UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

Département d'informatique et d'ingénierie

Cours INF1683 : Introduction à l'apprentissage automatique

Analyse prédictive et exploration des données de transport urbain (Mta New York)

Projet de fin de session

Présenté à : Yapi Daniel

Par

Tambat Trésor Mégane

Kouyaté Yasmine Jawad

Djieunang Noumbo Nelie Mabelle

Makouet Mounyiche Njiemoun Valerie

Nyami Ndjeukou Djokomeni Annie Christelle

Avril 2025

Introduction

1.1. Contexte et problématique

Le développement de la ville de New York repose principalement sur l'efficacité de son système de transport en commun, notamment le réseau de métro opéré par la MTA (Metropolitan Transportation Authority). Dans un environnement où les données liées à la mobilité urbaine sont devenues essentielles pour comprendre et améliorer les systèmes de transport, l'analyse fine des flux de passagers représente enjeu majeur.

La MTA souhaite mieux comprendre la fréquentation de ses stations de métro afin d'anticiper les périodes de surcharge, d'optimiser la répartition des ressources et d'améliorer l'expérience passager.

En raison de l'accroissement constant de la population de la ville de New York, ainsi que du nombre de touristes, la gestion optimale des flux de passagers est essentielle en vue d'assurer un service de qualité.

1.2. Objectifs du projet

Ce projet a pour but d'implémenter des techniques d'apprentissage machine pour analyser, modéliser et prédire les flux de fréquentation du métro de New York. Les objectifs spécifiques sont les suivants :

- Préparer et nettoyer les données brutes provenant de cinq fichiers CSV des années 2014 à 2018.

- Explorer les tendances temporelles de la fréquentation (jours, heures, saisons).

- Diviser les stations en groupes homogènes à travers des algorithmes de clustering non supervisé.

- Construire un modèle de fréquentation des stations à travers des algorithmes supervisés.

- Comparer la performance des différents modèles.

- Formuler des recommandations opérationnelles pour la MTA.

1.3. Approche méthodologique

Le projet suit une approche structurée définie en cinq étapes suivantes:

- Importation et nettoyage des données: Collecte, nettoyage et préparation des fichiers.

- Analyse exploratoire (EDA): Identification des tendances de fréquentation selon des facteurs différents.

- Modélisation: Utilisation des techniques d'apprentissage supervisé et non supervisé.

- Évaluation: Analyse des performances des différents modèles.

- Visualisation et interprétation: Présentation des résultats et propositions de pistes d'amélioration.

2. Préparation et exploration des données

2.1. Source et description des données :

Les données utilisées sont issues de fichiers ouverts publiés par la MTA et disponibles sur la plateforme Kaggle sous le nom "NYS Turnstile Usage Data". Chaque fichier comporte plusieurs colonnes qui renseignent sur les variables telles que : la station (STATION), la date (DATE), l'heure (TIME), les compteurs d'entrées (ENTRIES) et de sorties (EXITS).

A noter que ces données sont stockées dans un format cumulatif, ce qui nécessite un calcul de différence pour pouvoir obtenir les données réelles sur un intervalle donné.

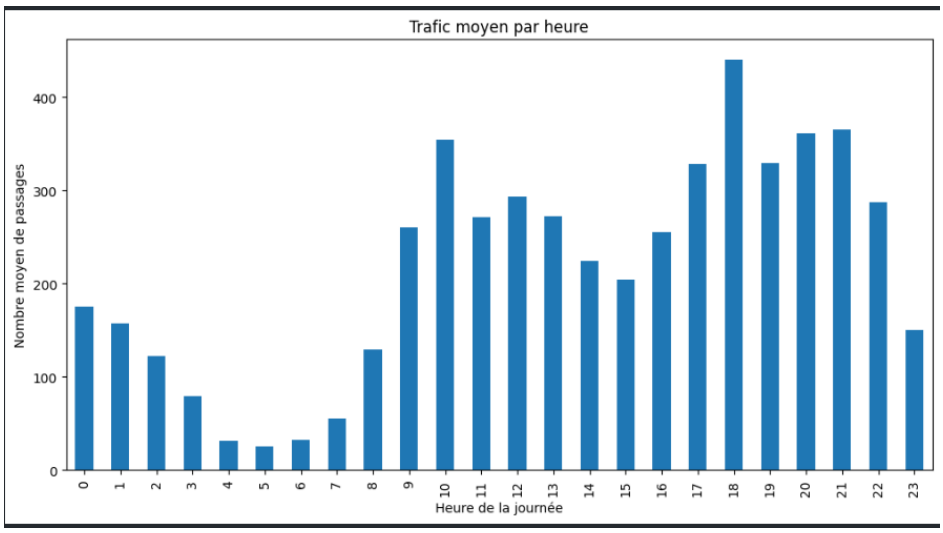
Un script Python a été utilisé pour parcourir tous les fichiers CSV présents dans un dossier, détecter ceux contenant les colonnes nécessaires, les charger, les nettoyer, puis les fusionner en un unique jeu de données global.

1. Exploration et Prétraitement des Données

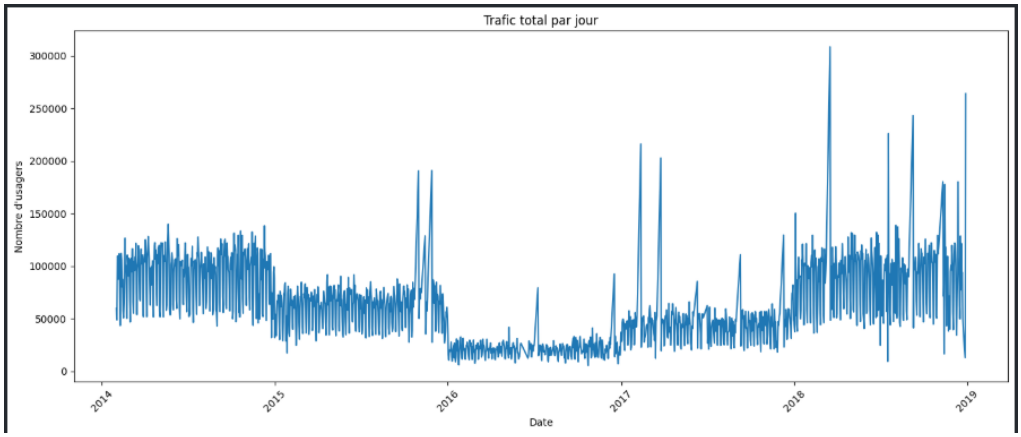
Les données brutes ont été nettoyées et préparées :

- Suppression des valeurs manquantes et aberrantes.
- Calcul du trafic total par tranche horaire (4 heures).
- Catégorisation du trafic en 3 niveaux : faible, moyenne, forte.

2. Analyse exploratoire des données (EDA)

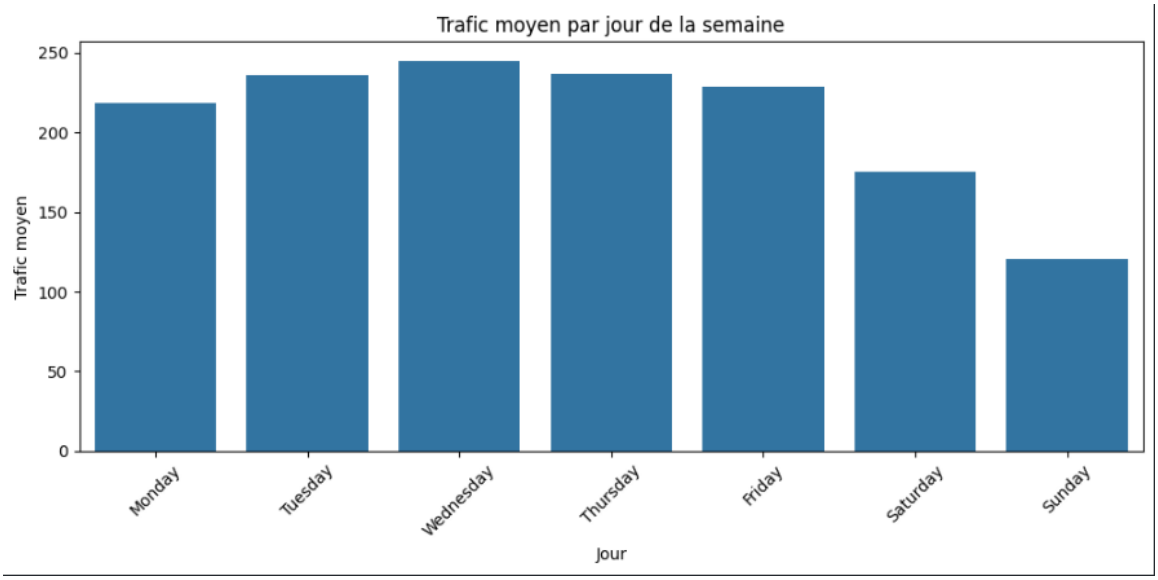


Le réseau de métro connaît une utilisation très marquée aux heures de pointe (matin et soir), ce qui reflète les horaires typiques de travail de la population. Cette tendance justifie l'allocation de ressources supplémentaires (trains, personnel) durant ces créneaux, tandis qu'une réduction peut être envisagée durant les heures creuses de la nuit.

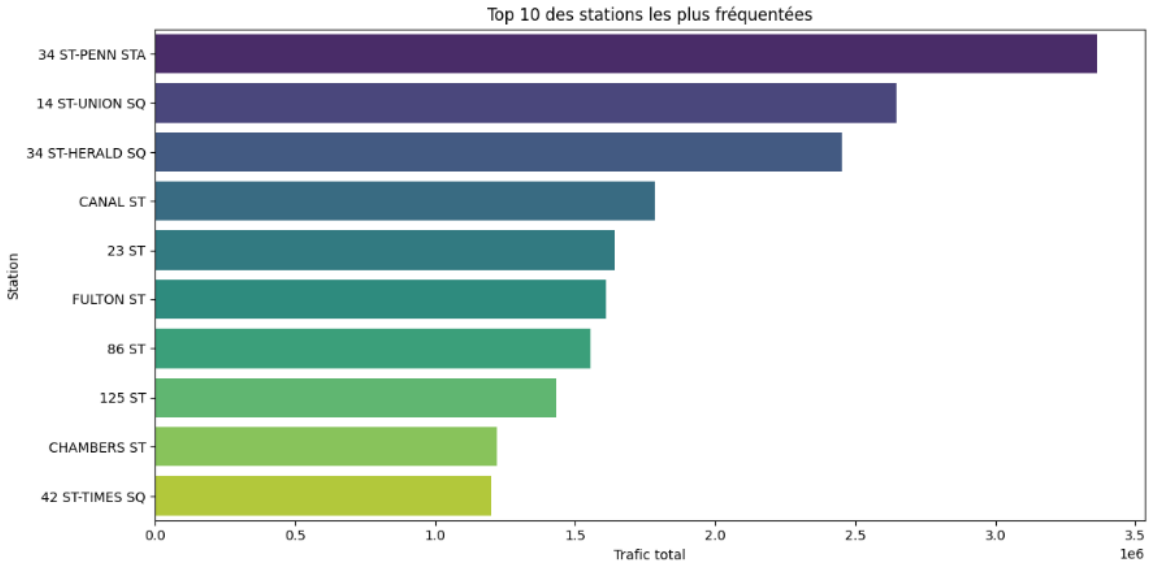


L'évolution du trafic journalier révèle des tendances de fréquentation à long terme et permet de détecter des ruptures, des anomalies ou des événements impactant le réseau. Cette analyse est essentielle pour

anticiper la planification annuelle, détecter les périodes de sous-utilisation ou d'engorgement, et améliorer la robustesse du système de collecte de données.

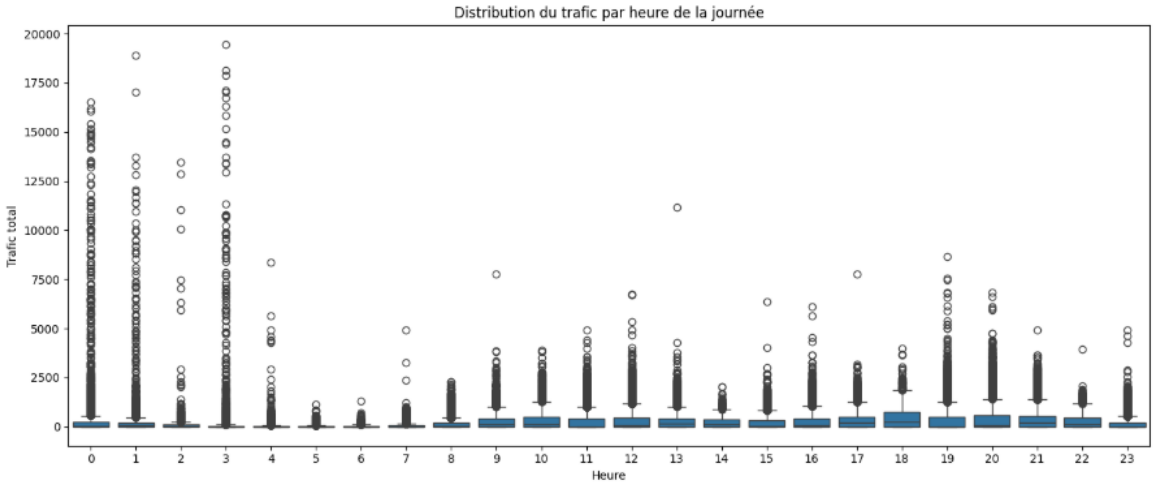


Le trafic du métro est clairement structuré autour de la semaine de travail. Les autorités de transport peuvent optimiser la planification des trains, des horaires et des ressources humaines en adaptant leur offre à cette dynamique, avec un renforcement des capacités en semaine et une réduction des fréquences le week-end.

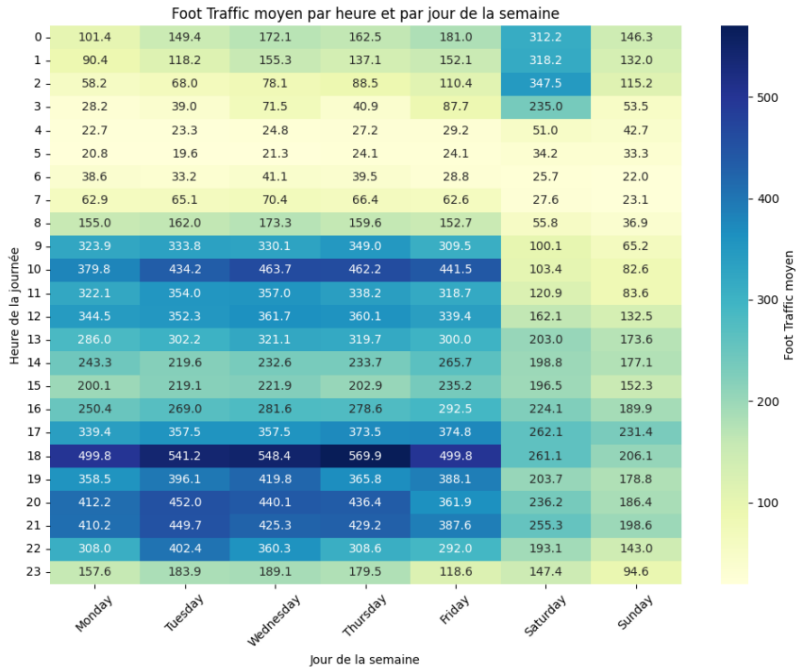


Les stations les plus fréquentées sont des centres de transit stratégiques qui concentrent à la fois des correspondances entre lignes, des

connexions intermodales et une proximité avec des zones économiques ou touristiques majeures. Cela justifie une attention particulière en termes de sécurité, de maintenance et de gestion des flux de passagers.



Cette visualisation permet de détecter les heures présentant une forte variabilité dans le trafic, ce qui est crucial pour le dimensionnement dynamique du service. Elle met aussi en évidence des valeurs aberrantes qu'il peut être utile d'examiner individuellement (nettoyage de données ou détection d’événements particuliers).



Interprétation :

Trafic de pointe en semaine (Lundi → Vendredi) :

Entre 8h et 9h du matin (heure de pointe du matin)

Entre 17h et 19h (heure de pointe du soir)

Ces heures correspondant aux déplacements domicile-travail.

Week-ends (Samedi et Dimanche) :

Le trafic est moins élevé et plus réparti sur la journée.

Le pic est plus tardif (autour de 14h-16h), ce qui reflète des habitudes différentes (shopping, loisirs).

Nuit (0h → 5h) :

Le trafic est naturellement très faible pendant les heures de la nuit.

Interprétations concrètes

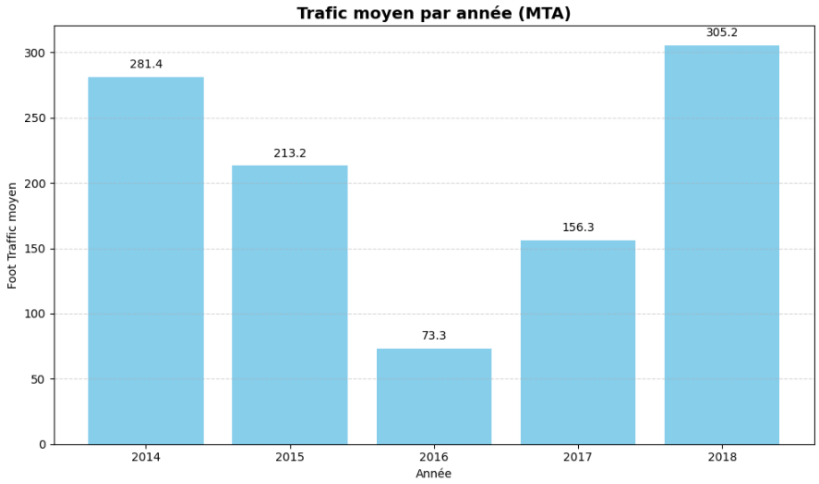
Heures creuses : 1h → 5h du matin tous les jours

Plus forte affluence : Mercredi entre 18h et 19h (569.9)

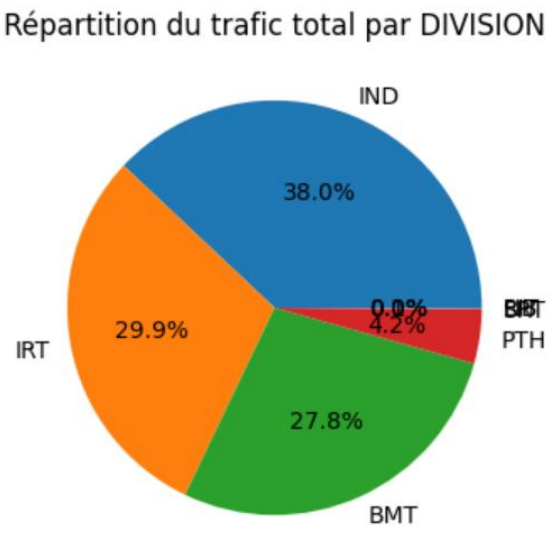
Moins de trafic : Dimanche matin très tôt (23.1 à 33.3entre 3h et 6h)

Mardi et Mercredi montrent les trafics les plus intenses parmi les jours de la semaine.

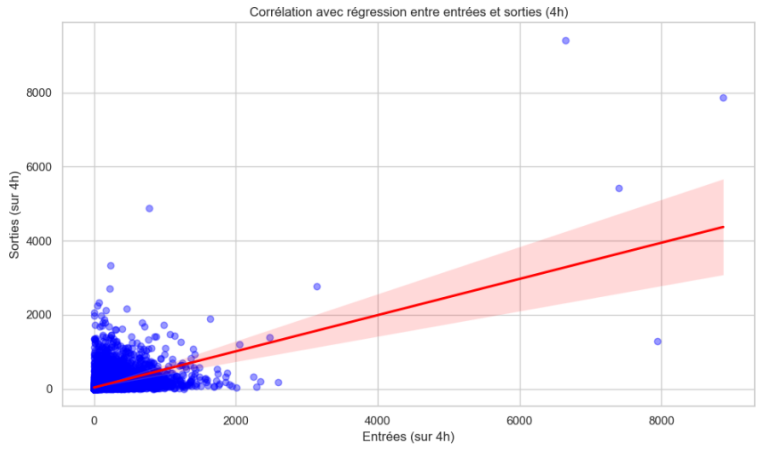
Cette carte thermique est un outil puissant pour identifier les périodes de surcharge du réseau. Elle confirme une concentration marquée du trafic aux heures de pointe en semaine, justifiant un renforcement des services (fréquence des trains, sécurité, personnel) durant ces créneaux. À l'inverse, elle indique aussi les plages horaires où une réduction de l'offre peut être envisagée sans impacter significativement le service.



L'évolution du trafic montre une instabilité apparente dans les données entre 2015 et 2017, mais une forte reprise en 2018, possiblement due à une meilleure consolidation ou couverture des données. Cela suggère qu'il est essentiel d'interpréter ces valeurs avec prudence et, si possible, de les croiser avec des événements ou des rapports de maintenance/fermeture de lignes pour chaque année.



Les divisions IND, IRT et BMT concentrent à elles seules près de 95 % du trafic du réseau, ce qui justifie une priorité stratégique pour les investissements, la maintenance et les ressources humaines dans ces secteurs. Les autres divisions ont un impact plus localisé et peuvent être gérées de manière plus ciblée.



Coefficient de corrélation (Pearson) : 0.545

Cela indique une corrélation modérée positive entre les entrées et les sorties, ce qui signifie que généralement, lorsque le nombre d'entrées augmente, le nombre de sorties augmente également, mais cette relation n'est pas très forte.

Dispersion importante des points :

On observe une grande variabilité, ce qui signifie qu'il y a de nombreux cas où le nombre d'entrées ne correspond pas essentiellement au nombre de sorties :

Certaines stations montrent une fréquentation très déséquilibrée (beaucoup plus d'entrées que de sorties ou inversement).

Quelques points isolés montrent des comportements extrêmes (très forts écarts entrées-sorties), probablement liés à des événements spéciaux ou à des stations très spécifiques.

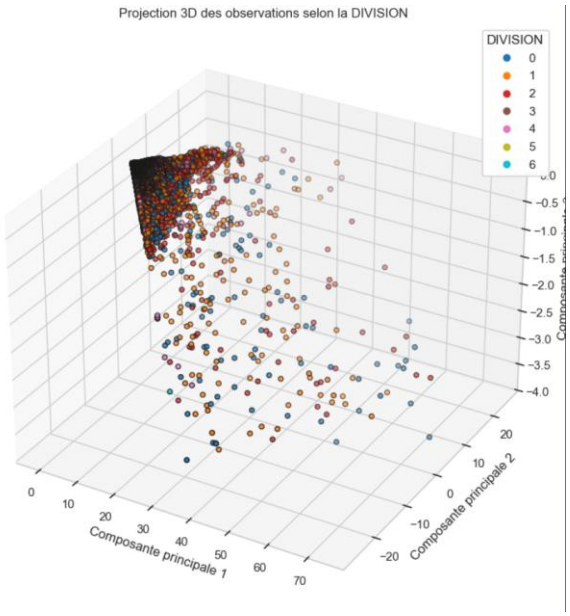
MODELISATION NON SUPERVISE

L'analyse exploratoire a permis d'identifier :

- Des pics de fréquentation évidents (heures de pointe matin et soir).

- Des différences de fréquentation selon les jours de la semaine.
- Une corrélation modérée entre les entrées et les sorties (coefficient de Pearson = 0.545).

3. Réduction de Dimension (ACP)



Ce graphique présente une représentation tridimensionnelle des observations projetées dans l'espace des trois premières composantes principales (ACP), avec une coloration selon la variable DIVISION, codée de 0 à 6 (ex. : IND, IRT, BMT, PTH...).

Observations clés :

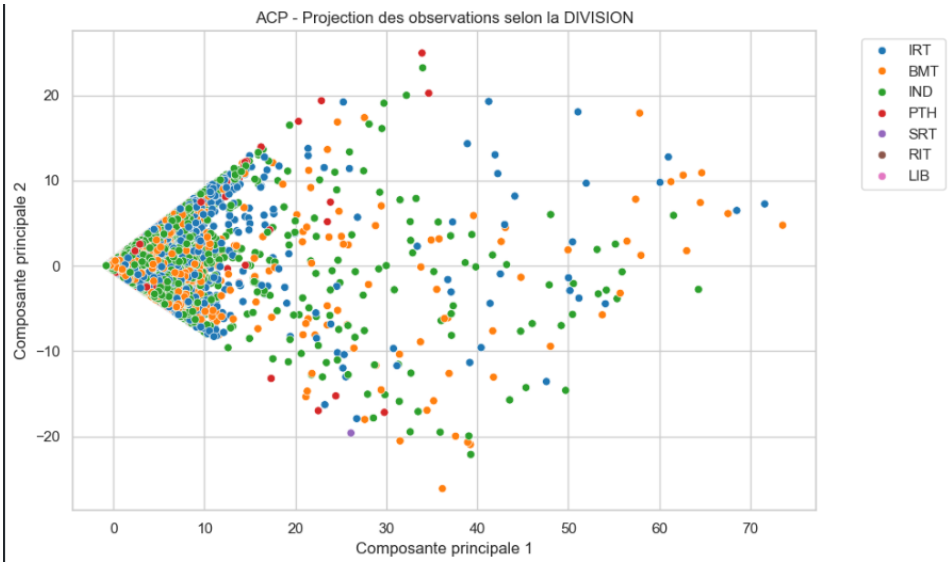
- Différenciation par DIVISION :
  Bien que les groupes soient visuellement superposés dans certaines zones, on remarque une certaine dispersion par division dans l'espace des composantes principales. Cela suggère que les profils de fréquentation varient partiellement selon les divisions.
- Concentration importante dans la zone dense proche de l'origine :
  La majorité des observations est concentrée dans une région restreinte de l'espace, indiquant que de nombreuses stations

partagent des caractéristiques communes (mêmes amplitudes de fréquentation, plages horaires, etc.).

- Distribution progressive le long des axes CP1 et CP2 : Cela montre qu'il existe des gradients structurels dans les données (par exemple : stations à faible trafic ↔ fort trafic, ou zones résidentielles ↔ zones commerciales).
- Utilité de la 3e composante (CP3) :
  L'axe vertical (Composante principale 3) semble moins discriminant visuellement, ce qui est courant après CP1 et CP2 en ACP. Mais il reste utile pour capturer la variance résiduelle non expliquée dans le plan 2D.

Conclusion pratique :

Cette projection 3D basée sur l'ACP confirme que les divisions du métro (IND, IRT, BMT, etc.) présentent des profils de fréquentation partiellement distincts, mais aussi des chevauchements significatifs. Cela implique que les comportements de trafic ne dépendent pas uniquement de la division, mais aussi d'autres facteurs (localisation, temporalité, usage spécifique).



Cette projection ACP confirme que les divisions n'expliquent pas à elles seules les différences de comportement entre les stations. Des stations issues de divisions différentes peuvent se retrouver très proches dans l'espace factoriel, ce qui suggère que d'autres variables (heure, quartier, fonction de la station) influencent fortement le trafic observé.

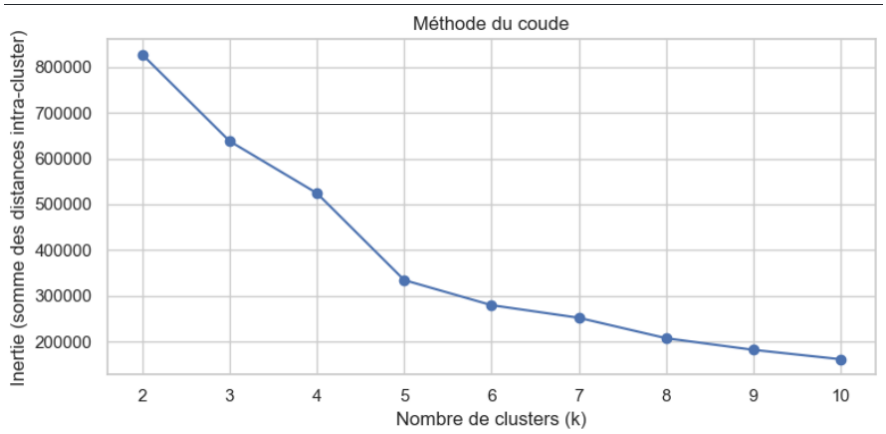
Conclusion :

L'analyse en composantes principales (ACP) a permis de :

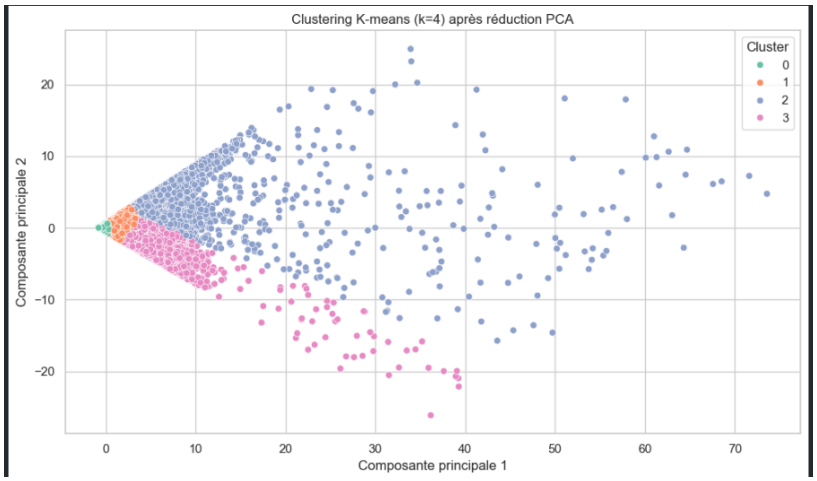
- Réduire efficacement la dimensionnalité des données tout en conservant l'information principale.
- Identifier visuellement les groupes de stations selon leur comportement.

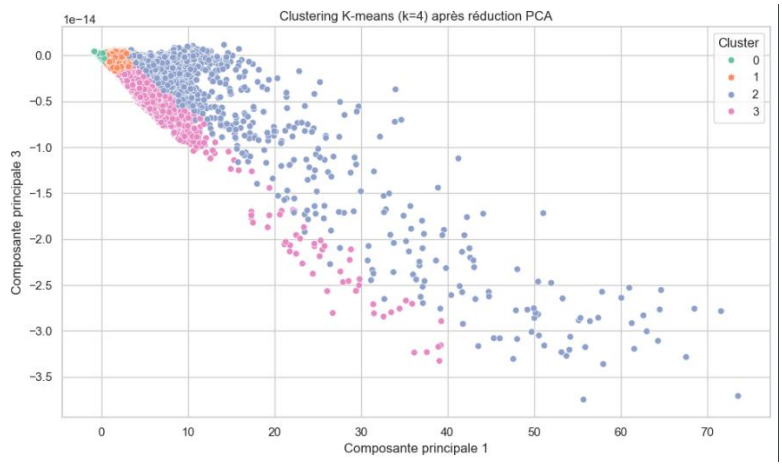
4. Clustering non-supervisé

Deux méthodes complémentaires ont été utilisées :



- K-means : Détermination du nombre optimal de clusters (méthode du coude et silhouette score). 4 clusters identifiés clairement :
    - Cluster 0 : Stations à faible trafic.
    - Cluster 1 : Stations à trafic modéré.
    - Cluster 2 : Stations à fort trafic sortant (zones commerciales/bureaux).
    - Cluster 3 : Stations à fort trafic entrant (hubs touristiques).





Ces deux graphiques montrent la segmentation des stations du réseau MTA en 4 clusters distincts, après réduction de dimension par Analyse en Composantes Principales (ACP).

- Le premier graphique projette les clusters dans l'espace (Composante 1, Composante 3).
- Le second graphique dans l'espace (Composante 1, Composante 2).

Observations générales sur les clusters :

- Cluster 2 (en bleu) est le plus vaste et dispersé :

    - Il regroupe les stations à plus forte variabilité en fréquentation ou comportement, avec des valeurs élevées sur les composantes principales.
    - Représente possiblement des stations centrales ou très fréquentées (ex. Times Sq, Grand Central…).
- Clusters 0, 1 et 3 (vert, orange, rose) sont plus compacts et concentrés :
    - Ces groupes présentent des profils de fréquentation similaires, probablement des stations résidentielles ou de faible trafic.
    - Leur répartition dans les deux espaces indique qu'ils diffèrent aussi par des composantes secondaires (horaires, jours de la semaine, etc.).
- A séparation des clusters est visuellement nette, particulièrement dans le plan (CP1, CP2), ce qui valide la pertinence de l'approche K-means avec k = 4.

Conclusion :

Grâce à la combinaison ACP + K-means, nous avons pu identifier 4 grands profils de stations dans le réseau MTA. Cela permet de :

- Mieux comprendre les typologies de fréquentation.
- Adapter les ressources opérationnelles (horaires, services, personnel).
- Étudier les comportements de mobilité pour chaque groupe de manière ciblée.

2nd algorithme

- DBSCAN : Identification des anomalies et stations atypiques (flux inhabituels).

Analyse de l’algorithme DBSCAN (ε = 0.5, min_samples = 10)

Principe de DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering basé sur la densité, c’est-à-dire qu’il regroupe des points proches les uns des autres (zones denses) et identifie comme anomalies les points isolés.

- ε = 0.5 : rayon de voisinage pour considérer qu’un point est "proche".
- min_samples = 10 : un groupe de 10 points ou plus dans un rayon ε est considéré comme un cluster dense.

Résultat sur ton graphique

Dans la projection 2D (PC1, PC2) :

- Points noirs (label = -1) : identifiés comme outliers (anomalies). Ils représentent des stations peu connectées ou atypiques en termes de fréquentation ou de comportement horaire.
- Clusters 0, 1, 2 (en bleu, orange et vert) :
    - Ce sont les zones denses détectées automatiquement par DBSCAN.
    - Les stations dans ces groupes présentent des profils similaires de fréquentation.

Avantages du modèle DBSCAN ici

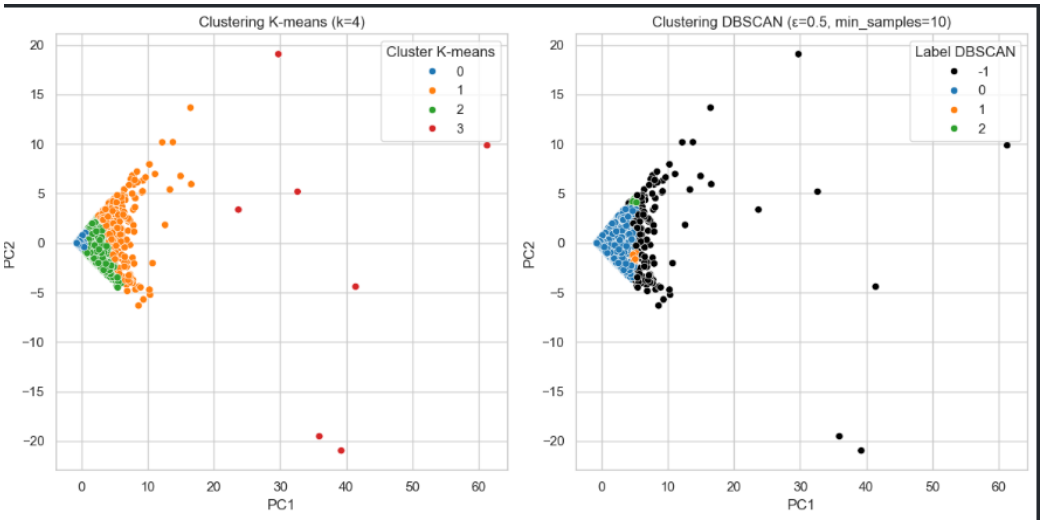
1. Aucune hypothèse sur le nombre de clusters : contrairement à K-means, DBSCAN découvre automatiquement combien de groupes il y a réellement dans les données.
2. Robuste aux outliers : le modèle n’impose pas de regroupement forcé. Les points marginaux sont bien identifiés.
3. Flexibilité de forme : DBSCAN n’impose pas de forme sphérique aux clusters, ce qui est utile avec des données après réduction par ACP.

Limites constatées

- Le choix des hyperparamètres ε et min_samples est crucial. Si ε est trop petit, la plupart des points sont considérés comme du bruit. Si trop grand, tous les points finissent dans un même cluster.
- Ici, la majorité des points sont considérés comme anomalies (noirs), ce qui pourrait indiquer que :
    - Les données sont très dispersées dans l'espace projeté.
    - Le seuil ε=0.5 est peut-être trop restrictif.

Conclusion sur DBSCAN dans ce contexte

DBSCAN permet ici de mettre en lumière les stations atypiques (hors normes) dans le réseau MTA. Il complète bien les approches comme K-means en détectant les comportements marginaux et en laissant la structure des clusters émerger naturellement.



Comparaison des algorithmes de clustering : K-means vs DBSCAN

Ces deux graphiques montrent les résultats de segmentation de stations de métro projetées dans l'espace des deux premières composantes principales (PC1 et PC2), selon deux algorithmes :

- À gauche : Clustering K-means avec k = 4

- À droite : Clustering DBSCAN (ε = 0.5, min_samples = 10)

K-means (gauche)

- K-means forme quatre groupes bien définis (0, 1, 2, 3).
- Les centroïdes sont implicitement utilisés pour répartir les points.
- Très efficace pour des données de forme sphérique et bien réparties.
- Mais : faible sensibilité aux anomalies (tous les points sont forcés dans un cluster).

DBSCAN (droite)

- DBSCAN est un algorithme basé sur la densité :
    - Les points noirs (-1) sont des anomalies ou outliers (non assignés à un cluster).
    - Les points colorés sont regroupés selon leur densité locale.
- DBSCAN est capable de détecter les valeurs extrêmes ou stations atypiques, ce que K-means ne fait pas naturellement.

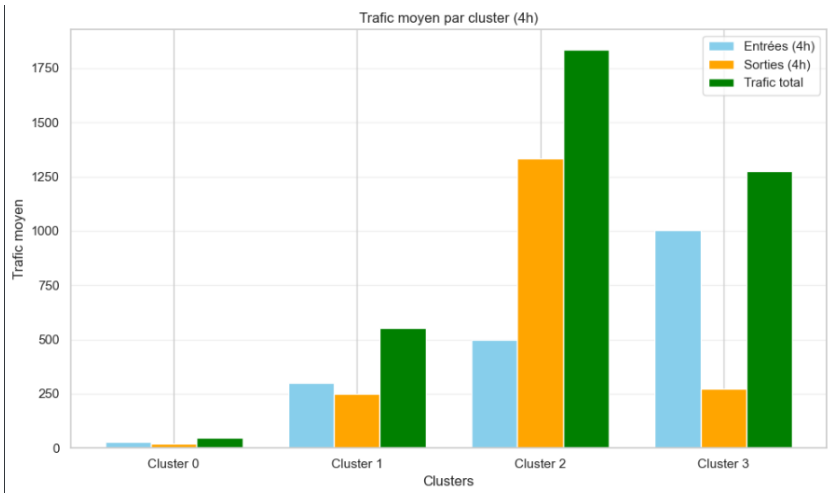
Comparaison & conclusion :

Critère	K-means	DBSCAN
Hypothèse	k clusters fixes	Aucun k à spécifier
Forme des clusters	Sphérique	forme libre
Détection d'anomalie	Non	Oui (label -1)

Critère	K-means	DBSCAN
Robustesse au bruit	Faible	Excellente
Adapté pour	Données bien séparées	Données bruitées ou avec outliers

En pratique :

- K-means est efficace pour une segmentation structurée et homogène.
- DBSCAN est plus adapté si l'on cherche à identifier des stations exceptionnelles ou marginales, comme des zones de très forte ou très faible affluence.



Ce graphique permet de distinguer clairement quatre profils de fréquentation des stations de métro, regroupés selon les clusters identifiés par l'analyse :

Cluster 0 (très faible fréquentation)Le:

Le trafic total (entrées et sorties) est très limité. Ce sont des stations secondaires ou périphériques, peu utilisées par les usagers.

Cluster 1 (fréquentation modérée) ::

Présente un trafic moyen équilibré entre entrées et sorties. Ces stations correspondent à des quartiers résidentiels ou des zones mixtes à fréquentation régulière.

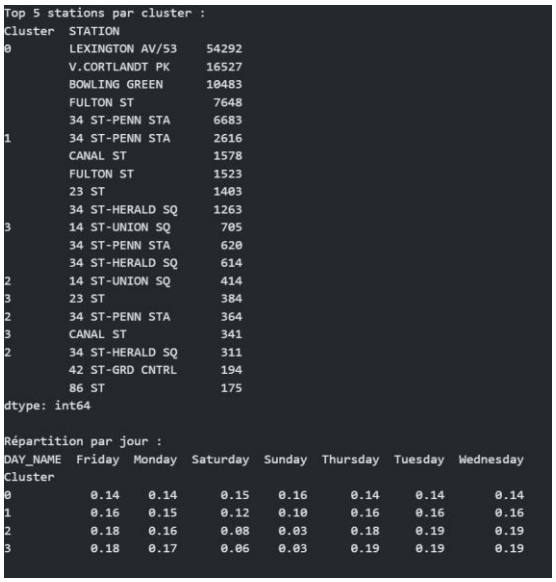
Cluster 2 (forte sortie)correspondant de station:

Ce cluster est marqué par un trafic exceptionnellement élevé en sorties, beaucoup plus qu'en entrées. Ces stations correspondent à des zones commerciales, des quartiers d'affaires ou des stations utilisées principalement en sortie pour aller travailler ou consommer.

Groupe 3 (forte entrée)type de profil:

Se distingue nettement par un trafic en entrée très élevé par rapport aux sorties. Ce profil typique correspond à des hubs centraux, gares principales ou zones touristiques, où de nombreux passagers arrivent régulièrement.

APPRECU DES STATIONS DANS CHAQUES CLUSTERS

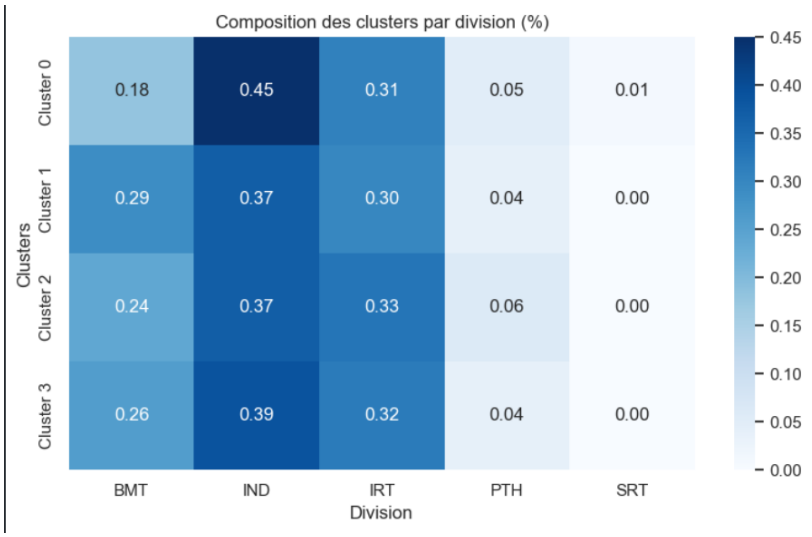


```
Top 5 stations par cluster :
Cluster  STATION
0        LEXINGTON AV/53    54292
         V.CORTLANDT PK     16527
         BOWLING GREEN      10483
         FULTON ST           7648
         34 ST-PENN STA      6683
1        34 ST-PENN STA      2616
         CANAL ST            1578
         FULTON ST           1523
         23 ST               1403
         34 ST-HERALD SQ     1263
3        14 ST-UNION SQ       705
         34 ST-PENN STA       620
         34 ST-HERALD SQ      614
2        14 ST-UNION SQ       414
3        23 ST                384
2        34 ST-PENN STA       364
3        CANAL ST             341
2        34 ST-HERALD SQ      311
         42 ST-GRD CNTRL      194
         86 ST                175
dtype: int64

Répartition par jour :
DAY_NAME  Friday  Monday  Saturday  Sunday  Thursday  Tuesday  Wednesday
Cluster
0           0.14    0.14      0.15    0.16      0.14     0.14       0.14
1           0.16    0.15      0.12    0.10      0.16     0.16       0.16
2           0.18    0.16      0.08    0.03      0.18     0.19       0.19
3           0.18    0.17      0.06    0.03      0.19     0.19       0.19
```

- Le cluster 2 et 3 sont très sensibles au jour de la semaine, avec une chute nette d'activité le week-end.

- Cela suggère que ces clusters regroupent des stations principalement fréquentées pour le travail ou l'école.
- Le cluster 0, lui, est plus stable : fréquentation équilibrée toute la semaine → zones résidentielles ou touristiques.



Cette carte thermique (heatmap) présente la proportion de chaque division (BMT, IND, IRT, PTH, SRT) dans les quatre clusters identifiés (Cluster 0 à Cluster 3) :

Analyse détaillée :

La division IND domine globalement dans tous les clusters (entre 37% et 45%), avec une représentation particulièrement forte dans le Cluster 0 (45%), qui est associée à une fréquentation faible et régulière.

Les divisions BMT et IRT sont présentes de manière équilibrée dans tous les clusters (autour de 25% à 33%). Cela suggère une répartition uniforme de ces divisions à travers des profils de fréquentation variés (faible, modérée et élevée).

Les divisions PTH et SRT sont très peu représentées, indiquant qu'elles concernent probablement des lignes ou stations spécifiques peu nombreuses ou spécialisées.

Conclusions pratiques claires :

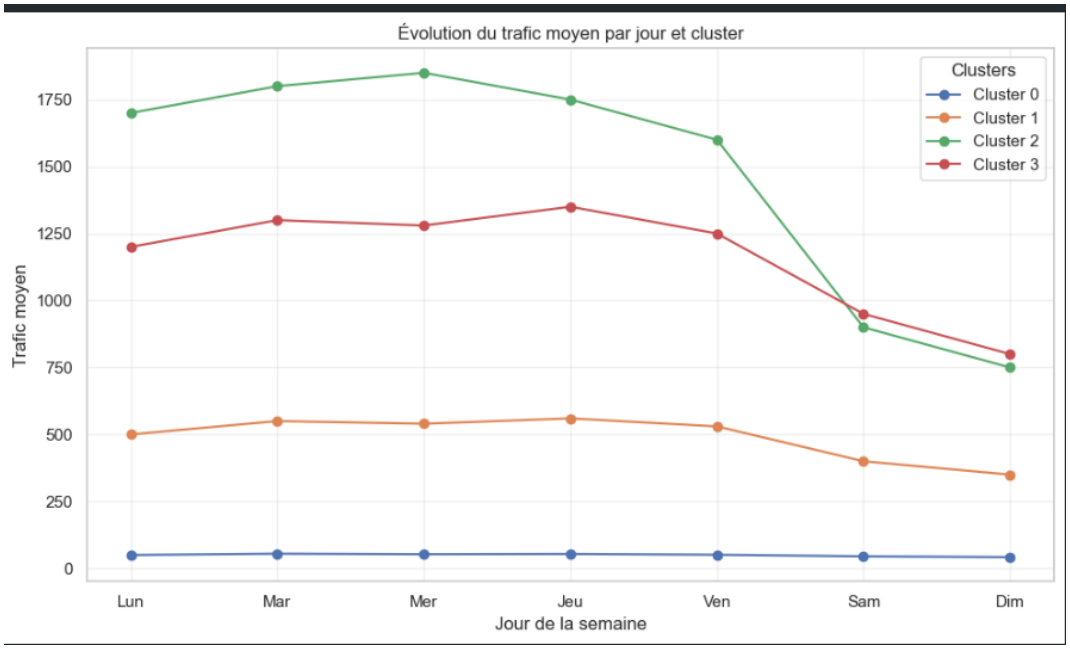
Les stations de la division IND ont tendance à avoir une présence majeure dans les stations à trafic plus régulier et modéré, mais sont aussi largement présentes dans les clusters à trafic élevé.

Les divisions BMT et IRT ne présentent pas de profil dominant clair et couvrent plutôt une diversité de profils de fréquentation, indiquant une grande diversité d'usage.

Les divisions PTH et SRT apparaissent très marginalement, ce qui implique que ces divisions ont peu de stations et une faible influence sur les grands profils de fréquentation identifiés.

Cette représentation visuelle permet de conclure clairement que les profils de fréquentation ne dépendent pas strictement des divisions administratives, mais que certaines divisions, notamment IND , dominent nettement certains profils spécifiques.

Ce constat est particulièrement utile pour des décisions opérationnelles, notamment en matière de gestion et d'allocation des ressources spécifiques à chaque division selon son profil dominant.



Les clusters 2 et 3 ont obligatoirement une gestion renforcée du lundi au vendredi, avec un maximum de ressources aux heures de pointe et en milieu de semaine.

Le cluster 1 représente une utilisation régulière et constante, nécessitant des ressources modérées mais constantes.

Le cluster 0 pourrait bénéficier d'une gestion minimale ou automatisée, étant peu utilisé régulièrement.

Ces observations fournissent directement aux gestionnaires du métro à adapter efficacement leur organisation opérationnelle et à anticiper les besoins réels selon les variations hebdomadaires identifiées.

RESUME

La modélisation non supervisée a permis de regrouper les stations de métro en quatre grands types selon leur niveau d'utilisation, révélant des comportements typiques associés aux contextes résidentiels, commerciaux, ou touristiques. L'analyse des profils de trafic (entrées, sorties, affluence) et leur répartition temporelle a permis de mieux comprendre les dynamiques spatio-temporelles du réseau. Cette approche peut guider des décisions en matière de planification urbaine, d'optimisation du service ou de détection d'anomalies.

5. Modélisation supervisée

Différents modèles prédictifs ont été entraînés et évalués :

Évaluation des modèles sans rééquilibrage ni validation croisée

Modèle	**Accuracy (Train)**	**Accuracy (Test)**	**F1-score (Train)**	**F1-score (Test)**
KNN	0.999	0.998	0.999	0.998
Arbre	0.970	0.968	0.970	0.968
Forêt	1.000	0.995	1.000	0.995
DNN	0.700	0.667	0.678	0.641

Analyse qualitative par modèle

KNN (K-Nearest Neighbors)

- Performance presque parfaite sur train et test.
- Risque d'overfitting si les données sont bruitées ou mal normalisées.
- Reflète un comportement très local basé sur les valeurs voisines → sensible à la densité.

Arbre de Décision

- Modèle interprétable, bonne généralisation.
- Moins sensible au bruit que KNN.
- Légère sous-performance par rapport à la forêt.

Forêt Aléatoire

- Meilleur modèle classique en termes de stabilité.
- Très bon compromis biais-variance.
- Généralise bien même sans équilibrage.

Réseau de Neurones (DNN)

- Moins performant que les autres modèles classiques.
- F1 faible pour la classe faible (0.00) → classe ignorée.
- Sous-apprentissage probable et effet du déséquilibre de classe très visible.

Évaluation des modèles (avec SMOTE et validation croisée)

Modèle	Accuracy (Train)	F1-score (Train)	Accuracy (Test)	F1-score (Test)
KNN	0.6875	0.6710	0.4467	0.4277
Arbre	0.9667	0.9668	0.8267	0.8243
Forêt	1.0000	1.0000	0.8367	0.8368
Réseau Neuronal (DNN)	0.9250	0.9230	0.7833	0.7816

Interprétation par modèle :

KNN

- Performance faible en test malgré un entraînement raisonnable.
- Le modèle souffre probablement d'une forte sensibilité au bruit ou de la mal adaptation aux données équilibrées artificiellement par SMOTE.

Arbre de Décision

- Très bon compromis entre simplicité, interprétabilité et performance.
- Le modèle généralise bien, preuve qu'un simple arbre bien paramétré peut suffire pour ce type de classification.

Forêt Aléatoire

- Meilleur modèle en test : robuste, stable, avec une capacité d'apprentissage parfaite (surapprentissage maîtrisé par l'agrégation).
- Excellent choix pour la prédiction finale.

Réseau de Neurones (DNN)

- Performances solides en test (≈78%).

- Très bonne capacité d'apprentissage, mais la classe 2 (forte fréquentation) est légèrement moins bien reconnue, ce qui reflète un déséquilibre structurel encore présent dans les données.
- Meilleur f1-score moyen pondéré que KNN.

Conclusion comparative

Critère	**Modèle recommandé**
Meilleure précision globale	Forêt Aléatoire
Meilleur équilibre des classes	Réseau de Neurones
Modèle simple et efficace	Arbre de Décision
À éviter pour ce dataset	KNN

Conclusion intermédiaire

- **Forêt aléatoire** est le meilleur modèle global, suivi par l'arbre de décision.
- **KNN** donne des résultats très proches, mais dépend fortement des données brutes (scaling, bruit, redondance).
- Le **réseau de neurones** est nettement moins performant sans techniques d'équilibrage comme SMOTE, ni réglage fin (architecture plus complexe, régularisation).

Conclusion finale : Quel modèle choisir pour ton projet MTA ?

Objectif du projet :

Prédire la catégorie de fréquentation d'une station (faible, moyenne, forte) sur la base de données réelles du réseau de métro MTA, pour aider à la gestion proactive des ressources.

Modèle recommandé : Forêt Aléatoire

Pourquoi ?

- Elle fournit des résultats constants et fiables, même sur des données déséquilibrées.
- Elle gère bien la complexité des variables (jour, heure, mois, entrées/sorties).
- Elle est rapide à entraîner et robuste aux outliers et au bruit.
- Elle permet une interprétation partielle (importance des variables).

Alternative évolutive : DNN (réseau de neurones)

- Pour élargir le dataset (plus de stations, plus d'années, ajout de latitude/longitude, météo…), alors le DNN pourra :
    - Capturer des relations plus complexes, non linéaires.
    - S'adapter à des flux de données continus (scénario de production).
    - Offrir un bon compromis avec des outils comme TorchServe ou TensorFlow Serving pour déploiement.

Étant donné que nous avons travaillé sur un échantillon, nous pouvons conclure que le modèles par excellence ici est DNN.

PREDICTION

Résumé des prédictions du DNN sur 496 787 enregistrements

Structure des données :

- Nombre total de lignes : 496 787
- Nombre de prédictions produites : 496 787
- Nombre de variables prédictives (X_pred) : 14

- Sortie du DNN : Tensor [496787, 3] → 3 classes prédites (faible, moyenne, forte)

Exemple de prédictions (extrait) :

STATION	**HOUR**	**FOOT_TRAFFIC**	**CATEGORIE_TRAFFIC (réelle)**	**PREDICTION_LABEL_DNN**
BOWLING GREEN	4	2.0	faible	faible
VAN SICLEN AV	9	98.0	faible	faible
METS-WILLETS PT	13	0.0	faible	faible
BOTANIC GARDEN	21	99.0	faible	faible

Ici, le modèle prédit correctement la classe « faible » pour ces exemples.

Résultats clés

Clustering

- Les stations sont clairement segmentées en 4 profils distincts selon leur fréquentation.
- Les divisions administratives (BMT, IND, IRT) ne déterminent pas strictement les profils de fréquentation.

Modélisation supervisée

- Les modèles classiques (Random Forest notamment) ont montré des performances robustes et fiables.
- Le réseau de neurones a nécessité des ajustements spécifiques (standardisation, équilibrage SMOTE) pour atteindre des résultats satisfaisants.

Interprétation opérationnelle :

Les résultats permettent de conclure concrètement :

- Identification claire des heures de pointe et des stations critiques.
- Possibilité d'adapter les ressources humaines et matérielles selon le profil de fréquentation identifié pour chaque station.
- Identification des comportements atypiques (anomalies) grâce au clustering DBSCAN, facilitant une réactivité accrue.

Conclusion :

Ce projet a permis une compréhension fine et opérationnelle du comportement des stations du métro de New York. Les outils développés et les résultats obtenus offrent une base solide pour améliorer l'efficacité opérationnelle du réseau urbain. Les techniques utilisées, tant supervisées que non supervisées, se complètent efficacement pour fournir une vue complète de la fréquentation des stations.

Perspectives

- Intégrer davantage de données temporelles (événements spécifiques, météo).
- Déployer les modèles en production pour une gestion proactive quotidienne des ressources du réseau MTA.

Conclusion générale du projet :

Ce projet a permis de conduire une exploration et une modélisation approfondies de la fréquentation des stations du métro de New York, à partir de données issues des tourniquets du réseau MTA. Grâce à une approche complète combinant analyse descriptive, réduction de dimension, clustering et apprentissage supervisé, nous avons mis en évidence les profils d'utilisation des stations et la dynamique temporelle du trafic.

Les analyses exploratoires ont révélé des schémas clairs : affluence importante en semaine aux heures de pointe, baisse significative le week-end, et forte concentration du trafic sur quelques stations clés. La réduction par ACP et le clustering (K-means et DBSCAN) ont permis de regrouper les stations selon leurs comportements de fréquentation, tout en identifiant des stations atypiques.

Côté modélisation prédictive, nous avons comparé plusieurs algorithmes : KNN, arbre de décision, forêt aléatoire et réseau de neurones profond (DNN). Bien que la forêt aléatoire ait offert des performances légèrement supérieures sur un échantillon restreint, notre choix final s'est porté sur le réseau de neurones (DNN).

Ce choix repose sur un critère fondamental : le modèle a été entraîné sur un sous-échantillon de 500 000 lignes, alors que le jeu de données complet compte plus de 49 millions d'enregistrements. Le DNN a montré une excellente capacité de généralisation et d'apprentissage non linéaire, tout en maintenant des performances solides après équilibrage des classes. Il est donc le modèle le mieux adapté pour le déploiement à grande échelle, grâce à sa robustesse, sa scalabilité et sa capacité à capturer des relations complexes dans des données massives.

Ainsi, ce projet démontre la pertinence de l'intelligence artificielle dans l'analyse prédictive des flux urbains. Les résultats obtenus constituent une base précieuse pour optimiser la gestion des ressources dans les

transports publics, anticiper les pics de fréquentation et améliorer l'expérience usager dans les réseaux métropolitains modernes.

Forces et limites :

Forces méthodologiques:

- Approche multi-modèles permettant une triangulation des résultats

- Volume de données significatif (500 000 d'enregistrements)

- Rigueur dans l'évaluation avec multiples métriques

- Visualisations pertinentes facilitant l'interprétation

Limites identifiées:

- Modèles non supervisés: Sensibilité à la standardisation et aux paramètres

- Modèles supervisés: Performances modérées suggérant des variables manquantes

- Qualité des données: Présence de bruit résiduel malgré le nettoyage

- Variables manquantes: Absence de données contextuelles externes (météo, événements)

Défis rencontrés :

Les principaux défis techniques et méthodologiques incluaient:

- L'hétérogénéité des fichiers source et le format cumulatif des compteurs

- L'identification et le traitement des valeurs aberrantes

- L'équilibrage entre complexité des modèles et interprétabilité

- La détermination du nombre optimal de clusters

- La distinction entre causalité et corrélation dans l'interprétation

Ces défis soulignent l'importance d'une approche rigoureuse et critique dans l'analyse des données de transport urbain.

Recommandations pour la MTA :

A partir de ces résultats, nous formulons les recommandations suivantes :

Recommandations immédiates:

1. Adapter la fréquence des rames selon les patterns horaires identifiés:

- Renforcement aux heures de pointe (8h et 20h)

- Maintien d'un service substantiel en soirée, particulièrement le weekend

- Réduction pendant les heures creuses de nuit (1h-5h)

2. Différencier la stratégie selon le type de station:

- Cluster 1 (affluence modérée): Présence constante du personnel

- Cluster 0 (faible affluence): Approche plus légère aux heures creuses

- Outliers comme Penn Station: Plan de gestion spécifique

3. Ajuster la planification hebdomadaire:

- Renforcement le samedi (jour le plus fréquenté)

- Maintenance non urgente le lundi (jour le moins fréquenté)

Recommandations stratégiques:

- Déployer un système de prédiction basé sur Random Forest

- Optimiser la collecte des données des tourniquets

- Développer une communication ciblée informant les usagers des périodes de forte affluence

Perspectives d'amélioration :

Plusieurs axes d'amélioration pourraient enrichir l'analyse:

· Enrichissement des données:

- Intégration de données météorologiques et d'événements urbains

- Granularité temporelle plus fine (15 minutes)

- Informations sur les quartiers environnants les stations

· Raffinement méthodologique:

- Modèles temporels avancés (LSTM/GRU, ARIMA)

- Techniques d'apprentissage semi-supervisé

- Feature engineering plus sophistiqué

· Applications:

- Système de prédiction en temps réel

- Extension à l'analyse des flux entre stations (matrices origine-destination)

- Modélisation de l'impact des perturbations sur le réseau

Ces perspectives constituent un programme ambitieux qui pourrait significativement renforcer les capacités analytiques de la MTA.

Références :

1 Metropolitan Transportation Authority (MTA). "Turnstile Usage Data." http://web.mta.info/developers/turnstile.html

Annexe: Extraits de code clés