Paris Smart-city !

- Contexte & Présentation générale du jeu de données
- Méthodologie d'analyse de données
- Synthèse & propositions Smart-city

Kévin Duranty

Contexte & Présentation générale du jeu de données

La Ville de Paris propose un **challenge Data** dans le cadre du programme "**Végétalisons la ville**" afin d'aider la capitale à devenir une Smart-city. L'objectif de ce challenge est d'**optimiser les tournées d'entretien des arbres de la ville**.

Pour cela nous disposons d'un jeu de données accessible à tous à l'adresse opendata.paris.fr.

Le jeu de donné est un fichier CSV de 27Mo qui contient les informations des arbres de la ville de Paris.

L'analyse de ces données s'effectuera grâce au langage **Python** et des différentes bibliothèques de data science :

- **Pandas**
- **Numpy**
- **Matplotlib & Seaborn**

Le projet à été réalisé dans un **jupiter notebook** en veillant à ce qu'un **environnement virtuel** soit créé pour assurer l'isolement du projet et la gestion des dépendances.

Méthodologie d’analyse de données

Notre analyse du jeu de données se fera en 3 phases :

1. **Exploration des données**.
2. **Traitement des données**.
3. **Analyse approfondie**, à travers 3 axes d’observation.

Méthodologie d'analyse de données

1.Exploration des données

	id	type_emplacement	domanialite	arrondissement	complement_addresse	numero	lieu	id_emplacement	libelle_francais	genre	
0	99874	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	19	Marronnier	Aesculus	hipp
1	99875	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	20	If	Taxus	
2	99876	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	21	If	Taxus	
3	99877	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	22	Erable	Acer	
4	99878	Arbre	Jardin	PARIS 17E ARRDT	NaN	NaN	PARC CLICHY-BATIGNOLLES-MARTIN LUTHER KING	000G0037	Arbre à miel	Tetradium	

(tableau 1)

Le tableau 1 nous présente les **200 137 lignes** et **18 colonnes** du jeu de données.

Chaque ligne correspond à un arbre parisien et chaque colonne révèle une information sur cette arbre.

#	Column	Non-Null Count	Dtype
0	id	200137 non-null	int64
1	type_emplacement	200137 non-null	object
2	domanialite	200136 non-null	object
3	arrondissement	200137 non-null	object
4	complement_addresse	30902 non-null	object
5	numero	0 non-null	float64
6	lieu	200137 non-null	object
7	id_emplacement	200137 non-null	object
8	libelle_francais	198640 non-null	object
9	genre	200121 non-null	object
10	espece	198385 non-null	object
11	variete	36777 non-null	object
12	circonference_cm	200137 non-null	int64
13	hauteur_m	200137 non-null	int64
14	stade_developpement	132932 non-null	object
15	remarquable	137039 non-null	float64
16	geo_point_2d_a	200137 non-null	float64
17	geo_point_2d_b	200137 non-null	float64

(tableau 2)

Le tableau 2 nous présente la liste des colonnes, le type de variable contenue dans celles-ci ainsi que le nombre de valeur non-nulle par colonne.

1.Exploration des données

	type_emplacement	domanialite	arrondissement	complement_addresse	lieu	id_emplacement	libelle_francais	genre	espece	variete	stade_developpement
count	200137	200136	200137	30902	200137	200137	198640	200121	198385	36777	132932
unique	1	9	25	3795	6921	69040	192	175	539	436	4
top	Arbre	Alignement	PARIS 15E ARRDT	SN°	PARC FLORAL DE PARIS / ROUTE DE LA PYRAMIDE	101001	Platane	Platanus	x hispanica	Baumannii'	A
freq	200137	104949	17151	557	2995	1324	42508	42591	36409	4538	64438

(tableau 3)

Le tableau 3 nous présente :

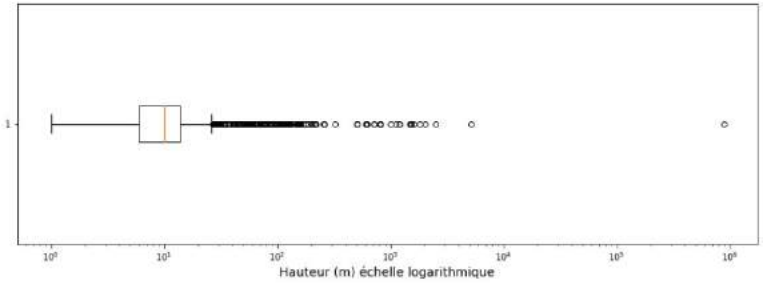
- le nombre de valeur non-nulle,
- le nombre de valeur unique par colonne,
- la valeur la plus présente parmis ces valeurs uniques et
- le nombre de fois que cette valeur apparaît.

Observations :

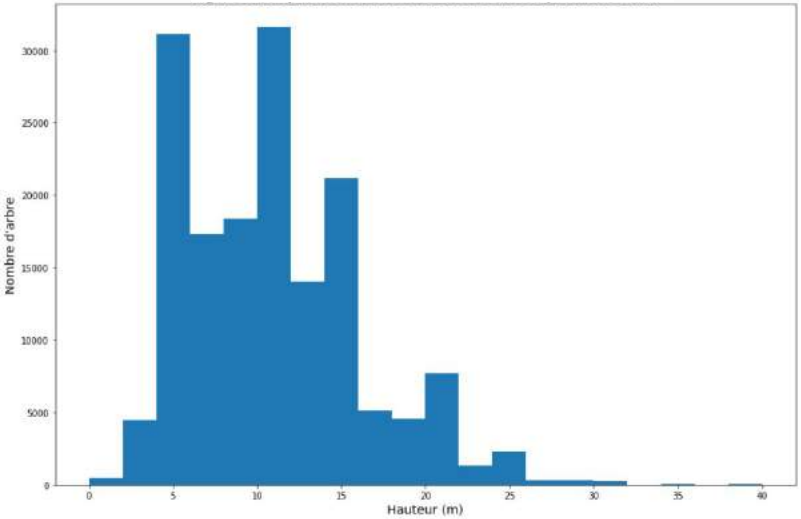
- l'arrondissement le plus peuplé en arbre : le 15e arr. avec 17 151 arbres,
- le stade de développement le plus atteint : le stade "Adulte" pour 63438 arbres,
- le nombre d'espèce différente : 175.
- la colonne "type_emplacement" ne contient qu'une valeur.

1.Exploration des données

(graphique 1 : répartition de la hauteur sur une échelle logarithmique)



(graphique 2 : histogramme de la hauteur en mètre)



count	160918.000000
mean	16.305808
std	2198.334745
min	1.000000
25%	6.000000
50%	10.000000
75%	14.000000
max	881818.000000

Name: hauteur_m, dtype: float64

(tableau 4)

Le tableau 4 présente une synthèse de la colonne “hauteur_m” avec notamment les quartiles de distribution, la moyenne et le nombre de valeur différente de zéro.

Bilan de l'exploration préliminaire :

- Présence de valeurs aberrantes et atypiques.
- Nettoyer le jeu de donnée des valeurs manquantes et colonnes inutiles.
- Nécessité de sélectionner une plage de valeur pour l’ étude approfondie.
- Observations similaires pour la colonne “circonférence_cm”

2. Traitement des données

Suppression des colonnes vides ou inutiles :

- la colonne "libellé_français",
- les colonnes "id", "id_emplacement" et “numéro”
- la colonne 'type_demplacement'

Suppression des valeurs dupliquées : 11 lignes.

Remplacement des valeurs nulles :

- Dans les colonnes “hauteur_m” par la moyenne par espèce
- Dans les colonnes “circonference_cm” par la moyenne par espèce

Définition de l'écart interquartile (EI) puis la sélection de nos valeurs

→ EI = Q3 - Q1

Critère de sélection :

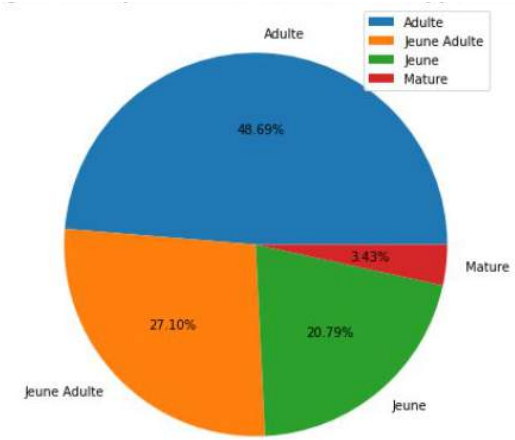
→ hauteur < Q3 + 1.5 * EI(hauteur)

→ circonférence < Q3 + 1.5 * EI(circonférence**)**

Méthodologie d'analyse de données

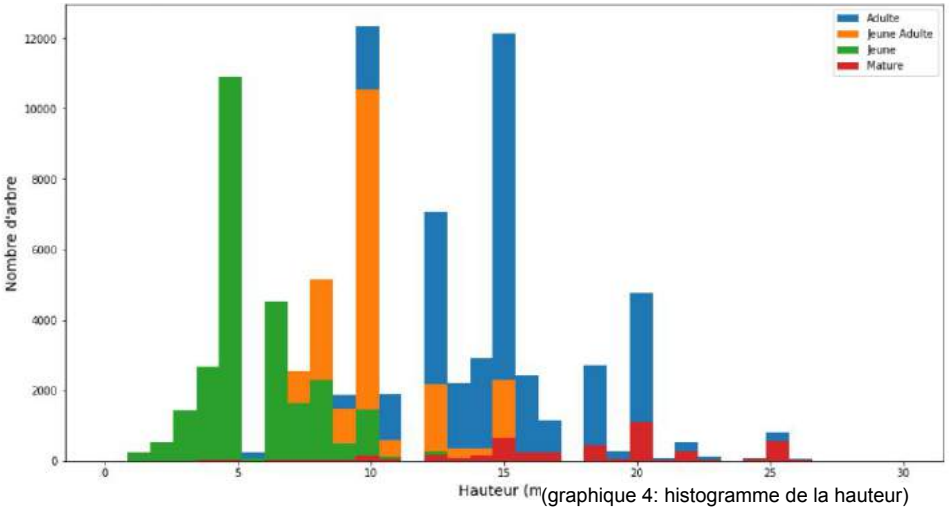
3. Analyse approfondie

→ par stade de développement

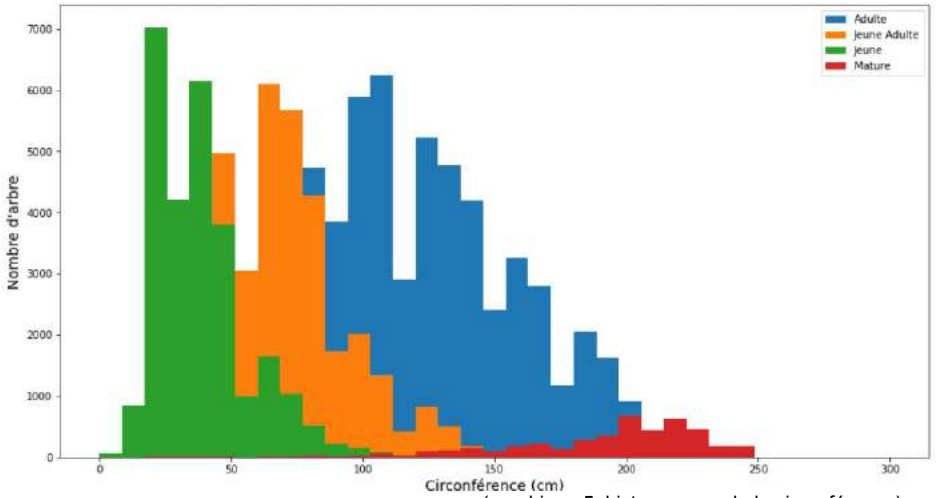


(graphique 3 : Répartition par stade de développement)

- Minorité d'arbre "Mature"
- Stade de développement progressif
- Hiérarchie du stade de développement.



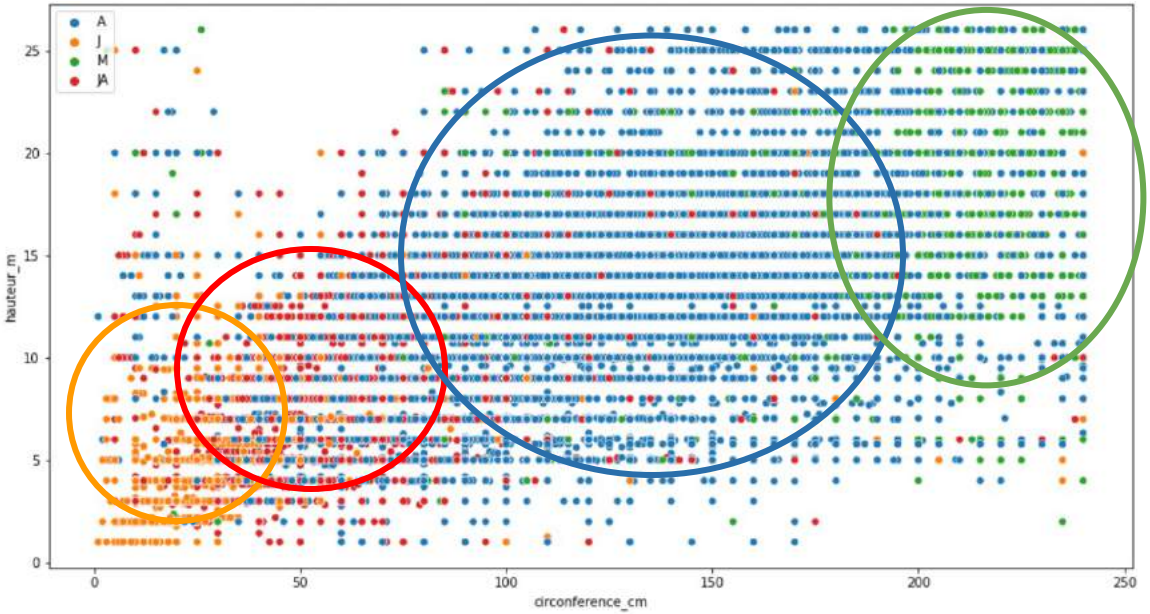
(graphique 4: histogramme de la hauteur)



(graphique 5: histogramme de la circonférence)

3. Analyse approfondie

→ par stade de développement



(graphique 6 : Répartition des arbres par stade de développement)

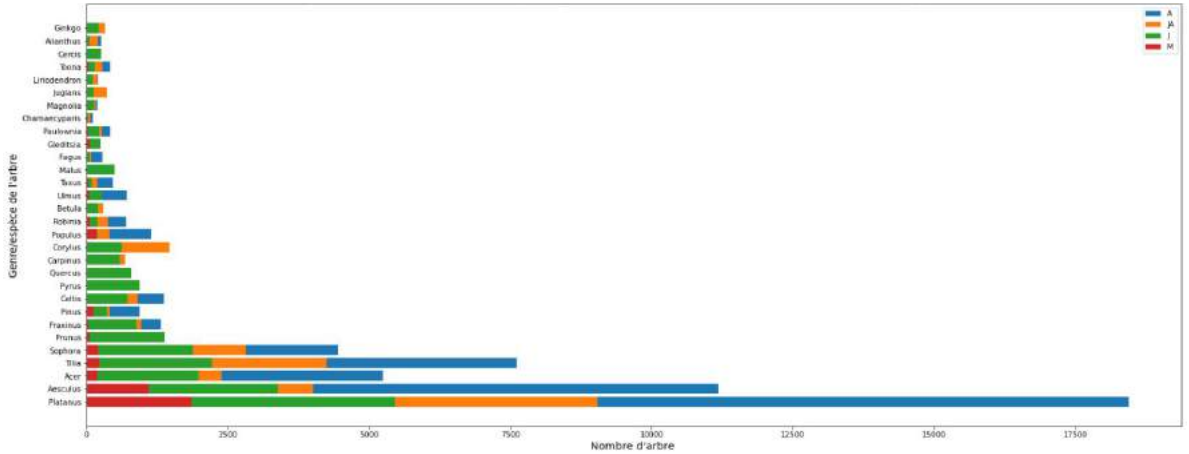
Sur le graphique 6 on observe des groupes d'arbre répartis en fonction de leur stade de développement.

Il permet également de vérifier si les informations sur certains arbres sont correctes.

3. Analyse approfondie

→ par espèce

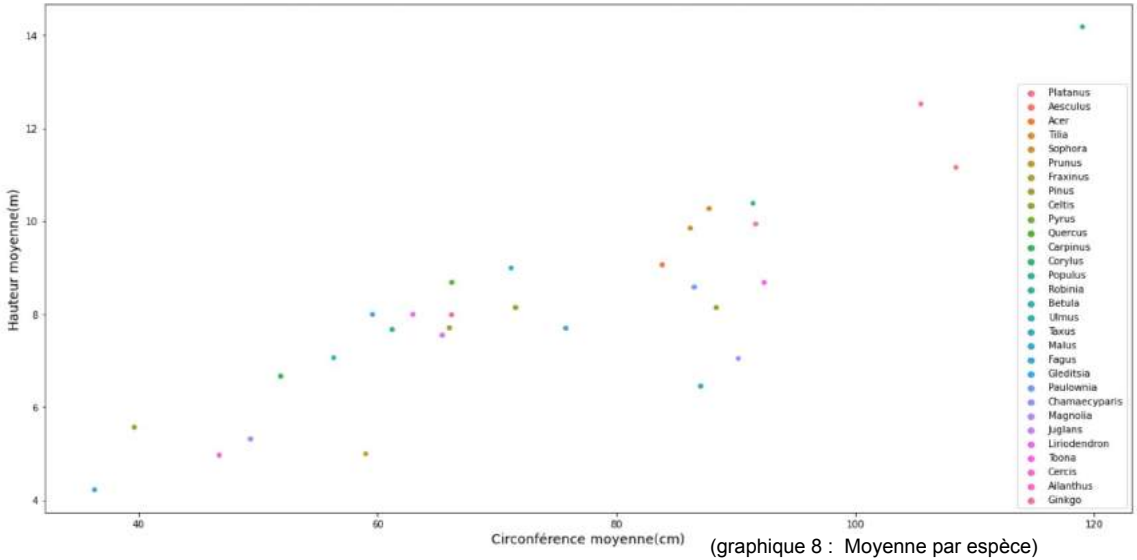
Le graphique 7 présente le nombre d'arbre par espèce en fonction des stades de développement.



(graphique 7 : Nombre d'arbre par espèce)

Le graphique 8 affiche par espèce la moyenne de la hauteur en fonction de la moyenne de la circonférence, pour les 28 espèces possédant plus de 1000 arbres à Paris.

Observations :
Il semble y avoir complémentarité dans l'offre des espèces implantées à Paris.



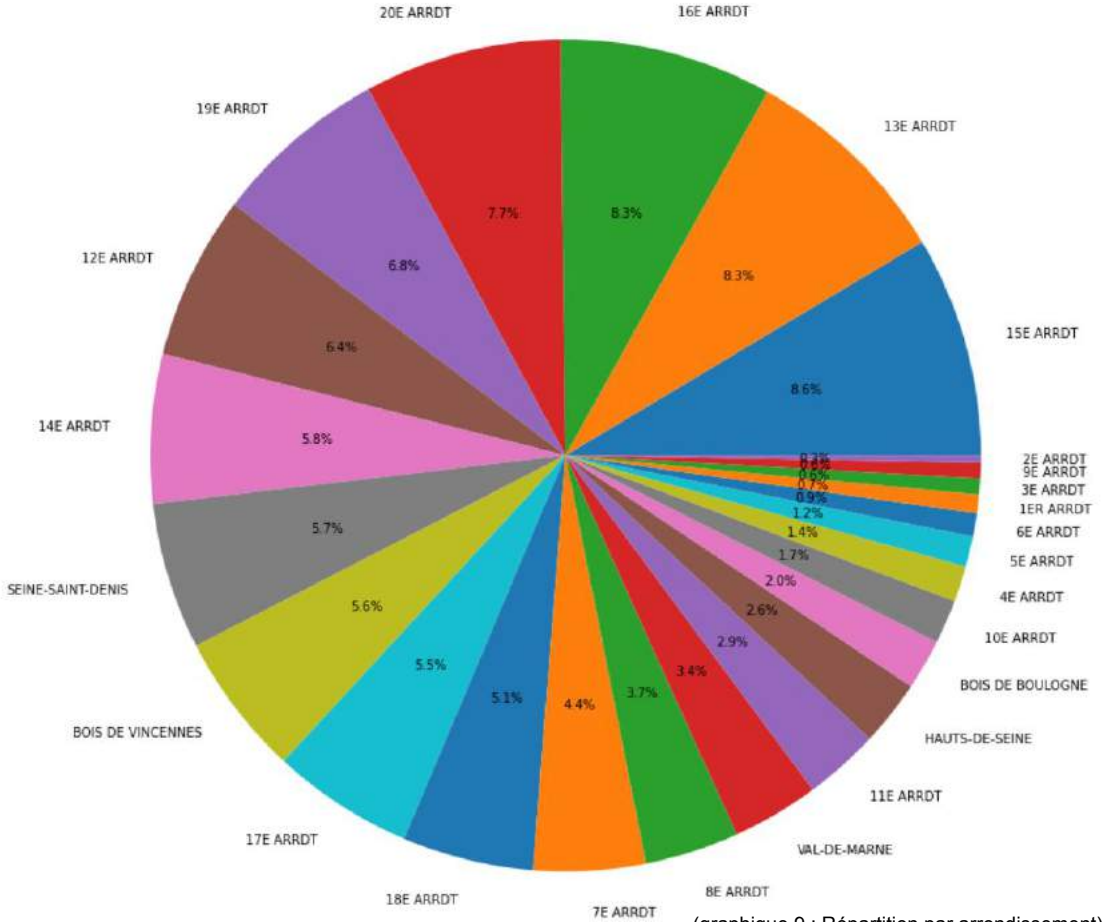
(graphique 8 : Moyenne par espèce)

3. Analyse approfondie
→ par arrondissement

Le graphique 9 nous montre la répartition en pourcentage des arbres par arrondissement.

Observations :

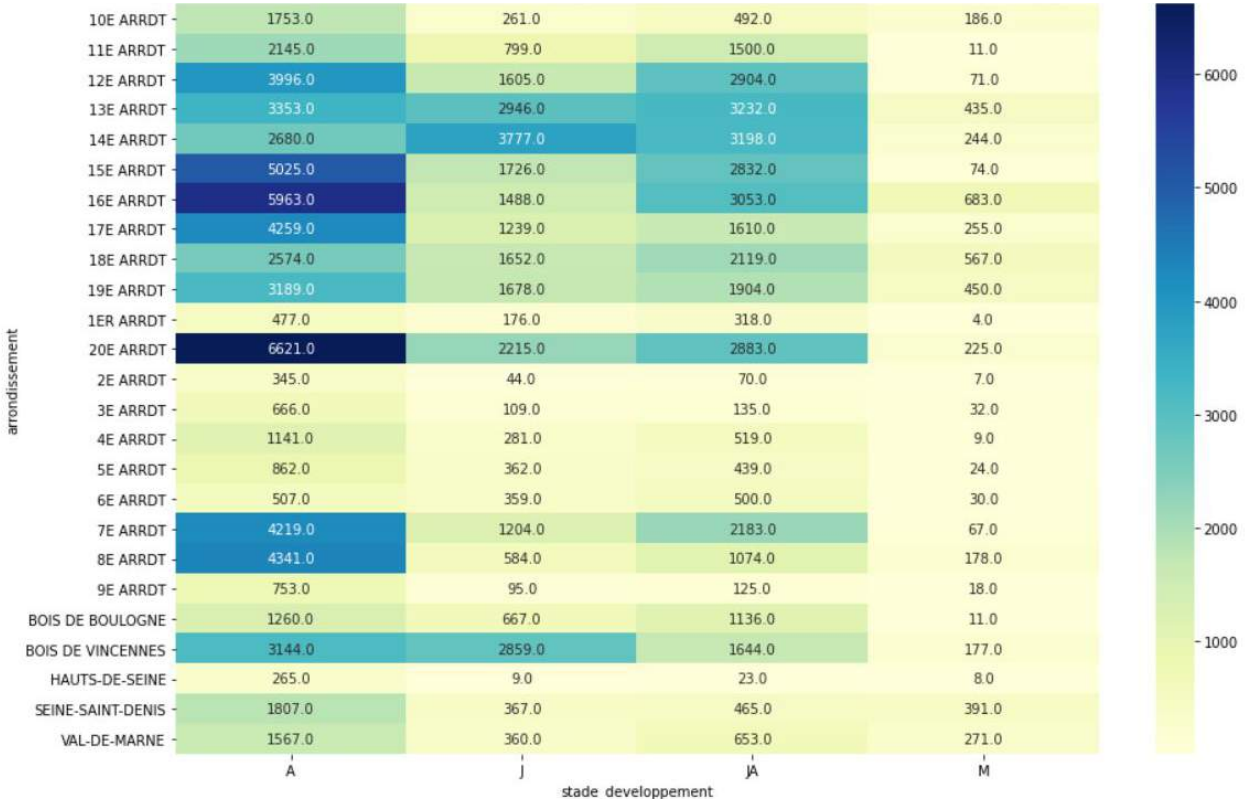
- le 15e possède autant d'arbre que 7 arrondissements,
- Il y a entre 30 et 130 espèce d'arbre par arrondissement.



(graphique 9 : Répartition par arrondissement)

3. Analyse approfondie

→ par arrondissement



(graphique 10 : Densité d'arbre par stade de développement et par arrondissement)

Observations :

- Faible présence du stade de développement "Mature".
- Le 20e est l'arrondissement le plus dense en arbre "Adulte".
- Peu d'arbre dans les 2e 3e, 4e 5e et 6e arrondissement.

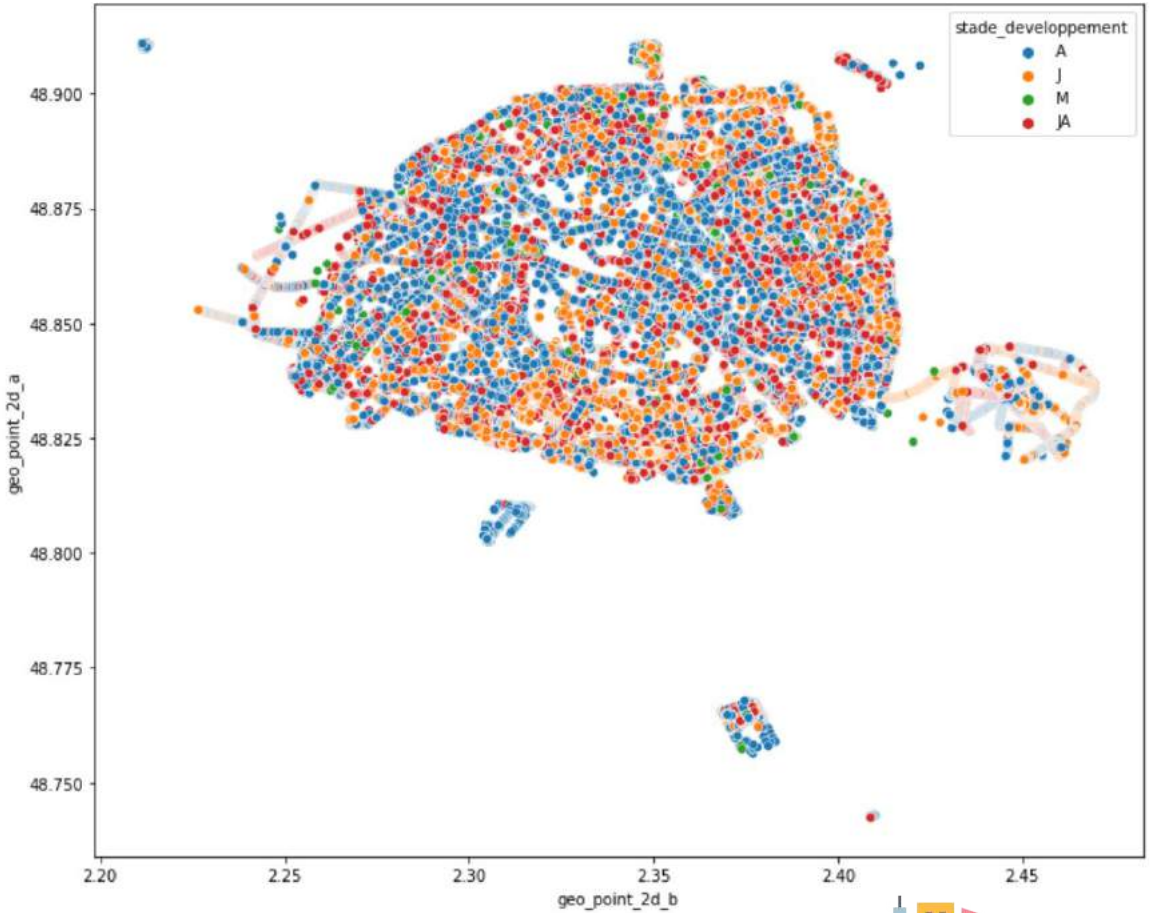
3. Analyse approfondie

→ par arrondissement

Le graphique 11 nous montre sur un plan la répartition des arbres différenciés par leur stade de développement.

Observations :

- Les arbres jeunes sont en périphérie.
- Les arbres adultes sont au centre.
- Il y a des zones vide au coeur de Paris.



(graphique 11 : Position des arbres en fonction de leur stade de développement)

Synthèse & propositions Smart-City

Synthèse de l'analyse approfondie

Stade de développement :

- Bonne répartition des arbres en fonction de leur stade de développement.
- Très peu d'arbre au stade mature, mais des arbres supérieur à 15m.
- Quatres stades de développement ne sont pas suffisant pour décrir l'état d'évolution des arbres.

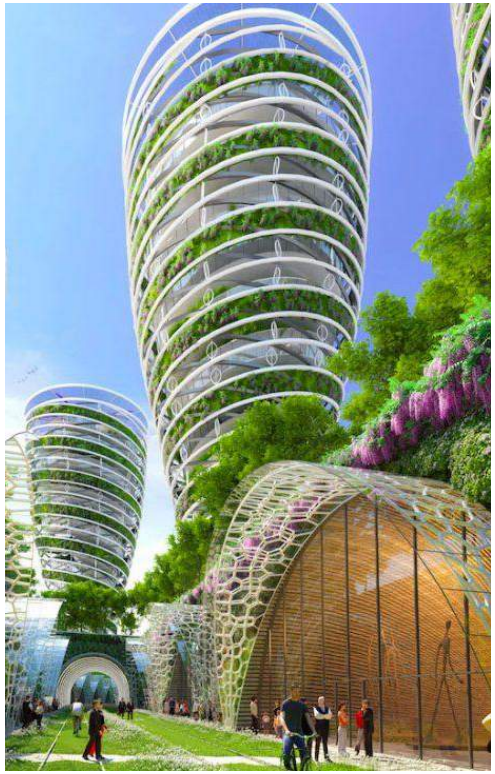
Espèce :

- 175 espèces mais 5 ressortent très largement du lot et représente 60% du parc.
- Répartition équivalente dans les stades de développement.
- Bonne complémentarité dans le choix des espèces.

Arrondissement

- Répartition non uniforme du stade de développement dans la ville.
- Répartition non uniforme par arrondissement.

Synthèse & propositions Smart-City



Proposition d'optimisation des tournées pour l'entretien des arbres de la ville

- Nous proposons à la Ville de Paris de se focaliser sur l'entretien des arbres 'Matures' qui ne sont que 3,4% du parc installé, mais qui sont essentiellement de grands arbres, plus de 14m de haut et 175 cm de circonférence. Ces arbres sont de plus essentiellement représentés par 4 à 5 espèces.
- Nous proposons d'effectuer les tournées d'entretien par arrondissement et par espèce.
- Nous proposons à la ville d'équiper les agents d'entretien de cet outil d'analyse.

Proposition concernant la diversité des arbres

- Nous proposons d'augmenter la diversité des arbres par arrondissement.
- Nous proposons d'augmenter la quantité d'arbres dans certains arrondissements.
- Nous proposons à la ville de renforcer l'implantation des espèces qui ont respectivement une hauteur et circonférence moyenne de 9 m, 80 cm et 11 m, 100 cm pour compléter l'offre d'arbre
- Nous proposons enfin à la ville de Paris d'implanter davantage d'arbres dans les lieux vides de la capitale.

Proposition liées à la collecte de données

- Nous proposons de modifier certaines colonnes du jeu de données afin d'améliorer la description des arbres notamment les données relatives au stade de développement.
- Nous proposons de développer un algorithme de traitement des données capable de repérer les outliers et corriger les valeurs manquantes grâce à un modèle prédictif.