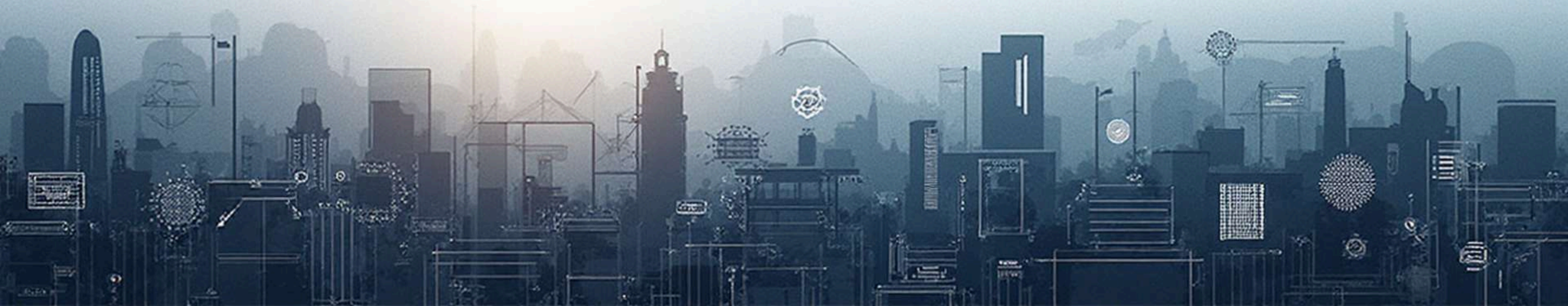
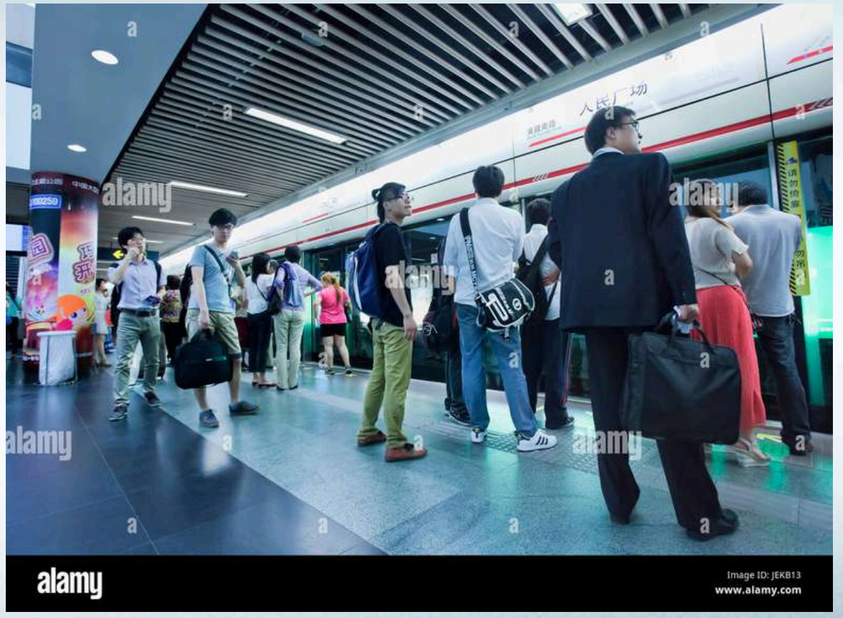
MEDDOUR Bisssem
MEGNOUX Julien
LAMBERT Vivien
RUBAN Roshanth
Loutfi Adam

Projet Scala & Apache Spark – Analyse de la Pollution à Pékin



Objectif du projet : Analyser, modéliser et prédire la pollution atmosphérique à Pékin avec Scala & Spark.

- 12 stations de mesure
- 5 ans de données (2013–2017)
- PM2.5, PM10, O_3, CO, NO_2,
- Différent traitement utilisés & graphe



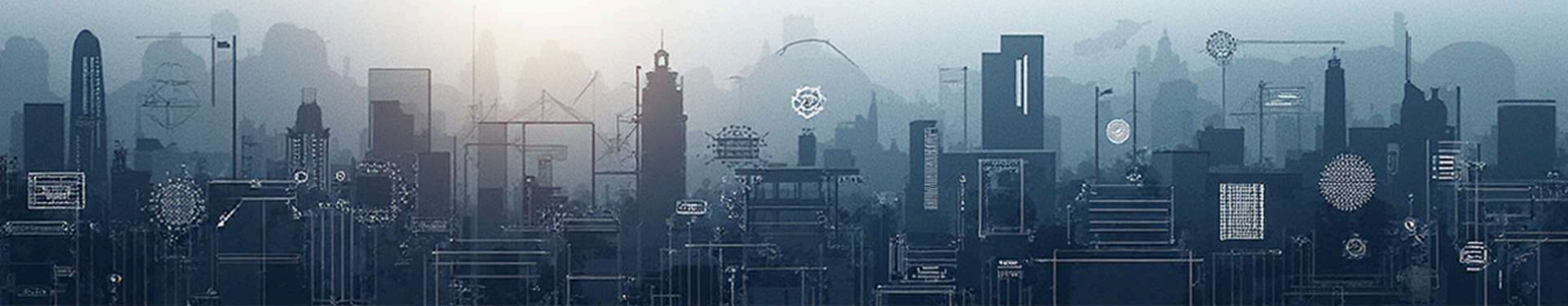
Dataset utilisé

Source : Beijing PM2.5 Data (UCI)

PRSA_Data_Aotizhongxin_20130301-201...	14/11/2025 10:20	Fichier CSV Micros...	2 736 Ko
PRSA_Data_Changping_20130301-20170...	14/11/2025 10:20	Fichier CSV Micros...	2 625 Ko
PRSA_Data_Dingling_20130301-2017022...	14/11/2025 10:20	Fichier CSV Micros...	2 579 Ko
PRSA_Data_Dongsi_20130301-20170228....	14/11/2025 10:20	Fichier CSV Micros...	2 541 Ko
PRSA_Data_Guanyuan_20130301-201702...	14/11/2025 10:20	Fichier CSV Micros...	2 599 Ko
PRSA_Data_Gucheng_20130301-2017022...	14/11/2025 10:20	Fichier CSV Micros...	2 559 Ko
PRSA_Data_Huairou_20130301-20170228...	14/11/2025 10:20	Fichier CSV Micros...	2 545 Ko
PRSA_Data_Nongzhanguan_20130301-20...	14/11/2025 10:20	Fichier CSV Micros...	2 739 Ko
PRSA_Data_Shunyi_20130301-20170228.c...	14/11/2025 10:20	Fichier CSV Micros...	2 525 Ko
PRSA_Data_Tiantan_20130301-20170228....	14/11/2025 10:20	Fichier CSV Micros...	2 559 Ko
PRSA_Data_Wanliu_20130301-20170228....	14/11/2025 10:20	Fichier CSV Micros...	2 563 Ko
PRSA_Data_Wanshouxigong_20130301-2...	14/11/2025 10:20	Fichier CSV Micros...	2 770 Ko

Contenu :

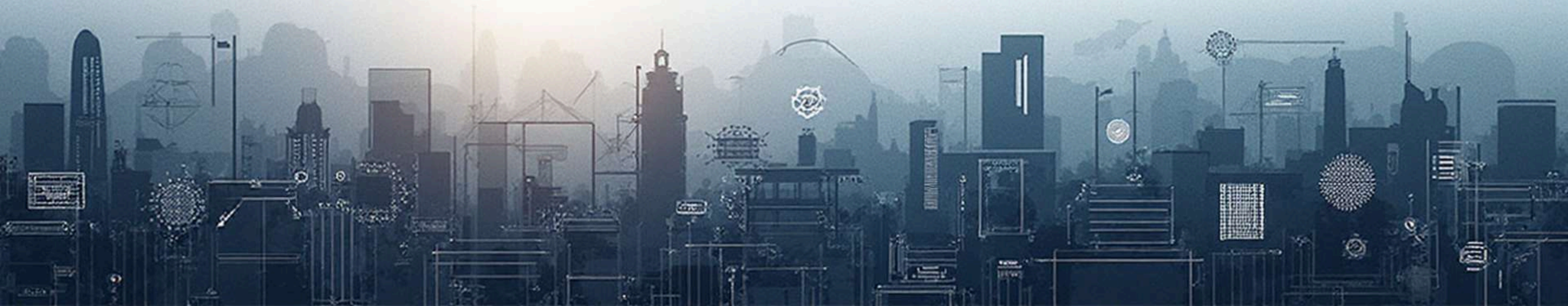
- **430 000 lignes**
- **Mesures horaires**
- **12 stations**
- **Variables : YEAR, MONTH, DAY, PM2.5, PM10, O3, etc.**



Pipeline global du projet

Étapes réalisées :

1. **Nettoyage des données**
2. **Feature engineering**
3. **Statistiques descriptives**
4. **GraphX : réseau des stations**
5. **MLlib : trois modèles prédictifs**
6. **Simulation temps réel & détection d’anomalies**



Partie 1 : Nettoyage

Opérations effectuées :

- **Suppression des doublons**
- **Suppression des lignes incomplètes**
- **Nettoyage de PM2.5 (“NA” → valeur numérique)**
- **Conversion en double**

```
=== Chargement des fichiers PRSA ===
Nombre total de lignes chargees : 420768
Lignes apres nettoyage : 412029
```

Partie 2 : Feature Engineering

But : Construire des variables utiles pour l’analyse & ML.

Features créées :

- **datetime**
- **season (winter, spring…)**
- **hour_category (morning, night…)**
- **is_weekend**
- **PM25 propre**

YEAR	MONTH	DAY	HOUR	PM25_clean	STATION
2013	3	26	10	194.0	Wanshouxigong
2013	4	21	21	102.0	Wanshouxigong
2013	4	21	22	71.0	Wanshouxigong
2013	4	27	7	70.0	Wanshouxigong
2013	5	1	15	46.0	Wanshouxigong
2013	5	3	23	98.0	Wanshouxigong
2013	5	10	5	97.0	Wanshouxigong
2013	5	11	10	71.0	Wanshouxigong
2013	5	14	5	15.0	Wanshouxigong
2013	6	4	23	71.0	Wanshouxigong

only showing top 10 rows

datetime	season	is_weekend	hour_category	PM25	STATION
2013-03-26 10:00:00	spring	0	morning	194.0	Wanshouxigong
2013-04-21 21:00:00	spring	1	evening	102.0	Wanshouxigong
2013-04-21 22:00:00	spring	1	evening	71.0	Wanshouxigong
2013-04-27 07:00:00	spring	1	morning	70.0	Wanshouxigong
2013-05-01 15:00:00	spring	0	afternoon	46.0	Wanshouxigong
2013-05-03 23:00:00	spring	0	evening	98.0	Wanshouxigong
2013-05-10 05:00:00	spring	0	night	97.0	Wanshouxigong
2013-05-11 10:00:00	spring	1	morning	71.0	Wanshouxigong
2013-05-14 05:00:00	spring	0	night	15.0	Wanshouxigong
2013-06-04 23:00:00	summer	0	evening	71.0	Wanshouxigong

only showing top 10 rows

Partie 3 : Statistiques descriptives

Objectif : Comprendre la pollution selon le temps et les stations.

Statistiques obtenues :

- **Moyenne PM25 par station**
- **Moyenne par heure**
- **Moyenne par mois / saison**
- **Tendances 2013–2017**

```
=== Statistiques descriptives ===
+-------------+-----------------+
|STATION      |PM25_mean        |
+-------------+-----------------+
|Dongsi       |86.19429678848283|
|Wanshouxigong|85.02413582402234|
|Nongzhanguan |84.83848298292484|
|Gucheng      |83.85208902318553|
|Wanliu       |83.37471599100398|
|Guanyuan     |82.93337203901532|
|Aotizhongxin |82.77361082632767|
|Tiantan      |82.16491115828656|
|Shunyi       |79.4916020028696 |
|Changping    |71.09974336541266|
|Huairou      |69.62636686112984|
|Dingling     |65.98949686451802|
+-------------+-----------------+
```

hour	PM25_mean
0	87.58842544316997
1	86.55975539067936
2	84.51625115955473
3	82.01466519414089
4	79.29976788719317
5	76.39536516039213
6	74.24095966620305
7	73.27224188961756
8	74.5333391284191
9	76.05975744211688
10	77.12279454609019
11	77.35168247981613
12	76.85832103429955
13	76.22007515706653
14	75.32682855299431
15	74.52381623780988
16	74.09924948695397
17	75.36720794392524
18	78.02209010899341
19	82.49701744186046
20	86.5555490959828
21	88.78055974041024
22	88.88793905559433
23	88.60675127491888

month	PM25_mean
1	93.66703682719546
2	87.57223442993471
3	94.66067775479556
4	72.73489020771513
5	63.10533821156949
6	69.09197095435685
7	71.74485251757115
8	53.4730143404799
9	61.478104613924806
10	91.72680042481127
11	93.33151859111256
12	104.5812437052171

year	PM25_mean
2013	80.04053649723235
2014	85.57570017602193
2015	79.626788228326 19
2016	71.930150943762 06
2017	92.67599234815877

Partie 3 : Stations les plus exposées

Objectif : Identifier les stations et les périodes les plus touchées par la pollution PM25

Cette analyse met en évidence les stations ayant les valeurs de PM25 les plus élevées (moyenne et pic maximum).
Elle inclut également l'étude des pics horaires et des variations saisonnières, permettant de comprendre quand et où la pollution est la plus importante.

=== 3.1 Stations les plus exposees ===

STATION	PM25_mean	PM25_max
Dongsi	86.19429678848283	737.0
Wanshouxigong	85.02413582402234	999.0
Nongzhanguan	84.83848298292484	844.0
Gucheng	83.85208902318553	770.0
Wanliu	83.37471599100398	957.0
Guanyuan	82.93337203901532	680.0
Aotizhongxin	82.77361082632767	898.0
Tiantan	82.16491115828656	821.0
Shunyi	79.4916020028696	941.0
Changping	71.09974336541266	882.0
Huairou	69.62636686112984	762.0
Dingling	65.98949686451802	881.0

=== 3.2 Pics horaires ===

hour	PM25_mean	PM25_max
22	88.88793905559433	770.0
21	88.78055974041024	685.0
23	88.60675127491888	737.0
0	87.58842544316997	809.0
1	86.55975539067936	881.0
20	86.5555490959828	685.0
2	84.51625115955473	999.0
19	82.49701744186046	670.0
3	82.01466519414089	857.0
4	79.29976788719317	801.0
18	78.02209010899341	684.0
11	77.35168247981613	705.0
10	77.12279454609019	661.0
12	76.85832103429955	844.0
5	76.39536516039213	770.0
13	76.22007515706653	741.0
9	76.05975744211688	640.0
17	75.36720794392524	689.0
14	75.32682855299431	718.0
8	74.5333391284191	610.0
15	74.52381623780988	708.0
6	74.24095966620305	720.0
16	74.09924948695397	687.0
7	73.27224188961756	712.0

season	PM25_mean
winter	95.48363017260837
autumn	82.33224553753918
spring	76.9735075454414
summer	64.67488111429427

Partie 3 : Détection automatique des anomalies dans les données.

Objectif :Repérer automatiquement les événements de pollution extrême à partir d'un indice global.

Nous avons construit un indice de pollution combinant plusieurs gaz (PM25, PM10, NO_2, CO, O_3) et normalisé sur l'ensemble des données.

Une valeur est considérée comme anomalie lorsqu'elle dépasse la moyenne + 3 écarts-types.

Cette étape permet d'isoler rapidement les épisodes de pollution grave

```
=== 3.3 Indice de pollution & anomalies ===
+-------------------+-------------+-----+----+---+---+------------------+
|datetime           |STATION      |PM25 |PM10|NO2|O3 |PollutionIndex    |
+-------------------+-------------+-----+----+---+---+------------------+
|2017-01-04 19:00:00|Gucheng      |571.0|679 |270|3  |0.5354310015520284|
|2014-02-15 00:00:00|Nongzhanguan |809.0|907 |137|7  |0.5242540989608206|
|2015-12-01 12:00:00|Gucheng      |733.0|733 |179|2  |0.522853688222319 |
|2014-01-16 00:00:00|Gucheng      |585.0|630 |195|29 |0.5174493792368285|
|2015-12-01 13:00:00|Gucheng      |741.0|741 |176|2  |0.5167147034487902|
|2017-01-03 23:00:00|Gucheng      |640.0|805 |185|4  |0.5161268612070459|
|2014-02-15 01:00:00|Nongzhanguan |781.0|888 |136|6  |0.5156806743011383|
|2014-01-16 04:00:00|Wanshouxigong|578.0|705 |192|4  |0.5121731598382518|
|2016-12-20 16:00:00|Gucheng      |634.0|664 |220|7  |0.5115553170369943|
|2015-12-01 15:00:00|Wanliu       |708.0|708 |179|8  |0.5113792397340443|
|2015-12-25 16:00:00|Aotizhongxin |635.0|647 |218|13 |0.5109956784789459|
|2016-12-20 15:00:00|Gucheng      |628.0|681 |214|9  |0.5105784637000378|
|2015-11-30 15:00:00|Wanshouxigong|687.0|704 |206|16 |0.5095013566731099|
|2013-03-08 14:00:00|Huairou      |534.0|835 |184|107|0.5094581255130017|
|2014-01-16 00:00:00|Wanshouxigong|623.0|768 |184|5  |0.507885702242436 |
|2014-02-26 13:00:00|Wanliu       |567.0|590 |259|2  |0.5071903900831385|
|2015-11-30 16:00:00|Wanshouxigong|687.0|704 |199|11 |0.5070840297153176|
|2017-01-01 12:00:00|Changping    |662.0|675 |203|5  |0.504893512751208 |
|2017-01-03 22:00:00|Gucheng      |626.0|720 |192|5  |0.5041036641163529|
|2014-01-16 03:00:00|Wanshouxigong|605.0|747 |183|3  |0.5039721546959491|
+-------------------+-------------+-----+----+---+---+------------------+
only showing top 20 rows


--- Anomalies detectees : 6604 ---
```

Partie 4 : GraphX – Modélisation du réseau

Objectif : Représenter les stations sous forme de graphe et étudier leurs interactions

Réseau GraphX construit :

- **Sommets = 12 stations**
- **Arêtes = connexions manuelles entre stations**
- **Poids = lien “1” pour symboliser une connexion simple**

```
=== Partie 4 avec GraphX  ===

=== Sommets (stations) ===
Sommet 4 : station=Dongsi, avgPM25=86.19429678848283
Sommet 11 : station=Guanyuan, avgPM25=82.93337203901532
Sommet 0 : station=Changping, avgPM25=71.09974336541266
Sommet 1 : station=Aotizhongxin, avgPM25=82.77361082632767
Sommet 6 : station=Nongzhanguan, avgPM25=84.83848298292484
Sommet 3 : station=Wanliu, avgPM25=83.37471599100398
Sommet 7 : station=Tiantan, avgPM25=82.16491115828656
Sommet 9 : station=Huairou, avgPM25=69.62636686112984
Sommet 8 : station=Dingling, avgPM25=65.98949686451802
Sommet 10 : station=Gucheng, avgPM25=83.85208902318553
Sommet 5 : station=Shunyi, avgPM25=79.4916020028696
Sommet 2 : station=Wanshouxigong, avgPM25=85.02413582402234

=== Aretes (connexions) ===
Arete 1 -> 4, poids=1.0
Arete 4 -> 11, poids=1.0
Arete 11 -> 2, poids=1.0
Arete 2 -> 7, poids=1.0
Arete 7 -> 3, poids=1.0
Arete 3 -> 10, poids=1.0
Arete 10 -> 6, poids=1.0
Arete 6 -> 5, poids=1.0
Arete 5 -> 9, poids=1.0
Arete 9 -> 0, poids=1.0
Arete 0 -> 8, poids=1.0
```

Partie 4 : Propagation & PageRank

- **PageRank (importance des stations) : Mesure de l'influence d'une station dans le graphe.**
- **Propagation de pollution: Idée est de simulée une propagation ou Chaque station envoie 10% de sa PM2.5 à ses voisines.**

```
=== Etude de la propagation de la pollution a travers le reseau ===

=== PageRank des stations (importance dans le reseau) ===
Sommet 8 (Dingling) : scorePageRank = 1.4417377130207998
Sommet 0 (Changping) : scorePageRank = 1.3995466176648939
Sommet 9 (Huairou) : scorePageRank = 1.34991003489324
Sommet 5 (Shunyi) : scorePageRank = 1.2915140551618824
Sommet 6 (Nongzhanguan) : scorePageRank = 1.2228129025367558
Sommet 10 (Gucheng) : scorePageRank = 1.1419880170954304
Sommet 3 (Wanliu) : scorePageRank = 1.046899916576224
Sommet 7 (Tiantan) : scorePageRank = 0.9350315630242164
Sommet 2 (Wanshouxigong) : scorePageRank = 0.8034217353159724
Sommet 11 (Guanyuan) : scorePageRank = 0.6485866438945087
Sommet 4 (Dongsi) : scorePageRank = 0.4664277128104338
Sommet 1 (Aotizhongxin) : scorePageRank = 0.25212308800563993

=== Pollution transmise aux stations voisines (10% de la PM25 moyenne) ===
Sommet 11 (Guanyuan) recoit  8.619429678848283 unites de pollution
Sommet 7 (Tiantan) recoit  8.502413582402234 unites de pollution
Sommet 5 (Shunyi) recoit  8.483848298292484 unites de pollution
Sommet 6 (Nongzhanguan) recoit  8.385208902318553 unites de pollution
Sommet 10 (Gucheng) recoit  8.337471599100398 unites de pollution
Sommet 2 (Wanshouxigong) recoit  8.293337203901533 unites de pollution
Sommet 4 (Dongsi) recoit  8.277361082632767 unites de pollution
Sommet 3 (Wanliu) recoit  8.216491115828656 unites de pollution
Sommet 9 (Huairou) recoit  7.94916020028696 unites de pollution
Sommet 8 (Dingling) recoit  7.109974336541267 unites de pollution
Sommet 0 (Changping) recoit  6.9626366686112984 unites de pollution
```

Partie 5 : Prédiction avec Spark MLlib

Objectif : prédire PM2.5 à partir de features temporelles + station

Régression linéaire

```
=== Partie 5 : Prediction de PM2.5 avec Spark MLlib ===

=== Modele 1 : Regression lineaire ===
RMSE = 80,78
```

Arbre de décision

```
=== Modele 2 : Arbre de decision ===
RMSE = 75,66
```

Forêt aléatoire

```
=== Modele 3 : Foret aleatoire ===
RMSE = 76,49
```

Partie 6 : Temps réel

Objectif : détecter des anomalies sur un flux simulé.

Étapes :

1. **Simulation de données en continu**
2. **Transformation : ajout d'un flag is_anomaly**
3. **Détection automatique (PM25 > 150)**

```
--- Flux transforme (avec detection des anomalies) ---
+----------------+-------------+-----+----------+
|timestamp       |station      |PM25 |is_anomaly|
+----------------+-------------+-----+----------+
|2025-01-01 12:00|Aotizhongxin |80.0 |false     |
|2025-01-01 12:00|Changping    |45.0 |false     |
|2025-01-01 12:00|Dingling     |95.0 |false     |
|2025-01-01 12:00|Dongsi       |180.0|true      |
|2025-01-01 12:00|Guanyuan     |160.0|true      |
|2025-01-01 12:00|Gucheng      |200.0|true      |
|2025-01-01 12:00|Huairou      |50.0 |false     |
|2025-01-01 12:00|Nongzhanguan |70.0 |false     |
|2025-01-01 12:00|Shunyi       |60.0 |false     |
|2025-01-01 12:00|Tiantan      |55.0 |false     |
|2025-01-01 12:00|Wanliu       |40.0 |false     |
|2025-01-01 12:00|Wanshouxigong|170.0|true      |
+----------------+-------------+-----+----------+
```

```
=== Partie 6 : Traitement en temps reel (simulation de flux) ===

--- Flux simule recu  ---
+----------------+-------------+-----+
|timestamp       |station      |PM25 |
+----------------+-------------+-----+
|2025-01-01 12:00|Aotizhongxin |80.0 |
|2025-01-01 12:00|Changping    |45.0 |
|2025-01-01 12:00|Dingling     |95.0 |
|2025-01-01 12:00|Dongsi       |180.0|
|2025-01-01 12:00|Guanyuan     |160.0|
|2025-01-01 12:00|Gucheng      |200.0|
|2025-01-01 12:00|Huairou      |50.0 |
|2025-01-01 12:00|Nongzhanguan |70.0 |
|2025-01-01 12:00|Shunyi       |60.0 |
|2025-01-01 12:00|Tiantan      |55.0 |
|2025-01-01 12:00|Wanliu       |40.0 |
|2025-01-01 12:00|Wanshouxigong|170.0|
+----------------+-------------+-----+
```

```
=== Anomalies detectees automatiquement ===
+----------------+-------------+-----+----------+
|timestamp       |station      |PM25 |is_anomaly|
+----------------+-------------+-----+----------+
|2025-01-01 12:00|Dongsi       |180.0|true      |
|2025-01-01 12:00|Guanyuan     |160.0|true      |
|2025-01-01 12:00|Gucheng      |200.0|true      |
|2025-01-01 12:00|Wanshouxigong|170.0|true      |
+----------------+-------------+-----+----------+
```

Conclusion

Résumé du travail :

- **Projet complet couvrant toutes les briques Spark**
- **Analyse avancée + graphe + ML modèle de prédiction + traitement en temps réel**
- **Mise en place d'une architecture professionnelle**

Merci pour votre attention !