

ANALYSE D'UN JEU DE DONNÉES

Introduction

Le jeu de données d'Airbnb analysé ici comprend 102 599 enregistrements représentant des annonces de locations dans diverses zones géographiques. Composé de 26 colonnes, ce jeu de données fournit une large variété d'informations : des caractéristiques générales de l'annonce aux spécificités géographiques, en passant par les attributs des logements et des données relatives aux réservations et aux évaluations. L'objectif principal de cette analyse est de segmenter ces annonces en groupes homogènes en utilisant des techniques de clustering, en prenant en compte des variables clés telles que le prix, la localisation et d'autres caractéristiques importantes pour la catégorisation.

Afin de déterminer le nombre optimal de clusters, nous appliquerons la méthode de la coudée ainsi que le score de silhouette, deux approches complémentaires permettant d'identifier le point optimal de séparation. Ce travail vise à obtenir une compréhension approfondie des différents segments de logements et de leur répartition géographique, avec un intérêt particulier pour l'analyse des prix et des caractéristiques propres à chaque type de logement.

Nous débuterons par une analyse préliminaire des données, puis nous approfondirons l'étude avec un jeu de données centré sur les utilisateurs, construit à partir des informations initiales.

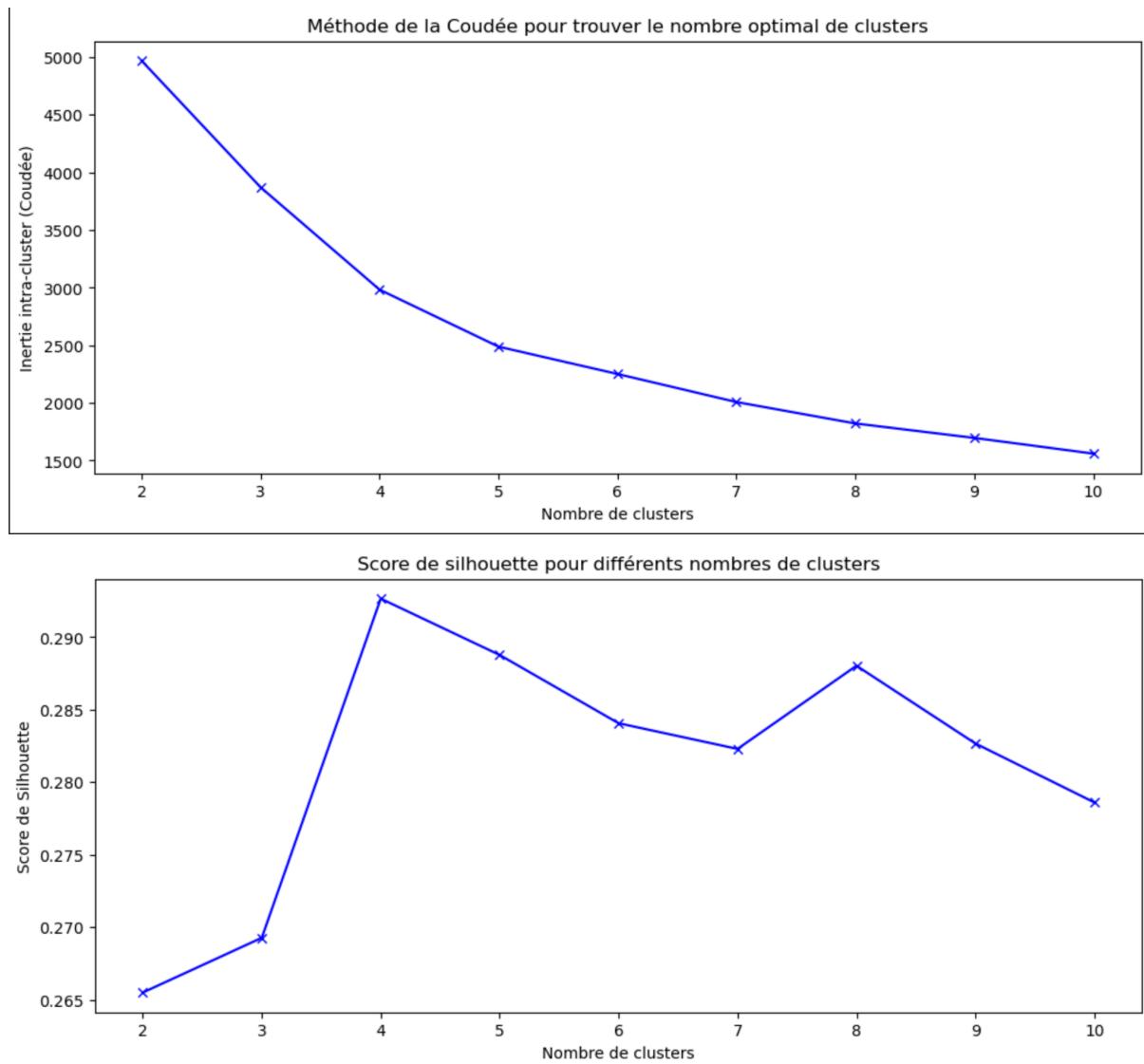
Analyse du jeu de données (Réalisée par Céline)

Pour commencer l'analyse, nous avons choisi de segmenter les logements en fonction du prix, en appliquant un clustering sur les différents types de logements. Les annonces ont été classées en quatre catégories principales : chambre privée (45 556 enregistrements), logement entier (53 701 enregistrements), chambre partagée (2 226 enregistrements) et chambre d'hôtel (116 enregistrements).

Avant de lancer l'algorithme de clustering k-means, il était essentiel de déterminer le nombre optimal de clusters. Pour cela, nous avons utilisé la méthode de la coudée (Elbow Method) couplée au score de silhouette. La méthode de la coudée nous offre une première estimation visuelle du nombre idéal de clusters en identifiant un "coude" dans la courbe d'inertie, tandis que le score de silhouette fournit une validation quantitative en mesurant la qualité de la séparation des clusters. Cette double approche nous permet ainsi de

sélectionner un nombre de clusters optimal, basé sur un équilibre entre une analyse visuelle et une mesure quantitative.

Pour “chambre partagée” :

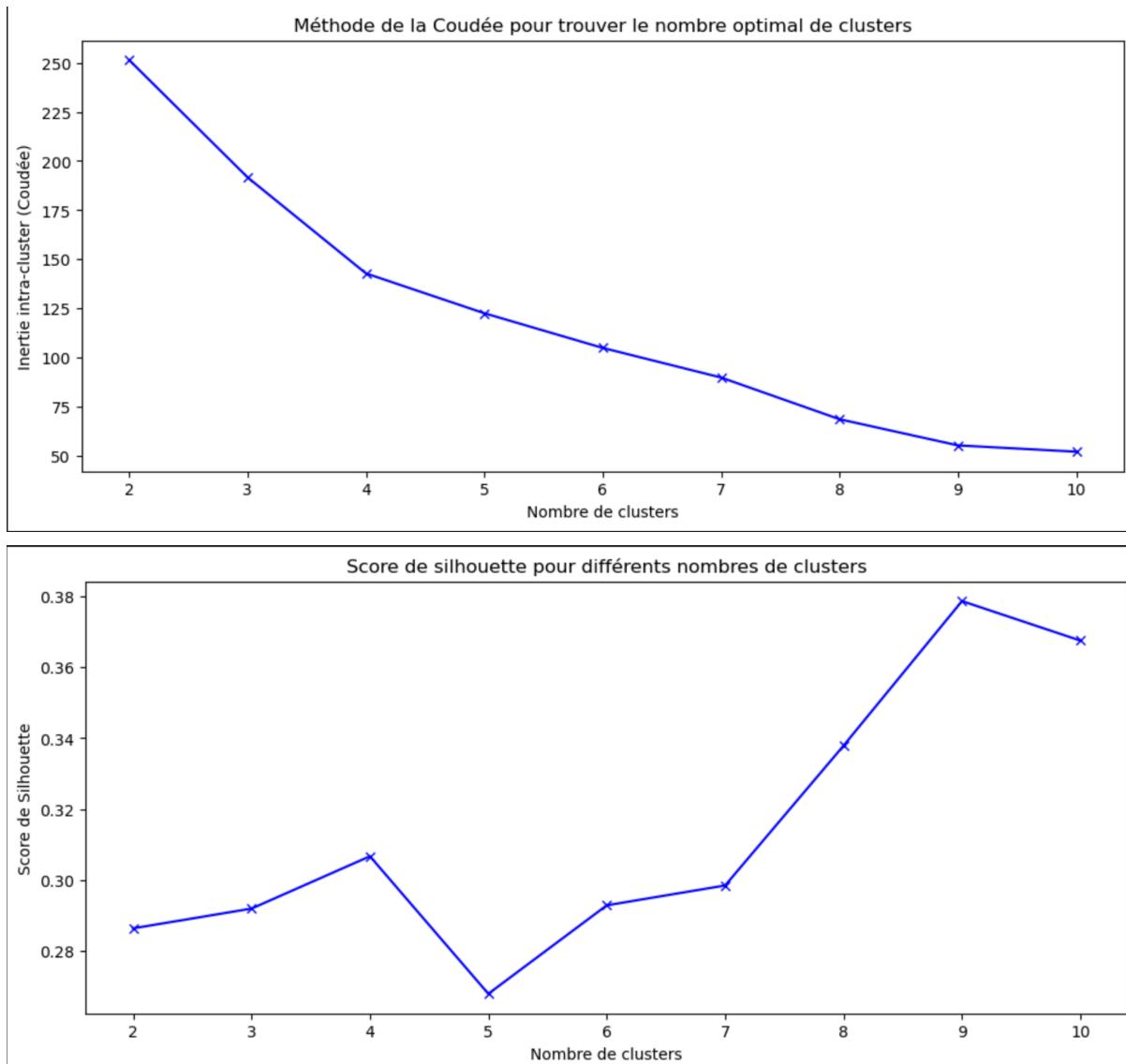


Sur le graphique représentant l'application de la méthode de la coudée, on observe que l'inertie diminue de façon importante pour les petits nombres de clusters (surtout entre 2 et 4), puis la baisse se ralentit progressivement au-delà de 4 clusters. La coudée semble donc se situer autour de 4 clusters. Cela pourrait indiquer que 4 est un bon choix pour le nombre de clusters, car après ce point, l'amélioration en inertie devient moindre.

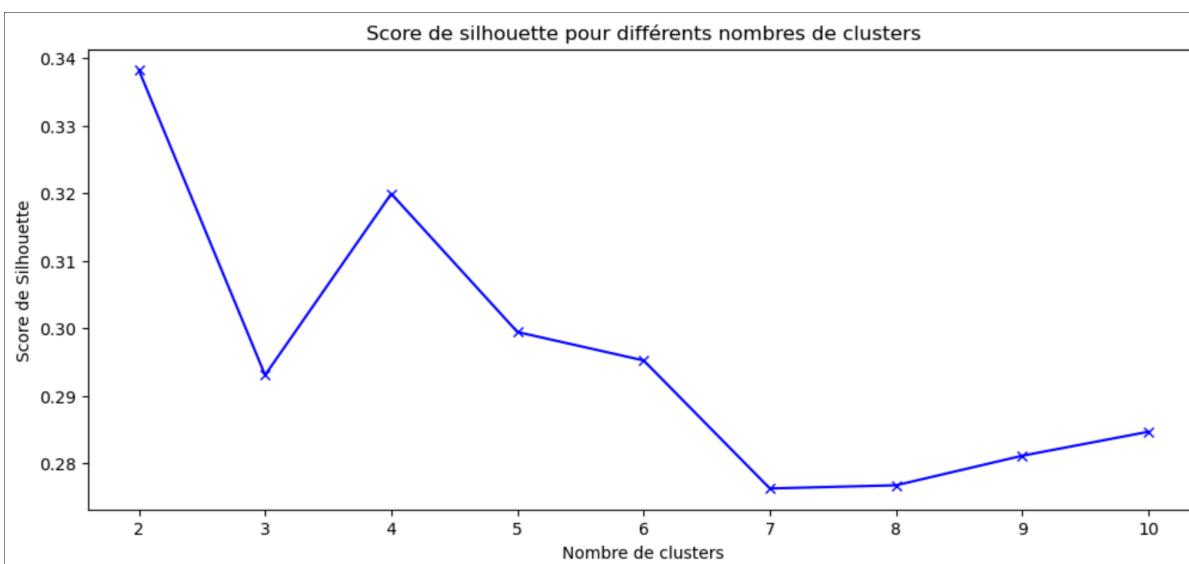
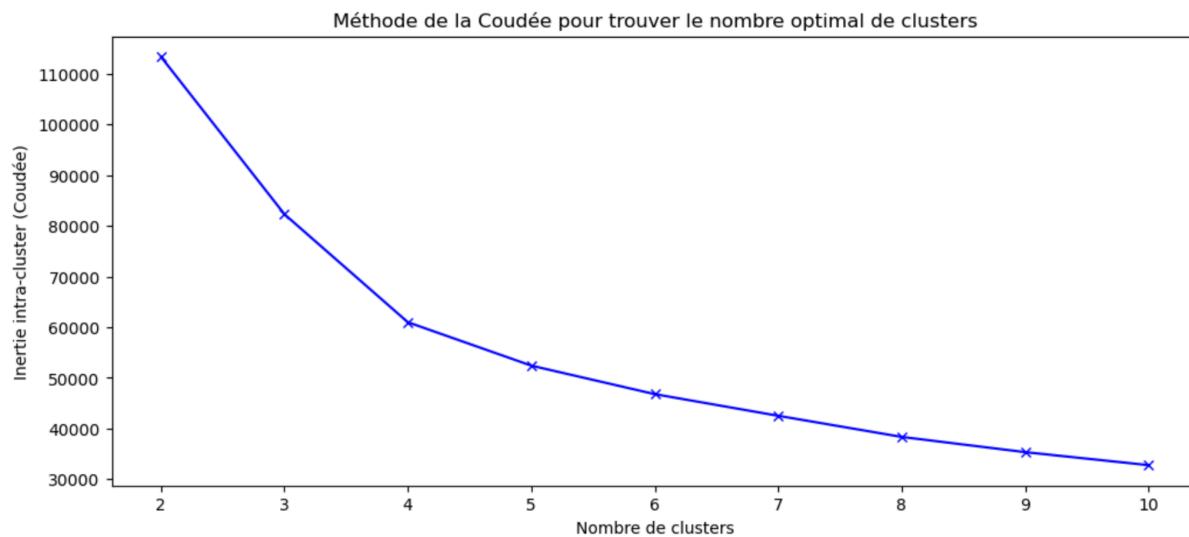
Puis, sur le deuxième graphique, on remarque un pic autour de 4 clusters, qui est le score le plus élevé. Cela renforce l'idée que 4 clusters est un choix optimal, car le score de silhouette est élevé à ce niveau, ce qui suggère que la qualité des clusters est meilleure avec 4 clusters.

Nous avons réalisé le même genre d'analyse pour les différents types de logements (chambre privée, logement entier et chambre d'hôtel).

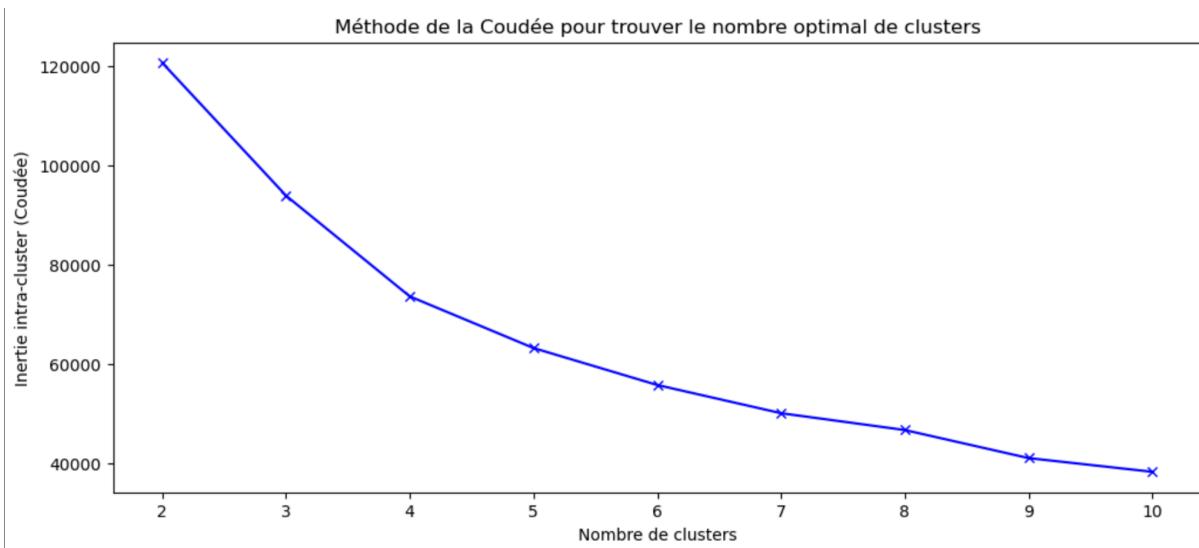
Pour “chambre d’hôtel” :

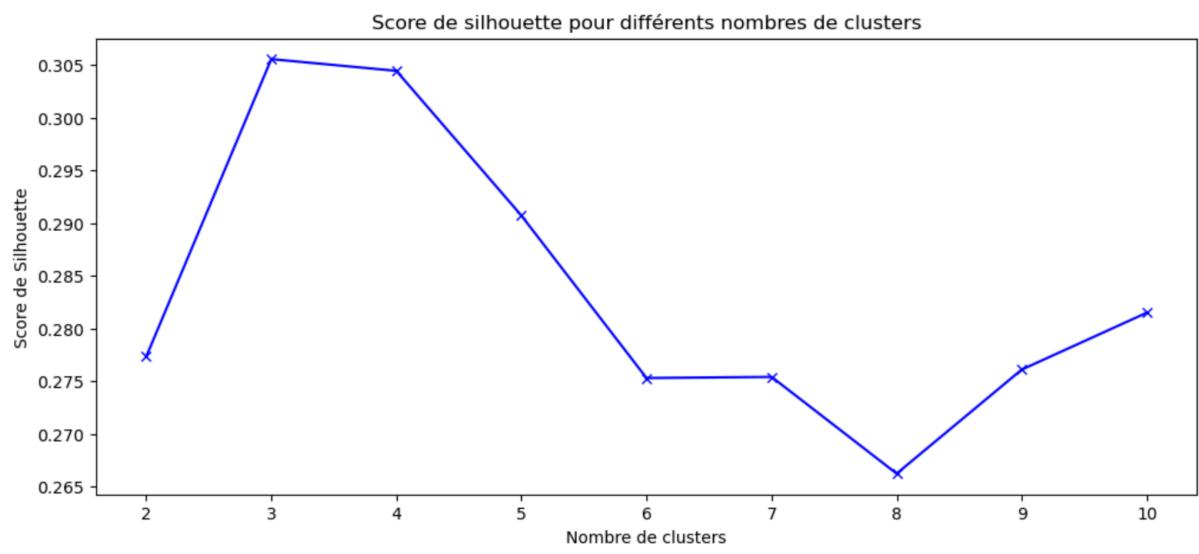


Pour “chambre privée” :

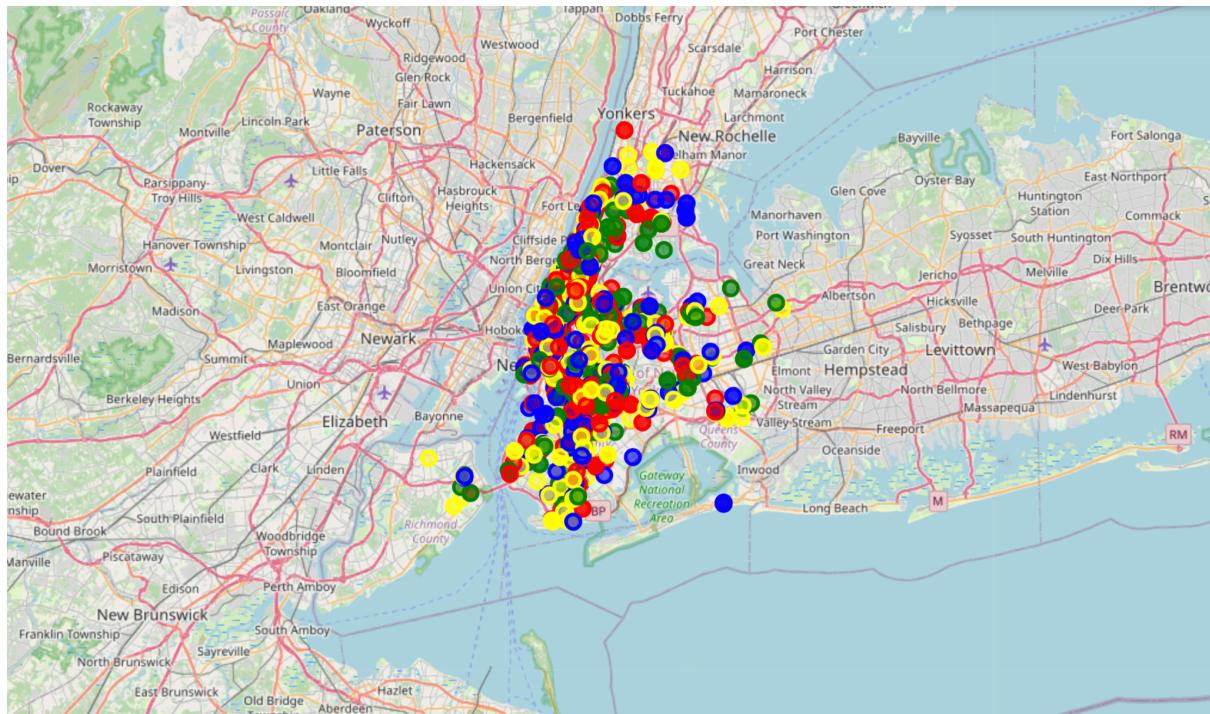


Pour “logement entier” :





Dans la suite de notre analyse, nous nous sommes concentrés sur le type "chambre partagée". Nous pourrions réaliser les mêmes analyses sur les autres types de logement. Nous avons continué en appliquant la méthode de clustering de k-means sur nos données en visualisant les clusters sur une carte Folium. **Notre questionnement principal étant de savoir si la localisation influence directement les prix des logements à New York ?**

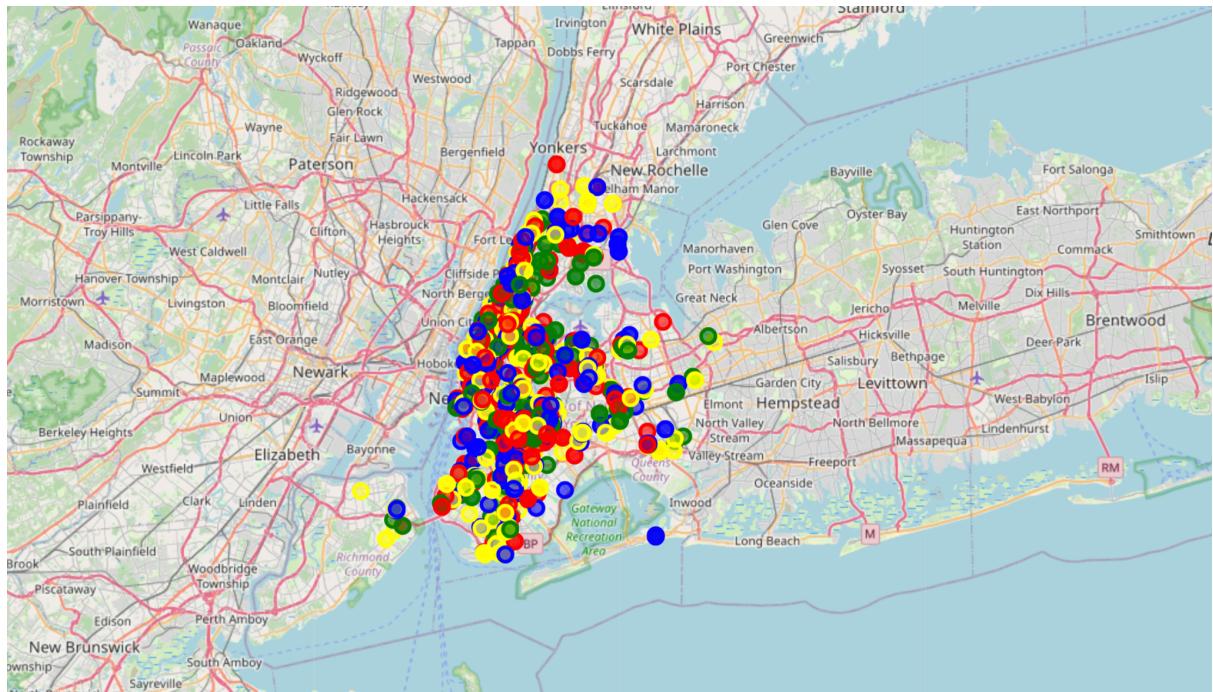


Application de l'algorithme Kmeans (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix

Les clusters obtenus ne semblent pas refléter fidèlement les caractéristiques des quartiers. Cela pourrait s'expliquer, en partie, par l'absence d'informations clés comme la superficie des logements, qui est probablement un facteur déterminant du prix. De plus, nous ne

disposons pas non plus de l'année de construction du logement, qui est une caractéristique qui pourrait également avoir une influence significative sur les prix.

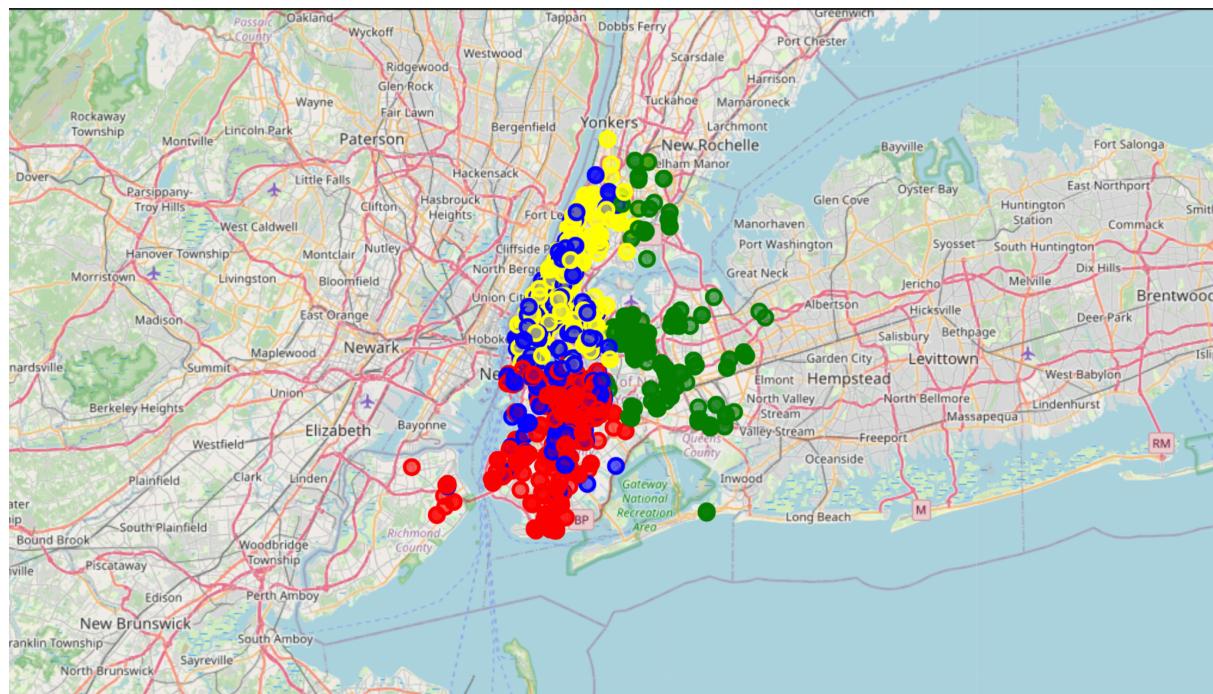
De plus, étant donné que les clusters basés uniquement sur le prix ne montraient pas de schémas spatiaux évidents, nous avons choisi d'ajouter les coordonnées géographiques (latitude et longitude) pour tenter de capturer des regroupements plus liés à des régions spécifiques. Cette approche vise à rendre les clusters plus pertinents d'un point de vue géographique.



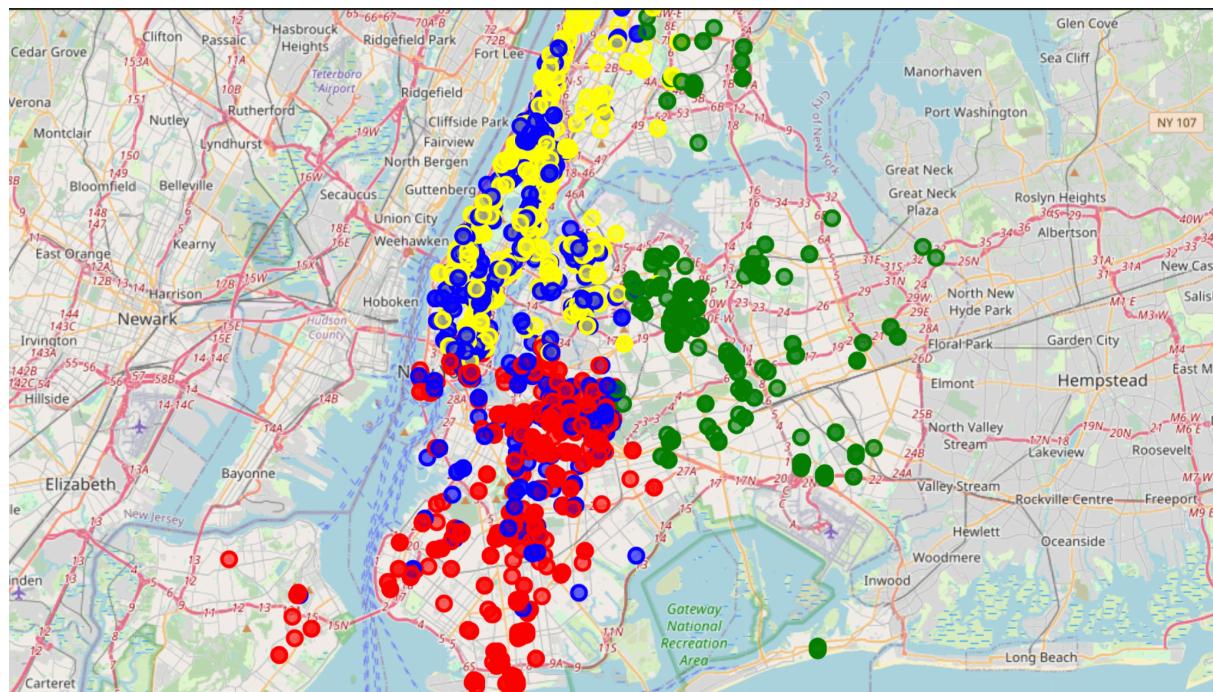
Application de l'algorithme Kmeans (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude et longitude

Les clusters semblent systématiquement être influencés par la variable *prix*, probablement en raison de l'échelle des différences entre les coordonnées géographiques (latitude et longitude), qui sont très faibles (de l'ordre de 0,01). Pour résoudre ce problème et rendre les différences géographiques plus significatives dans le calcul des distances, nous avons réalisé plusieurs ajustements. Premièrement, nous avons standardisé les variables (*prix*, *latitude* et *longitude*) à l'aide d'un *StandardScaler*. Cette approche permet de normaliser l'influence de chaque variable dans le modèle. Ainsi, aucune variable ne domine le calcul des distances, notamment le *prix*, qui risquait de biaiser le clustering. Cela favorise également la formation de clusters plus homogènes, incluant des dimensions géographiques et non seulement économiques. Deuxièmement, étant donné la faible amplitude des variations en latitude et longitude, nous avons appliqué un facteur d'échelle en multipliant ces deux variables par 1 000. Cette transformation rend les distances géographiques plus significatives sans modifier l'échelle de la variable *prix*. Cette méthode garantit un meilleur équilibre entre les dimensions géographiques et économiques lors de l'analyse des clusters.

Résultat après ces modifications :



Application de l'algorithme Kmeans (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude normalisée et longitude normalisée



Application de l'algorithme Kmeans (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude normalisée et longitude normalisée

Afin, de rendre notre analyse plus précise, nous avons calculé plusieurs caractéristiques :

- **Les valeurs moyennes et les écarts-types des prix par cluster** : Ces données permettent de vérifier si chaque cluster correspond à un certain niveau de prix (élevé, moyen, bas) et s'il reflète des différences significatives.
- **La répartition des clusters par borough** : Permet de voir si chaque cluster se concentre dans un borough spécifique ou est réparti sur plusieurs.
- **Centroïdes des clusters** : Ils indiquent la position centrale de chaque cluster et aident à vérifier leur correspondance avec les principales zones géographiques.
- **Le nombre de logements par cluster et par quartier** : Cette donnée permettrait de savoir si certains clusters sont concentrés dans certains quartiers spécifiques de New York, confirmant une correspondance avec les quartiers. Par exemple, un cluster qui couvre majoritairement le quartier de Williamsburg à Brooklyn pourrait être associé à des logements de prix moyen-élevé, tandis qu'un cluster couvrant la zone sud de Manhattan pourrait indiquer des logements très chers.
- **Les distances moyennes aux principaux points d'intérêt** (Times Square, Central Park, aéroports) : Ces distances aident à évaluer l'influence de la proximité à des lieux majeurs (Times Square, Central Park, aéroports) sur les clusters, en distinguant par exemple les zones centrales (prix élevés) des périphériques (prix plus bas).

Calcul des différentes caractéristiques mentionnées ci-dessus pour la catégorie "chambre partagée" avec 4 clusters :

Valeurs moyennes et écarts-types des prix par cluster :

cluster	price_mean	price_std
0	469.838141	260.112177
1	981.736034	142.557544
2	676.022388	322.576286
3	375.785246	203.930213

Centroïdes géographiques :

Cluster	Latitude	Longitude
0	40.6655371657508	-73.95707865137379
1	40.738991405168065	-73.95744986710083
2	40.73877682835821	-73.84280782089553
3	40.78393716396721	-73.95463577455737

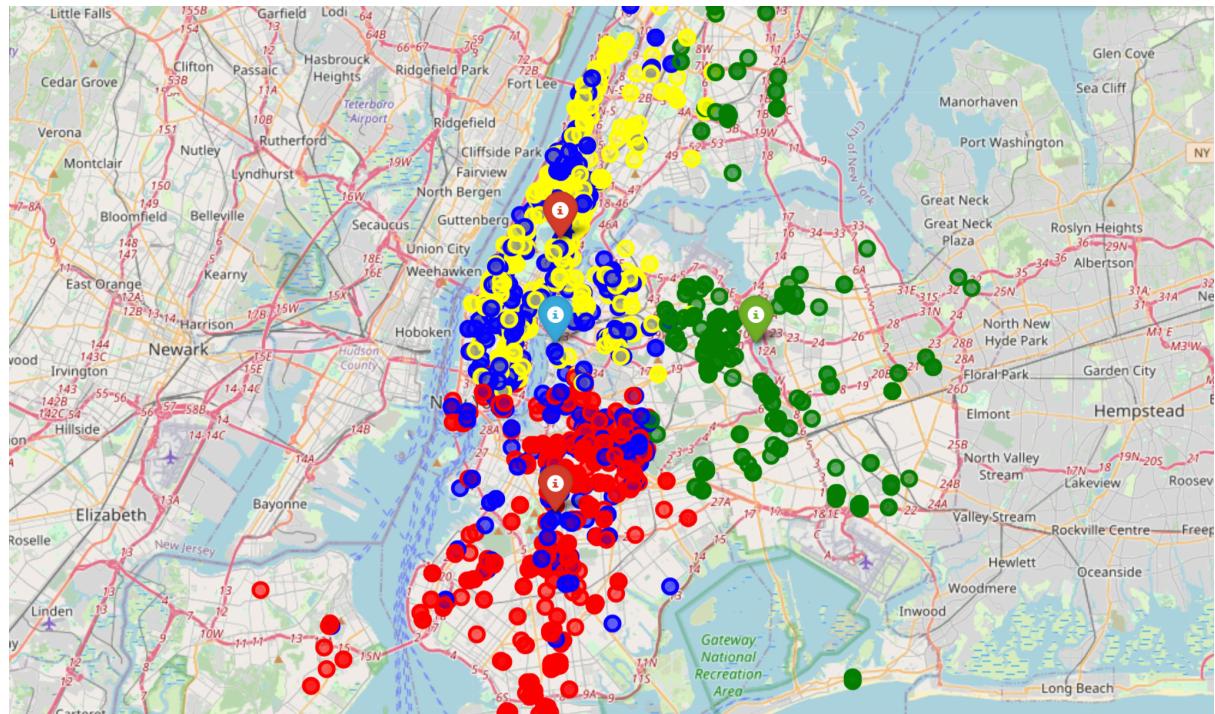
Distances moyennes aux points d'intérêt pour chaque cluster :

Cluster	Times Square	Central Park	JFK Airport
0	10.54 km	13.31 km	15.37 km
1	3.17 km	5.20 km	18.64 km
2	12.24 km	11.78 km	12.13 km
3	3.88 km	1.16 km	21.76 km

Répartition des clusters par quartier :

cluster	Bronx	Brooklyn	Manhattan	Queens	Staten Island
0	0	551	53	7	13
1	14	240	387	73	2
2	37	21	0	210	0
3	66	8	464	72	0

Nous avons également affiché les points d'intérêts sur la carte folium :



Application de l'algorithme Kmeans (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude normalisée et longitude normalisée. Affichage des points d'intérêts de New York

Analysons maintenant les résultats obtenus en traitant caractéristique par caractéristique pour chaque cluster.

Prix moyens et écarts-types par cluster

Cluster 0 : Prix élevé (\$469.83), grande dispersion (\$260.11), correspondant à des zones de Brooklyn et des parties abordables.

Cluster 1 : Prix très élevé (\$981.73), faible dispersion (\$142.55), probablement Manhattan.

Cluster 2 : Prix intermédiaire (\$676.02), grande dispersion (\$322.57), zones proches de Queens.

Cluster 3 : Prix bas (\$375.78), dispersion modérée (\$203.93), zones abordables de Manhattan ou Brooklyn.

Centroïdes géographiques

Cluster 0 (Brooklyn) : Les coordonnées sont cohérentes avec une localisation dans le sud-ouest de Brooklyn.

Cluster 1 (Manhattan, Midtown) : Correspond à une zone centrale, probablement Midtown Manhattan, où les prix sont les plus élevés.

Cluster 2 (Queens) : Situé dans l'est de Queens, suggérant un regroupement de logements proches des zones résidentielles ou plus éloignées.

Cluster 3 (Manhattan, Upper East) : Localisation proche de Central Park et de l'Upper East Side.

Distances moyennes aux points d'intérêt

Les clusters 1 et 3 sont les plus proches des lieux touristiques majeurs comme Times Square et Central Park, ce qui correspond aux logements chers et bien situés dans Manhattan.

Le cluster 2, bien que plus éloigné, est cohérent avec des logements situés dans Queens, souvent moins centraux.

Le cluster 0 est situé à des distances intermédiaires, reflétant des logements à Brooklyn, souvent bien desservis mais éloignés des attractions principales.

Répartition des clusters par quartier

Clusters 1 et 3 : Majoritairement Manhattan, selon les prix.

Cluster 0 : Majoritairement Brooklyn, prix intermédiaires.

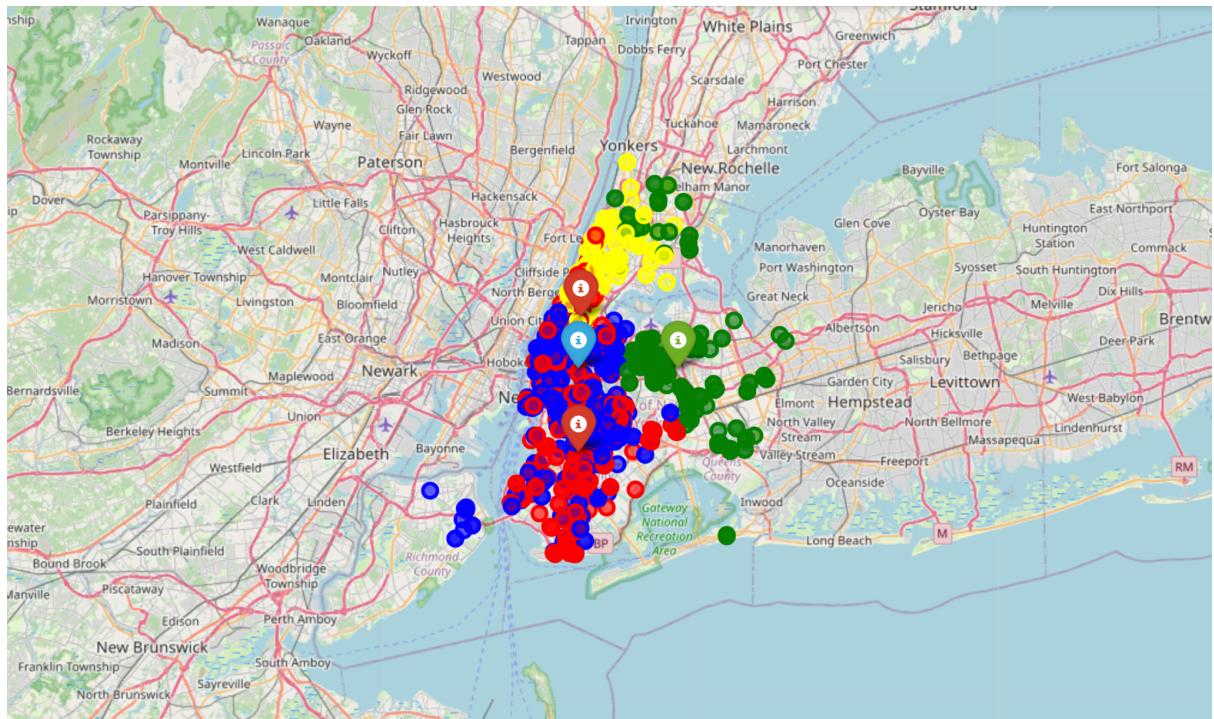
Cluster 2 : Principalement Queens, zone résidentielle à prix moyens.

Staten Island : Faible représentation.

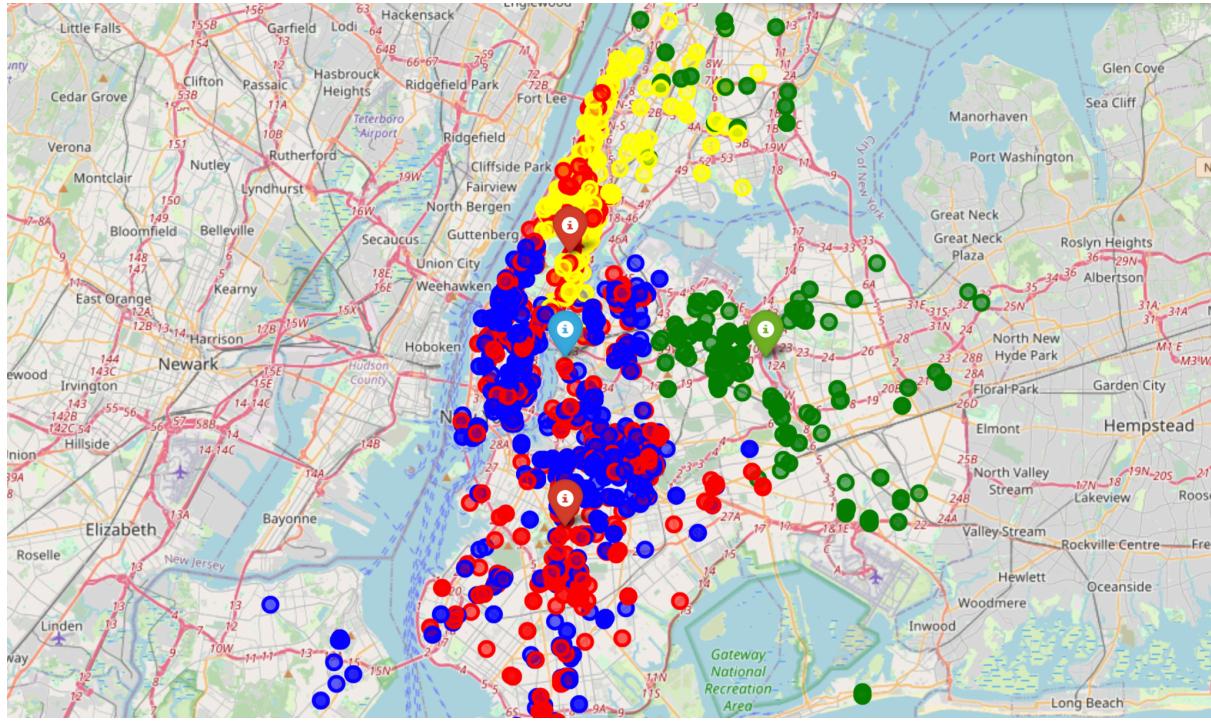
Ainsi, nous avons pu remarquer que nos clusters semblent bien refléter les quartiers de New York. Pour résumer :

- **Cluster rouge (Cluster 0)** : Concentre des logements à Brooklyn et dans certaines parties du Bronx, correspondant à des prix moyens.
- **Cluster jaune et bleu (Clusters 1 et 3)** : Représentent Manhattan, avec une distinction probable entre Midtown et l'Upper East Side.
- **Cluster vert (Cluster 2)** : Bien défini dans Queens, avec quelques points épars dans les autres boroughs.

Nous avons également utilisé l'algorithme Agglomerative Clustering :



Application de l'algorithme Agglomerative Clustering (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude normalisée et longitude normalisée. Affichage des points d'intérêts de New York



Application de l'algorithme Agglomerative Clustering (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude normalisée et longitude normalisée. Affichage des points d'intérêts de New York

En réalisant le même genre d'analyse que précédemment (caractéristique par caractéristique pour chaque cluster), nous obtenons le résultat global suivant :

- **Cluster 0** (cluster rouge) : Logements haut de gamme situés dans des quartiers périphériques ou exclusifs.
- **Cluster 1 et 3** (cluster bleu et jaune) : Logements proches de Manhattan, abordables, et destinés à des touristes ou visiteurs cherchant une localisation centrale.
- **Cluster 2** (cluster vert) : Mélange de logements intermédiaires à variés, situés dans des quartiers résidentiels éloignés comme le Queens ou le Bronx.

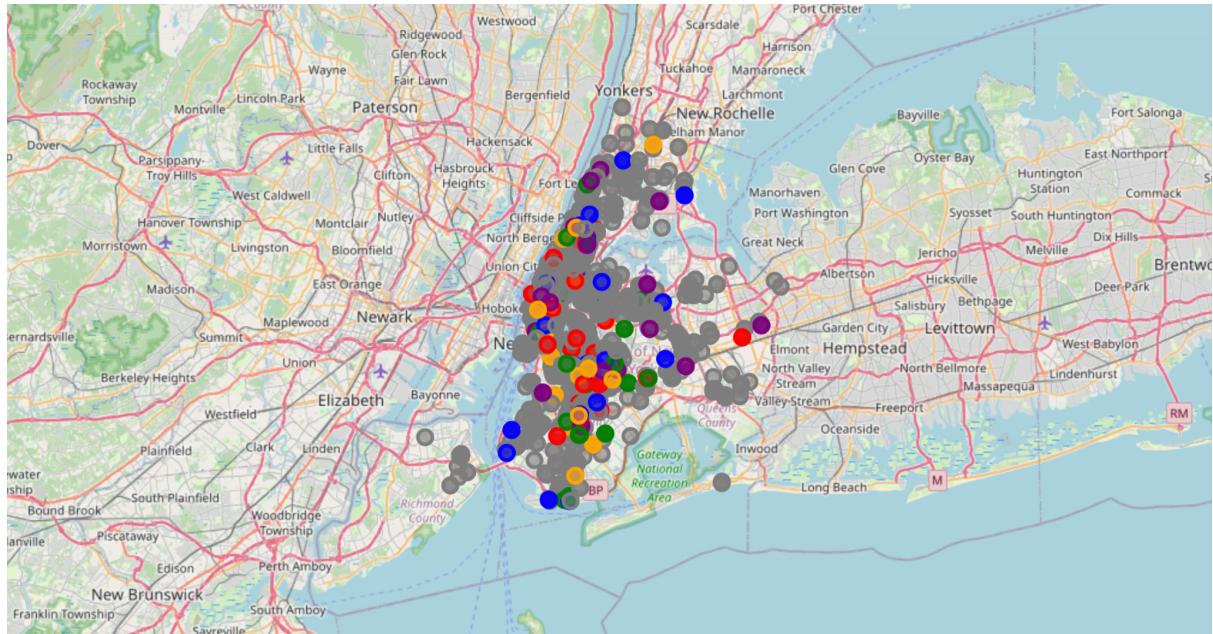
Ainsi, comme pour l'application de l'algorithme Kmeans, les clusters semblent refléter les spécificités géographiques et socio-économiques de New York.

Les analyses basées sur Agglomerative Clustering et K-means produisent des résultats différents. Dans l'analyse 2, le cluster 0 (rouge) est vu comme haut de gamme, tandis que dans l'analyse 1, il correspond à Brooklyn avec des prix moyens. Les clusters 1 et 3 (jaune et bleu) sont abordables et proches de Manhattan dans l'analyse 2, sans distinction précise, alors que l'analyse 1 différencie Midtown (jaune) et l'Upper East Side (bleu). Enfin, le cluster 2 (vert) est hétérogène en prix et zones dans l'analyse 2, mais principalement localisé dans le Queens dans l'analyse 1.

Ces différences s'expliquent par les algorithmes utilisés. Agglomerative Clustering (analyse 2) privilégie les distances locales, regroupant des zones éloignées mais similaires en prix, avec des écarts-types élevés. K-means (analyse 1) produit des clusters plus compacts et

homogènes, optimisant la variance intra-cluster et offrant une meilleure cohérence géographique.

Nous avons également réalisé une analyse en utilisant l'algorithme DBSCAN :



Application de l'algorithme DBSCAN (4 clusters) pour les logements de type "chambre partagée" en prenant en compte comme features : prix, latitude normalisée et longitude normalisée.

Nous obtenons les résultats suivants :

cluster 1 : ce cluster couvre des logements situés au cœur de Manhattan (Midtown, Hell's Kitchen).

cluster 2 : ce cluster est probablement situé dans l'Upper East Side ou l'Upper West Side.

cluster 3 : ce cluster semble inclure des logements situés dans des zones moins centrales comme le Queens ou le Bronx.

Nous avons considéré que l'utilisation de l'algorithme DBSCAN n'était pas optimal dans notre cas pour plusieurs raisons :

- La continuité géographique n'est pas garantie : des logements éloignés mais aux prix similaires peuvent être regroupés, ce qui ne reflète pas la réalité des quartiers.
- Choix des paramètres complexe : choix sensible des paramètres eps et min_samples

A l'inverse, K-means était facile à paramétrier en ayant trouvé le nombre de clusters préalablement avec la méthode du coude, définit des clusters homogènes et ces centroïdes permettent d'obtenir une synthèse des caractéristiques moyennes (prix et localisation).

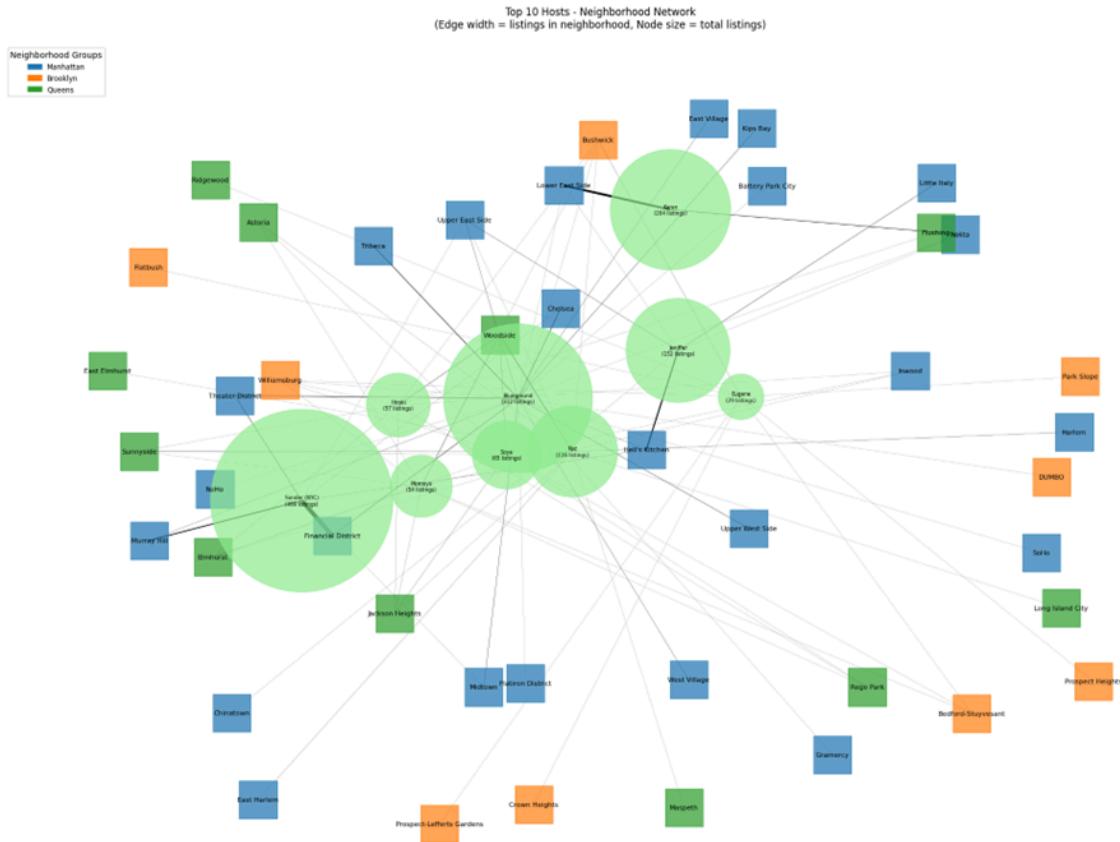
Mais encore, l'algorithme Agglomerative Clustering permet d'obtenir des clusters qui reflètent des regroupements réalistes et contigus. De plus, en utilisant une distance géographique, l'algorithme génère des clusters naturels adaptés à l'analyse des quartiers.

Analyse des dynamiques et des inégalités dans le réseau Airbnb à New York : une approche par analyse de réseau (Réalisé par Nathan)

L'utilisation d'une approche par réseaux permet d'analyser les relations entre les hôtes, les quartiers et les annonces Airbnb à New York. En représentant ces éléments sous forme de graphes, on peut identifier les acteurs les plus influents et comprendre comment l'activité est distribuée à travers la ville. Cette méthode permet de visualiser la structure du marché, de repérer les zones les plus connectées et de mettre en évidence les inégalités dans la répartition des annonces, ce qui est essentiel pour comprendre l'impact d'Airbnb sur l'économie locale et le marché du logement.

La centralité de degré, qui mesure le nombre de connexions d'un nœud dans le réseau, reflète l'influence directe des hôtes et des quartiers. Blueground, un acteur majeur sur le marché d'Airbnb, se distingue comme l'hôte ayant la centralité de degré la plus élevée, avec un score de 0,404. Cela indique que Blueground a établi de nombreuses connexions avec divers quartiers, illustrant ainsi sa position dominante dans la gestion des annonces à l'échelle de la ville. Kaz, Soya, Momoyo et Hiroki suivent, bien que leur influence soit moins marquée. Ces hôtes reflètent une concentration disproportionnée de l'activité du marché entre quelques opérateurs qui gèrent des annonces à grande échelle.

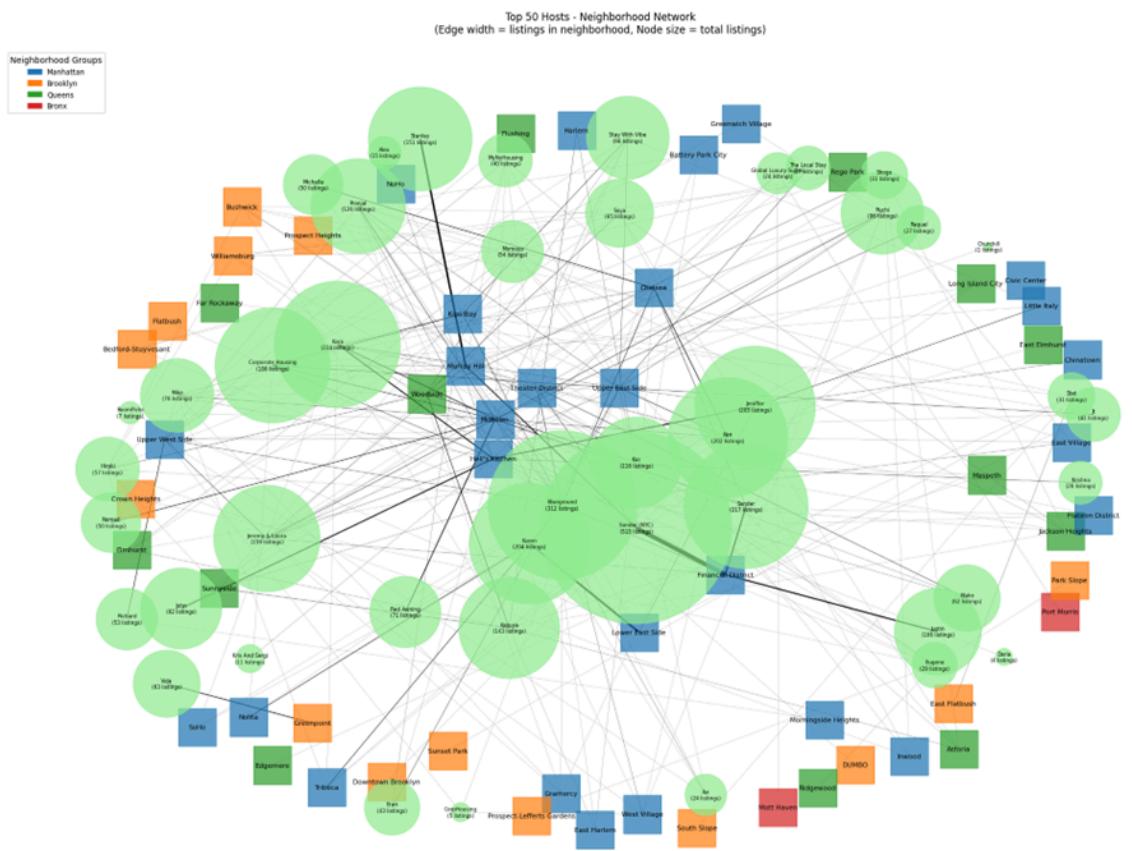
Le graphique ci-dessous montre les 10 hôtes les plus influents, soulignant l'ampleur de la concentration de l'activité entre les principaux acteurs. Blueground apparaît clairement comme le leader du marché, suivi de quelques hôtes de moindre envergure mais néanmoins significatifs.



Du côté des quartiers, Mid Town, Hell's Kitchen et l'Upper East Side se démarquent comme les quartiers les plus connectés. Mid Town, Hell's Kitchen, connus pour leur ambiance culturelle et leur proximité avec Manhattan, attirent fortement à la fois les hôtes et les touristes, renforçant leur centralité dans le réseau. De même, la centralité de l'Upper East Side reflète son attrait en raison de son caractère aisément accessible et de son accès à des attractions emblématiques comme Central Park. C'est d'ailleurs la raison pour laquelle les prix sont généralement plus élevés dans le quartier de Manhattan. Les quartiers périphériques comme Ridgewood et East Harlem montrent une centralité de degré nettement inférieure, mettant en lumière une répartition spatiale inégale des annonces Airbnb.

La centralité vectorielle met en lumière les nœuds qui sont bien connectés à d'autres nœuds influents, soulignant la qualité plutôt que la quantité des connexions. De manière intéressante, des hôtes de moindre envergure comme Soya et Momoyo atteignent les scores de centralité vectorielle les plus élevés, surpassant de grands opérateurs tels que Blueground. Cela indique que ces hôtes, bien qu'ils ne dominent pas en termes de connexions directes, sont intégrés dans des clusters de quartiers très influents. Leur influence provient de leur positionnement stratégique dans le réseau, plutôt que de leur taille globale.

Le graphique ci-dessous présente les 50 premiers hôtes selon la centralité de degré. Ce graphique montre que la concentration de l'activité se poursuit bien au-delà des 10 premiers hôtes, mais avec une répartition plus étendue.



Après avoir examiné le deuxième graphique montrant les 50 hôtes les plus influents, il apparaît clairement que les acteurs les plus puissants du marché sont majoritairement concentrés dans les zones les plus riches de la ville, notamment à Manhattan. Ces hôtes, comme Blueground, possèdent un grand nombre d'annonces dans des quartiers prisés tels que l'Upper East Side et Financial District, où la demande touristique et locative est forte. En revanche, on remarque une faible présence de ces grands acteurs dans des quartiers moins aisés comme le Bronx. Cette concentration d'influence dans les zones plus huppées renforce les inégalités géographiques du marché Airbnb, avec des conséquences directes sur la disponibilité des logements et l'accès au marché locatif pour les résidents des quartiers moins privilégiés.

Ces conclusions révèlent plusieurs tendances générales et implications pour Airbnb à New York. Une observation clé est la concentration de l'influence entre un petit groupe d'hôtes. Des opérateurs comme Blueground dominent le marché, gérant des centaines d'annonces dans plusieurs quartiers. Ce niveau de concentration soulève des préoccupations concernant l'équité et la concurrence, les petits hôtes indépendants ayant du mal à rivaliser. Il souligne également la nécessité de cadres réglementaires pour empêcher la monopolisation du marché et garantir des conditions de concurrence équitables.

La domination de certains quartiers et hôtes a des implications significatives pour le marché du logement à New York. Les hôtes qui gèrent un volume élevé d'annonces réduisent la disponibilité des logements à long terme, contribuant à l'augmentation des loyers et aux pénuries de logements dans les quartiers très demandés. Cette dynamique affecte particulièrement les zones centrales, où les marchés immobiliers sont déjà sous tension. À l'inverse, l'activité Airbnb limitée dans les quartiers périphériques met en lumière des opportunités manquées de diversification économique et de croissance liée au tourisme.

Pour relever ces défis, une approche multifacette est nécessaire. Des mesures réglementaires, comme des plafonds sur le nombre d'annonces qu'un hôte peut gérer, pourraient freiner la concentration du marché et promouvoir une concurrence équitable. Les décideurs pourraient également envisager des restrictions de zonage pour limiter les locations de courte durée dans les zones résidentielles, en particulier dans les quartiers où la demande en logement est élevée. Pour encourager une répartition plus équitable de l'activité Airbnb, des incitations pourraient être offertes aux hôtes qui répertorient des propriétés dans des quartiers moins représentés. Ces incitations pourraient prendre la forme de réductions de frais ou d'avantages fiscaux, contribuant à équilibrer la structure du réseau et à réduire sa dépendance envers les acteurs dominants.

La promotion de la diversité parmi les hôtes est une autre stratégie essentielle. Soutenir les hôtes indépendants et à petite échelle pourrait favoriser un réseau plus décentralisé, réduisant l'influence des grands opérateurs comme Blueground. Encourager la diversité parmi les hôtes renforcerait non seulement la concurrence, mais créerait également plus d'opportunités pour les individus et les familles de participer au marché des locations de courte durée.

Enfin, des efforts pour étendre l'activité Airbnb aux quartiers moins représentés pourraient bénéficier à l'ensemble du réseau. Des campagnes de marketing mettant en avant les attractions uniques de ces quartiers pourraient attirer à la fois des touristes et des hôtes. Des politiques incitant les hôtes à lister des propriétés dans des zones périphériques pourraient promouvoir un écosystème Airbnb plus équilibré, tout en répartissant plus équitablement les bénéfices économiques du tourisme à travers la ville.

En conclusion, cette analyse du réseau d'Airbnb à New York met en lumière l'influence significative d'un petit nombre d'hôtes et de quartiers. Des acteurs comme Blueground dominent le marché, tandis que des quartiers comme Hell's Kitchen et Midtown servent de hubs centraux. La répartition inégale de l'activité Airbnb soulève des préoccupations en matière d'équité, d'accessibilité au logement et de dynamique des quartiers. Des mesures réglementaires, un soutien aux hôtes indépendants et des efforts pour diversifier l'activité Airbnb à travers les quartiers pourraient aider à relever ces défis et à créer un marché plus équitable et durable. En examinant les relations complexes entre les hôtes et les quartiers, cette analyse fournit une base pour des politiques éclairées et des recherches futures sur les impacts socio-économiques d'Airbnb en milieu urbain.