1. Cf GitHub Q1.R
2. Il ne serait pas intéressant d’appliquer une ACP sur ce dataset car la majeure partie de ses variables sont qualitatives.
3. Cf GitHub Q3.R
4. Cf GitHub Q4.R
5. Cf GitHub Q5.R

Analyse du clustering (K-means)

Nous avons appliqué un algorithme de K-means sur les stations de métro franciliennes, en utilisant trois variables :

* le niveau de pollution (transformé en variable numérique),
* la longitude (stop\_lon),
* la latitude (stop\_lat).

Résultats du clustering

* Le modèle a identifié 3 clusters, mais les centres de ces clusters sont très proches géographiquement (tous autour de Paris intra-muros).
* Le cluster représentant la pollution élevée contient seulement 13 stations, contre 242 pour la pollution moyenne et 111 pour la faible.
* La répartition spatiale est donc fortement déséquilibrée, avec une concentration importante de stations dans des zones proches.

Interprétation

* Le clustering ne révèle pas de lien clair entre la localisation géographique des stations et leur niveau de pollution.
* Cela s'explique par :
  + la faible variation spatiale (stations concentrées dans Paris),
  + la rareté des stations très polluées,
  + l'absence de standardisation initiale qui réduit le poids de la variable niveau\_pollution.

Conclusion

Le clustering met en évidence une limite des données : la concentration géographique des stations et la répartition inégale des niveaux de pollution empêchent de tirer des conclusions fortes sur une éventuelle corrélation entre la position géographique et la pollution dans le réseau de métro.

Analyse du modèle des K plus proches voisins

Nous avons appliqué l’algorithme k-plus proches voisins (k-NN) avec k = 5 sur les données filtrées du métro, en utilisant uniquement les coordonnées géographiques (stop\_lat, stop\_lon) comme variables explicatives, et en ciblant la variable niveau\_pollution.  
Les classes "pas de données" et "station aérienne" ont été exclues afin de ne conserver que les niveaux significatifs : pollution faible, pollution moyenne, pollution élevée.

#### Résultats :

* Exactitude du modèle : 63,64 %
* Matrice de confusion :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pollution élevée | Pollution faible | Pollution moyenne |
| Prévu : élevée | 0 | 1 | 0 |
| Prévu : faible | 0 | 16 | 21 |
| Prévu : moyenne | 3 | 15 | 54 |

#### Analyse :

* Le modèle arrive globalement à distinguer les zones à pollution moyenne, mais a plus de mal à bien classer les zones à pollution faible ou élevée, qui sont souvent confondues avec la classe "pollution moyenne".
* Cela peut s’expliquer par le fait que seules les coordonnées géographiques sont utilisées, ce qui est sans doute insuffisant pour expliquer à elles seules les différences de pollution.  
  D’autres variables comme la densité de trafic, la fréquentation ou le type de station (ouverte/fermée) pourraient améliorer le modèle.

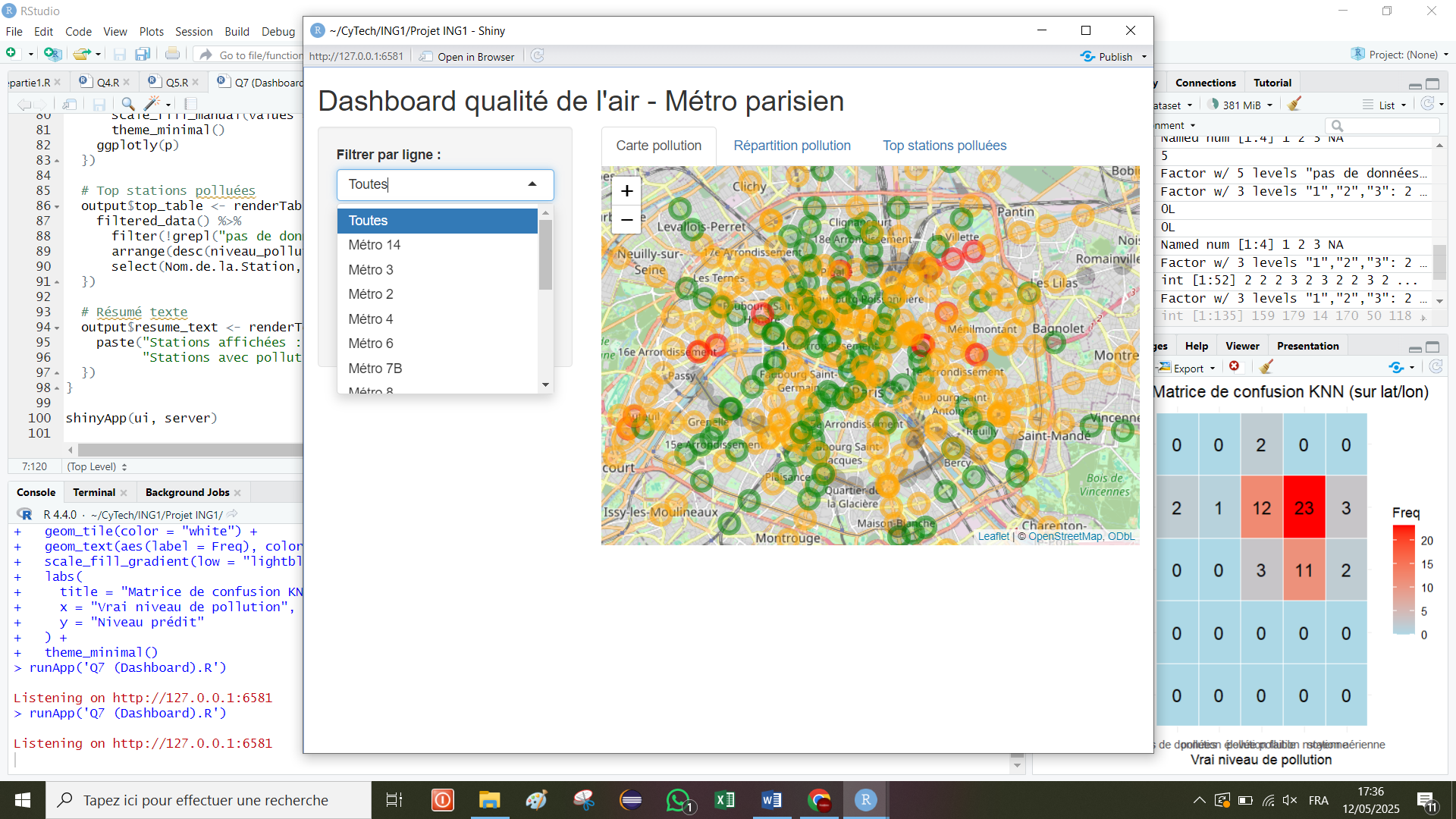
#### Conclusion :

Le modèle k-NN offre une performance correcte mais limitée (63 % de précision), probablement à cause de la faible quantité de variables explicatives. Il serait pertinent d'enrichir les données avec des informations contextuelles pour améliorer les résultats.

1. Ce dashboard sur la qualité de l'air dans le métro parisien présente un grand intérêt en tant qu'outil de visualisation et d'analyse des données de pollution. Il permet aux utilisateurs de suivre en temps réel les niveaux de pollution dans les stations de métro, d'identifier les zones les plus touchées, et d'avoir une vue d'ensemble de la répartition de la pollution en fonction des lignes. Ce type de dashboard est essentiel pour sensibiliser les citoyens à la qualité de l'air qu'ils respirent, particulièrement dans les environnements urbains denses comme le métro.

Les acteurs principaux auxquels ce dashboard est destiné incluent :

* 1. Les autorités publiques et les gestionnaires de transports (comme la RATP et les collectivités locales par exemple) qui peuvent utiliser ces informations pour prendre des décisions éclairées sur les mesures à adopter pour améliorer la qualité de l'air dans les stations de métro et prévenir la pollution.
  2. Les usagers du métro, qui peuvent bénéficier d'informations sur les niveaux de pollution afin d'adapter leur comportement, comme choisir des heures de transport moins polluées.
  3. Les chercheurs et les experts en environnement qui peuvent analyser les tendances de la pollution dans le métro et proposer des solutions pour réduire son impact.
  4. Les décideurs politiques qui peuvent utiliser ces données pour orienter leurs politiques publiques en matière de gestion de la qualité de l'air et d'aménagement urbain.
  5. En résumé, ce dashboard est un outil essentiel pour une gestion proactive de la qualité de l'air et permet de prendre des actions ciblées en fonction des zones les plus polluées.



Sources Partie 1 :

<https://pkgs.rstudio.com/flexdashboard/articles/shiny.html>

<https://rstudio.github.io/shinydashboard/>

<https://www.datacamp.com/fr/tutorial/k-means-clustering-r>

<https://www.datacamp.com/fr/tutorial/k-nearest-neighbors-knn-classification-with-r-tutorial>