Rapport sur les données spaciales d'Airbnb

Mars 2023

Par :

Awa Syr DIAGNE

&

Lama ZIDAN

Professeur:

Julien Ah-Pine

Table des matières

Rapport sur l’analyse des données spatiales de la plateforme d’AirBnB sur les logements de Lyon

Introduction

Nous disposons d’un jeu de données fourni par la plateforme d’AirBnb concernant les locations de logement de particuliers dans la ville de Lyon. Notre étude portera sur la variable review_scores_rating qui représentent les notes données par les clients après leur séjour. Nous allons tout d’abord commencer par présenter le contexte de notre étude en élaborant nos hypothèses, puis nous développerons les méthodes statistiques qui nous permettront de mener à bien notre analyse et enfin, nous présenterons notre modèle de prédictions utilisé.

I. Présentation du contexte

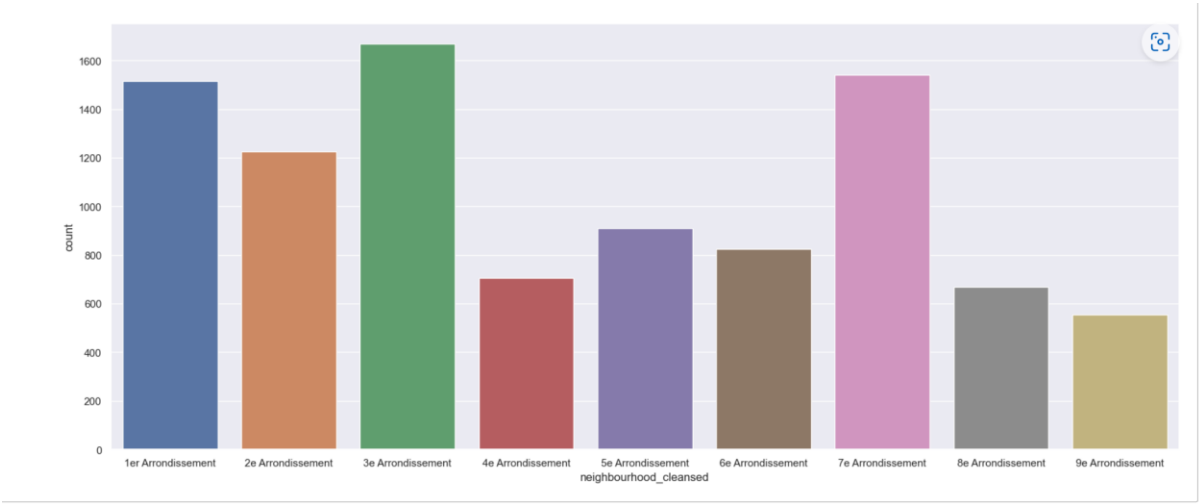
Notre étude va s'intéresser à savoir quels sont les arrondissements à Lyon qui ont un review_scores_rating de biens élevé et quels sont les arrondissements à Lyon qui ont un review_scores_rating de bien faible.
Notre étude portera essentiellement à savoir quelles sont les zones (arrondissements), où les notes obtenues sont bien élevées et celles qui ne le sont pas.
Pour ce faire, nous émettons deux hypothèses : H0 et H1 où H0, notre hypothèse de départ représente l’hypothèse selon laquelle les observations sont totalement aléatoires et qu'il n'y a pas d’autocorrélation spatiale entre les arrondissements lyonnais et H1, le contraire.

1. Présentation générale des données

Tout d’abord, nous disposons d’un jeu de données à hauteur d’une dimension de 9614 lignes et 75 observations.

```
Entrée [6]: data['neighbourhood_cleansed'].value_counts()

Out[6]: 3e Arrondissement     1668
        7e Arrondissement     1541
        1er Arrondissement    1516
        2e Arrondissement     1225
        5e Arrondissement      911
        6e Arrondissement      826
        4e Arrondissement      704
        8e Arrondissement      669
        9e Arrondissement      554
        Name: neighbourhood_cleansed, dtype: int64
```



Ci-dessus se trouvent tous les logements répartis en arrondissements.
Nous pouvons clairement deviner que les logements des particuliers les plus prisées se situent dans le troisième arrondissement lyonnais. Nous avons 1668 logements. Ce qui peut s'expliquer car ce quartier est très attractif. Il y a le plus grand centre commercial de France et même un des plus grands d'Europe, nous avons la gare également.

```
Entrée [7]: data['room_type'].value_counts()

Out[7]: Entire home/apt    7413
        Private room       2112
        Shared room          63
        Hotel room           26
        Name: room_type, dtype: int64
```

Pareil, la plupart des logements prisés sont des maisons et des appartements.
Ils ont pour la plupart des logements avoisinant les 50$, ce qui est l'équivalent d'environ 47 €. De ce fait, ils attirent tous types de clients par son accessibilité.

Cependant, ces loyers peuvent aller jusqu'à 993$ soit 936€ même si elles sont rares. Ces biens peuvent servir d'endroits pour des événements ou autres.

Tous ces biens répertoriés ont reçu des notes de la part de leurs clients après chaque visite. Les avis sont pour la plupart très favorables. 2030 logements ont reçu l'excellente note 5/5. Ce qui révèle que la majorité des clients ont été très séduite et très satisfaite de leur passage dans ces biens.

Pour une meilleure approche et éviter une analyse biaisée, il va falloir procéder à un prétraitement.

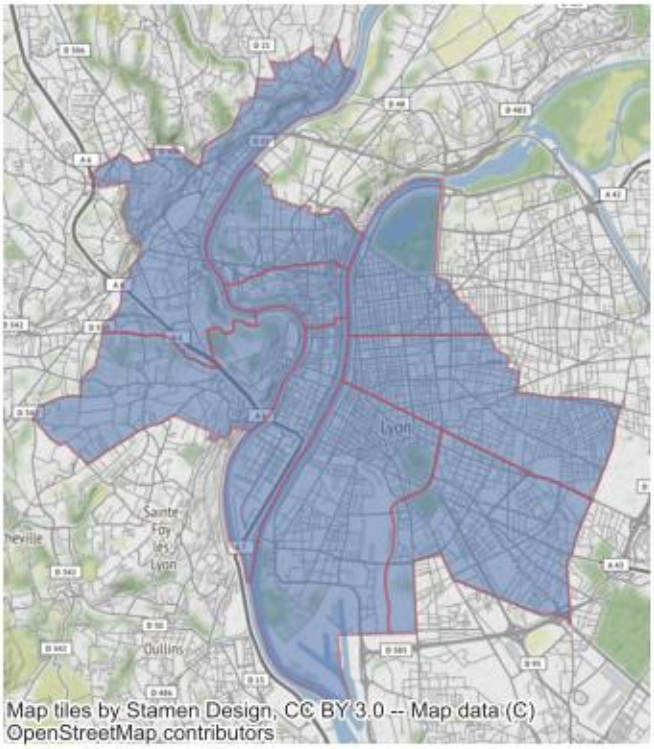
2. Prétraitement des données

Nous observons que nous n'avons pas des valeurs doubles, et que à la base nous avons 88550 valeurs manquantes en totale. Nous allons sélectionner les variables qui nous intéressent pour réaliser notre étude, et nous allons choisir les variables suivantes :
neighbourhood_cleansed","latitude","longitude","accommodates","price","review_scores_rating","number_of_reviews","calculated_host_listings_count","review_scores_communication","review_scores_cleanliness","review_scores_location","minimum_nights","maximum_nigh"
Notre nouveau dataset contient désormais 9614 observations et 13 variables. Nous remarquons que nous avons des valeurs manquantes présentées dans les variables review_scores_rating, review_scores_cleanliness, review_scores_communication, et review_scores_location, pour cela, nous allons les remplacer par la médiane de la colonne. Nous pensons que cette méthode de remplacement est mieux que de supprimer les NA, puisque le nombre des NA's et importants, et la suppression des NA's pourra biaisé nos données.

Ensuite, nous passons à ajouter des variables jugées importantes pour notre étude, en ajoutant une variable qui stock le prix sans le dollar « price_int ». Ensuite, créer une variable « geometry » à partir des deux colonnes « longitude » et « latitude », et obtenir alors notre dataset en format Geo data frame.
En dernière étape, nous pensons à ajouter une variable « insécurité », cette variable binaire est égale à 1 si le bien est dans un quartier sensible de Lyon et 0 sinon. Cette variable va être utilisé comme hypothèse de la prédiction dans la dernière partie de ce projet.
Nous allons supposer que le fait que le bien soit situé dans un quartier sensible ou dangereux va jouer dans la note de ce bien, c'est à dire, si le bien est dans un quartier sensible ou dangereux cela va conduite à une rate faible pour ce bien et vice versa.

3. Le dataset des 9 arrondissements de Lyon

La ville de Lyon a été représentée en polygone. Ce dernier est une surface qui a au moins une LineString qui débute et se termine au même point. Chaque point est relié à l'autre.

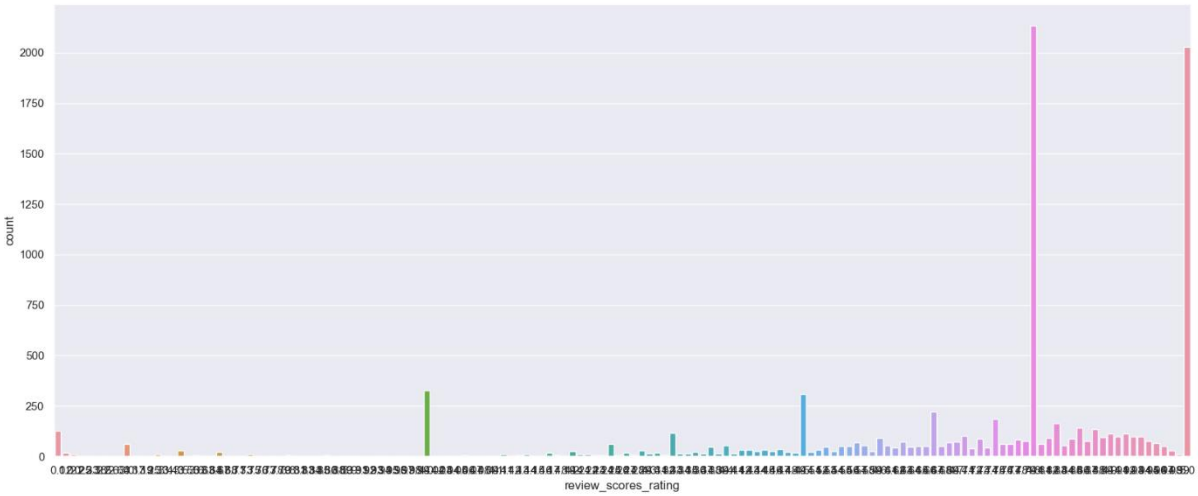


4. Analyse de Motifs de points (La densité des biens à louer à Lyon)

En premier temps, nous allons représenter le nombre de biens noté avec chaque rate :

```
#afficher pour chaque rate le nombre des biens noté avec:
data_final["review_scores_rating"].value_counts()
```

```
4.80    2133
5.00    2030
4.00     327
4.50     309
4.67     221
        ...
3.93       1
3.88       1
3.95       1
3.84       1
3.25       1
Name: review_scores_rating, Length: 148, dtype: int64
```

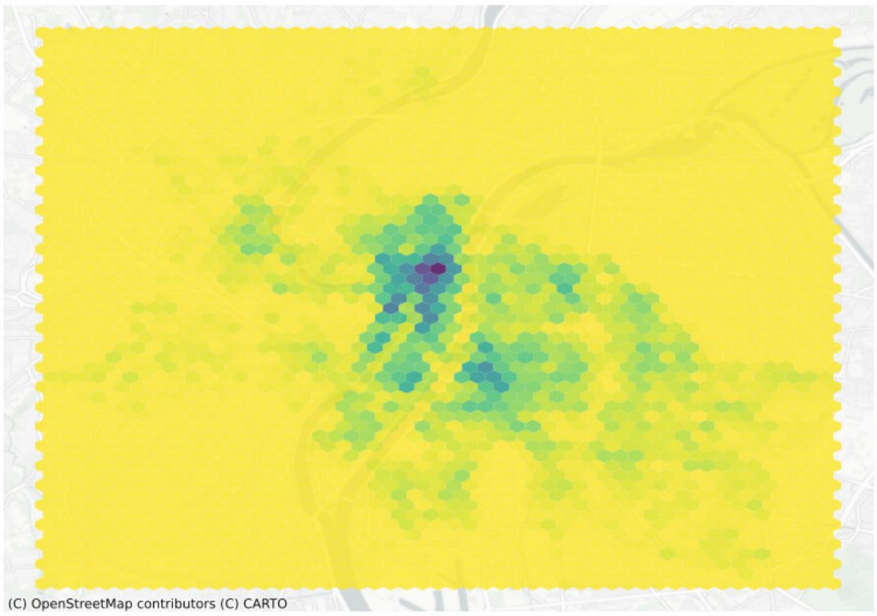


Comme indiqué dans le tableau précèdent : la note 4.80 est donnée à 2133 biens, et la note 5.00 est donnée à 2030 biens

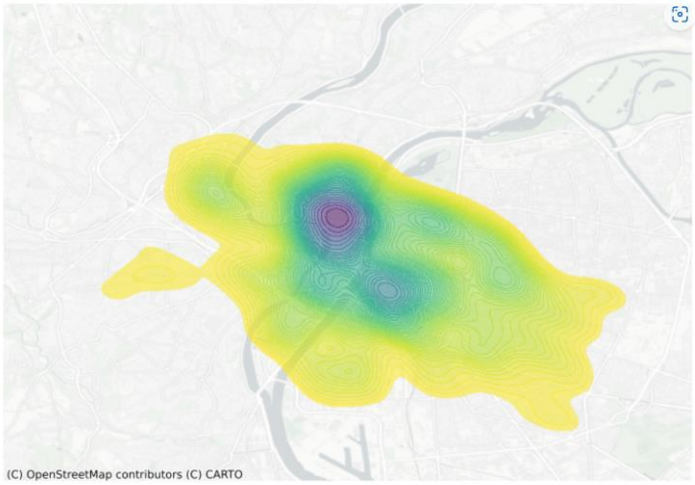
Nous allons à présent représenter le motifs de points Lyon à partir de longitude et latitude:

Ce type de représentation permet de mieux situer les zones à forte concentration sur Lyon. Nous remarquons qu'il y a une forte concentration dans le 1er,2e, 3e,4e et un peu du 7e arrondissements. La particularité de ces zones c'est qu'elles sont animées par des commerces et des lieux touristiques. Dans le 1er, nous avons des hôtels et des lieux touristiques, dans le 2e arrondissement nous avons la Place Bellecour et son Hôtel, La Confluence, etc… Dans le 3e arrondissement, nous avons la Part Dieu, son Centre Commercial et sa gare. Pour le 7e, nous avons essentiellement des commerces mais nous avons l'Université de Lyon 2 également.

Pour ces 3 zones, le secteur du commerce est le facteur dominant de la forte concentration des biens.
Le point commun de ces endroits c'est qu'ils sont tous situés au bord des quais du Rhône, qui fait partie des endroits mythiques de Lyon.
Les clients auront beaucoup plus de facilité à se déplacer car ils auront tout à proximité.

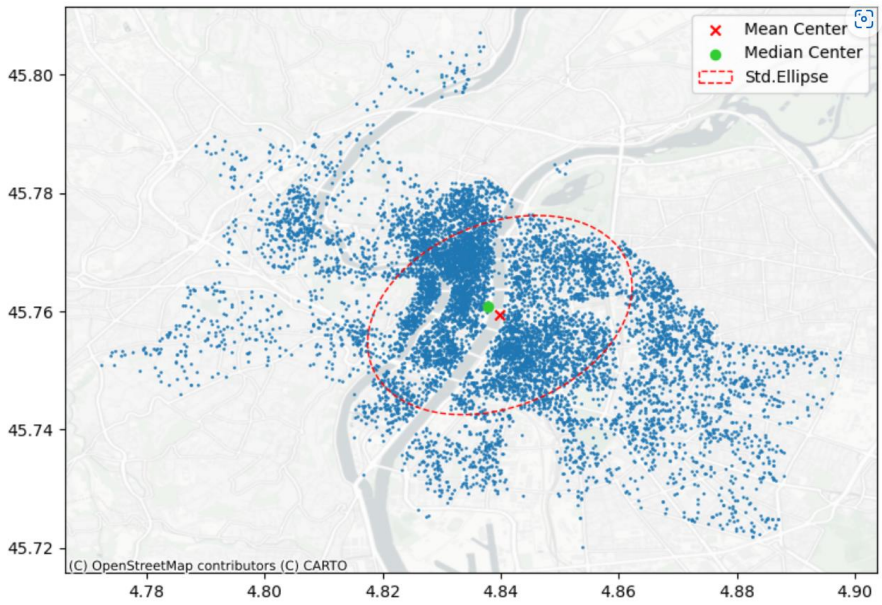


L'agrégation locale et les nuances de couleurs permettent de voir plus clairement les localisations où la majorité des biens sont localisés.
Nous avons une forte concentration dans le 1er arrondissement.



Cette représentation est obtenue grâce à l'estimateur du noyau. Le but est de calculer la probabilité que les biens se trouvent à tel ou tel endroit. Nous voyons que les biens sont également concentrés sur les mêmes zones. Ce qui nous réconforte dans notre analyse.

Pour appuyer encore plus notre analyse nous avons déterminé le point médian et le point moyen. Ces derniers sont à 0. 02 mètres du point de tendance central. Le point médian est la médiane de chaque distribution de coordonnées. Nous avons autant de points de part et d'autre de la médiane. Le point médiane quant à elle représente la moyenne de chaque distribution de coordonnées parmi les observations.

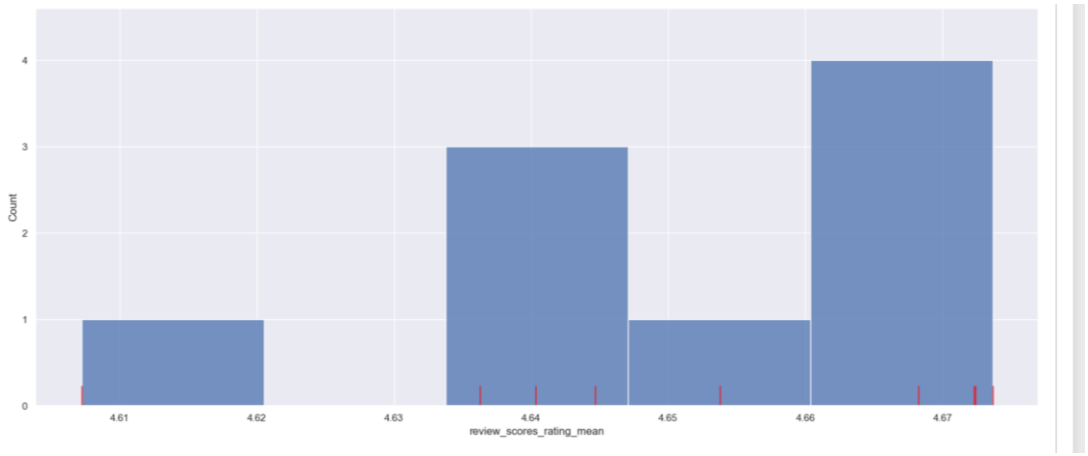


II. Utilisations des méthodes statistiques restituées par des cartes choroplèthes

Nous allons étudier la discrétisation de la variable review_scores_rating.
Tout d'abord, nous allons ajouter une colonne contenant la moyenne de la note par arrondissement de notre variable qui est « review_scores_rating_moyenne »

1. Histogramme sur les *review_scores_rating*

L'histogramme est une des techniques de statistique qui nous permet de pouvoir déterminer l'amplitude des variables mais aussi de voir leur distribution sur l'ensemble des données.
Pour cet histogramme nous avons les différentes notes reçues données aux hôtes.

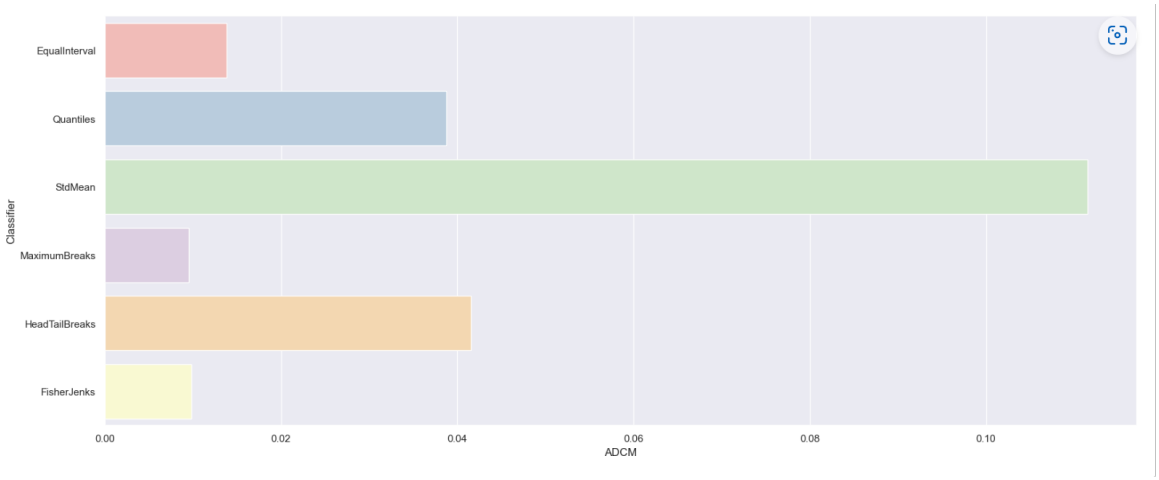


La distribution est négativement asymétrique, la queue de distribution pointe vers la gauche : leur valeur d'asymétrie est inférieure à 0.
La distribution est positivement asymétrique, la queue de distribution pointe vers la droite : la valeur d'asymétrie est supérieure à 0.
La médiane est en dessous de la moyenne. En moyenne, la plupart des arrondissements ont des notes (review_scores_rating) > 4. 63
Il est aussi possible de discrétiser la variable à l'aide de la libraire mapclassify.
Elle implémente une famille de schémas de classification pour les cartes choroplèthes. Il se concentre sur la détermination du nombre de classes et l'attribution d'observations à ces classes.
Nous avons la classe 5 qui comporte le plus d'observations.
Les intervalles sont définis de sorte à ce que le nombre d'éléments dans chacun d'eux est approximativement égale (d'où l'utilisation de percentiles).
Donc nous avons plusieurs méthodes : EqualInterval, Quantiles, StdMean, MaximumBreaks, HeadTailBreaks et FisherJenks
La différence entre eux est que :

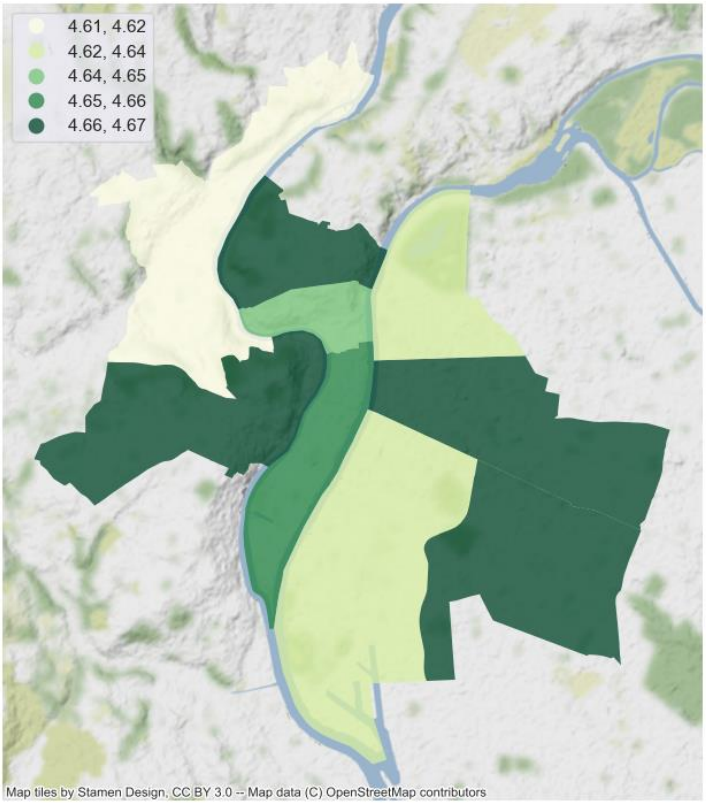
- **MaximumBreacks** trie les données par ordre croissant et définit des valeurs de séparation qui engendrent des intervalles.
- **Head Tail Breaks** trie les données par ordre décroissant, elle affiche sa Head et sa Tail et la valeur de séparation qui représente la moyenne des observations des classes.
- **Fisher Jenks** quant à elle, résout optimalement le problème de la minimisation de la somme des écarts à la moyenne (locale) en valeur absolue en utilisant la programmation dynamique.

2. Comparaison entre les différentes méthodes

Afin de pouvoir sélectionner la meilleure modalité de discrétisation, nous allons observer la part qu'elle occupe à l'aide d'un critère courant qui est la moyenne des écarts en valeurs absolues.



MaximumsBreacks est la méthode la plus pertinente à nos yeux pour notre jeu de données.



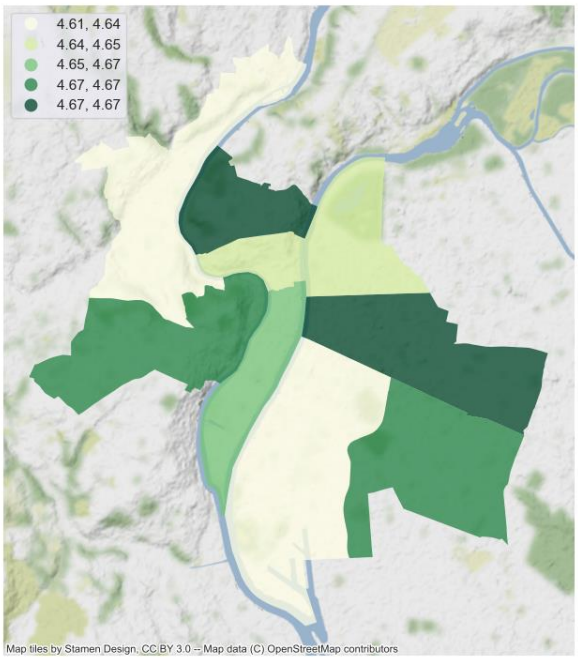
Nous avons la carte de la ville de Lyon. Nous voyons différentes zones en fonction de notre variable auquel on a rajouté la moyenne. On doit indiquer que les rates se varient entre 4.61 et 4.67 en moyenne pour les arrondissements, pour cela on ne peut pas dire que 4.61 est une mauvaise note, mais on va considérer que cette note est un peu loin de 5 par rapport aux autres notes.
Donc la note de 4. 61 qui représente le 9e arrondissement est considéré comme étant la plus faible note. Contrairement au 3e,4e,5e et 8e arrondissements qui ont les meilleures notes. Ils sont suivis du 2e arrondissement puis du 1er et enfin du 6e et 7e arrondissement.
Nous pouvons dire que ces notes diffèrent d'un arrondissement à un autre, et que les observations sont indépendantes, on remarque alors que les observations sont totalement aléatoires.

III. Autocorrélations spatiales entre les arrondissements

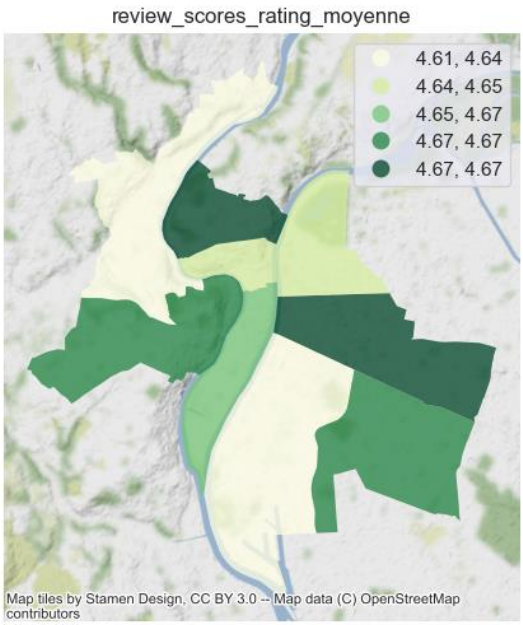
1. Observations et Analyses

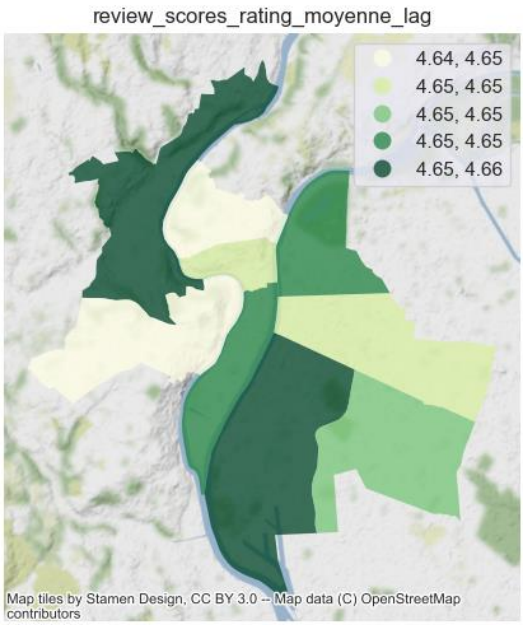
Nous voulons étudier maintenant l'autocorrélation spatiale entre ces arrondissements. Le but est de savoir s'il existerait un lien qui pourrait expliquer cette répartition inégale des notes c'est à dire nous chercherons à savoir si les arrondissements possédant un review_scores_rating élevé auront aussi des review_scores_rating élevé dans les arrondissements voisins.

La différence entre cette carte ci-dessous et la précédente est tout simplement au niveau de la méthode. En effet, sur celle-ci, nous avons utilisé la méthode quantiles.



Pour cette carte, nous voyons que le 9e et le 7e arrondissement ont les plus faibles review_scores_rating contrairement au 4e et 3e arrondissement qui ont des review_scores_rating plus élevés. Ils sont suivis du 5e et 8e puis du 2e et enfin du 1er et 6e arrondissement.
Nous retrouvons pour certains les mêmes comportements en termes de classification. En effet, le 9e et le 3e arrondissement restent toujours à la même classification.
Il existe des statistiques permettant de caractériser le degré de clustering spatial associé à une carte. Avant de les introduire, nous présentons un concept central, le décalage spatial/spatial lag.
Le spatial lag operator est l'une des applications les plus directes et classiques des spatial weights (SW)

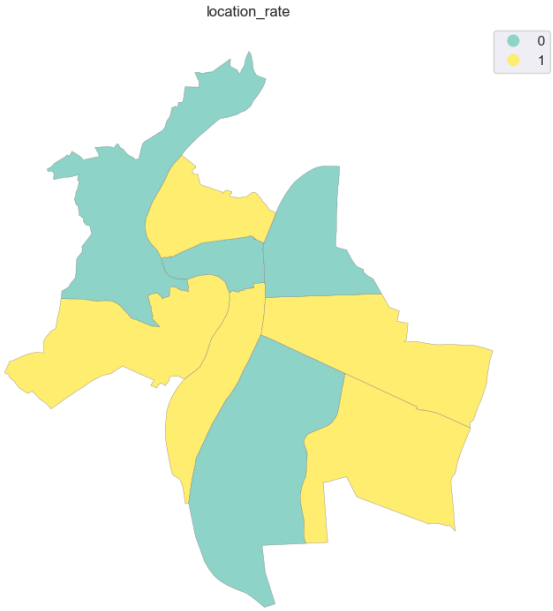




Il permet d'avoir un lissage local des valeurs selon les poids des relations spatiales qu'entretient chaque objet avec son voisinage. Nous pouvons conclure à l'aide de la carte ci-dessus que cette fois-ci, ce sont le 9ème et le 7ème arrondissement où nous retrouvons les locations les mieux notés. Les 4ème et 5ème arrondissements sont parmi les locations les moins bien notés.

2. Variable binaire

Avec la méthode quantiles, nous allons séparer notre variable en deux clusters. Le but est de représenter d'une part les arrondissements possédant un review_scores_rating supérieure à 4,65 par 1 et les autres par 0.



Nous pouvons très vite voir que les arrondissements qui sont les mieux notés et ceux qui ne le sont pas.
Les arrondissements les mieux notés sont : 4,5,2,3,8 arrondissements, et les moins bien notés sont : 9,1,6,7

“La statistique de test est basée sur la comparaison des nombre de 00, 11 et 01 empiriques (c à d dans l’échantillon) versus les nombre de 00, 11 et 01, théoriques (c à d que l’on aurait observée si une structure spatiale totalement aléatoire générait les données). En statistique inférentielle. Cela revient à poser l'hypothèse nulle suivante : “les valeurs sont générées par une structure spatiale totalement aléatoire. Une façon d’approximer ces nombres théoriques sous l'hypothèse nulle est de permuter les valeurs de la variable. Parmi les objets et de calculer pour cette permutation le nombre de 00, 11 et 01. On répète cette opération plusieurs fois et on peut alors calculer la moyenne des nombre 00, 11 et 01 sur l’ensemble des permutations réalisées. Intuitivement, si la moyenne des nombres de 00 et 11 sur l’ensemble des permutations est en dessous des valeurs empiriques alors on aura tendance à rejeter l’hypothèse nulle”.

```
from pysal.explore import esda
from numpy.random import seed
seed (1234)
jc = esda.join_counts.Join_Counts(arr_lyon["rate_binaire"] , w)

print (jc.bb)
print (jc.ww)
print (jc.bw)
```

```
5.0
3.5
14.0
```

Nous avons obtenu 5 pour le nombre de 00, 3. 5 pour le nombre de 11 et 14 pour le nombre de 01 ou 10.

Nous allons calculer la p-value qui nous donne la probabilité d’observer une statistique de test de valeur au moins aussi extrême que celle que l’on observerait si l’hypothèse de départ H0 était vérifiée.

```
print (jc.p_sim_bb)
```

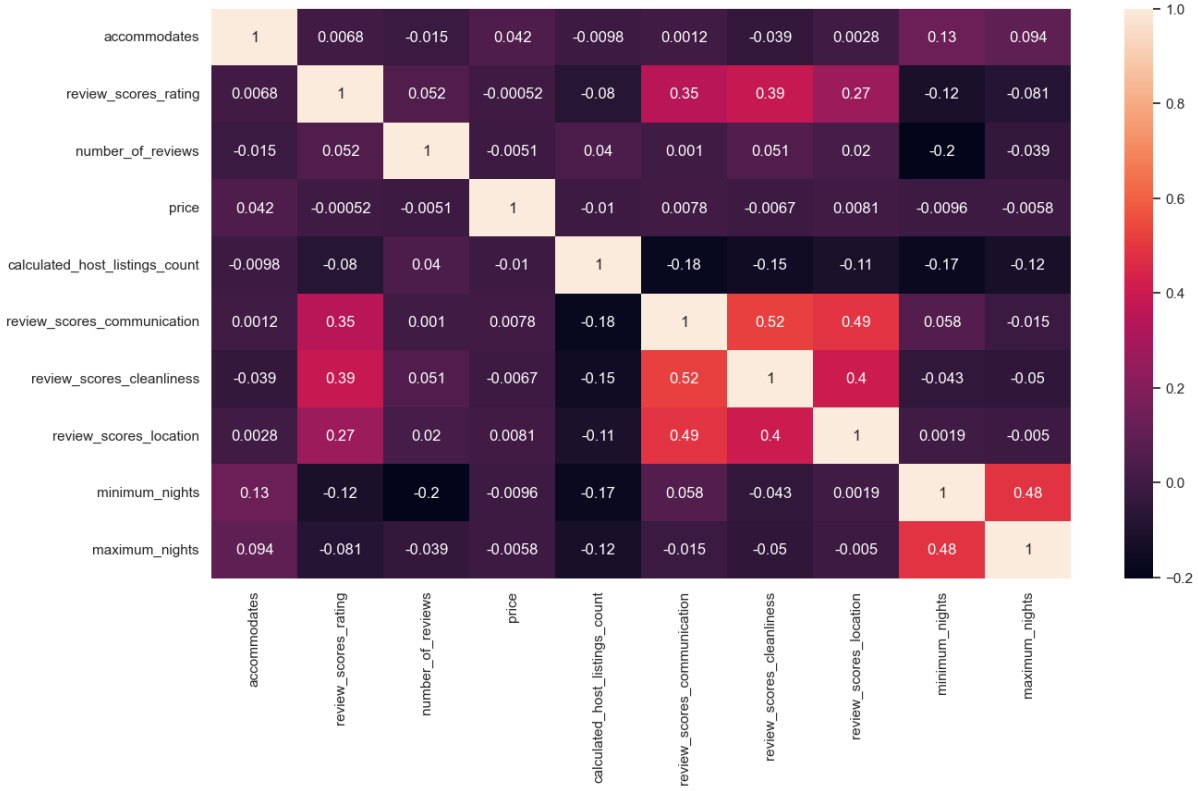
```
0.928
```

Nous obtenons un score de 0,928. La p-value est très grande, donc on peut en déduire que l'hypothèse nulle H0 sera acceptée : les observations sont totalement aléatoires.

IV. Modèle de prédictions de notre variable review_scores_rating : application de la régression linéaire multiple

La régression linéaire multiple nous permettra de pouvoir observer le lien entre notre variable cible review_scores_rating et les autres variables.

1. Corrélation entre la variable review_scores_rating et les autres variables



Nous voyons que les variables les plus corrélées avec la variable "review_scores_rating sont review_scores_comunication et review_scores_cleanliness et review_scores_location.
Nous avons une corrélation de 0.39 pour la note de propreté, plus une location est propre et meilleure sera sa notation. Également, la variable expliquée à une corrélation de 0.35 pour les notes de communications, plus la communication du propriétaire est satisfaisante et plus la note sera élevée.

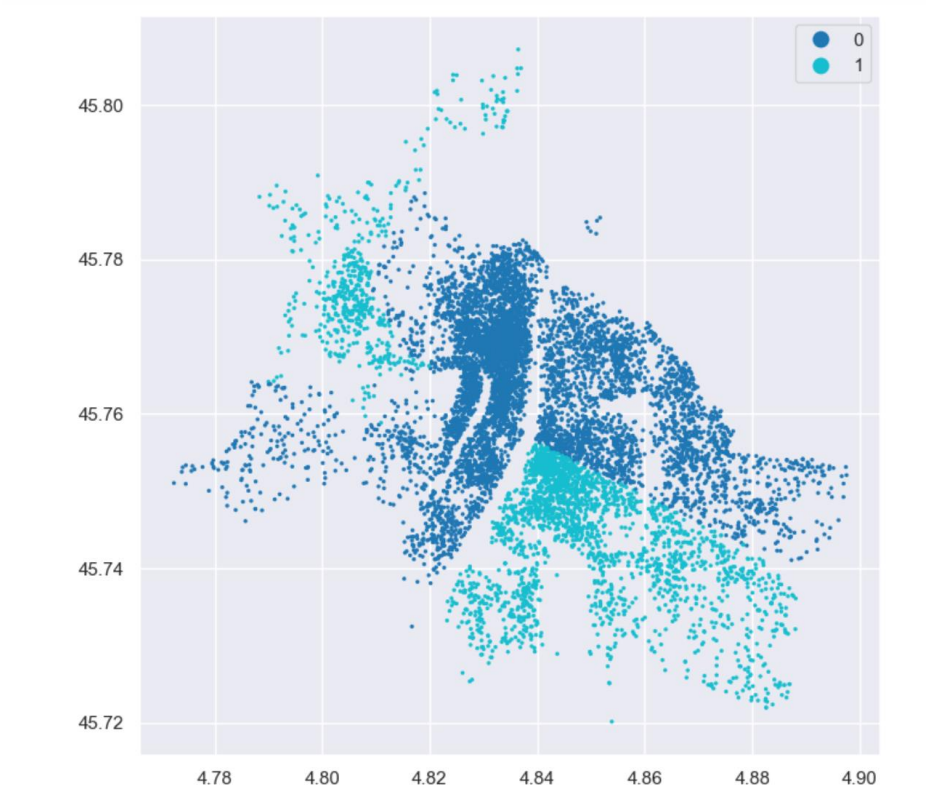
La localisation de bien a aussi une influence sur la variable note, avec une corrélation plus faible de 0.27. Nous voulons prédire le « review_scores_rating » d'une location à partir des caractéristiques de celle-ci. Nous n'intégrons pas dans un premier temps des variables associées à une information géographique quelconque. La liste de variables sélectionnée est alors la suivante :
"accommodates","number_of_reviews","price","calculated_host_listings_count","review_scores_communication","review_scores_cleanliness","review_scores_location","minimum_nights","maximum_nights".

2. Interprétation des résultats du modèle de prédiction

Notre modèle de prédiction nous affiche plusieures résultats, parmi celles-ci, la p-value de prob(F-statistic) qui permet d'évaluer la signification de toutes les variables ensembles, avec une hypothèse nulle que tous les coefficients de régression soit égal à 0. Dans notre modèle avec un grand nombre de données, nous avons un échantillon qui suit une loi normale avec un niveau de confiance de 95, p-value est égal à 0, alors nous rejetons H0 qui suppose que le coefficient des variables explicatives soit nul. De plus, R2, la variation en pourcentage de la dépendance expliquée par des variables indépendantes, est de 19%, ce qui signifie que 19% des variations de « review_scores_rating » sont expliqués par les variables explicatives ce qui n'est pas significatif. Également, les coefficients de « review_scores_cleanliness » et de « « review_scores_communication » sont de 0.37 et 0.39, ces deux variables sont explicatives de la variable « review_scores_rating » ce qui confirme nos hypothèses de départ.

3. Intégrer une variable spatiale (Indicateur de Sécurité)

Pour approfondir notre étude, nous pouvons supposer qu'une location se trouvant dans un arrondissement sensible de Lyon et si ce dernier aurait une influence sur les notes des biens. Donc, nous allons répartir les arrondissements en deux clusters : ceux qui sont dans des zones sensibles par 1 et 0 sinon.

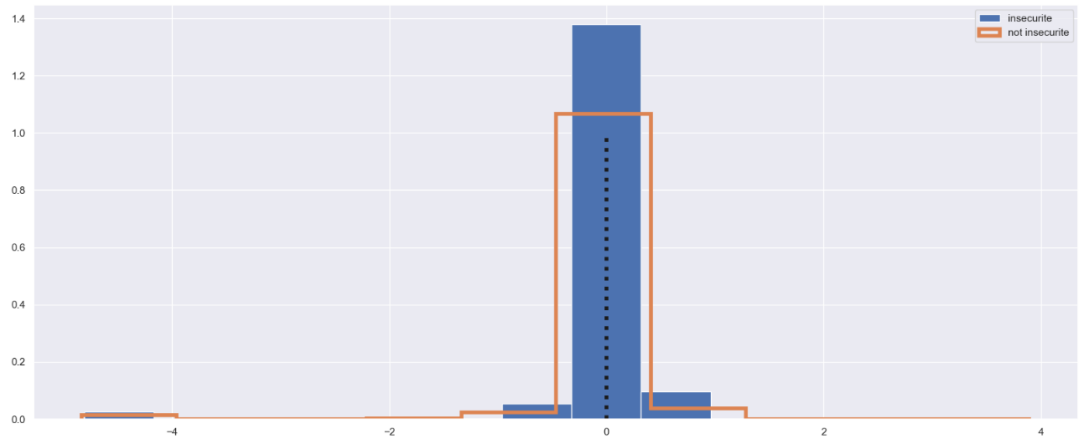


Selon l'indicateur de sécurité.
Nous supposons que les notes prennent bien en compte le paramètre de la sécurité autrement dit nous supposons que les zones ayant des notes faibles seront des zones sensibles.
Parmi les 9 arrondissements de la ville de Lyon, ceux qui présentent une plus forte délinquance, un taux plus élevé de crimes et de délits, et un sentiment d'insécurité plus important sont principalement les 7ème, 8ème et 9ème arrondissement.
L'idée est de considérer dans notre hypothèse ces zones aussi comme étant des endroits sensibles.
Le 9ème arrondissement de Lyon en particulier les quartiers Vergoin, Vaise, Gorge de Loup, Montribloud et surtout son quartier de la Duchère. En effet, ce dernier serait devenu le quartier le plus dangereux de Lyon.

A noter que le 8ème arrondissement fait maintenant partie des nouvelles zones de sécurité prioritaires.
Le 7ème arrondissement est le plus vaste arrondissement de Lyon. Sa réputation d'arrondissement à éviter est surtout liée au quartier de la Guillotière et au quartier de Gerland.
Notre hypothèse de départ H0 : les erreurs de la variable cible « review_scores_rating» des locations situant dans des quartiers dangereux sont plus faible que les erreurs de cette variable pour des locations non situées dans des quartiers dangereux.

4. Etudier les résidus de notre modèle :

Nous avons créé une variable booléenne qui est True si le logement se situe dans une zone dangereuse et False sinon. Autrement dit, la variable booléenne est True si le logement se trouve au 7,8 et 9 et 0 sinon.



Nous avons ci-dessus un histogramme de la distribution des résidus de sécurité et de non-sécurité. Nous constatons une structure de clustering spatiale des erreurs :il semble que les distributions des locations situant dans des quartiers dangereux versus non-dangereux ont des allures différentes

5. Application du test de Student

Pour encore appuyer notre hypothèse, nous utiliserons le test de Student et comparerons les moyennes de ces deux distributions empiriques (insecurite et not insecurite). Notre hypothèse H0 que l'on appellera H0_student sera que les deux distributions sont identiques.
Nous obtenons une p-value de 73% ce qui veut dire que nous ne rejetterons pas H0_student. Donc, en conclusion, nous pouvons dire que les deux distributions sont proches l'une de l'autre.
Par conséquent, notre hypothèse départ qui était que les erreurs de la variable review_scores_rating seraient plus ou moins élevés selon le niveau de sécurité de la zone(arrondissement) est rejetée. Il n'y a pas de différence entre les deux.

CONCLUSION

En définitive, nous avons conclu que les logements ayant reçu un review_scores_rating élevé autrement dit les hôtes qui ont reçu de bonnes notes pour leur logement, concernent plus les biens situés dans des zones qui sont accessibles qui sont pour la plupart des lieux touristiques. On a également compris que ces notes ne dépendaient pas de l'indicateur de sécurité. A priori, grâce aux tests effectués, il n'y a aucun lien entre les notes reçues et l'indicateur de sécurité. Ceci peut sans doute s'expliquer par le fait que la réputation compte beaucoup pour les clients. De ce fait, ces derniers ont tendance à se renseigner avant de pouvoir choisir un bien sur la plateforme d'Airbnb