description
city_id	City name/code
zone_id	Area/ward of city
date	YYYY-MM-DD
pm25	PM2.5 value
pm10	PM10 value
no2	Nitrogen dioxide
water_quality_index	Overall water quality
reservoir_level	% storage capacity

🗂 DATASET 2 — Weather Data (`weather_daily.csv`)

Column Name	Description
city_id	
zone_id	
date	
humidity	
wind_speed	
rainfall	
average_temperature	*(optional)*

🗂 DATASET 3 — Violation Records (`violations_log.csv`)

Column Name	Description
city_id	
zone_id	
date	
violation_type	air/water/waste/noise
severity	1–5 scale
offender_id	

🗀 DATASET 4 — Urban & Social Indicators (`urban_profile.csv`)

(Usually static or monthly updated)

Column Name	Description
city_id	
zone_id	
population_density	
industrial_density	
green_cover_percentage	
drainage_quality_index	
social_vulnerability_index	

🧠 ENGINEERED ML DATASET (`ml_features_daily.csv`)

This is what goes into the models.

Each row = one zone for one day.

☑ FINAL FEATURE COLUMNS FOR ML MODEL

🖂 Air Quality

```
pm25
pm10
no2
pollution_trend_3days
```

🌩️ Weather

```
humidity
wind_speed
rainfall_last_3_days
```

💧 Water

```
water_quality_index
reservoir_level
```

⚠️ Violations

```
violations_last_7_days
avg_violation_severity
repeat_offender_rate
```

🏙️ Urban Structure

```
population_density
industrial_density
green_cover_percentage
drainage_quality_index
```

🧍‍♂️ Social Risk

```
social_vulnerability_index
```

🎯 Target Column (for training)

```
risk_score
```

(or `risk_level` if classification)

📌 FINAL ML DATASET COLUMN ORDER (READY TO USE)

```
city_id
zone_id
date

pm25
```

```
pm10
no2
pollution_trend_3days

humidity
wind_speed
rainfall_last_3_days

water_quality_index
reservoir_level

violations_last_7_days
avg_violation_severity
repeat_offender_rate

population_density
industrial_density
green_cover_percentage
drainage_quality_index

social_vulnerability_index

risk_score
```

🎯 Which features are actually used in models?

🔘 Pollution Forecast Model uses:

```
pm25, pm10, no2
humidity, wind_speed, rainfall_last_3_days
violations_last_7_days
industrial_density
```

🔘 Risk Score Model uses:

👉 ALL engineered feature columns (except date, city_id, zone_id)

🧠 OVERALL SYSTEM = 3 SMART MODELS

Your backend will run **three AI brains**:

1️⃣ Pollution Forecast Model
2️⃣ Environmental Risk Score Model
3️⃣ Hotspot Detection Model

Each gives different actionable feedback.

🟢 MODEL A — POLLUTION FORECAST MODEL

🎯 Purpose

Predict how bad air pollution will be in coming days.

📥 INPUT FEATURES

Example input for Zone A (today):

```
pm25 = 130
pm10 = 210
no2 = 45
humidity = 70%
wind_speed = 3 km/h
rainfall_last_3_days = 5 mm
violations_last_7_days = 6
industrial_density = 0.8
```

📤 OUTPUT (Feedback)

```
Predicted PM2.5 tomorrow = 165
Predicted PM2.5 after 3 days = 190
```

🧠 Interpretation:

👉 “If current trend continues, air will become very dangerous in next 3 days.”

💡 How it helps city:

• Issue health alerts
• Restrict vehicles
• Inspect factories

🟢 MODEL B — ENVIRONMENTAL RISK SCORE MODEL (MAIN)

🎯 Purpose

Combine everything to predict overall environmental danger.

📥 INPUT FEATURES (full engineered row)

Example:

```
pm25 = 150
pm10 = 220
no2 = 50
pollution_trend_3days = rising

humidity = 72
wind_speed = 2
rainfall_last_3_days = 3

water_quality_index = 55
reservoir_level = 40%

violations_last_7_days = 7
avg_violation_severity = 4
repeat_offender_rate = 0.6

population_density = high
industrial_density = high
green_cover_percentage = low
drainage_quality_index = poor

social_vulnerability_index = high
```

📤 OUTPUT (Feedback)

Option 1 (Score):

```
risk_score = 84
```

Option 2 (Label):

```
risk_level = HIGH
```

□ Interpretation:

👉 “This zone is likely to become environmentally unsafe very soon.”

💡 How it helps city:

• Focus inspections here
• Improve drainage
• Reduce industrial emissions
• Target vulnerable communities

□ MODEL C — HOTSPOT DETECTION MODEL

🎯 Purpose

Find zones that are always risky over time.

(Not day prediction, but pattern detection)

📥 INPUT (historical feature patterns)

Example:

```
Zone A → high pollution + many violations + poor drainage
Zone B → moderate
Zone C → low
```

📤 OUTPUT (Feedback)

```
Cluster 1 → High-risk zones: Zone A, Zone D, Zone F
Cluster 2 → Medium-risk: Zone B, Zone E
Cluster 3 → Low-risk: Zone C
```

□ Interpretation:

👉 “These areas are structurally dangerous and need permanent solutions.”

💡 How it helps city:

• Long-term planning
• Infrastructure improvement
• Law enforcement focus

📊 SUMMARY TABLE (FOR YOUR TEAM)

Model	Input	Output	Meaning
Pollution Forecast	Air + weather + violations	Future PM2.5/PM10	How pollution will change
Risk Score Model	All features	Risk score (0–100)	Overall danger level
Hotspot Model	Historical patterns	Risk clusters	Long-term risky zones

---

🎯 REALISTIC DAY IN SYSTEM

City uploads today’s data →

🧠 Model A:
“Pollution rising sharply in 3 days”

🧠 Model B:
“Zone A risk = 84 (HIGH)”

🧠 Model C:
“Zone A belongs to high-risk hotspot cluster”

👉 Final feedback to authorities:

⚠ Zone A will become highly dangerous soon
⚠ Take preventive action now