



## RAPPORT PROJET “FOUILLE DE DONNÉES”

AZAR Gloria, LARIBI Fatma et FAUCHON Marilou  
4IF4



Janvier 2023

|  |           |
|--|-----------|
| <b>I- Introduction</b>   | <b>2</b>  |
| <b>II- Nettoyage des données</b>   | <b>3</b>  |
| 1- Analyse des données à disposition                                       | 3         |
| 2- Filtres appliqués sur ces données                                       | 3         |
| A- Traitement des données corrompues                                       | 4         |
| a- Traitement des valeurs nulles   | 4         |
| b- Traitement des données incohérentes                                     | 5         |
| B- Traitement des doublons   | 6         |
| C- Données filtrées  | 6         |
| <b>III- Découverte de points d'intérêt grâce au clustering</b>             | <b>8</b>  |
| 1- K-means   | 8         |
| Explications sur le Workflow   | 8         |
| Résultats du clustering  | 8         |
| 2- Clustering Hiérarchique   | 11        |
| Explications sur le Workflow   | 12        |
| Résultats du clustering  | 16        |
| 3- L'algorithme DBSCAN   | 20        |
| A- Prise en compte de l'espace   | 20        |
| Explications sur le Workflow   | 20        |
| Résultats du clustering  | 21        |
| B- Prise en compte de l'espace et du temps                                 | 25        |
| Explications supplémentaires sur le Workflow                               | 25        |
| Résultats du clustering  | 26        |
| <b>IV- Manipulation des tags pour la découverte des centres d'intérêts</b> | <b>30</b> |
| 1- Extraction et nettoyage des mots-clés                                   | 30        |
| Explications sur le Workflow   | 30        |
| Résultats  | 32        |
| 2- Extraction des règles   | 33        |
| Explications sur le Workflow   | 33        |
| Résultats  | 34        |
| <b>V- Informations destinées aux clients et aux équipes futures</b>        | <b>35</b> |
| <b>VI- Bibliographie et références</b>                                     | <b>38</b> |

# Rapport projet Fouille de Données

## I- Introduction

Lyon est une ville française située dans le département du Rhône qui est de plus en plus visitée par des touristes français ou étrangers. Il s'agit d'une ville qui accueille de nombreux évènements (sportifs, culturels,...) et qui est dotée d'un patrimoine historique, culturel (fête des Lumière en Décembre), gastronomique avec des paysages attrayants. De plus, cette ville renferme des chercheurs et laboratoires qui organisent des séminaires afin de partager la recherche scientifique au reste du monde. Ainsi, pour permettre aux touristes comme aux habitants de Lyon de se déplacer facilement à travers la ville, la métropole propose un vaste réseau de transport composé, en 2019, de 140 lignes dont 2 funiculaires, 4 lignes de métro, 6 lignes de tram et 128 lignes de bus. Cette même année, 496 millions de voyages ont été effectués sur ces lignes. Quoique le réseau de transport en commun lyonnais TCL est parmi les meilleurs réseaux de transports en commun européens, il est dans son intérêt de s'adapter à cette masse de touristes. Pour cela, il faut étudier et analyser les moments et les endroits que fréquentent les touristes à Lyon afin de rendre leurs séjours plus agréables et le voyage des passagers dans les transports moins contraignant. C'est dans ce but que le Grand Lyon a lancé un appel d'offre public auquel nous avons répondu.



Dans un premier temps, à l'aide d'un capteur social nous avons réalisé une collecte de médias géolocalisés à travers l'API du service Flickr. Ainsi, nous allons pouvoir procéder à une étude du jeu de données afin de trouver de manière non-intrusive les zones à forte densité de touristes à moindre coût.

Dans cette étude, nous avons travaillé sur la détection de centres d'intérêt dans l'agglomération lyonnaise à l'aide de données de photos téléchargées sur

Flickr sur lesquelles nous avons appliqué un clustering. Ce projet a été mené en trois temps :

- L'Analyse et le nettoyage des données
- Clustering hors DBSCAN: K-means et hiérarchique
- Clustering DBSCAN

## II- Nettoyage des données

### 1- Analyse des données à disposition

Le fichier CSV donné dans le cadre de cette étude comportait 18 colonnes et environ 42 000 lignes. Elles correspondent à des données issues de Flickr décrivant des photos prises sur Lyon entre 2010 et 2019. Les informations sur chacune des photos sont : son id, la date à laquelle la photo a été prise (année, mois, jour, heure et minute), la date à laquelle la photo a été mise en ligne sur Flickr (année, mois, jour, heure et minute), ses tags et ses coordonnées géographiques (longitude et latitude). Nous précisons que dans la première partie de cette étude consistant à former des clusters, les attributs exploités pour la détection de centres d'intérêt ne doivent en aucun cas être révélateur de la nature du lieu en tant que centre d'intérêt (par exemple, l'attribut *tag*), comme le but ici est pouvoir les détecter en exploitant la spatio-temporalité des photos. Dans ce cas-là, seules la latitude et la longitude doivent contribuer à la détection de ces lieux et non pas les "tags" par exemple.

### 2- Filtres appliqués sur ces données



## A- Traitement des données corrompues

#### a- Traitement des valeurs nulles

Après avoir regarder de plus près les valeurs des attributs de chaque tuple sur KNIME, nous avons établis différents constats à leur sujet :

- La colonne 17 (nommée Col 17) semble être vide : nous avons utilisé l'outil *value counter* afin de compter les valeurs qui y sont non nulles et il n'y en avait aucune. Nous avons regardé de plus près les colonnes 16,17 et 18 afin de trouver des explications à ce phénomène.
  - Toujours à l'aide de l'outil *value counter*, la colonne 18 ne comportait que deux valeurs non nulles et dans la colonne 16 environ 420 046 valeurs sont nulles ( $420240-420046 = 194$  lignes non vides).
  - A partir de l'outil *Row Filter*, on observe les données telles que la colonne 16 ne soit pas vide, et on constate que ces données comportent un problème de décalage. En effet, certaines valeurs de date sont décalées par rapport à la colonne correspondante puisque dans la colonne prévue à cet effet se trouve une valeur qui n'est pas de type integer, entraînant le fait de se trouver sur ces trois autres colonnes (16, 17 et 18). Nous pouvons observer ce phénomène dans l'image suivante :

Dans cette capture ne sont représentées que les lignes comportant des valeurs non nulles dans la colonne 16. Nous pouvons y constater que la colonne `date_upload_minute` contient des données qui ne sont pas de type integer, et c'est la colonne d'après qui contient la valeur de la minute de la date de téléchargement, et toutes les valeurs des colonnes suivantes se retrouvent décalées dans la colonne de droite.

Finalement, nous avons appliqué un *Row Filter* sur les attributs de dates (minute et heure) afin de ne récupérer que les lignes où les dates sont de type **Integer**. Ensuite, nous avons appliqué un *Column Filter* afin d'éliminer les colonnes 16, 17 et 18.

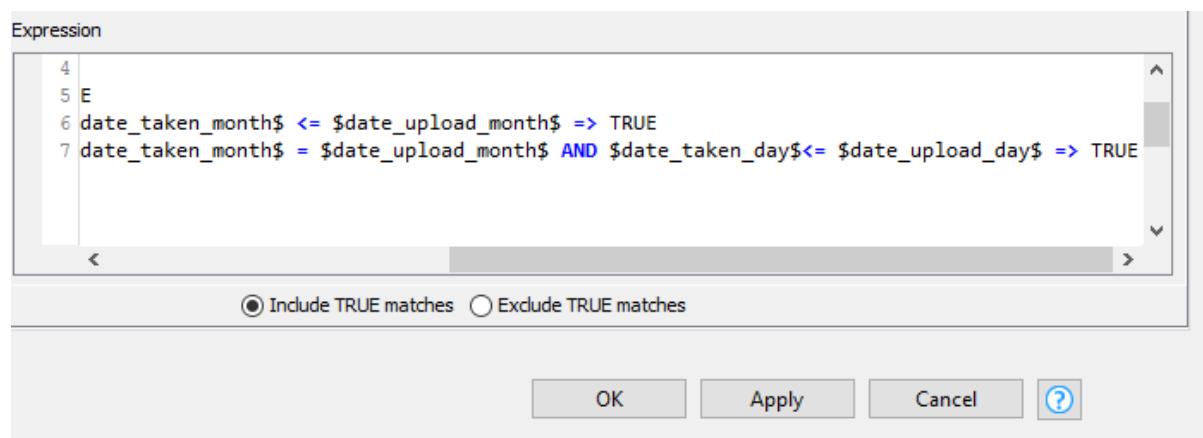
#### b- Traitement des données incohérentes

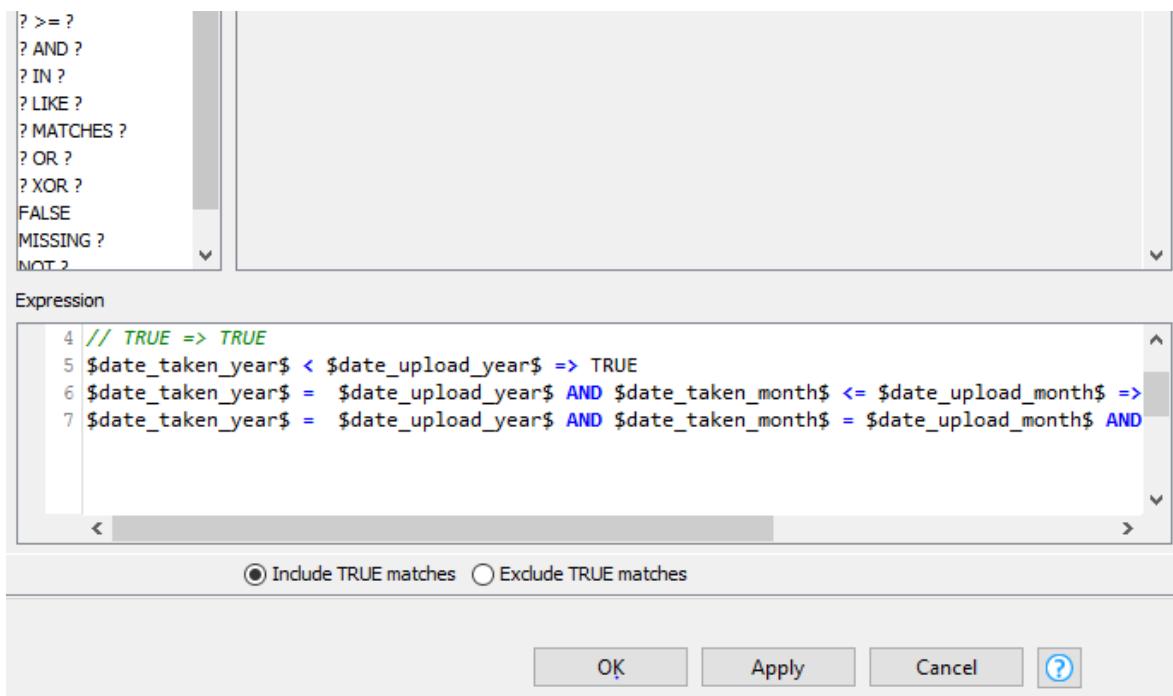
Nous avons voulu éliminer de notre étude les tuples qui sont, par les valeurs de leurs attributs, incohérents. Par exemple, lorsque la date de téléchargement d'une photo est inférieure à sa date de création, ou lorsque les attributs de date sont des valeurs non conformes (par exemple : une valeur de mois supérieure à 12, etc). Nous avons donc conçu deux filtres :

- Par la succession de l'utilisation de plusieurs *Row Filter*, nous avons mis en place un filtre de vérification de conformité de valeurs des attributs des dates de création et des dates de téléchargement :

$1 \leq \text{month} \leq 12$   
 $2010 \leq \text{year} \leq 2019$   
 $1 \leq \text{day} \leq 31$   
 $0 \leq \text{minute} \leq 59$

- A l'aide de l'outil *Rule Based Row Filter*, nous avons construit un filtre qui vérifie que la date de création d'une photo est inférieure à sa date de téléchargement à l'aide des formules suivantes :





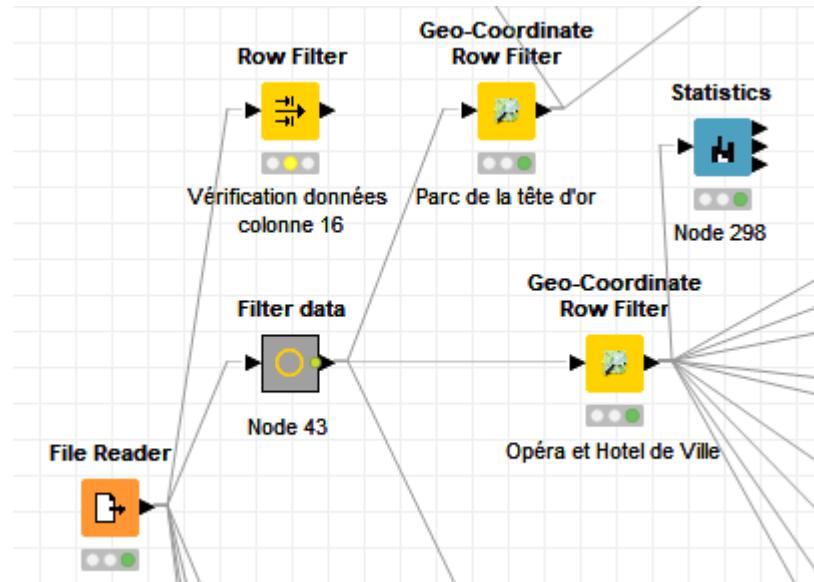
## B- Traitement des doublons

Dans le but d'éliminer les données redondantes d'information, nous avons appliqué un filtre éliminant les doublons, simulé par un *groupBy* sur l'ensemble des attributs existant. Ainsi, les données ayant les mêmes valeurs d'attributs ne seront représentées qu'une seule fois.

## C- Données filtrées

Suite à l'application de tous ces filtres, **165 976** lignes persistent et seront exploitées dans notre étude. Elles étaient initialement de l'ordre de **420 240** lignes, le nombre de données qui ont été retenues par les filtres nous a fait prendre conscience de l'importance de la partie de traitement des données issues du monde réel, le nombre de données aberrantes pouvant fausser nos mesures n'étant pas négligeable.

Visualisons quelques statistiques sur ces données :



### Capture d'écran des statistiques obtenus pour la zone Opéra et Hotel de Ville

On constate bien que pour le lieu touristique constitué de l'Opéra de Lyon et de l'Hôtel de Ville, les photos sont en moyenne prises en période d'été ( $mean(date\_taken\_month) = 7.6$ ) dans l'après-midi ( $mean(date\_taken\_hour)=15.5$ ).

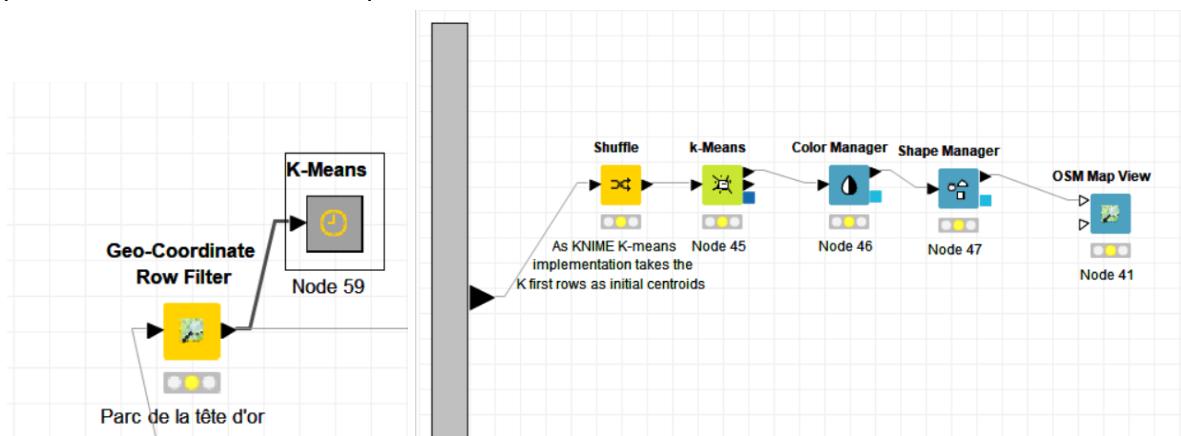
Ce qui paraît assez cohérent puisqu'il s'agit du moment auquel le plus d'animations en face de ces deux monuments ont lieu.

## III- Découverte de points d'intérêt grâce au clustering

### 1- K-means

K-means est un algorithme de clustering qui vise à trouver  $k$  clusters. En prenant  $k$  centroïdes, chaque point est affecté à un cluster selon le centroïde le plus proche. Les centroïdes sont ensuite recalculés jusqu'à la convergence.

Nous avons testé cet algorithme sur nos données. Dans la partie qui suit, nous allons présenter le workflow ainsi que les résultats trouvés.



### Explications sur le Workflow

Après l'application des filtres grâce au métanode “Filter data” de la partie nettoyage de données:

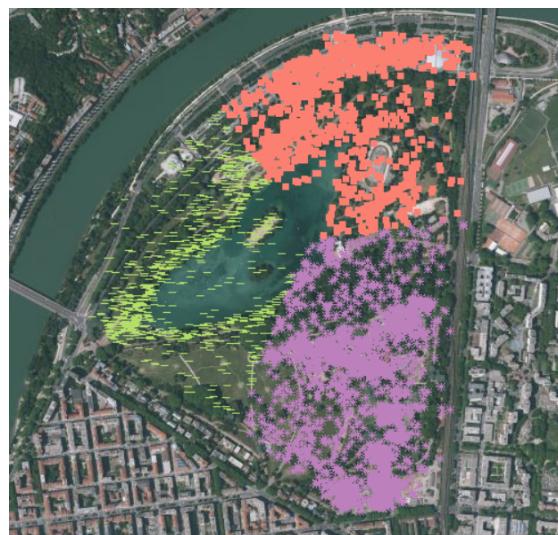
1. On a utilisé le composant *Geo-Coordinate Row filter* pour filtrer les résultats et afficher que ceux de la zone géographique qui correspond au “parc tête d’or”
2. Dans le métanode K-means, on utilise le composant *Shuffle* pour que K-means choisisse les premiers centroïdes de manière plus aléatoire (et pas les premières  $K$  lignes). On exécute l'algorithme et on affiche les résultats sur la carte avec des couleurs et des formes personnalisées afin de mieux visualiser le résultat grâce aux noeuds respectifs *k-Means*, *Color Manager*, *Shape Manager* et *OSM Map View*.

### Résultats du clustering

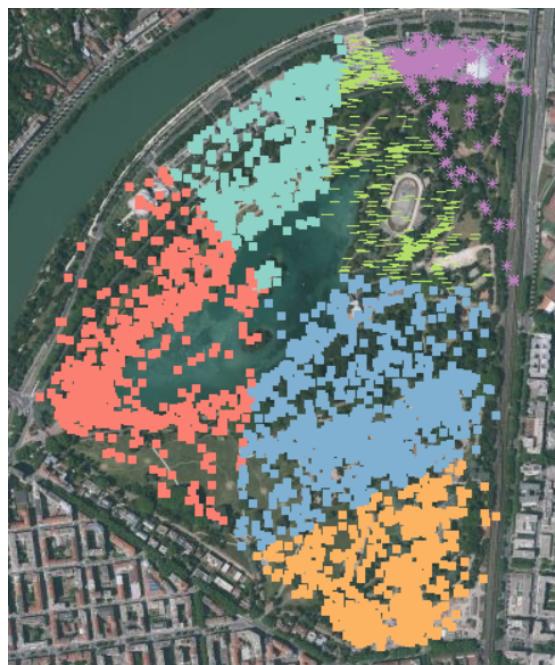
La zone étudiée: Parc tête d'or



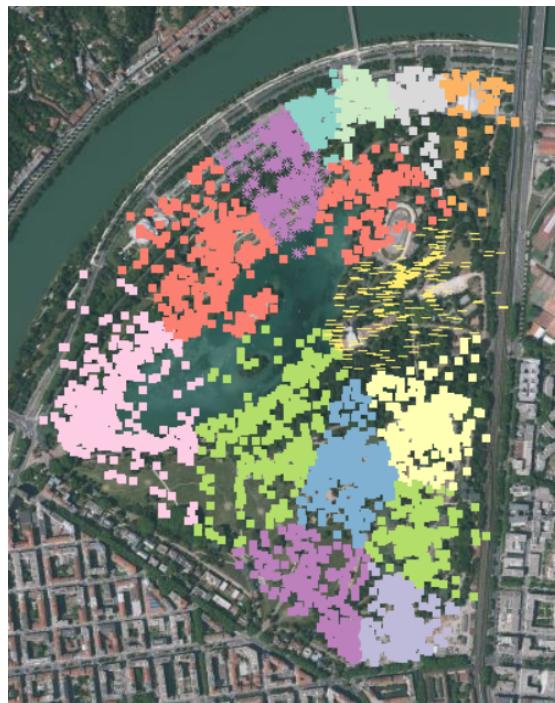
Pour  $k = 3$



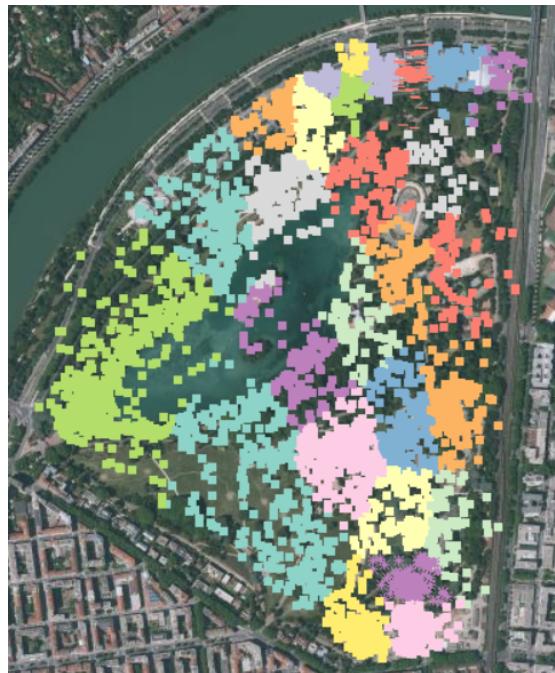
Pour k= 6



Pour k = 15



Pour k = 30



L'exécution de K-means est rapide. Les clusters ont à peu près la même taille et la forme des clusters est toujours globulaire. K-means regroupe les points selon leur localisation dans l'espace et non pas les centres d'intérêt.

Nous avons aussi constaté que le choix de k pose problème parce qu'il n'y a pas de formule pour le calculer. C'est un choix qui dépend des données. Il faut expérimenter pour déterminer le nombre de clusters qui semblent pertinents.

## 2- Clustering Hiérarchique

Les algorithmes de clustering hiérarchique se divisent en 2 catégories : top-down ou bottom-up. Les algorithmes ascendants traitent chaque point de données comme un seul cluster au départ, puis fusionnent (ou agglomèrent) successivement des paires de clusters jusqu'à ce que tous les clusters aient été fusionnés en un seul cluster contenant tous les points de données. Le clustering hiérarchique ascendant est donc appelé clustering agglomératif hiérarchique ou HAC. Cette hiérarchie de clusters est représentée sous forme d'arbre (ou dendrogramme). La racine de l'arbre est le cluster unique qui rassemble tous les échantillons, les feuilles étant les clusters avec un seul échantillon.

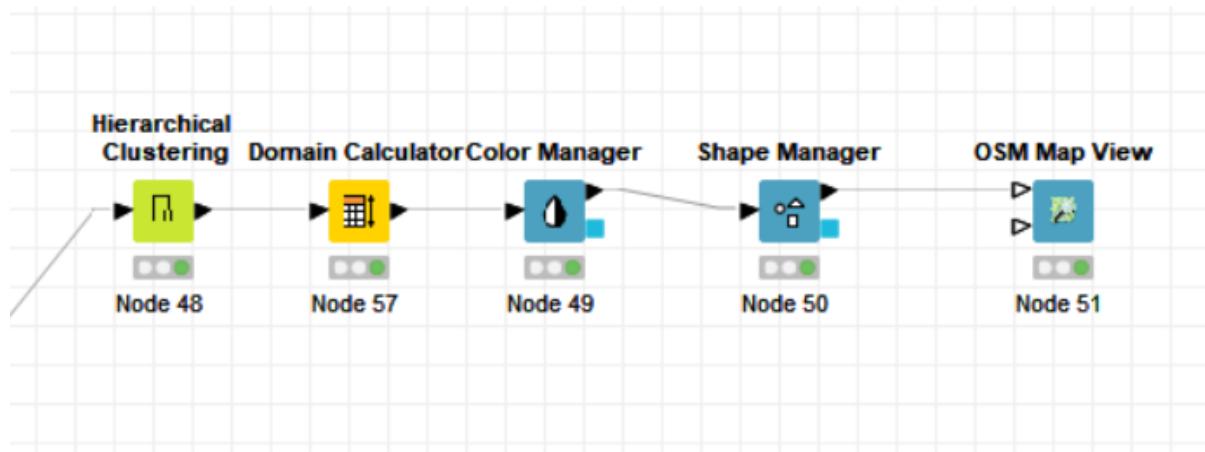
On choisit ici de limiter la zone géographique à l'hôtel de ville et l'opéra afin d'essayer de réduire le temps d'exécution du clustering hiérarchique.



Le choix pour la fonction de distance est Euclidean parce qu'on cherche la distance à vol d'oiseau.

3 types de linkage:

- Average Linkage: La distance entre les clusters est définie par la distance moyenne.
- Single linkage: La distance entre les clusters est la distance minimale.
- Complete linkage: La distance entre les clusters est la distance maximale.



#### Explications sur le Workflow

Ici, le workflow ressemble à celui de K-means, on utilise le noeud *Hierachial clustering*, suivi de *Domain Calculator* pour la mise à jour des informations sur les domaines des attributs et on affiche le résultat.

- K= 3, SINGLE



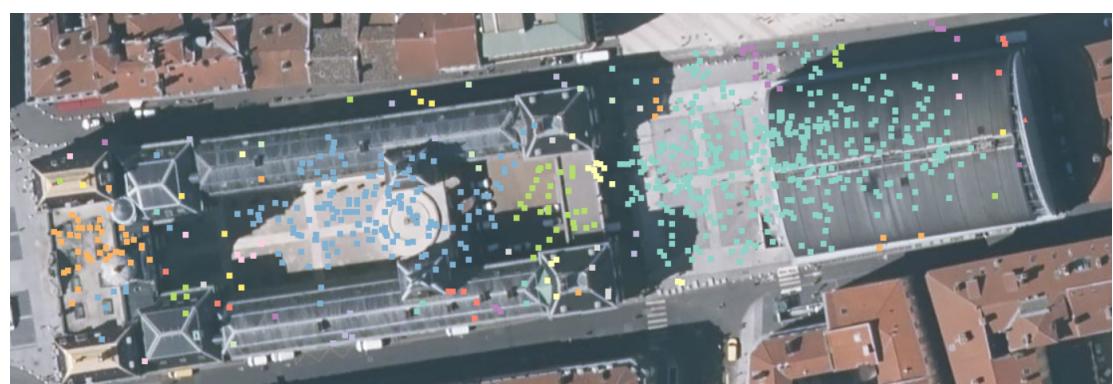
- K= 15, SINGLE



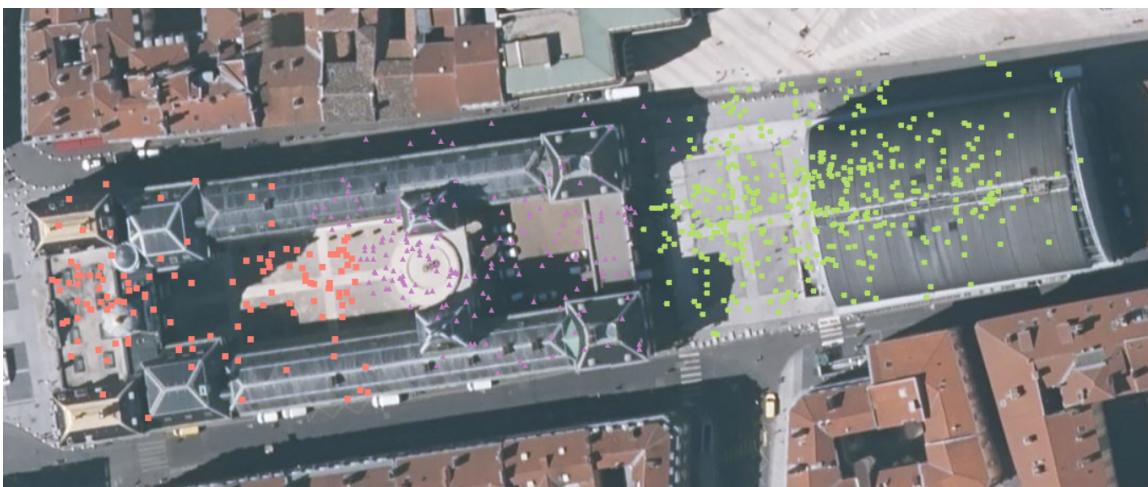
- K=30, SINGLE



- K = 70 SINGLE



- K= 3, COMPLETE



- K= 15, COMPLETE



- K=30, COMPLETE



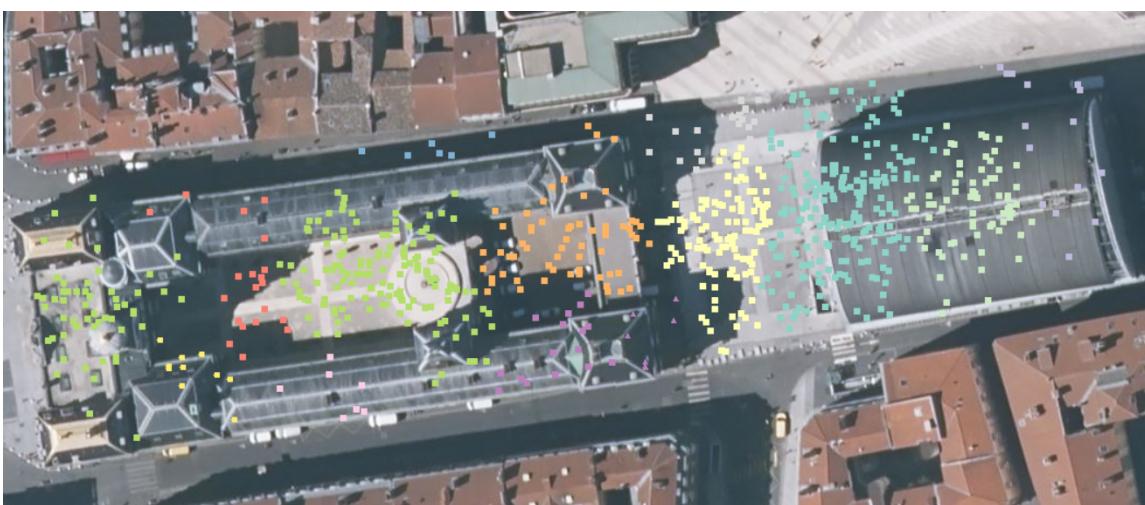
- K = 70 COMPLETE



- K= 3, AVERAGE



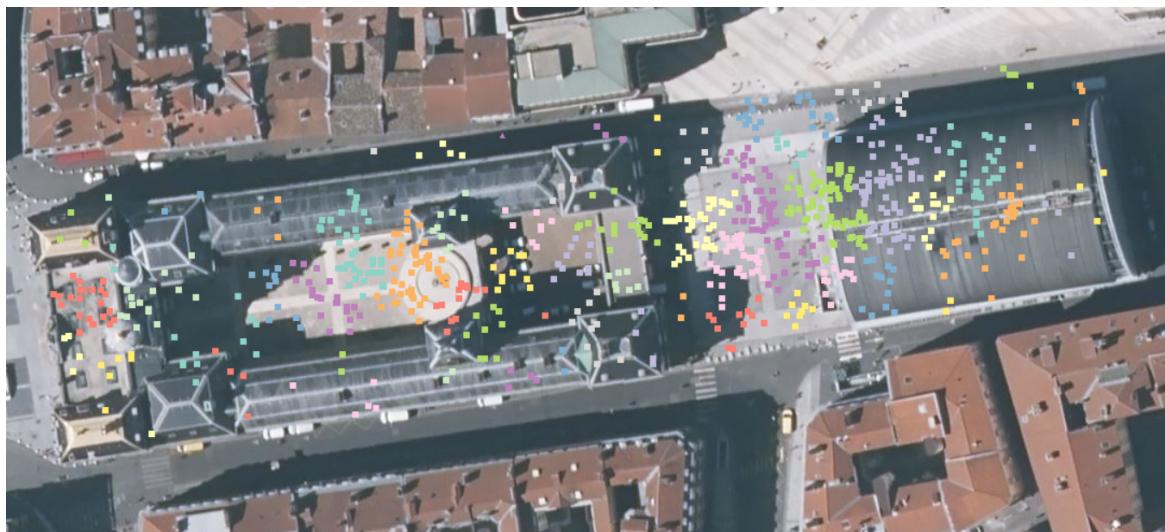
- K= 15, AVERAGE



- K=30, AVERAGE



- K = 70 AVERAGE



### Résultats du clustering

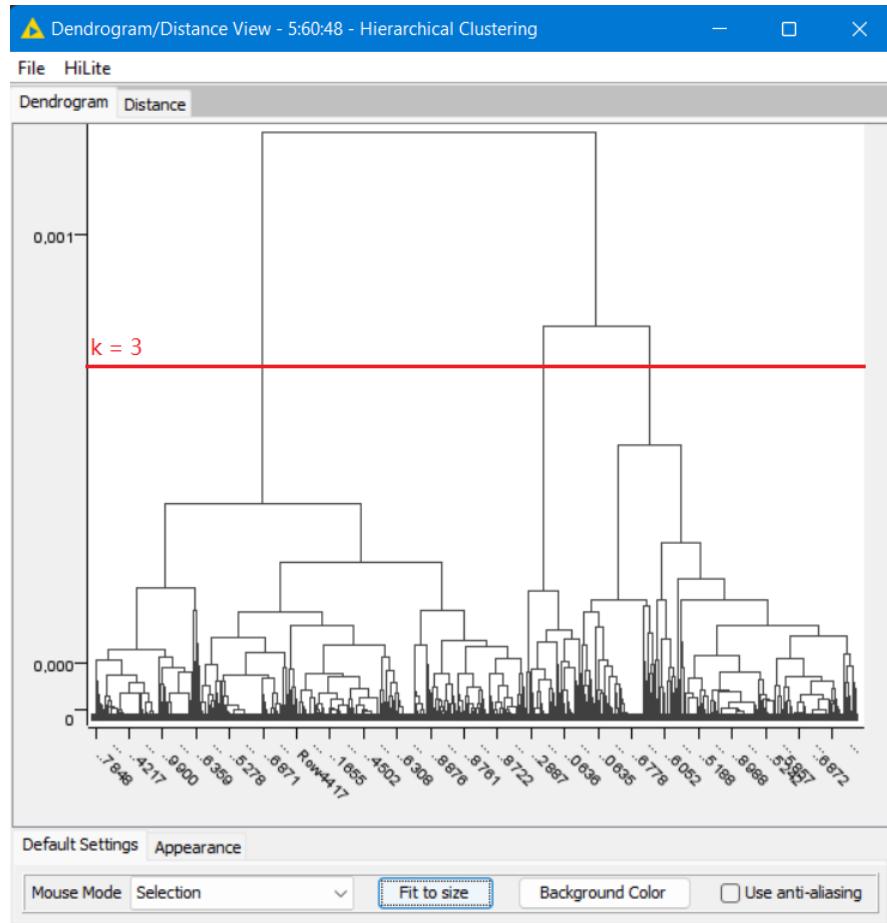
Les clusters n'ont approximativement pas la même taille contrairement à K-means comme on remarque pour SINGLE, COMPLETE et AVERAGE pour K=70.

On remarque parfois que le clustering hiérarchique tente de mettre le maximum de points proches dans un même cluster et le reste dans le nombre de clusters restants. Ceci s'est manifesté par des résultats non pertinents pour le clustering SINGLE linkage pour K=3,15 et 30. Les points sont assez proches que le clustering hiérarchique met la majorité dans un seul gros cluster central et distribue le reste parfois dans des clusters constitués d'un seul point. Le clustering avec SINGLE linkage est rapide et peut bien fonctionner sur des données non globulaires, mais cette méthode est sensible au bruit.

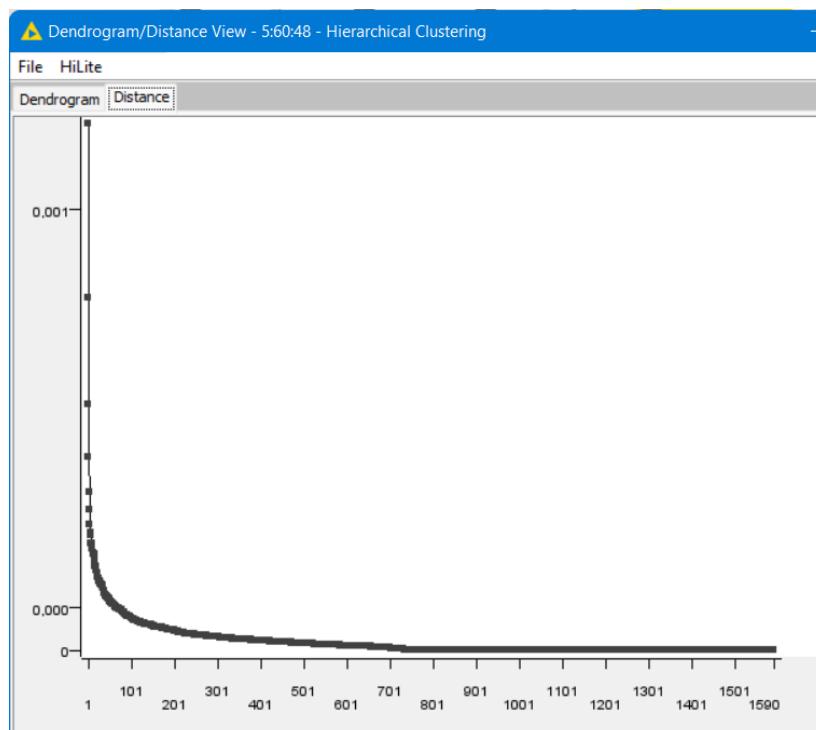
En effet, on remarque par comparaison des dendrogrammes de AVERAGE, COMPLETE et SINGLE pris pour la valeur K=3, ce dernier a tendance à ajouter un point à la fois au cluster créant ainsi des clusters filandreux.

Pour K=3, les dendrogrammes et les courbes de distances de AVERAGE et COMPLETE sont respectivement représentés et sont comparables.

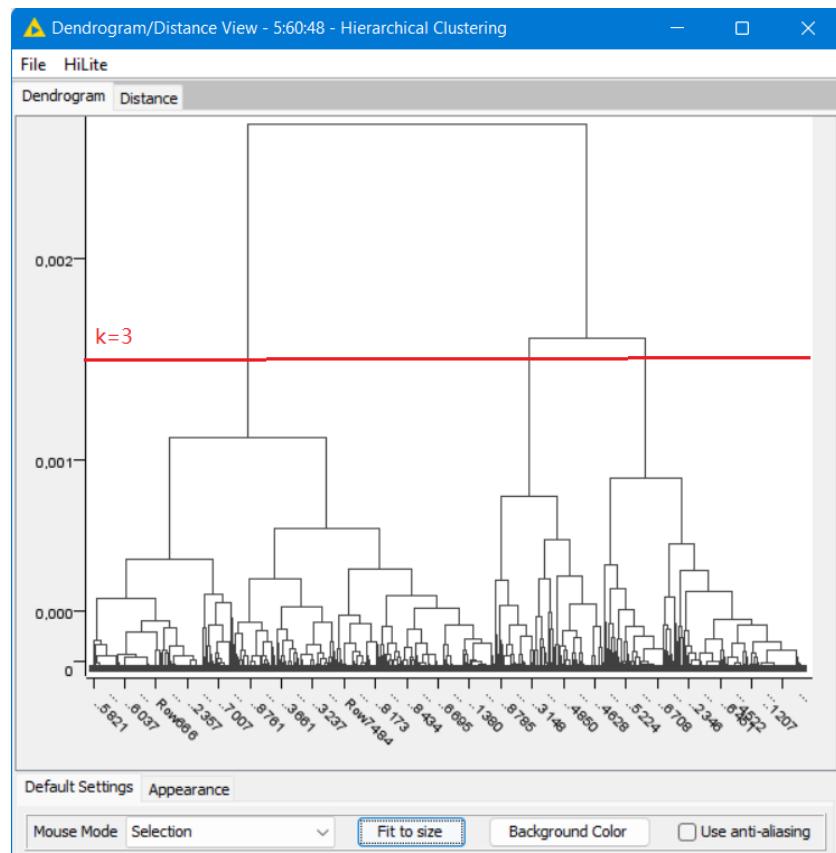
### Dendrogramme de AVERAGE K=3



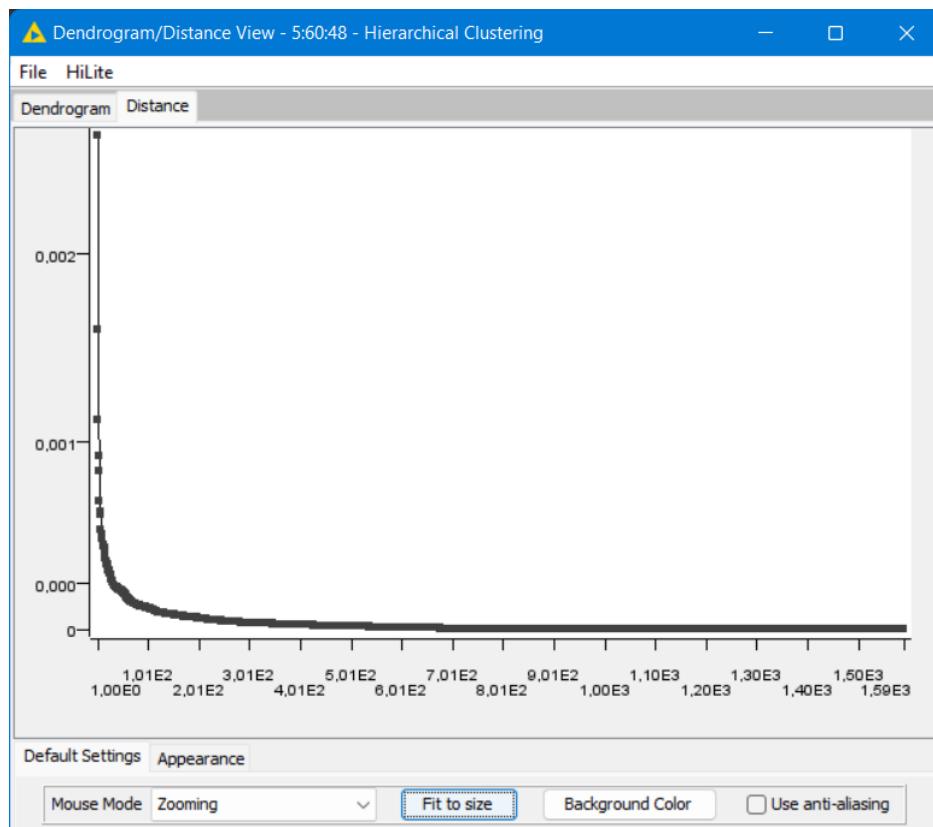
Courbe des distances de AVERAGE pour K=3



## Dendrogramme de COMPLETE pour K=3

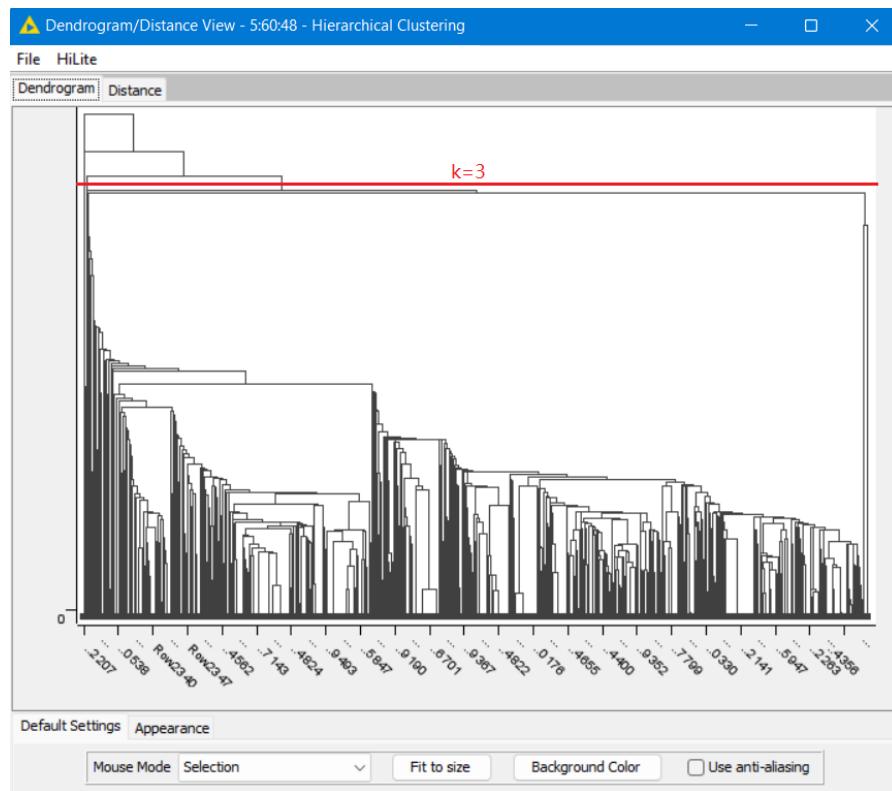


## Courbe des distances de COMPLETE pour K=3

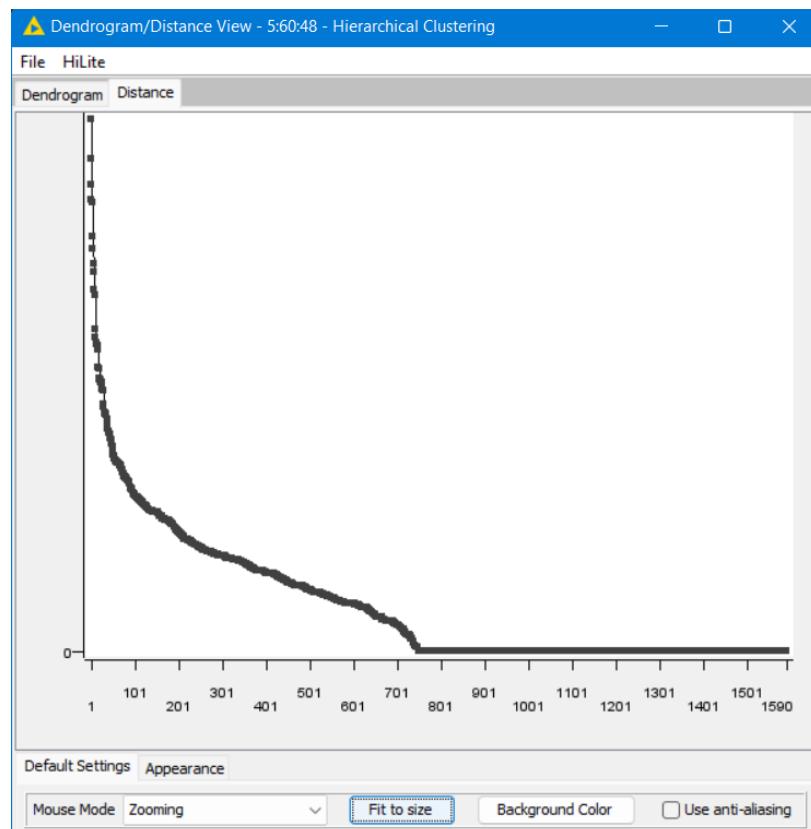


Alors que pour SINGLE pour la même valeur de K, on remarque la différence dans la façon de construire les clusters.

Dendrogramme de SINGLE pour K=3



Courbe des distances de SINGLE pour K=3



Avec AVERAGE, COMPLETE et SINGLE pour des K plus importants, il constitue des clusters un peu plus équilibrés et cohérents mais pas nécessairement de même taille comme on remarque avec K= 70 SINGLE (le gros cluster a été divisé, d'où un résultat plus pertinent), K= 3 COMPLETE et K= 3 AVERAGE. On distingue un cluster pour l'opéra national de ville et le reste des clusters caractérisent l'hôtel de ville.

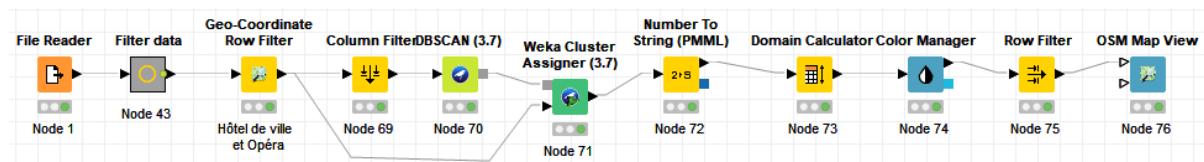
Pour des valeurs plus importantes de K, prenant AVERAGE pour K=15, on remarque qu'il a tendance à organiser les clusters en colonnes. Les clusterings avec AVERAGE et COMPLETE fonctionnent mieux que SINGLE et ne se limitent pas à des clusters globulaires comme on remarque par exemple pour K=15 et K=30. Les résultats des 2 méthodes sont comparables.

Finalement, on conclut que le clustering hiérarchique prend plus de temps que K-means (moins performant) puisque sa complexité en  $n^3$  est supérieure à celle de K-means.

### 3- L'algorithme DBSCAN

L'algorithme DBSCAN est un algorithme qui définit des clusters en regardant la densité locale des points pris en compte. Le nombre de clusters et le nombre de points de bruit d'un résultat de clustering obtenu avec DBSCAN va dépendre des valeurs des paramètres epsilon  $\epsilon$  (les points d'un même cluster se situent dans un certain rayon de proximité) et minPoints minPts (nombres de points dans le rayon de proximité d'un point central).

#### A- Prise en compte de l'espace



#### Explications sur le Workflow

1. Nous utilisons le nœud *Geo-Coordinate Row Filter* pour pouvoir sélectionner des données appartenant à une zone géographique réduite (une zone comportant l'Hôtel de Ville et l'Opéra).
2. Nous utilisons le nœud *Column Filter* pour sélectionner uniquement les attributs intéressants pour l'analyse courante : la latitude et la longitude (*lat* et *long*).
3. Nous utilisons le nœud *DBSCAN (3.7)* en faisant varier les paramètres epsilon  $\epsilon$  et minPoints minPts qui ont un impact sur les clusters résultant de l'exécution de l'algorithme.
4. Nous utilisons le nœud *Weka Cluster Assigner (3.7)* pour pouvoir lier les données générées par *DBSCAN* et par *Geo-Coordinate Row Filter*.

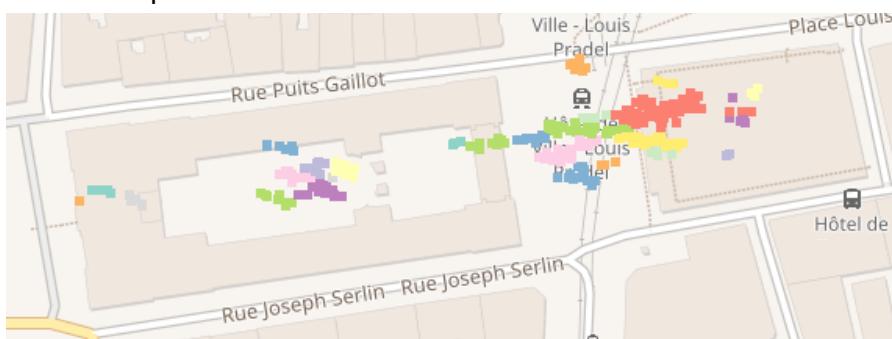
5. Nous utilisons le nœud *Number To String* pour pouvoir convertir les valeurs de la colonne Winner Cluster de integer à string. DBScan créé un label de cluster qui est par défaut un numérique, ce n'est pas idéal. Ce composant nous permet de passer de string à numérique et vice-versa.
6. Nous utilisons le nœud *Domain Calculator* pour pouvoir enlever la restriction sur le nombre possible de valeurs de clusters fixé à 60, pour pouvoir avoir plus de 60 clusters.
7. Nous utilisons le nœud *Color Manager* pour pouvoir attribuer une couleur à chaque cluster et les identifier sur la carte.
8. Nous utilisons le nœud *Row Filter* pour filtrer le bruit, c'est-à-dire enlever les points qui n'appartiennent à aucun cluster.
9. Nous utilisons le nœud *OSM Map View* pour pouvoir visualiser le résultat du clustering sur une carte.

### Résultats du clustering

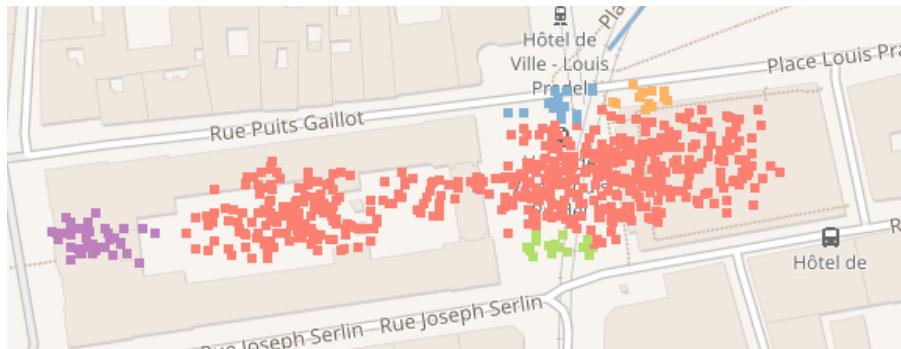
Dans un premier temps, nous avons testé plusieurs valeurs de `minPts` et  $\epsilon$  pour l'exécution de DBSCAN pour pouvoir retenir des valeurs qui donnent des résultats de clustering pertinents. Parmi les couples de valeurs peu pertinents nous en avons choisi deux à présenter :



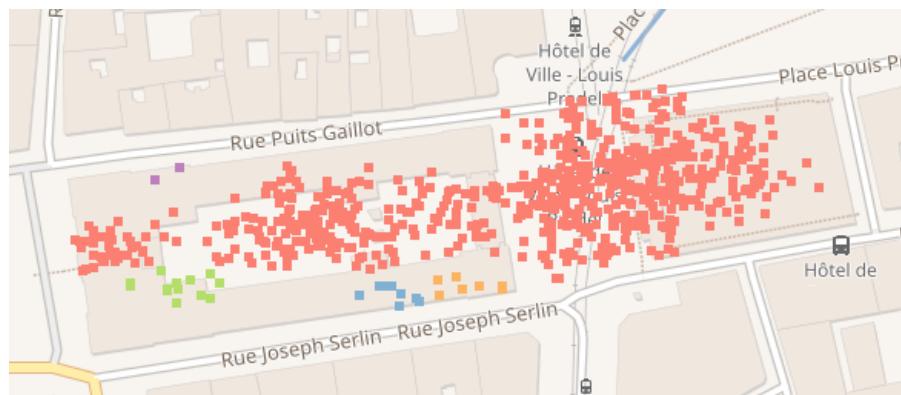
Pour  $\epsilon = 0.01$  et `minPts` = 10 : nous observons des clusters qui semblent trop petits pour pouvoir correspondre à un centre d'intérêt et beaucoup de points ont été enlevés car ils correspondaient à des points de bruit.



Pour  $\epsilon = 0.02$  et `minPts` = 10 : nous observons un meilleur clustering que pour  $\epsilon = 0.01$  mais certains clusters semblent toujours trop petits pour pouvoir correspondre à un centre d'intérêt et nous avons toujours beaucoup de points qui ont été enlevés car ils correspondaient à des points de bruit.



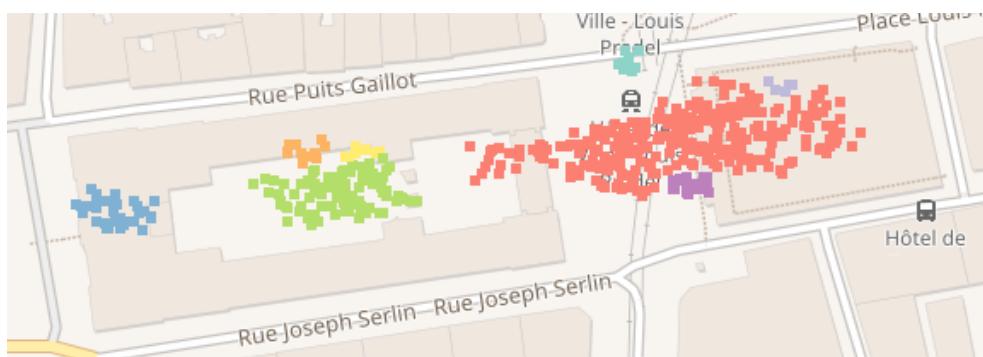
Pour  $\varepsilon = 0.05$  et  $minPts = 16$  : nous observons un gros cluster principal ce qui nous semble être un résultat de clustering pas assez précis et donc pas assez pertinent par rapport aux observations que nous avons pu faire par la suite.



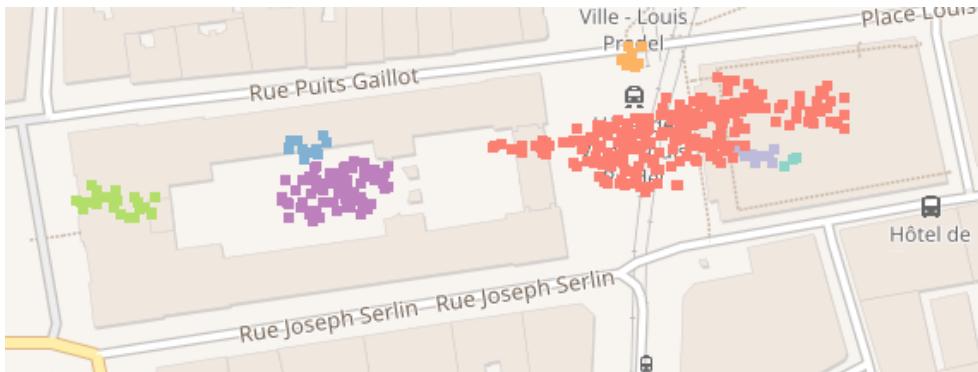
Pour  $\varepsilon = 0.06$  et  $minPts = 10$  : mêmes observations que pour le résultat précédent de clustering pour  $\varepsilon = 0.05$  et  $minPts = 16$ .

Dans un second temps, nous avons décidé de faire varier la valeur du paramètre  $\varepsilon$  tel que :  $\varepsilon \in \{0.03, 0.04\}$  et pour chacune de ces valeurs de  $\varepsilon$  nous faisons varier la valeur du paramètre  $minPts$  tel que :  $minPts \in \{10, 12, 14, 16\}$ . Ainsi, nous avons obtenu 8 résultats de clustering. Cependant, les résultats pour des valeurs de  $minPts$  de 10 et 12, et de 14 et 16 étant fortement similaires nous avons fait le choix de représenter uniquement les clusterings pour  $minPts = 10$  et pour  $minPts = 16$ . Les résultats sont les suivants :

- Pour  $\varepsilon = 0.03$  :

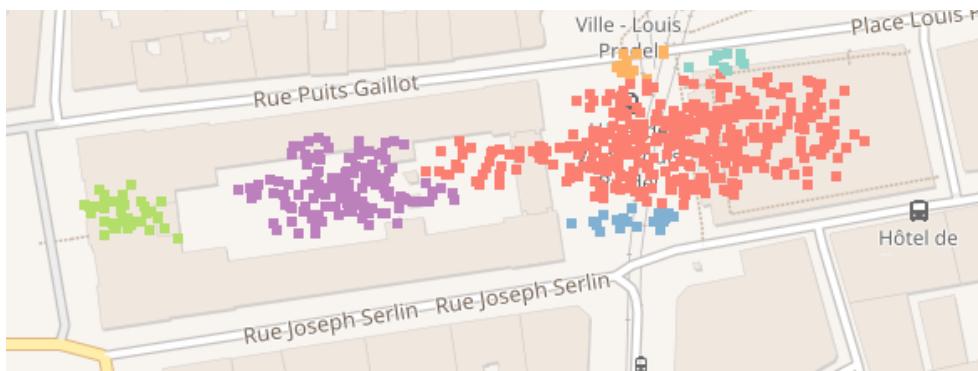


Pour  $\text{minPts} = 10$ .

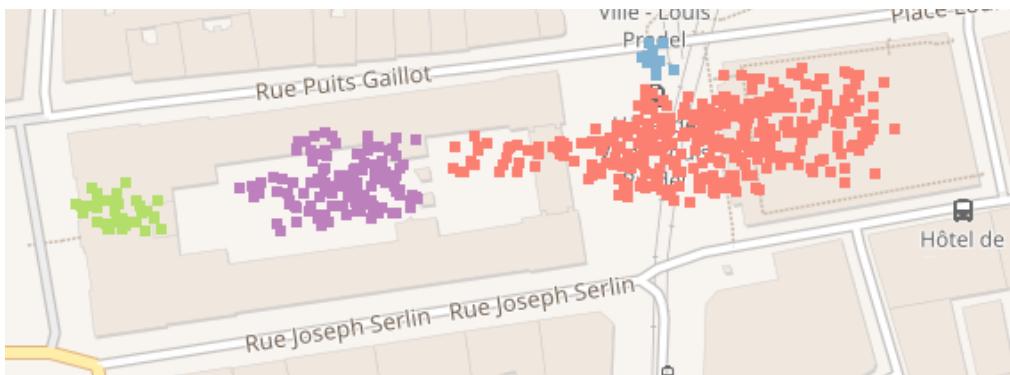


Pour  $\text{minPts} = 16$ .

- Pour  $\varepsilon = 0.04$  :



Pour  $\text{minPts} = 10$ .



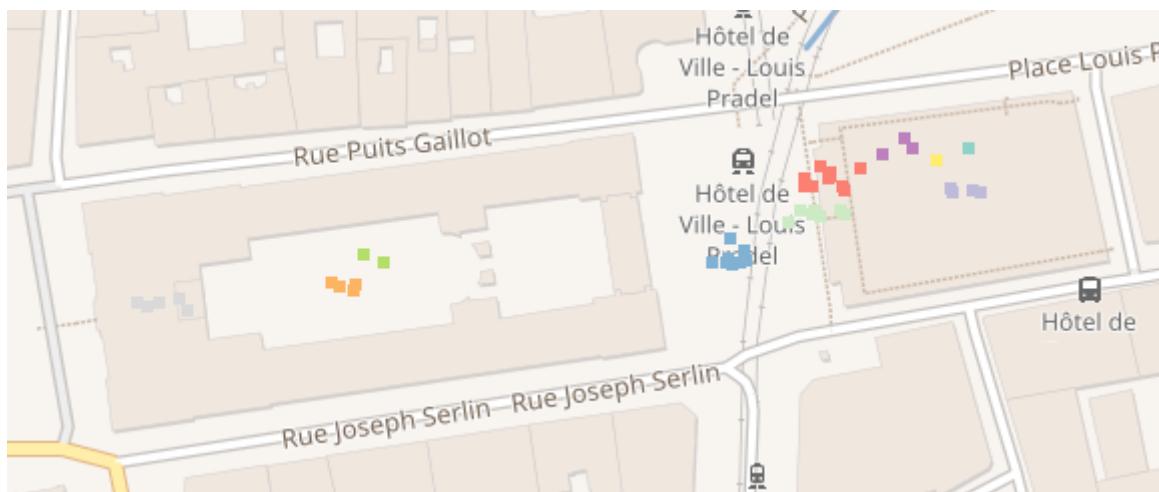
Pour  $\text{minPts} = 16$ .

Nous observons un clustering intéressant pour  $\varepsilon = 0.03$  et  $\varepsilon = 0.04$ . En effet, pour ces valeurs de  $\varepsilon$  on arrive à distinguer deux clusters principaux au niveau de l'Hôtel de ville et de l'Opéra. La valeur de  $\varepsilon$  est donc ni trop petite (des clusters sont considérés seulement pour des points très proches géographiquement), ni trop grande (des clusters sont considérés pour des points plutôt éloignés géographiquement). Nous pouvons également observer que pour  $\text{minPts} = 16$  dans les deux cas le clustering observé est moins satisfaisant que pour  $\text{minPts} = 10$  car nous avons plus de points qui sont considérés comme du bruit et qui ne sont pas pris en compte dans les clusters.

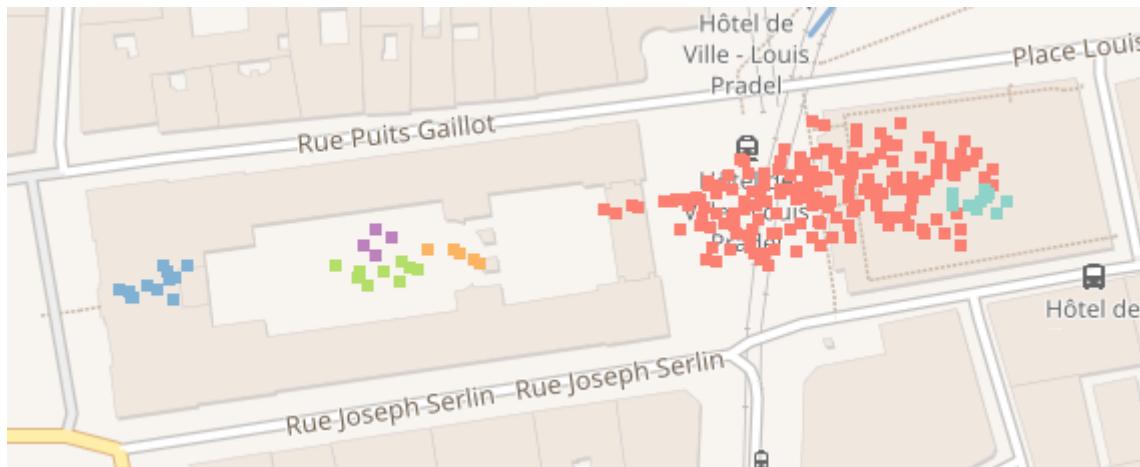
Remarquons que ces valeurs de  $\varepsilon$  et  $\text{minPts}$  sont pertinentes ici pour la zone géographique que nous avons utilisé mais ces paramètres pourront ne pas être pertinents sur une autre zone qui possède une densité de points différentes.

De plus, dans le cas de certains résultats nous pouvons remarquer que parfois nous observons un cluster qui semble correspondre à un unique point mais si nous regardons le nombre de points qui le constituent nous remarquons qu'il y a plusieurs points qui correspondent à des photos qui ont été prises au même endroit. La carte n'est donc pas l'unique visualisation qui nous permet d'analyser le clustering mais nos yeux y sont habitués.

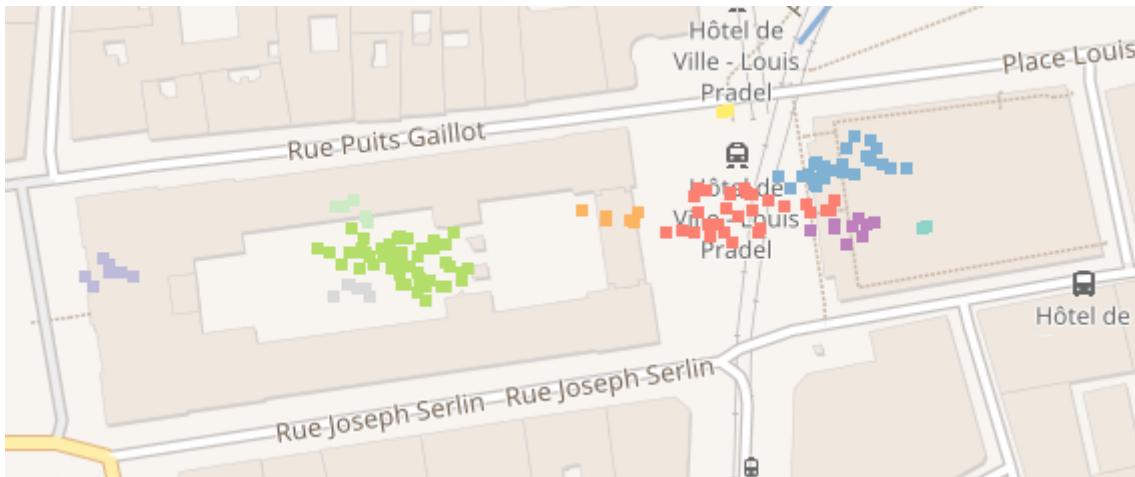
Enfin, toujours en utilisant un clustering spatial nous avons fait le choix d'observer les résultats de plusieurs clusterings en filtrant les résultats pour sélectionner uniquement une partie de la journée (matin, après-midi ou soirée). Pour cela, nous avons utilisé un noeud *Row Filter* et nous avons filtré les valeurs de *date\_taken\_hour*. Nous obtenons les résultats suivants en fixant  $\epsilon = 0.04$  et  $minPts = 10$ :



Pour le matin :  $6 \leq date\_taken\_hour \leq 12$ .



Pour l'après-midi :  $13 \leq date\_taken\_hour \leq 18$ .

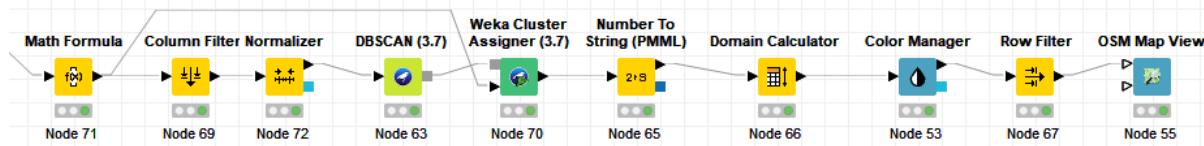


Pour la soirée :  $19 \leq date\_taken\_hour \leq 23$ .

Pour les 3 périodes, nous observons des clusters de tailles différentes mais les zones qu'ils couvrent sont similaires. Nous pouvons donc observer que l'Opéra, de même que l'Hôtel de ville et la place entre les 2 monuments, est une zone qui est couverte par un ou plusieurs clusters dans la journée ce qui est une information intéressante. Nous retrouvons l'information que nous avions trouvé lors du calcul de statistiques sur les données de cette zone géographique : plus de photos sont prises dans l'après-midi. Durant cette période, on observe un cluster principal qui couvre la zone de l'Opéra. Pour la période du matin on observe plusieurs clusters qui semblent être petits et pour la période du soir on observe plusieurs clusters qui semblent être de tailles moyennes.

## B- Prise en compte de l'espace et du temps

Nous avons pris la décision d'élargir notre analyse des clusters en prenant en compte une nouvelle dimension : la dimension temporelle. Cela va nous permettre de visualiser d'autres informations. Par exemple, nous allons pouvoir voir l'effet de différents évènements annuels, mensuels, quotidiens... ayant lieu à Lyon. Pour prendre en compte la dimension temporelle dans notre clustering, nous avons repris une configuration similaire à celle pour le clustering DBSCAN qui prend en compte uniquement la dimension spatiale des données et nous avons créé une nouvelle variable temps.



### Explications supplémentaires sur le Workflow

Certains éléments utilisés dans le Workflow ci-dessus ont la même fonction que lorsqu'ils ont été utilisés dans le Workflow de la partie III.3.A. Les éléments qui se situent à gauche du Workflow sont les nœuds et métanodes : *File Reader*, *Filter Data* et *Geo-Coordinate Row Filter* (de gauche à droite).

1. Nous utilisons le nœud *Math Formula* pour créer la variable `temps` qui convertit les valeurs correspondant à `date_taken` (date à laquelle la photo a été prise) en secondes selon la formule suivante :

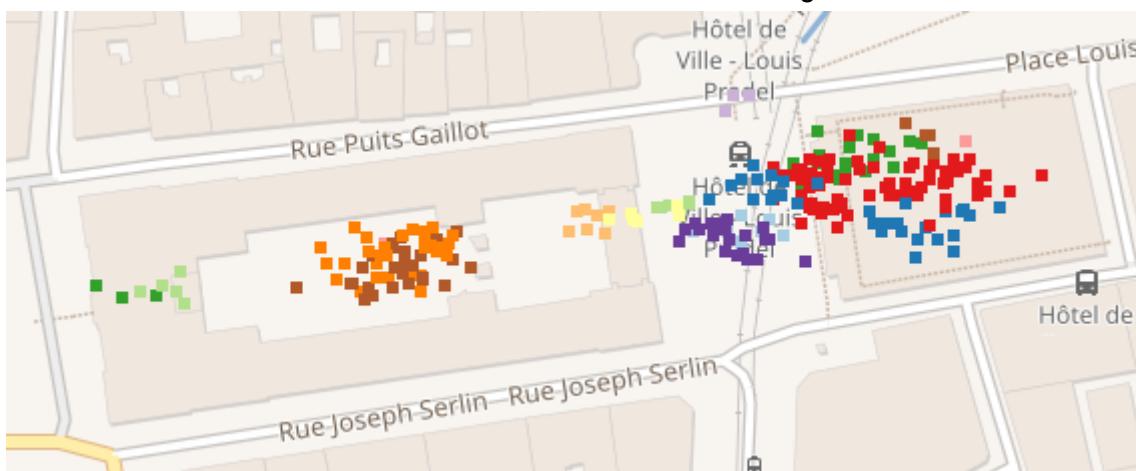
$\$date\_taken\_minute\$ * 60 + \$date\_taken\_hour\$ * 3600 + \$date\_taken\_day\$ * 24 * 3600$   
 $+ \$date\_taken\_month\$ * 31 * 24 * 3600 + \$date\_taken\_year\$ * 365 * 24 * 3600$

2. Nous utilisons ensuite le nœud *Column Filter* pour sélectionner uniquement les attributs intéressants (temps, latitude et longitude) pour l'analyse courante : `temp`, `lat` et `long`.
3. Nous avons utilisé le nœud *Normalizer* pour pouvoir ramener toutes les valeurs de ces 3 attributs entre 0 et 1. Cette normalisation nous permet de donner autant de poids donc d'importance aux 3 dimensions dans le clustering. Remarquons qu'on pourrait vouloir donner plus de poids à la dimension temporelle plutôt qu'à la dimension spatiale et inversement et donc utiliser un DBSCAN paramétrique.

Remarquons qu'ici nous ne prenons pas en compte la date à laquelle la photo a été mise en ligne (`date_upload`) sur Flickr mais bien la date à laquelle elle a été prise car c'est cette dernière qui a de la valeur et qui porte de l'information utile pour notre analyse.

## Résultats du clustering

Après avoir regardé plusieurs résultats de clustering pour différentes valeurs de  $\varepsilon$  et `minPts`, nous avons décidé de retenir le résultat du clustering obtenu avec  $\varepsilon = 0.06$  et `minPts = 10`. Nous observons d'abord les résultats du clustering sur une carte :



Remarquons que la visualisation des clusters sur une carte lorsque nous utilisons un clustering sur un timestamp n'est pas la visualisation la plus pertinente car nous pouvons avoir 2 points au même endroit géographiquement mais qui appartiennent à 2 clusters différents du fait du timestamp. Nous avons donc décidé d'observer les données obtenus pour ce clustering dans un tableau qui contient toutes les informations de chaque photo :

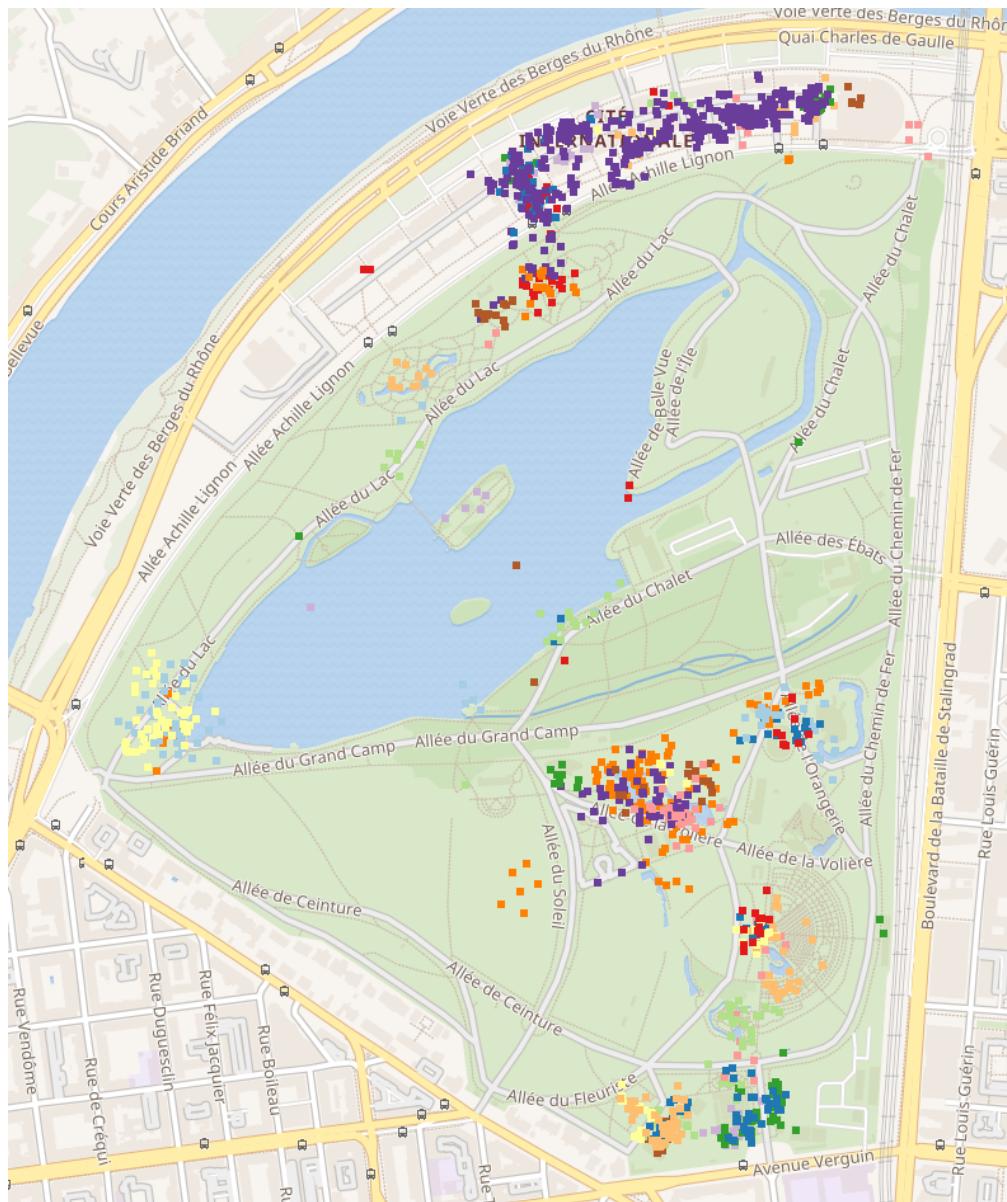
|           |                |              |        |       |                                  |                 |    |    |    |    |      |
|-----------|----------------|--------------|--------|-------|----------------------------------|-----------------|----|----|----|----|------|
| Row124548 | 3.150984261... | 27476137@N04 | 45.768 | 4.837 | 2016,fdf,lyon,n1,supermilimil    | P1030160        | 25 | 23 | 9  | 12 | 2016 |
| Row124551 | 3.151000930... | 27476137@N04 | 45.768 | 4.837 | 2016,fdf,lyon,n1,supermilimil    | P1030130        | 14 | 23 | 9  | 12 | 2016 |
| Row124576 | 3.154103413... | 27476137@N04 | 45.768 | 4.837 | 2016,fdf,lyon,n1,supermilimil    | Opéra . lyon    | 24 | 23 | 9  | 12 | 2016 |
| Row124677 | 3.154113330... | 27476137@N04 | 45.768 | 4.837 | 2016,fdf,lyon,n1,supermilimil    | P1030140        | 16 | 23 | 9  | 12 | 2016 |
| Row125254 | 3.175096791... | 91783102@N07 | 45.768 | 4.837 | ?                                | Orchestra Pit   | 52 | 12 | 29 | 1  | 2017 |
| Row125255 | 3.175099064... | 91783102@N07 | 45.768 | 4.837 | ?                                | Public Hall     | 51 | 12 | 29 | 1  | 2017 |
| Row125256 | 3.175102417... | 91783102@N07 | 45.768 | 4.837 | ?                                | Great Hall      | 43 | 12 | 29 | 1  | 2017 |
| Row125297 | 3.178266942... | 91783102@N07 | 45.768 | 4.837 | ?                                | Seats           | 32 | 14 | 29 | 1  | 2017 |
| Row125298 | 3.178268244... | 91783102@N07 | 45.768 | 4.837 | ?                                | Steel Stairs    | 57 | 12 | 29 | 1  | 2017 |
| Row126707 | 3.221536777... | 91783102@N07 | 45.768 | 4.837 | ?                                | Seats           | 29 | 14 | 29 | 1  | 2017 |
| Row127089 | 3.235207690... | 27476137@N04 | 45.768 | 4.837 | 2016,fdf,lyon,n1,supermilimil    | P1030147        | 18 | 23 | 9  | 12 | 2016 |
| Row127354 | 3.244185163... | 91783102@N07 | 45.768 | 4.837 | ?                                | Great Hall ...  | 48 | 12 | 29 | 1  | 2017 |
| Row127454 | 3.247186382... | 91783102@N07 | 45.768 | 4.837 | ?                                | Great Hall ...  | 47 | 12 | 29 | 1  | 2017 |
| Row127799 | 3.255413234... | 91783102@N07 | 45.768 | 4.837 | ?                                | Access to th... | 50 | 12 | 29 | 1  | 2017 |
| Row127800 | 3.255418924... | 91783102@N07 | 45.768 | 4.837 | ?                                | Rehearsal Pi... | 36 | 11 | 29 | 1  | 2017 |
| Row127840 | 3.257422141... | 35971282@N00 | 45.768 | 4.837 | 2017,france,rhone,lyon,hotel,... | Glasses         | 2  | 22 | 31 | 1  | 2017 |
| Row127890 | 3.259475680... | 91783102@N07 | 45.768 | 4.837 | ?                                | Great Hall      | 43 | 12 | 29 | 1  | 2017 |
| Row127891 | 3.259479062... | 91783102@N07 | 45.768 | 4.837 | ?                                | Rehearsal Pi... | 34 | 11 | 29 | 1  | 2017 |
| Row127980 | 3.261938148... | 35971282@N00 | 45.768 | 4.837 | 2017,france,rhone,lyon,hotel,... | Hotel de Ville  | 52 | 21 | 31 | 1  | 2017 |
| Row133661 | 3.458883167... | 39415781@N06 | 45.768 | 4.836 | lyon,france,burgundy,bourgog...  | Opéra Nouv...   | 10 | 14 | 4  | 6  | 2017 |

Dans ce tableau, et plus précisément dans l'encadré marron qui comporte les informations concernant la date à laquelle la photo a été prise, nous pouvons remarquer que les photos appartenant à ce cluster ont été prise dans la même zone et correspondent toutes à une certaine période temporelle (décembre 2016 - janvier 2017) et plusieurs de ces photos ont été prise pendant la même journée. Ce clustering semble donc intéressant et peut contenir de l'information concernant un évènement.

|          |                |              |        |       |                                       |                                     |    |    |    |    |      |
|----------|----------------|--------------|--------|-------|---------------------------------------|-------------------------------------|----|----|----|----|------|
| Row37154 | 6.90291533E9   | 70627699@N00 | 45.768 | 4.835 | ?                                     | IMG_1374s                           | 35 | 11 | 13 | 2  | 2012 |
| Row37295 | 6.91268031E9   | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 57 | 14 | 8  | 4  | 2012 |
| Row37296 | 6.912685518E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar roquette au Street day         | 58 | 14 | 8  | 4  | 2012 |
| Row37300 | 6.912712492E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 13 | 15 | 8  | 4  | 2012 |
| Row37301 | 6.912714718E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar marin au Street day            | 13 | 15 | 8  | 4  | 2012 |
| Row37302 | 6.912720086E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 15 | 15 | 8  | 4  | 2012 |
| Row37303 | 6.912725784E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 16 | 15 | 8  | 4  | 2012 |
| Row37304 | 6.912736462E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 17 | 15 | 8  | 4  | 2012 |
| Row37310 | 6.912776254E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 4  | 17 | 8  | 4  | 2012 |
| Row37311 | 6.91277874E9   | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Kube de Knar au Street day          | 5  | 17 | 8  | 4  | 2012 |
| Row37312 | 6.912784894E9  | 21031983@N03 | 45.768 | 4.835 | statue,nikon,lyon,hoteldeville,...    | Knar au Street day                  | 9  | 17 | 8  | 4  | 2012 |
| Row37313 | 6.912788498E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 13 | 17 | 8  | 4  | 2012 |
| Row37314 | 6.912793344E9  | 21031983@N03 | 45.768 | 4.835 | statue,nikon,lyon,hoteldeville,...    | Knar au Street day                  | 21 | 17 | 8  | 4  | 2012 |
| Row37324 | 6.912668598E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar aviateur au Street day         | 29 | 18 | 8  | 4  | 2012 |
| Row37325 | 6.912876088E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar rose au Street day             | 29 | 18 | 8  | 4  | 2012 |
| Row38988 | 7.058758899E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 57 | 14 | 8  | 4  | 2012 |
| Row38987 | 7.058765433E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar sioux au Street day            | 58 | 14 | 8  | 4  | 2012 |
| Row38988 | 7.058777060E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar biscuit au Street day          | 59 | 14 | 8  | 4  | 2012 |
| Row38994 | 7.058792323E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar aviateur au Street day         | 12 | 15 | 8  | 4  | 2012 |
| Row38995 | 7.058799521E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar sioux au Street day            | 14 | 15 | 8  | 4  | 2012 |
| Row38996 | 7.058805695E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 15 | 15 | 8  | 4  | 2012 |
| Row38997 | 7.058808795E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 15 | 15 | 8  | 4  | 2012 |
| Row38998 | 7.058815527E9  | 21031983@N03 | 45.768 | 4.835 | statue,nikon,lyon,hoteldeville,...    | Knar au Street day                  | 17 | 15 | 8  | 4  | 2012 |
| Row39007 | 7.058863875E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 7  | 17 | 8  | 4  | 2012 |
| Row39008 | 7.058872555E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 15 | 17 | 8  | 4  | 2012 |
| Row39009 | 7.058880325E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 22 | 17 | 8  | 4  | 2012 |
| Row39010 | 7.058883049E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar au Street day                  | 23 | 17 | 8  | 4  | 2012 |
| Row39026 | 7.058947725E9  | 21031983@N03 | 45.768 | 4.835 | nikon,hoteldeville,nikkor,d500        | Hotel de ville                      | 15 | 18 | 8  | 4  | 2012 |
| Row39027 | 7.058954807E9  | 21031983@N03 | 45.768 | 4.835 | nikon,lyon,hoteldeville,nikkor,...    | Knar souris verte au Street day     | 29 | 18 | 8  | 4  | 2012 |
| Row43421 | 7.77922504E9   | 29262809@N08 | 45.768 | 4.835 | france,hôpital,villages,tourisme,...  | Lyon - La Garonne de Bartholdi (... | 11 | 15 | 9  | 8  | 2012 |
| Row39765 | 7.168776865E9  | 37290448@N04 | 45.768 | 4.836 | city,windows,blackandwhite,france,... | L'Opéra                             | 42 | 3  | 2  | 6  | 2012 |
| Row39766 | 7.168779769E9  | 37290448@N04 | 45.768 | 4.836 | city,france,water,glass,stone,...     | Artificial Creek                    | 39 | 3  | 2  | 6  | 2012 |
| Row40701 | 7.353980386E9  | 37290448@N04 | 45.768 | 4.836 | door,wood,city,france,french,...      | Ornate Wooden Door                  | 44 | 3  | 2  | 6  | 2012 |
| Row48446 | 8.264567298E9  | 90054633@N02 | 45.768 | 4.836 | de,lyon,des,fête,fontaine,ville,...   | Fontaine de l'Hôtel de ville Lyon   | 50 | 14 | 11 | 12 | 2012 |
| Row50795 | 8.457402219E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 49 | 20 | 16 | 8  | 2012 |
| Row50796 | 8.457402361E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 49 | 20 | 16 | 8  | 2012 |
| Row50797 | 8.457402523E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 48 | 20 | 16 | 8  | 2012 |
| Row50798 | 8.457402691E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 48 | 20 | 16 | 8  | 2012 |
| Row50799 | 8.457402983E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Lyon, Hôtel de Ville                | 46 | 20 | 16 | 8  | 2012 |
| Row50800 | 8.457403265E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 44 | 20 | 16 | 8  | 2012 |
| Row50801 | 8.457403461E9  | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 44 | 20 | 16 | 8  | 2012 |
| Row50815 | 8.45879471E9   | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Opéra National de Lyon              | 48 | 20 | 16 | 8  | 2012 |
| Row50816 | 8.4587954958E9 | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Lyon, Place Louis Pradel            | 45 | 20 | 16 | 8  | 2012 |
| Row50817 | 8.45850548E9   | 40695821@N00 | 45.768 | 4.836 | travel,panorama,france,canon,...      | Lyon, Place Louis Pradel            | 43 | 20 | 16 | 8  | 2012 |

Nous faisons les mêmes observations pour cette partie du tableau. Cependant ici les photos composant ces 2 clusters ont toutes été prises respectivement par 2 utilisateurs : le cluster "rose" est composé majoritairement de photos prises par la même personne le 8 avril 2012, le jour de "Street Day" une exposition à l'hôtel de ville de l'artiste Knar. Le cluster "violet" correspond majoritairement à des photos de l'opéra et de la place Louis Pradel prises par la même personne le 16 août 2012.

Nous avons décidé de regarder les résultats d'un clustering qui prend en compte l'espace et le temps sur une autre zone géographique : le Parc de la Tête d'Or. Nous avons fixé :  $\epsilon = 0.03$  et  $minPts = 16$ . Le résultat observé est le suivant :



Nous observons plusieurs clusters sur cette carte mais nous décidons d'observer les résultats dans un tableau pour mieux les analyser et les interpréter.

|           |                 |               |        |       |                   |                          |    |    |    |    |      |
|-----------|-----------------|---------------|--------|-------|-------------------|--------------------------|----|----|----|----|------|
| Row88794  | 1.634636738...  | 58672892@N06  | 45.775 | 4.856 | ?                 | Parc de la Tête d'Or...  | 2  | 9  | 29 | 3  | 2015 |
| Row89500  | 1.651782147...  | 58672892@N06  | 45.776 | 4.857 | lyon,parcdel...   | Parc de la Tête d'Or...  | 40 | 21 | 13 | 4  | 2015 |
| Row90334  | 1.670934881...  | 58672892@N06  | 45.776 | 4.857 | lyon,magnoli...   | Parc de la Tête d'Or...  | 22 | 12 | 22 | 3  | 2015 |
| Row90503  | 1.675352840...  | 58672892@N06  | 45.776 | 4.857 | lyon,crocus...    | Parc de la Tête d'Or...  | 58 | 11 | 8  | 3  | 2015 |
| Row90650  | 1.67812539E10   | 58672892@N06  | 45.775 | 4.856 | ?                 | Parc de la Tête d'Or...  | 7  | 9  | 29 | 3  | 2015 |
| Row91559  | 1.699438675...  | 66401985@N05  | 45.775 | 4.856 | france,flow...    | ?                        | 22 | 14 | 12 | 4  | 2015 |
| Row91863  | 1.708921046...  | 66401985@N05  | 45.776 | 4.856 | flowers,plan...   | ?                        | 32 | 9  | 26 | 4  | 2015 |
| Row92320  | 1.724944021...  | 66401985@N05  | 45.775 | 4.856 | plants,franc...   | ?                        | 36 | 9  | 26 | 4  | 2015 |
| Row92782  | 1.75504844E10   | 58672892@N06  | 45.775 | 4.857 | fleur,lyon,n...   | Parc de la Tête d'Or...  | 11 | 11 | 16 | 5  | 2015 |
| Row92791  | 1.7555393937... | 10986181@N05  | 45.775 | 4.856 | waterlily,né...   | ?                        | 9  | 11 | 11 | 5  | 2015 |
| Row92792  | 1.755540684...  | 10986181@N05  | 45.775 | 4.856 | waterlily,né...   | ?                        | 11 | 11 | 11 | 5  | 2015 |
| Row94871  | 1.852854011...  | 58672892@N06  | 45.776 | 4.857 | lyon,néroph...    | Parc de la Tête d'Or...  | 41 | 10 | 6  | 6  | 2015 |
| Row97155  | 1.956676341...  | 128069291@... | 45.775 | 4.857 | macro,natur...    | DSC00569                 | 28 | 15 | 24 | 5  | 2015 |
| Row97324  | 1.966380096...  | 58672892@N06  | 45.775 | 4.857 | lyon,libellule... | Parc de la Tête d'Or...  | 17 | 12 | 8  | 7  | 2015 |
| Row98982  | 2.100468580...  | 25407932@N08  | 45.775 | 4.857 | flowers,lyon...   | Flowers in the Parc ...  | 30 | 13 | 13 | 8  | 2015 |
| Row138625 | 3.602287345...  | 112075157@... | 45.775 | 4.857 | filmborn,fra...   | Wall of Flowers          | 53 | 14 | 17 | 6  | 2015 |
| Row82200  | 1.534044426...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 12 | 20 | 6  | 12 | 2014 |
| Row82215  | 1.534309239...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 50 | 20 | 6  | 12 | 2014 |
| Row82218  | 1.534309275...  | 23607756@N03  | 45.779 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 22 | 20 | 6  | 12 | 2014 |
| Row82219  | 1.534309289...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 17 | 20 | 6  | 12 | 2014 |
| Row82220  | 1.534309292...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 17 | 20 | 6  | 12 | 2014 |
| Row82317  | 1.535692187...  | 48551155@N05  | 45.778 | 4.846 | light,france...   | Jardin d'hiver           | 34 | 22 | 6  | 12 | 2014 |
| Row82394  | 1.536565883...  | 37420732@N08  | 45.778 | 4.846 | lyon,lumière...   | Jardin d'hiver           | 11 | 20 | 9  | 12 | 2014 |
| Row82403  | 1.536708547...  | 83212364@N05  | 45.778 | 4.846 | ?                 | Parc de la Tête d'Or     | 33 | 20 | 7  | 12 | 2014 |
| Row82404  | 1.536708855...  | 83212364@N05  | 45.778 | 4.846 | ?                 | Parc de la Tête d'Or     | 33 | 20 | 7  | 12 | 2014 |
| Row82405  | 1.536708960...  | 83212364@N05  | 45.778 | 4.846 | ?                 | Parc de la Tête d'Or     | 34 | 20 | 7  | 12 | 2014 |
| Row82440  | 1.537200326...  | 96968841@N07  | 45.778 | 4.847 | ?                 | IMG_0582                 | 25 | 20 | 6  | 12 | 2014 |
| Row82455  | 1.537406194...  | 41610421@N05  | 45.778 | 4.846 | illumina...       | Jardin d'hiver           | 36 | 19 | 7  | 12 | 2014 |
| Row82456  | 1.537407166...  | 41610421@N05  | 45.778 | 4.846 | illumina...       | Jardin d'hiver           | 42 | 19 | 7  | 12 | 2014 |
| Row82473  | 1.537661914...  | 41610421@N05  | 45.778 | 4.846 | illumina...       | Jardin d'hiver           | 17 | 19 | 7  | 12 | 2014 |
| Row82474  | 1.537663598...  | 41610421@N05  | 45.778 | 4.846 | illumina...       | Jardin d'hiver           | 30 | 19 | 7  | 12 | 2014 |
| Row83845  | 1.550437867...  | 32799292@N00  | 45.778 | 4.846 | france,lyon       | LYG 148                  | 36 | 10 | 7  | 12 | 2014 |
| Row83858  | 1.550700057...  | 32799292@N00  | 45.778 | 4.847 | france,lyon       | LYG 151                  | 37 | 10 | 7  | 12 | 2014 |
| Row83868  | 1.550929305...  | 129383859@... | 45.778 | 4.846 | festival_lights   | IMG_4875                 | 57 | 19 | 6  | 12 | 2014 |
| Row83869  | 1.550929416...  | 129383859@... | 45.778 | 4.846 | festival_lights   | IMG_4872                 | 40 | 19 | 6  | 12 | 2014 |
| Row83886  | 1.551190815...  | 129383859@... | 45.778 | 4.846 | festival_lights   | IMG_4888                 | 4  | 20 | 6  | 12 | 2014 |
| Row83887  | 1.551190858...  | 129383859@... | 45.778 | 4.846 | festival_lights   | IMG_4882                 | 3  | 20 | 6  | 12 | 2014 |
| Row84226  | 1.559400625...  | 101490213@... | 45.778 | 4.846 | park,street...    | Lyon - Parc de la Tê...  | 46 | 16 | 1  | 1  | 2015 |
| Row85026  | 1.575561433...  | 41746313@N04  | 45.778 | 4.847 | lyon,parctét...   | Résistance               | 32 | 14 | 10 | 11 | 2014 |
| Row85107  | 1.577543948...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 16 | 20 | 6  | 12 | 2014 |
| Row85117  | 1.577670539...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 54 | 20 | 6  | 12 | 2014 |
| Row85118  | 1.577670607...  | 23607756@N03  | 45.779 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 20 | 20 | 6  | 12 | 2014 |
| Row85119  | 1.577700357...  | 23607756@N03  | 45.778 | 4.846 | canon,lyon,...    | Lyon - Fête des lumi...  | 15 | 20 | 6  | 12 | 2014 |
| Row85241  | 1.579151393...  | 9822280@N00   | 45.778 | 4.846 | lyon,parcdel...   | Fête de lumières, Parc   | 52 | 19 | 8  | 12 | 2014 |
| Row85242  | 1.579173748...  | 9822280@N06   | 45.778 | 4.847 | lyon,parcdel...   | Fête de lumières, le ... | 50 | 19 | 8  | 12 | 2014 |
| Row85259  | 1.579285023...  | 37985894@N00  | 45.778 | 4.846 | lyon,festival...  | Fête des Lumières 2...   | 9  | 21 | 8  | 12 | 2014 |
| Row85265  | 1.579343961...  | 9822280@N05   | 45.778 | 4.847 | canon5diii,ly...  | Fête de lumières, le ... | 28 | 19 | 8  | 12 | 2014 |
| Row85266  | 1.579345939...  | 48551155@N05  | 45.778 | 4.846 | light,france...   | Jardin d'hiver           | 43 | 22 | 6  | 12 | 2014 |
| Row85311  | 1.579738812...  | 96968841@N07  | 45.778 | 4.847 | ?                 | Parc de la tête d'Or     | 27 | 20 | 6  | 12 | 2014 |
| Row85347  | 1.579927476...  | 37420732@N08  | 45.779 | 4.846 | lyon,lumière...   | Jardin d'hiver           | 11 | 20 | 9  | 12 | 2014 |
| Row85348  | 1.579927788...  | 37420732@N08  | 45.779 | 4.846 | lyon,lumière...   | Jardin d'hiver           | 12 | 20 | 9  | 12 | 2014 |
| Row85375  | 1.579944695...  | 83212364@N05  | 45.778 | 4.847 | ?                 | Parc de la tête d'Or     | 38 | 20 | 7  | 12 | 2014 |
| Row85383  | 1.579954820...  | 37420732@N08  | 45.779 | 4.846 | lyon,lumière...   | Jardin d'hiver           | 12 | 20 | 9  | 12 | 2014 |

Ces deux tableaux, qui correspondent à 2 clusters que nous observons sur la carte (le jaune clair qui se trouve au niveau de la "Porte des Enfants du Rhône" et le orange clair qui se trouve au niveau de la "Serre des plantes carnivores d'Afrique"), nous donnent des informations intéressantes sur ces 2 clusters. Le cluster "jaune clair" correspond à des photos qui ont été prises en début décembre 2014, soit pendant la Fête des Lumières, une fête très touristique à Lyon. Le cluster "orange clair" correspond à des photos de plantes prises en été 2015.

Remarquons que lorsque nous prenons en compte la dimension temporelle et la dimension spatiale pour effectuer un clustering alors on obtient des résultats pertinents avec des clusters qui correspondent à un centre d'intérêt situé dans le temps et l'espace. Cependant, nous obtenons aussi des résultats moins pertinents avec des clusters qui contiennent presque uniquement des photos prises par une seule personne et pendant une période de la journée courte.

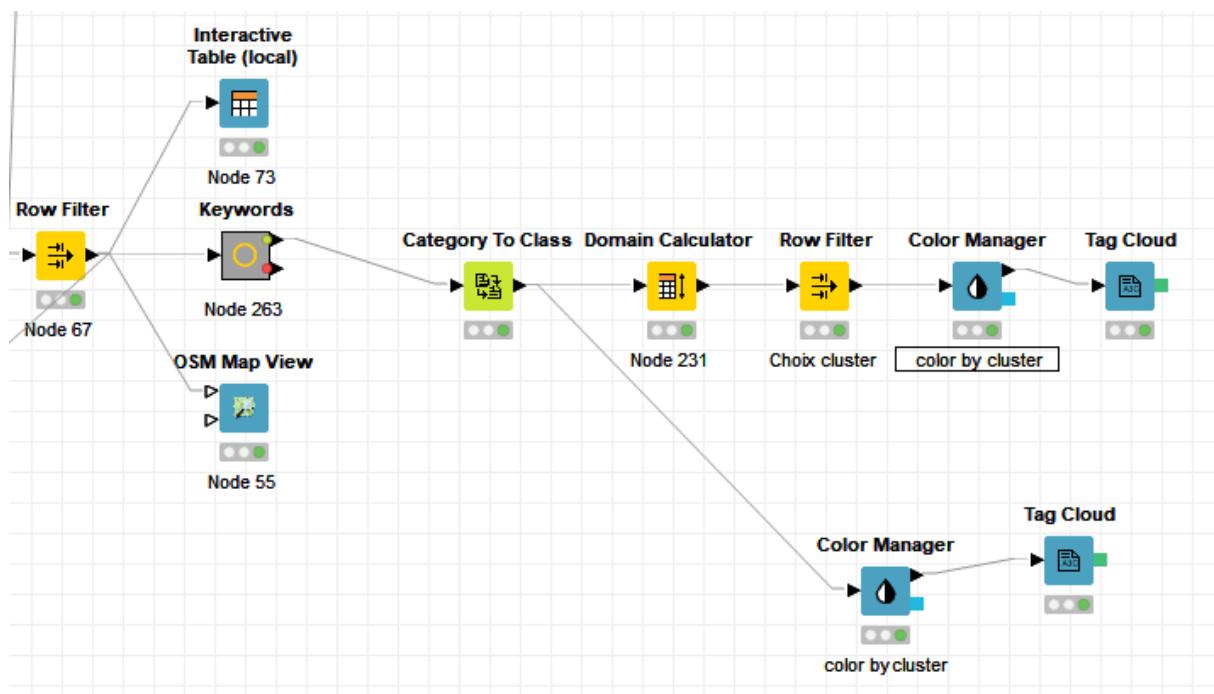
## IV- Manipulation des tags pour la découverte des centres d'intérêts

### 1- Extraction et nettoyage des mots-clés

Explications sur le Workflow

- Workflow Tagcloud pour la zone géographique de l'Opéra et de l'Hôtel de ville

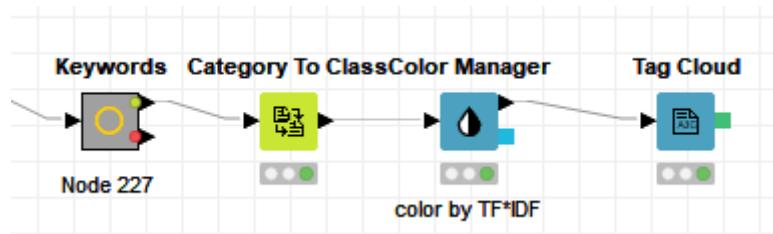
*Ici, le métanode Keywords fait suite à l'enchaînement de nœuds utilisés pour la méthode de clustering DBSCAN.*



Dans ce workflow nous utilisons les noeuds suivants :

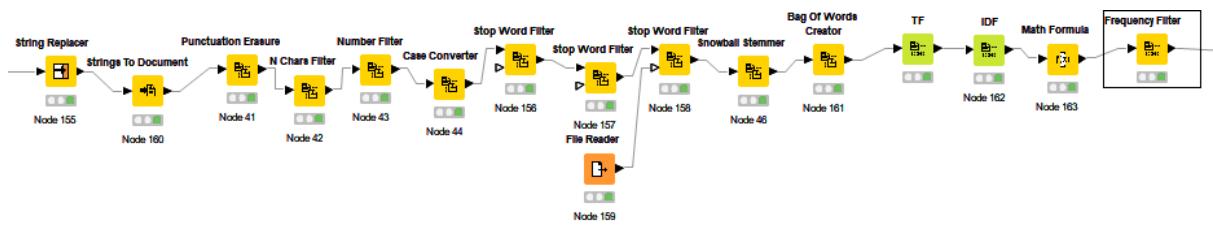
- Le métanode *Keywords* nous permet d'effectuer plusieurs actions sur les tags des différents tuples
- Le nœud *Category To Class* permet de convertir les chaînes spécifiées en documents
- Les noeuds *Domain Calculator*, *Row Filter* et *Color Manager* nous permettent de sélectionner uniquement un cluster parmi tous pour pouvoir observer le Tag Cloud d'un seul cluster
- Le noeud *Color Manager* nous permet d'attribuer une couleur à chaque cluster
- Le noeud *Tag Cloud* nous permet d'observer les tags présents dans les clusters sous forme de nuages de mots

- Workflow Tagcloud pour la totalité des données



Ici, le métanode Keywords fait suite au nœud Group By au début de la suite de nœuds qui permet de supprimer les doublons en prenant en compte toutes les données.

#### Métanode Keywords :



Dans ce métanode, nous avons appliqué certains filtres et effectué certaines modifications au niveau des tags afin de faciliter leur exploitation:

- remplacement des virgules par des espaces
- conversion des strings en document
- supprimer la ponctuation du document
- les tags de moins de 3 caractères sont supprimés
- les tags composés uniquement de nombre(s) sont supprimés
- La rédaction est entièrement en minuscule
- les termes suivants sont éliminés des tags : france, lyon, europe, iphone et geotagged. D'où les 2 noeuds Stop Word Filter, un pour des mots en français, l'autre en anglais
- vérifier le radical du mot PORTER
- *Bag of words creator* : faire une liste des termes uniques, créer autant de lignes qu'il y a des termes individuels dans chaque photo : ex si une photo a 10 termes alors elle sera sur 10 lignes
- calculer la TF (Term Frequency): calcule la fréquence relative des termes (tf) de chaque terme en fonction de chaque document et ajoute une colonne contenant la valeur tf. La valeur est calculée en divisant la fréquence absolue d'un terme selon un document par le nombre de tous les termes de ce document.
- calculer la valeur de l'IDF (Inverse Document Frequency): Calcule trois variantes de la fréquence de document inverse (idf) pour chaque terme en fonction de l'ensemble de documents donné et ajoute une colonne contenant la valeur idf. Idf lisse, normalisé et probabiliste. La variante par défaut est l'idf lisse spécifiée comme suit :  $\text{idf}(t) = \log(1 + (f(D) / f(d, t)))$ , où  $f(D)$  est le

nombre de tous les documents et  $f(d, t)$  est le nombre de documents contenant le terme  $t$ .

- calculer le produit de IF\*IDF à l'aide de MATH FORMULA
  - On filtre par IF\*IDF

## Résultats

En prenant en compte uniquement les clusters de la zone géographique étudiée dans le clustering DBSCAN qui comprend l'Opéra et l'Hôtel de ville on obtient le nuage de mots suivant :



En observant les tags de grande taille dans ce nuage de mots on peut effectuer plusieurs observations. Tout d'abord, on observe des tags comme "fetedeslumier", "opera", "streetdai" qui correspondent à des centres d'intérêt que l'on a pu observer dans nos précédentes analyses. De plus, nous pouvons également observer des tags que nous n'avions pas observé précédemment comme "gaypridelyon". Enfin, nous remarquons que dans ce nuage de tags on observe des tags plus ou moins pertinents dans l'analyse de centres d'intérêt.

En prenant en compte juste un cluster de la zone géographique de l'Opéra et de l'Hôtel de Ville on observe le nuage de mots suivant :



On peut effectuer les mêmes observations que pour le nuage de mots précédent en remarquant que parmi tous ces tags il y en a que nous avions déjà repérés mais également d'autres que nous n'avions pas détecté comme "breakdanc". Enfin, parmi ces tags il existe donc des éléments intéressants pour notre analyse et qui apporte de l'information en plus de celle apportée par un clustering spatial et temporel. Néanmoins, il existe également des tags qui ne sont pas pertinents dans notre analyse comme "martinesodaigu" qui est le nom d'une utilisatrice de Flickr.

En prenant en compte toutes les données après avoir séparé les doublons on obtient le nuage de mots suivant :

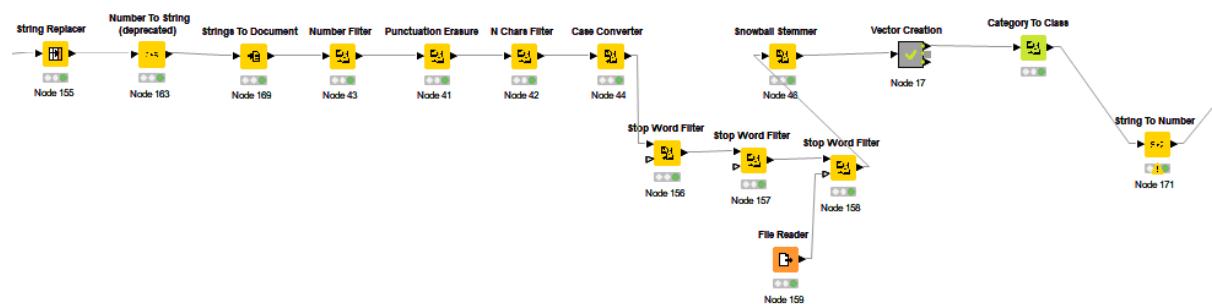


En regardant à l'échelle de la ville on peut distinguer plusieurs tags qui pourraient représenter un centre d'intérêt comme "canalslyon", "leshallesdelyonpaulbocuse". Néanmoins, en regardant l'ensemble des données on risque de ne pas détecter de multiples centres d'intérêt pertinents que l'on peut détecter en regardant des zones plus réduites géographiquement. Donc, en effectuant les deux analyses on peut croiser les données avec les clusterings effectués précédemment et peut être arriver à établir des conclusions.

## 2- Extraction des règles

Explications sur le Workflow

L'extraction des règles a été faite sur le Parc de la Tête d'Or



## Résultats

Association Rules - 0:162 - Association Rule Learner (Borgelt)

File Hilitc Navigation View

Table "default" - Rows: 81 Spec - Columns: 11 Properties Flow Variables

| Row ID | S Consequent     | [...] Antecedent            | I ItemSe... | D Relativ... | D RuleCo... | D Absolut... | D Relativ... | D RuleLift | D RuleLift% | D Absolut... | D Relativ... |
|--------|------------------|-----------------------------|-------------|--------------|-------------|--------------|--------------|------------|-------------|--------------|--------------|
| Row0   | lat (#1)         | [55]                        | 1           | 4.545        | 100         | 1            | 4.55         | 1          | 100         | 22           | 100          |
| Row1   | lat (#1)         | [sido,internetdesobjet]     | 1           | 4.545        | 100         | 1            | 4.55         | 1          | 100         | 22           | 100          |
| Row2   | internetdesobjet | [sido,lat (#1)]             | 1           | 4.545        | 100         | 1            | 4.55         | 11         | 1,100       | 2            | 9.091        |
| Row3   | sido             | [internetdesobjet,lat (#1)] | 1           | 4.545        | 50          | 2            | 9.09         | 11         | 1,100       | 1            | 4.545        |
| Row4   | internetdesobjet | [sido]                      | 1           | 4.545        | 100         | 1            | 4.55         | 11         | 1,100       | 2            | 9.091        |
| Row5   | sido             | [internetdesobjet]          | 1           | 4.545        | 50          | 2            | 9.09         | 11         | 1,100       | 1            | 4.545        |
| Row6   | lat (#1)         | [sido]                      | 1           | 4.545        | 100         | 1            | 4.55         | 1          | 100         | 22           | 100          |
| Row7   | lat (#1)         | [internetdesobjet]          | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row8   | lat (#1)         | [parc,delatéedor,lyon]      | 1           | 4.545        | 100         | 1            | 4.55         | 1          | 100         | 22           | 100          |
| Row9   | lyon             | [parc,delatéedor,lat (#1)]  | 1           | 4.545        | 50          | 2            | 9.09         | 1.1        | 110         | 10           | 45.455       |
| Row10  | lyon             | [parc,delatéedor]           | 1           | 4.545        | 50          | 2            | 9.09         | 1.1        | 110         | 10           | 45.455       |
| Row11  | lat (#1)         | [parc,delatéedor]           | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row12  | franc            | [franc,lyon]                | 1           | 4.545        | 100         | 1            | 4.55         | 1          | 100         | 22           | 100          |
| Row13  | lyon             | [franc,lat (#1)]            | 1           | 4.545        | 50          | 2            | 9.09         | 1.1        | 110         | 10           | 45.455       |
| Row14  | lyon             | [franc]                     | 1           | 4.545        | 50          | 2            | 9.09         | 1.1        | 110         | 10           | 45.455       |
| Row15  | lat (#1)         | [franc]                     | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row16  | lat (#1)         | [dor,tét,parc,...]          | 1           | 4.545        | 100         | 1            | 4.55         | 1          | 100         | 22           | 100          |
| Row17  | lyon             | [dor,tét,parc,...]          | 1           | 4.545        | 50          | 2            | 9.09         | 1.1        | 110         | 10           | 45.455       |
| Row18  | parc             | [dor,tét,lyon,...]          | 1           | 4.545        | 50          | 2            | 9.09         | 1.375      | 137.5       | 8            | 36.364       |
| Row19  | tét              | [dor,parc,lyon,...]         | 1           | 4.545        | 50          | 2            | 9.09         | 1.375      | 137.5       | 8            | 36.364       |
| Row20  | dor              | [tét,parc,lyon,...]         | 1           | 4.545        | 50          | 2            | 9.09         | 1.375      | 137.5       | 8            | 36.364       |
| Row21  | lyon             | [dor,tét,parc]              | 1           | 4.545        | 50          | 2            | 9.09         | 1.1        | 110         | 10           | 45.455       |
| Row22  | parc             | [dor,tét,lyon]              | 1           | 4.545        | 50          | 2            | 9.09         | 1.375      | 137.5       | 8            | 36.364       |
| Row23  | tét              | [dor,parc,lyon]             | 1           | 4.545        | 50          | 2            | 9.09         | 1.375      | 137.5       | 8            | 36.364       |
| Row24  | dor              | [tét,parc,lyon]             | 1           | 4.545        | 50          | 2            | 9.09         | 1.375      | 137.5       | 8            | 36.364       |
| Row25  | lat (#1)         | [dor,tét,parc]              | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row26  | parc             | [dor,tét,lat (#1)]          | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row27  | tét              | [dor,parc,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row28  | dor              | [tét,parc,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row29  | parc             | [dor,tét]                   | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row30  | tét              | [dor,parc]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row31  | dor              | [tét,parc]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row32  | lat (#1)         | [dor,tét,lyon]              | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row33  | lyon             | [dor,tét,lat (#1)]          | 2           | 9.091        | 50          | 4            | 18.2         | 1.1        | 110         | 10           | 45.455       |
| Row34  | tét              | [dor,lyon,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row35  | dor              | [tét,lyon,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row36  | lyon             | [dor,tét]                   | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row37  | tét              | [dor,lyon]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row38  | dor              | [tét,lyon]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row39  | lat (#1)         | [dor,tét]                   | 4           | 18.182       | 100         | 4            | 18.2         | 1          | 100         | 22           | 100          |
| Row40  | tét              | [dor,lat (#1)]              | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row41  | dor              | [tét,lat (#1)]              | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row42  | tét              | [dor]                       | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row43  | dor              | [tét]                       | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row44  | lat (#1)         | [dor,parc,lyon]             | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row45  | lyon             | [dor,parc,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.1        | 110         | 10           | 45.455       |
| Row46  | parc             | [dor,lyon,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row47  | dor              | [parc,lyon,lat (#1)]        | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row48  | lyon             | [dor,parc]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.1        | 110         | 10           | 45.455       |
| Row49  | parc             | [dor,lyon]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row50  | dor              | [parc,lyon]                 | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row51  | lat (#1)         | [dor,parc]                  | 4           | 18.182       | 100         | 4            | 18.2         | 1          | 100         | 22           | 100          |
| Row52  | parc             | [dor,lat (#1)]              | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row53  | dor              | [parc,lat (#1)]             | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row54  | parc             | [dor]                       | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row55  | dor              | [parc]                      | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row56  | lat (#1)         | [dor,lyon]                  | 4           | 18.182       | 100         | 4            | 18.2         | 1          | 100         | 22           | 100          |
| Row57  | lyon             | [dor,lat (#1)]              | 4           | 18.182       | 50          | 8            | 36.4         | 1.1        | 110         | 10           | 45.455       |
| Row58  | lyon             | [dor]                       | 4           | 18.182       | 50          | 8            | 36.4         | 1.1        | 110         | 10           | 45.455       |
| Row59  | lat (#1)         | [dor]                       | 8           | 36.364       | 100         | 8            | 36.4         | 1          | 100         | 22           | 100          |
| Row60  | lat (#1)         | [tét,parc,lyon]             | 2           | 9.091        | 100         | 2            | 9.09         | 1          | 100         | 22           | 100          |
| Row61  | lyon             | [tét,parc,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.1        | 110         | 10           | 45.455       |
| Row62  | parc             | [tét,lyon,lat (#1)]         | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row63  | tét              | [parc,lyon,lat (#1)]        | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row64  | lyon             | [tét,parc]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.1        | 110         | 10           | 45.455       |
| Row65  | parc             | [tét,lyon]                  | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row66  | tét              | [parc,lyon]                 | 2           | 9.091        | 50          | 4            | 18.2         | 1.375      | 137.5       | 8            | 36.364       |
| Row67  | lat (#1)         | [tét,parc]                  | 4           | 18.182       | 100         | 4            | 18.2         | 1          | 100         | 22           | 100          |
| Row68  | parc             | [tét,lat (#1)]              | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row69  | tét              | [parc,lat (#1)]             | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row70  | parc             | [tét]                       | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row71  | tét              | [parc]                      | 4           | 18.182       | 50          | 8            | 36.4         | 1.375      | 137.5       | 8            | 36.364       |
| Row72  | lat (#1)         | [tét,lyon]                  | 4           | 18.182       | 100         | 4            | 18.2         | 1          | 100         | 22           | 100          |

Nous constatons que l'étude des tags et des règles qui en sont déduites dans le cas de cette zone géographique nous rapporte généralement que des informations sur le nom du parc et la ville où il est situé. Nous ne parvenons pas obtenir des informations sur les centres d'intérêt internes au parc (jardin botanique, zoo, etc).

## V- Informations destinées aux clients et aux équipes futures

Maintenant que nous avons analysé différentes méthodes dans le but de découvrir des centres d'intérêts au moyen de workflow KNIME sur des enregistrements Flickr, nous allons analyser les résultats de ces méthodes afin d'essayer de conclure sur leur efficacité. La découverte de centres d'intérêts au moyen de l'étude des données issues d'images disponibles sur Flickr est possible mais connaît cependant certaines contraintes.

Il s'agit de trouver tout d'abord une méthode de clustering adaptée, englobant aussi bien l'algorithme que ses paramètres. Après avoir testé le clustering avec les trois algorithmes *K-Means*, *Hierarchical Clustering* et *DBSCAN*, nous avons pu conclure que l'algorithme DBSCAN était le plus performant en termes de temps de réponse et de résultats obtenus.

En effet, *K-Means* calcule des clusters de tailles presque égales et de formes globulaires indépendamment du nombre de centres d'intérêts présents dans une zone précise, il n'est pas retenu. Quant au *Hierarchical Clustering*, pour un nombre élevé de clusters et une zone assez réduite, nous obtenons des résultats pertinents. Dans notre étude, nous avons constaté que *Average* était le mode de calcul de distance pour le *Hierarchical clustering* le mieux adapté parmi ceux qui ont été testés (il est possible que *Ward* soit plus performant mais avec KNIME on n'a pas pu tester cette méthode). Cependant, cet algorithme avec le nombre de clusters à calculer étant important a un délai de réponse important et cela pour une zone réduite de la ville de Lyon, il ne s'agit donc pas de la méthode retenue pour le clustering. L'algorithme de *DBSCAN* contrairement aux algorithmes de *K-Means* et de *Hierarchical Clustering*, a un temps de réponse raisonnable avec des résultats pertinents. Il permet aussi une bonne gestion des bruits. Cependant, cet algorithme nécessite de donner des valeurs à ses paramètres (*epsilon* et *minPts*) avant d'effectuer le clustering. Si nous voulons tourner cet algorithme sur l'ensemble de la ville de Lyon, la divergence de densités sur l'ensemble de l'agglomération risque fortement de rendre les résultats moins pertinents.

Afin de s'affranchir de cette limite imposée par l'algorithme de *DBSCAN*, il est possible de diviser la ville en plusieurs zones réduites afin de favoriser l'utilisation de paramètres différents sur chacune de ces zones. Un traitement est donc nécessaire afin de définir les paramètres adaptés à chaque zone géographique de la ville afin de visualiser des résultats pertinents en termes de clusters représentant bien des centres d'intérêts distincts.

Une phase de post-traitement demeure toutefois nécessaire afin de visualiser avec plus d'attention les clusters: on peut avoir un cluster qui correspond à deux centres d'intérêts assez similaires. De plus, certains clusters pourraient ne pas correspondre à des centres d'intérêts intéressants (photos prises lors d'une manifestation, etc...). Il n'est également pas impossible que deux clusters représentent en réalité un même centre d'intérêt.

Ceci dit, avec un workflow KNIME et des données téléchargées sur Flickr, nous pouvons obtenir une idée de la localisation des centres d'intérêts sur la ville lyonnaise, à condition de réaliser les deux phases de traitement avant et après clustering avec *DBSCAN*.

Une fois l'algorithme de clustering choisi, nous pouvons porter une attention plus approfondie au niveau des attributs utilisés pour le clustering.

Il s'agit également de trouver les paramètres que l'on souhaite étudier parmi toutes les informations que nous pouvons récupérer sur des enregistrements Flickr. Dans un premier temps, nous avons effectué un clustering en prenant en compte uniquement la dimension spatiale. Cette méthode nous apporte de l'information sur la densité de photos prises par les individus à un certain emplacement géographique. Par exemple, lorsque nous nous sommes concentrés sur la zone géographique comprenant l'Opéra et l'Hôtel de ville, nous avons observé plusieurs clusters aux endroits symboliques de cet endroit à Lyon : l'Opéra, la Place de la Comédie ou encore l'Hôtel de ville. En utilisant un clustering qui se concentre uniquement sur la dimension géographique il faut analyser les données "à la main" en s'assurant de la pertinence des données si l'on reste sur des enregistrements Flickr qui possèdent certaines failles (cluster contenant uniquement des photos prises par un unique utilisateur, visualisation sur une carte 2D parfois peu représentative). Cela ne facilite pas le développement d'une méthode de détection des centres d'intérêts automatique. Ce type de clustering nous permet donc d'obtenir des informations à valeur ajoutée mais ce n'est pas le plus optimal.

Un clustering qui prend en compte la dimension spatiale et temporelle nous apporte plus d'information à valeur ajoutée car les événements saisonniers apparaissent et sont visibles lors de l'analyse des clusters. Par exemple dans le cas de Lyon la fête des lumières qui est un événement majeur pour la ville de Lyon, du gay pride ou encore de congrès annuel dans la cité internationale. En détectant ce genre d'événements, si ces derniers se reproduisent périodiquement dans le futur, la ville de Lyon sera en mesure d'augmenter la fréquence des lignes TCL pour assurer aux visiteurs un confort de déplacement, d'assurer l'hébergement et de s'organiser pour assurer le meilleur déroulement de ces événements. Ce type de clustering nous apporte donc de la valeur ajoutée exploitable par les clients. Cependant, en effectuant ce type de clustering nous avons également obtenu des résultats peu pertinents pour la découverte de centres d'intérêts. En effet, nous avons pu obtenir des clusterings qui comprennent uniquement des photos prises par un même individu or selon nous cela n'est pas caractéristique d'un centre d'intérêt. Si les clients souhaitent contrer ce phénomène et ne pas prendre en compte ce type de cluster il pourront créer une mesure qui indiquera un rapport entre le nombre de photographes et le nombre de photos ou une tout autre mesure similaire. On pourrait aussi suggérer de grouper par user et tag (ou ensemble de tags) pour limiter les données à une photo par utilisateur pour le même centre d'intérêt et ainsi les tags les plus fréquents seront des centres d'intérêt globaux pour lesquels on peut faire des conclusions qui ne concernent pas seulement une seule personne. Comme par exemple dans cette capture le user "Thierry Ehrmann" a pris plusieurs photos de "cyberpunk". Ces 2 tags apparaîtront sur le word cloud même si ceci ne concerne qu'un seul user et que "Thierry Ehrmann" ne représente pas un centre d'intérêt.

Une autre suggestion peut être le fait d'utiliser des SparkCloud (nuage de tags qui incorpore des sparklines) pour voir le changement des tags par rapport au temps (et repérer les tags qui sont utilisés annuellement comme fête des lumières et ceux utilisés une seule fois et donc pour un évènement ponctuel ou un seul utilisateur).



Fig. 1. SparkClouds showing the top 25 words for the last time point (12th) in a series. 50 additional words that are in the top 25 for the other time points can be (top) filtered out or (bottom) shown in gray at a smaller fixed-size font. (bottom) is used in the study.

Le filtrage des mots clés pose aussi des difficultés. Bien qu'on ait filtré les mots génériques, on aurait toujours des tags qui apparaissent qui n'auront pas de valeur dans le cas de notre recherche. Ce qui est fréquent n'est pas forcément discriminant. Il faut donc

veiller à faire attention à ce genre de phénomènes et nettoyer les données avant de commencer. On pourrait nous baser sur les tags (produits par les utilisateurs), et formuler un nouveau framework d'évaluation basé sur l'intuition qu'une image peut être remplacée par un ou plusieurs tags qui véhiculent le même sens que l'image elle-même. Idéalement, il y a un seul tag qui « mieux » décrit l'image dans son ensemble, ou ne pas se baser sur les tags des utilisateurs et essayer le “[Automatic Image Tagging](#)”.

Concernant les outils, on a remarqué que, bien que KNIME soit facile à utiliser et intuitif, il n'est pas collaboratif. On pourrait utiliser Google Colab avec Python pour ce genre d'étude afin de collaborer entre l'équipe sur un même fichier. Par exemple pour visualiser le word cloud (remarque: ici on a visualisé le word cloud sans limiter la zone géographique):

[https://colab.research.google.com/drive/1Y3r5HLuW\\_aogQKTdcl7dL-oB5o3y9zRv?usp=sharing](https://colab.research.google.com/drive/1Y3r5HLuW_aogQKTdcl7dL-oB5o3y9zRv?usp=sharing)



En guise de conclusion, les qualités de cette approche incluent la possibilité de visualiser et de manipuler facilement les données, ainsi que la possibilité d'intégrer différents types de traitement des données. Cependant, les limites incluent la nécessité de disposer d'une grande quantité de données bien étiquetées pour obtenir des résultats significatifs.

## VI- Bibliographie et références

- Utilisation de Python/Google Colab:  
[https://colab.research.google.com/drive/1Y3r5HLuW\\_aogQKTdcl7dL-oB5o3y9zRv?usp=sharing](https://colab.research.google.com/drive/1Y3r5HLuW_aogQKTdcl7dL-oB5o3y9zRv?usp=sharing)  
<https://youtu.be/aM-zhxyVE54>
- Article sur “Text Mining for Automatic Image Tagging”:  
<https://aclanthology.org/C10-2074.pdf>