


Participez à un concours sur la Smart City

Data challenge pour la ville de Paris :
Végétons la ville

Xiaofan LEI

Table des matières



○	Contexte
○	Environnement : AWS EC2 et VPC
○	Présentation générale du jeu de données
○	Préparation des données
○	Démarche méthodologique d'analyse de données
○	Librairies utilisées
○	Synthèse de l'analyse de données

Contexte

Contexte général

- l'ONG a organisé des challenges de Data Science en ligne « Data is for Good »

Choix du challenge : Végétons la ville

- La ville de Paris a sponsorisé un challenge dans le cadre du programme “Végétons la ville”, afin d'aider Paris à devenir une smart-city en optimisant des tournées pour l'entretien du patrimoine arboré de la ville.
- Car les arbres sont fragilisés par les conditions citadines alors qu'ils :

rafraichissent
l'air en été

apportent de
l'ombre

améliorent la
qualité de
l'air

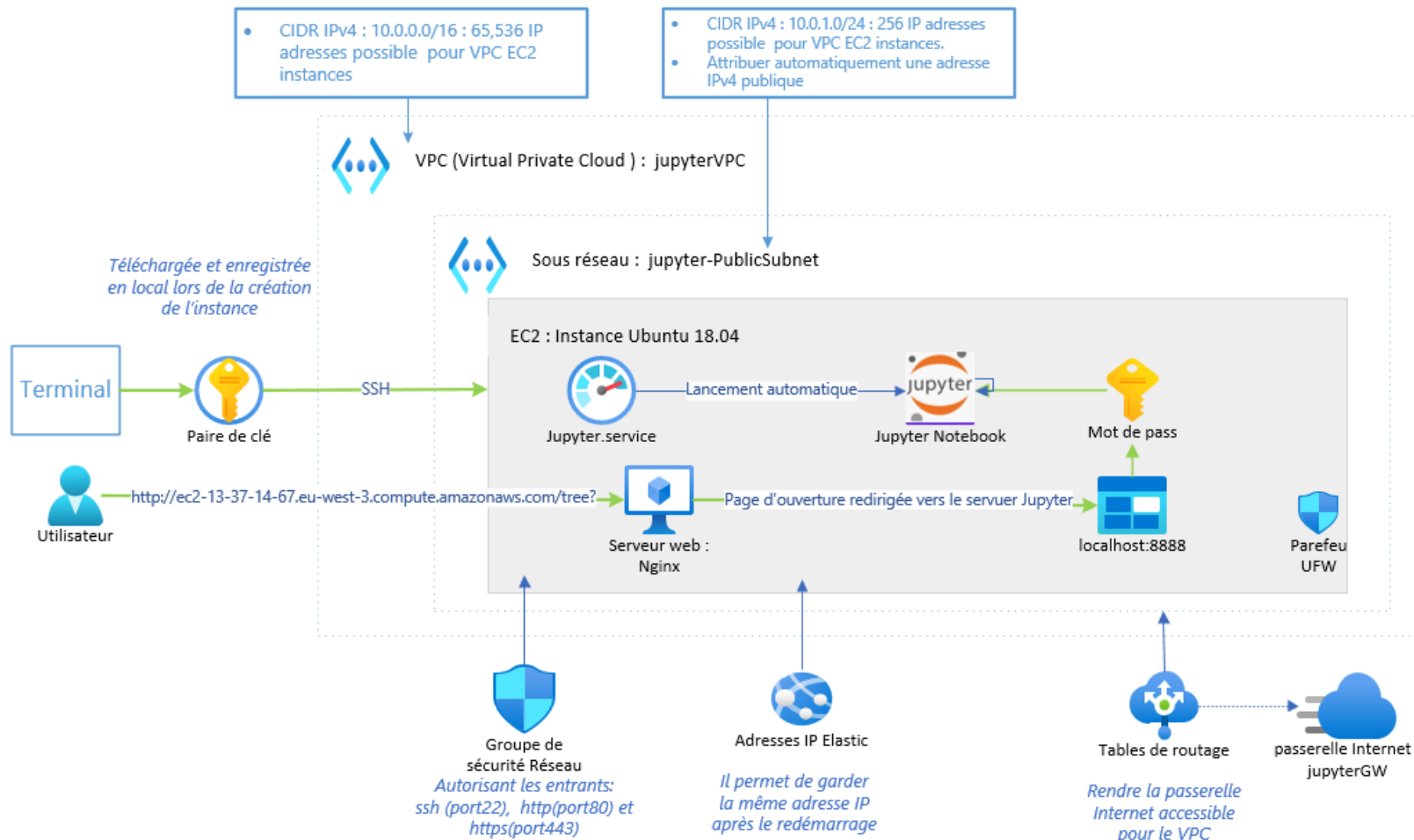
créent la
biodiversité

sont
esthétiques
marquent les
saisons

ont une
action
relaxante

...

Création d'un environnement cloud: AWS EC2 et VPC



► Possibilité de créer des Amazon Machine Images (AMI) pour faciliter le déploiement des instances similaires

Présentation générale du jeu de données

Ce dataset est composé de 200137 individus avec 18 variables

	mode	nb na	% na
type_emplacement	Arbre	0	0.01%
domanialite	Alignement	1	0.01%
arrondissement	PARIS 15E ARRD	0	0.01%
complement_adresse	SN°	169235	84.56%
numero	-	200137	100.00%
lieu	PARC FLORAL DE PARIS / ROUTE DE LA PYRAMIDE	0	0.01%
id_emplacement	101001	0	0.01%
libelle_francais	Platane	1497	0.75%
genre	Platanus	16	0.01%
espece	x hispanica	1752	0.88%
variete	Baumannii'	163360	81.62%
stade_developpement	A	67205	33.58%

- ▶ le genre, l'espèce et le libelle_francais sont bien renseignés (<1% na)
- ▶ « Libellé français » est le nom commun donné aux arbres.
- ▶ la plus part des variétés restent vacantes(>81% na).
- ▶ « Domanialité » a qu'1 seule valeur manquante et peut être comblée par le mode
- ▶ Les arrondissements et les lieux sont dûment renseignés
- ▶ Compléments d'adresse inexploitable (~85% na)
- ▶ Il manque des arbres non précisés en stade de développement (~34% na)

Présentation générale du jeu de données

	count	mean	std	min	25%	50%	75%	max	nb na	% na
remarquable	137,039.00	0.00	0.04	0.00	0.00	0.00	0.00	1.00	63098	31.53%
geo_point_2d_a	200,137.00	48.85	0.03	48.74	48.84	48.85	48.88	48.91	0	0.00%
geo_point_2d_b	200,137.00	2.35	0.05	2.21	2.31	2.35	2.39	2.47	0	0.00%
circonference_cm	200,137.00	83.38	673.19	0.00	30.00	70.00	115.00	250,255.00	0	0.00%
hauteur_m	200,137.00	13.11	1,971.22	0.00	5.00	8.00	12.00	881,818.00	0	0.00%

► Remarquables

- Les arbres remarquables sont marqués par « 1 », donc les arbres non marqués (~32%) seront considérés comme non remarquables.

► Données géographiques

- au complet
- Pas d'erreurs visibles (les écarts-type sont entre 0 et 0,1)

► Hauteurs & circonférences

- Des erreurs sont visiblement induites car leur écart-type sont trop grands. La différence entre max et le Q3 est relativement importante, les outliers doivent se trouver dans la tranche supérieure au Q3.

Préparation des données : Identification des outliers

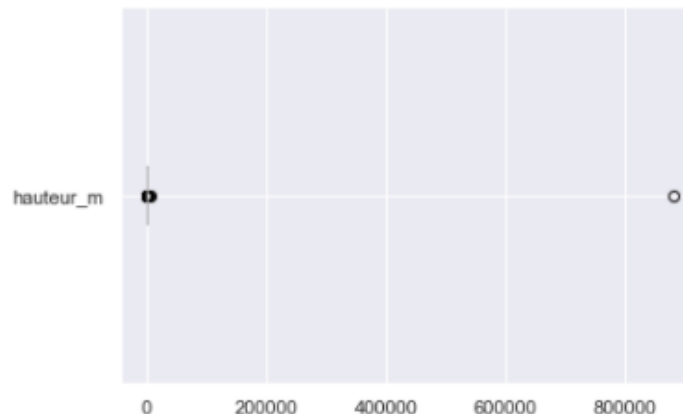
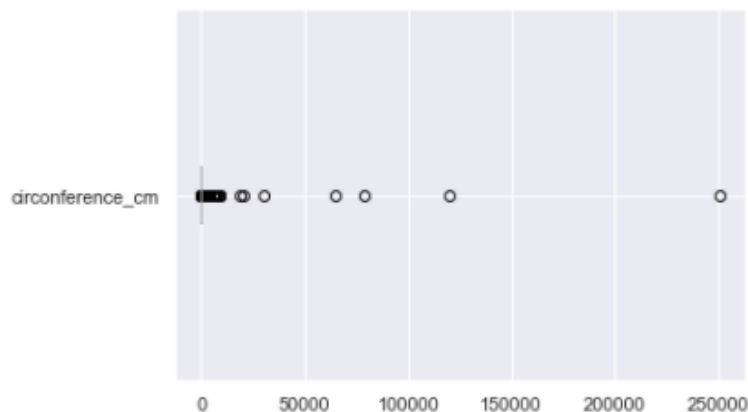
Méthode A : sur Internet

La platane d'Orient du Parc Monceau est le plus gros arbre de Paris. Planté en 1814, son tronc mesure en effet **7 mètres** de circonférence pour une hauteur de **31 mètres** d'environ.

Méthode B : la règle 1,5 x écart interquartile

La plus part des arbres parisiens mesurent moins de **22,5 m** de haut pour une circonférence de moins de **2,5m**.

La plus part des arbres parisiens ont une circonference de moins de 242.5 cm. La plus part des arbres parisiens ont une hauteur de moins de 22.5 m.



Préparation des données : Traitement

Vérification de l'impact

Si l'on se fie aux mesures trouvées sur internet (7m de circonférence et 31 m de haut), 637 sujets dont 10 spécimens remarquables doivent être revus.

Ces individus ne peuvent pas être simplement écartées de l'analyse à cause des informations importantes y sont associées, comme celle d'arbre remarquable.

La modification sur les hauteurs et les circonférences n'auront pas d'impact significatif sur l'analyse, car elle représentent uniquement 0.3 % sur l'ensemble des données.

Traitement

Etape 1 : Les erreurs commises sur les longueurs proviennent souvent de l'unité de mesure, donc les chiffres trop importants sont considérés comme saisis ou importés d'autres logiciels en utilité de mesure moindre (cm ou mm), ils seront reconvertis en m ou cm.

Etape 2 : les hauteurs et les circonférences dépassant les limites fixées sont ignorées (en na) tout en conservant les autres informations.

Démarche méthodologique d'analyse de données

Objectif :

- Cet étude est focalisée sur l'analyse des caractéristiques individuelles présentes dans ce dataset ainsi que leur représentation cartographique afin de répondre aux besoins d'optimisation des tournées.

Représentation des indicateurs statistiques

- Variable quantitative ou qualitative: : Les boîtes à moustache, graphiques en camembert, tuyaux d'orgue ou histogrammes sont utilisés pour visualiser la distribution d'une variable.
- Variables quantitative et qualitative: Plusieurs boîtes à moustache affichées dans un même graphique offrent une meilleure vue sur la distribution

Représentation cartographique

- Trois librairies géographique sont utilisées : géopandas, plotly et folium.
- Géopandas permet d'afficher les emplacements sur un fond de fichier shape, alors que folium et plotly offre une visualisation des points sur une carte interactive.

Librairies utilisées

Analyse des données sous forme de tableaux numériques

Numpy : calculs mathématiques

Pandas : manipulation et l'analyse des données

statsmodels.api.OLS : calcul de la régression linéaire

sklearn.cluster .KMeans : classification non supervisée

Visualisation des données sous formes de graphiques

Matplotlib : plotbox, camembert, histo, tuyaux d'orgue...

Visualisation des données sous formes de cartographie

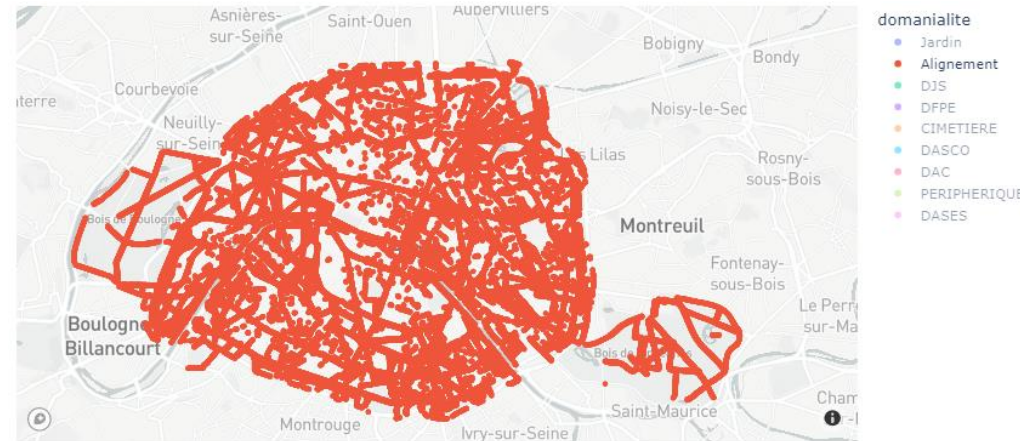
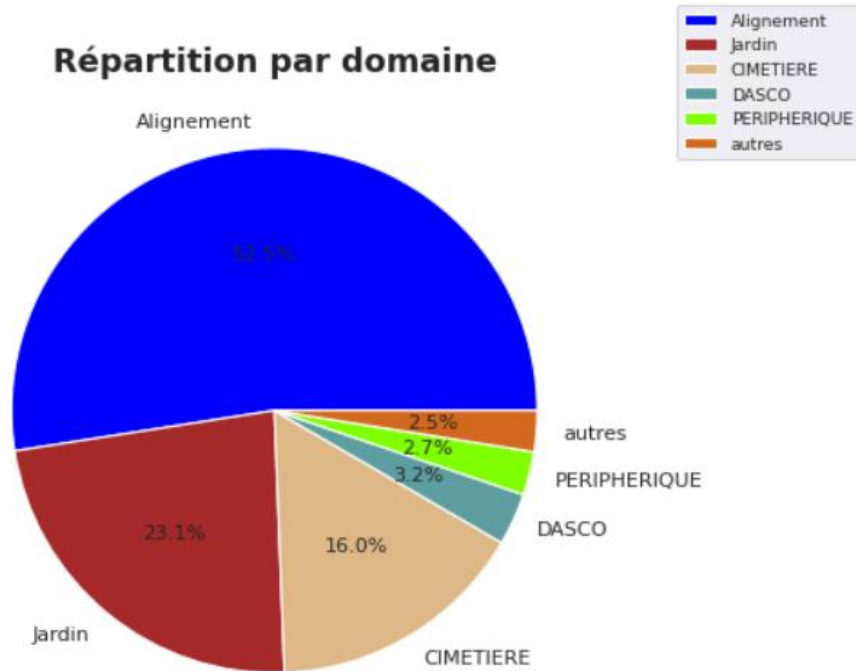
Geopandas & shapely.geometry : carto sur un shp de Paris

plotly.express : carte interactive en basant sur json de Paris

Folium : carte interactive avec Leaflet

Synthèse de l'analyse de donnée : domanialité

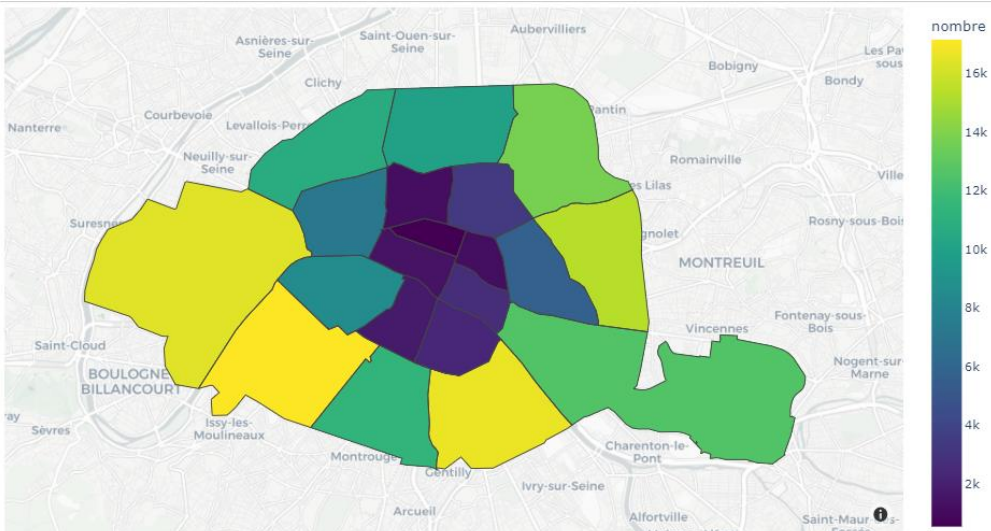
- Les arbres proches des chaussées représentent plus que la moitié (52,4%) du patrimoine arboré à Paris, suivis par les jardins et les cimetières.



- Nous pouvons observés qu'à Paris, la plupart des routes sont abrités.

Synthèse de l'analyse de donnée : arrondissement

- ▶ Le platane reste le plus répandu dans les arrondissements de Paris.
- ▶ Les arbres remarquables sont plus présents dans le 16ème arrondissement

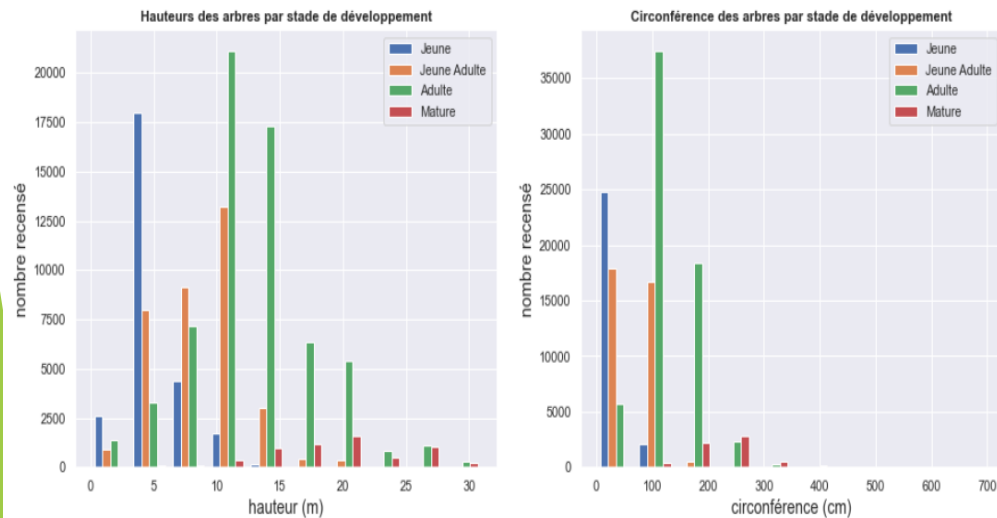
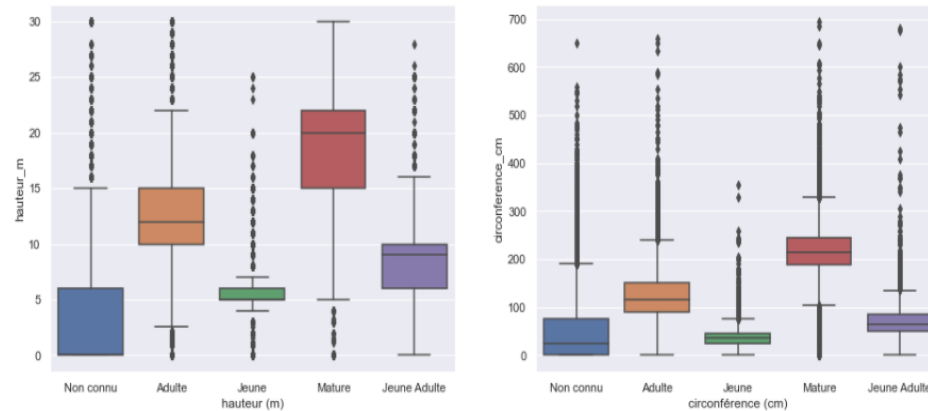


arrondissement	nombre	hauteur_m(moyenne)	circonference_cm(moyenne)	le plus présent	remarquables
BOIS DE BOULOGNE	3974	8.2	66.1	Aesculus	0
BOIS DE VINCENNES	11480	8.4	69.0	Tilia	0
HAUTS-DE-SEINE	5298	1.5	16.8	Acer	0
PARIS 10E ARD	3382	10.2	83.1	Platanus	1
PARIS 11E ARD	5645	10.2	81.7	Platanus	1
PARIS 12E ARD	12552	8.9	82.7	Platanus	23
PARIS 13E ARD	16476	7.7	72.4	Platanus	4
PARIS 14E ARD	11381	9.2	97.4	Platanus	10
PARIS 15E ARD	17121	6.3	69.8	Tilia	7
PARIS 16E ARD	16373	10.0	93.7	Platanus	49
PARIS 17E ARD	10751	7.8	80.6	Platanus	7
PARIS 18E ARD	9981	8.2	79.6	Platanus	10
PARIS 19E ARD	13672	8.5	90.8	Platanus	10
PARIS 1ER ARD	1406	8.4	82.3	Platanus	1
PARIS 20E ARD	15327	8.6	90.8	Acer	15
PARIS 2E ARD	546	9.4	76.6	Platanus	0
PARIS 3E ARD	1209	9.9	82.3	Platanus	3
PARIS 4E ARD	2735	10.3	90.2	Platanus	6
PARIS 5E ARD	2350	11.1	94.4	Platanus	6
PARIS 6E ARD	1755	11.5	89.0	Platanus	2
PARIS 7E ARD	8548	11.4	93.6	Platanus	11
PARIS 8E ARD	7238	11.4	104.4	Aesculus	6
PARIS 9E ARD	1167	10.0	81.9	Platanus	2
SEINE-SAINT-DENIS	11564	4.5	50.8	Aesculus	0
VAL-DE-MARNE	7569	7.2	68.8	Platanus	0
ALL	199500	-	-	-	174

- ▶ Plus qu'on s'approche du centre de Paris, moins on trouve d'espaces arborées.

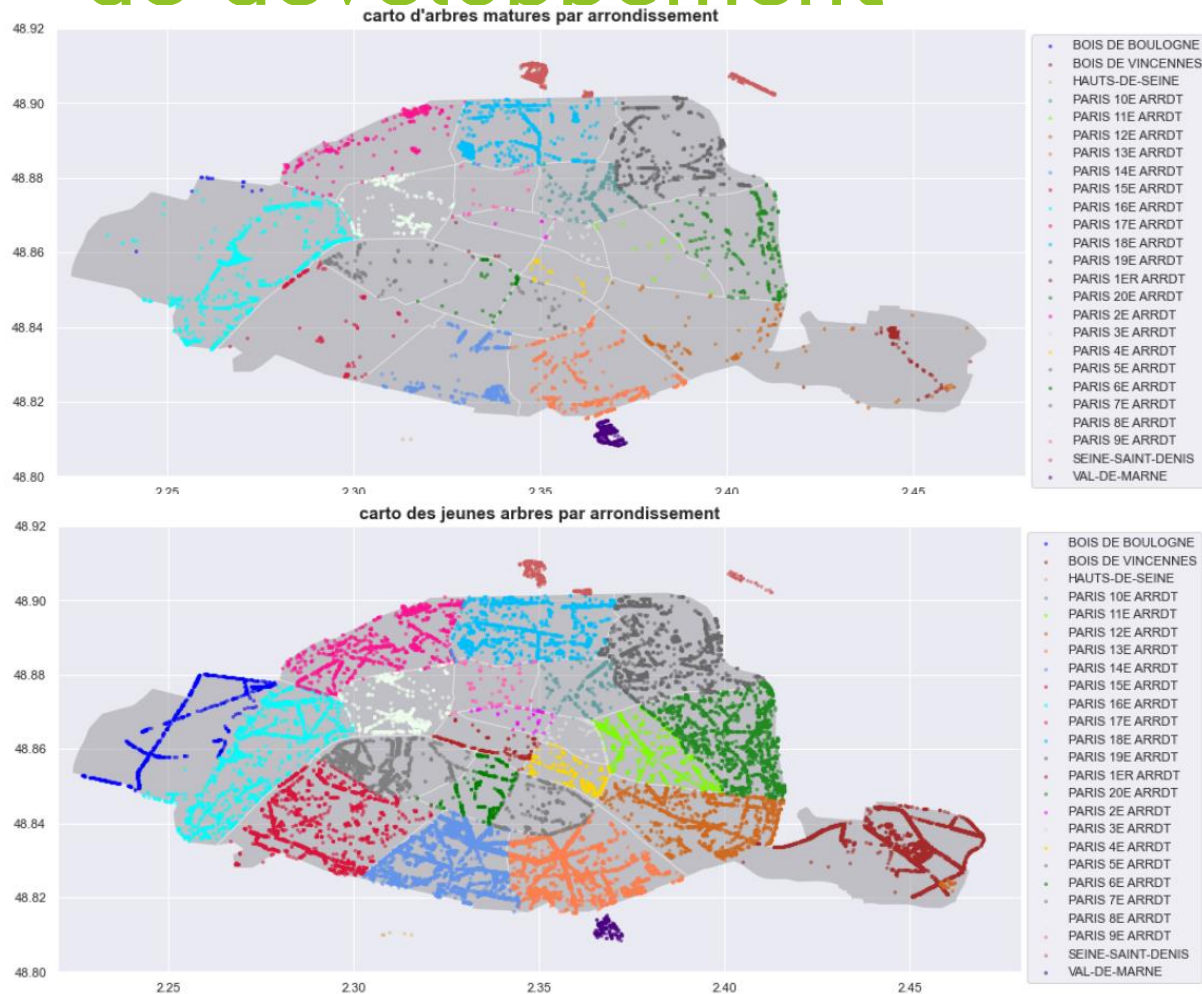
Synthèse de l'analyse de donnée : stade de développement

- Les arbres non identifiés sont probablement des arbres nouvellement plantés au vu de leur tailles.



- Relativement peu d'arbres ont atteint la maturité.
- La plupart des jeunes arbres mesurent environ 5m alors à l'âge adulte, ils peuvent attendre entre 10 et 15 mètres.
- Un effort de renouvellement du patrimoine sylvicole est visible à Paris, représenté par le nombre important de jeunes d'arbres.

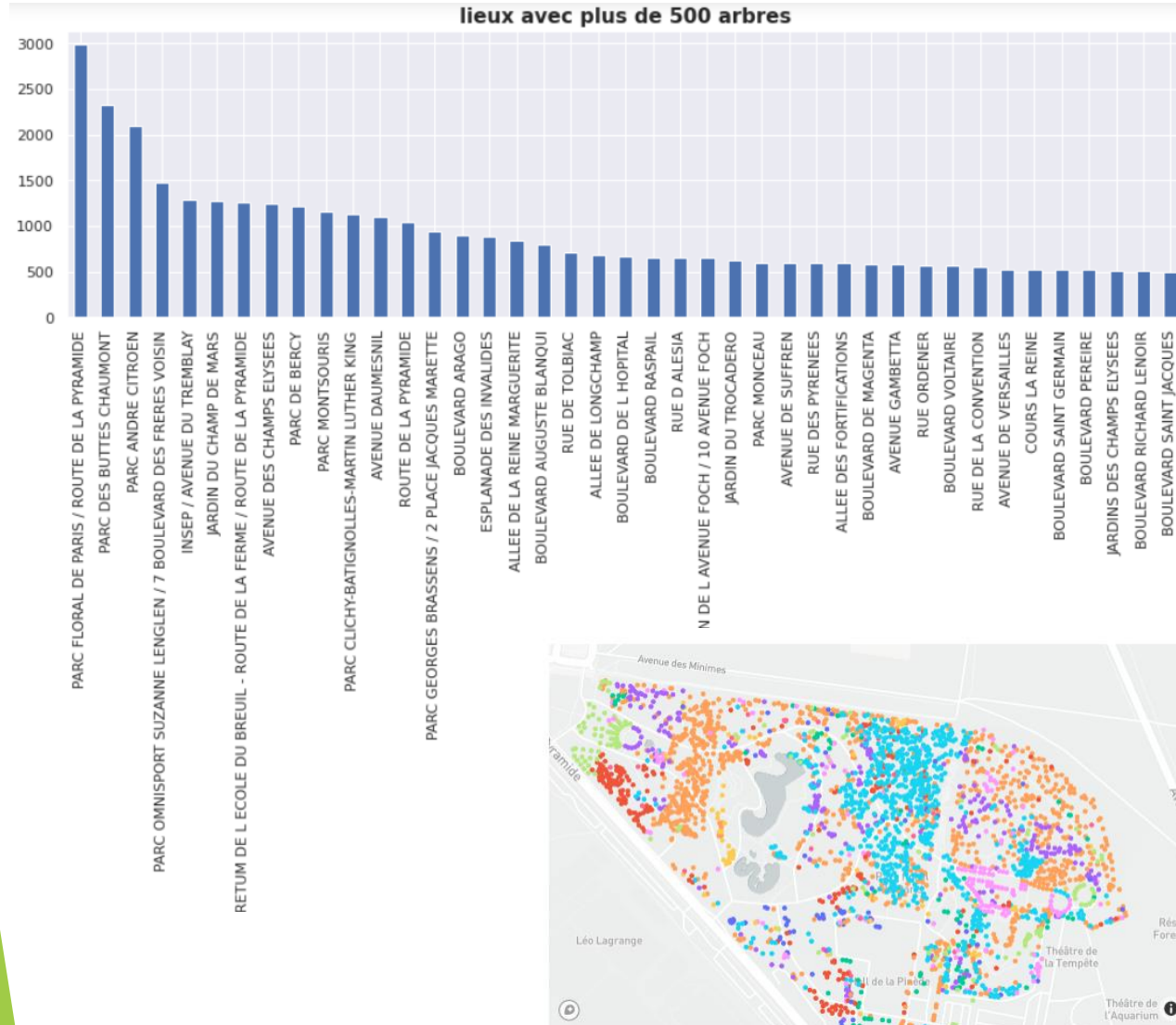
Synthèse de l'analyse de donnée : stade de développement



► Pour la sécurité des parisiens, les aïeux doivent faire l'objet d'une surveillance particulière et réaliser des diagnostics de santé réguliers.

► Les arbres de moins de 3 ans nécessitent des soins particuliers afin de garantir une bonne reprise et un développement harmonieux (Arrosage régulier, tuteur, tailles de formation...)

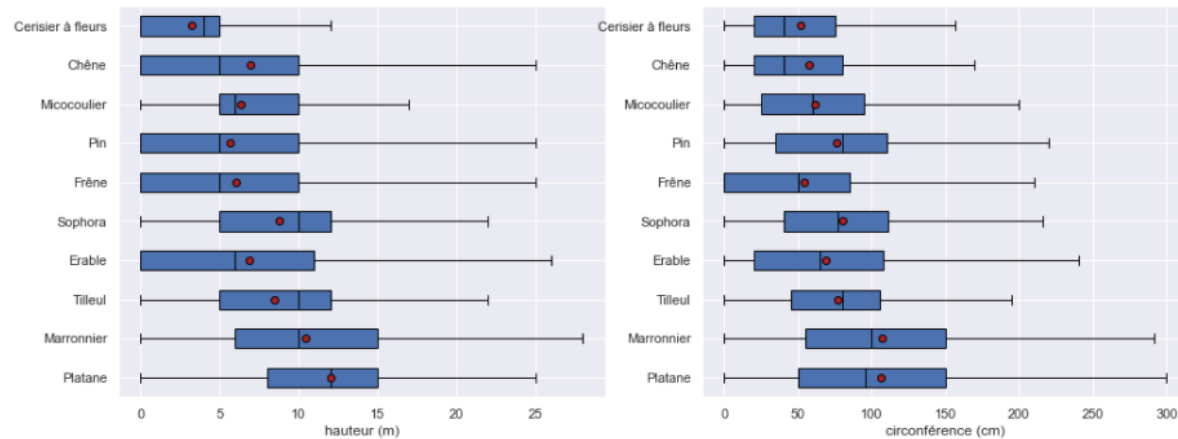
Synthèse de l'analyse de donnée : lieu



- Nous pouvons constater que les parcs et les boulevards ont une concentration importante d'arbres, ce qui confirme le résultat d'analyse par domanialité.
- Le parc floral de Paris possède un nombre impressionnant (~3000) d'arbres

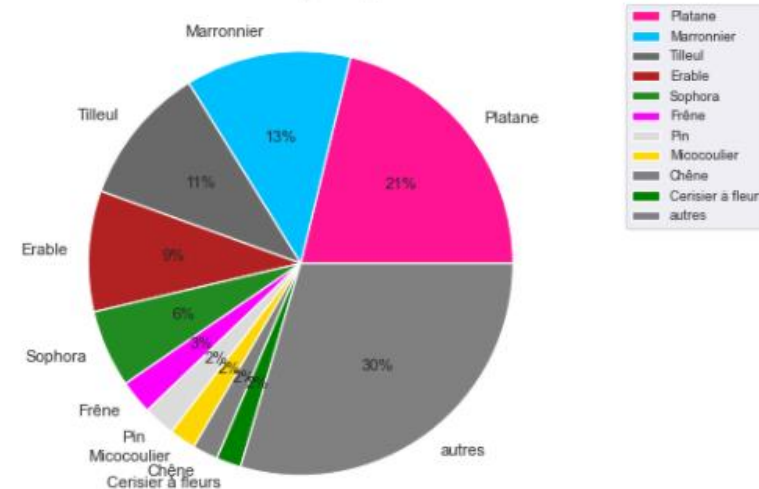
- Nous pouvons y trouver un petit bois de pins (bleu)

Synthèse de l'analyse de donnée : essence



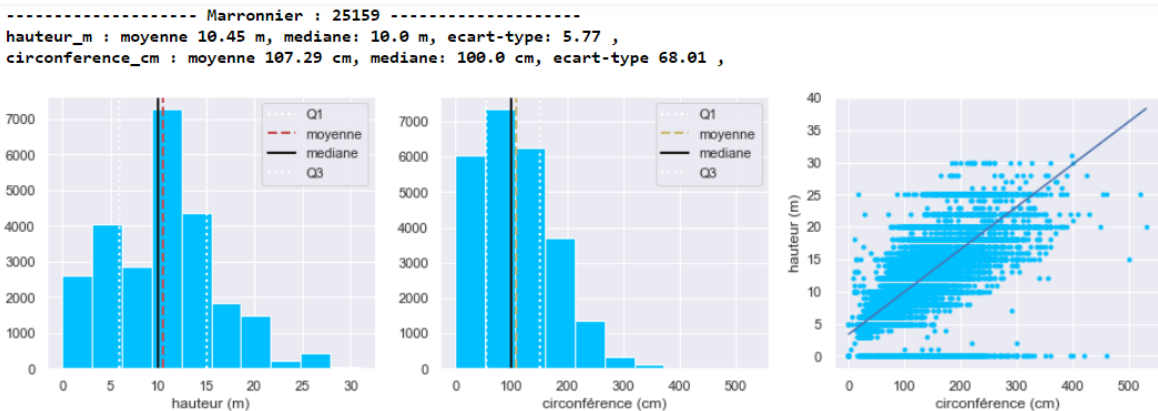
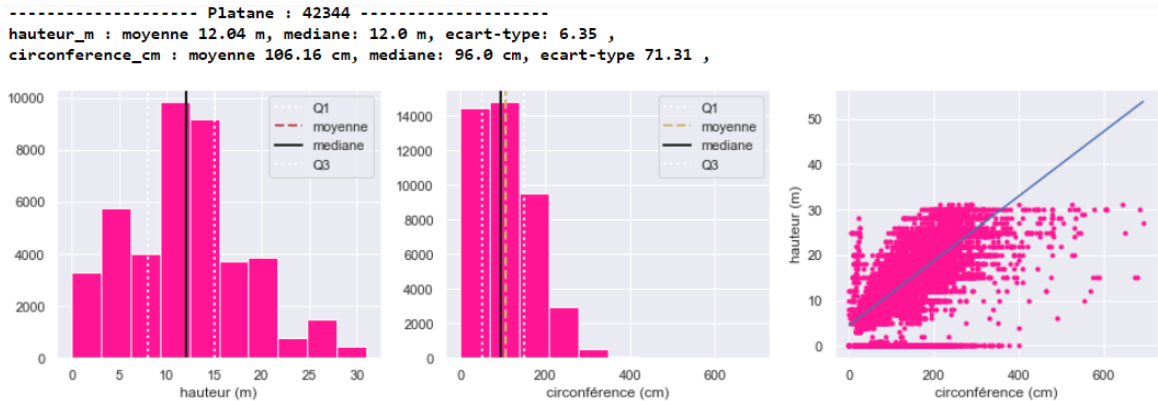
Nombre d'essences présent à Paris : 193

10 essences les plus répandues



- Nous pouvons croiser plus de platanes (42344), marronniers(25207), tilleuls(21305) et érables(18389) dans Paris, eux quatre représentent plus que la moitié d'espèces sylvicoles parisiennes.
- De manière générale, es cerisiers à fleurs sont plus petits de taille, alors les platanes se démarquent des autres par sa grandeur.

Synthèse de l'analyse de donnée : essence

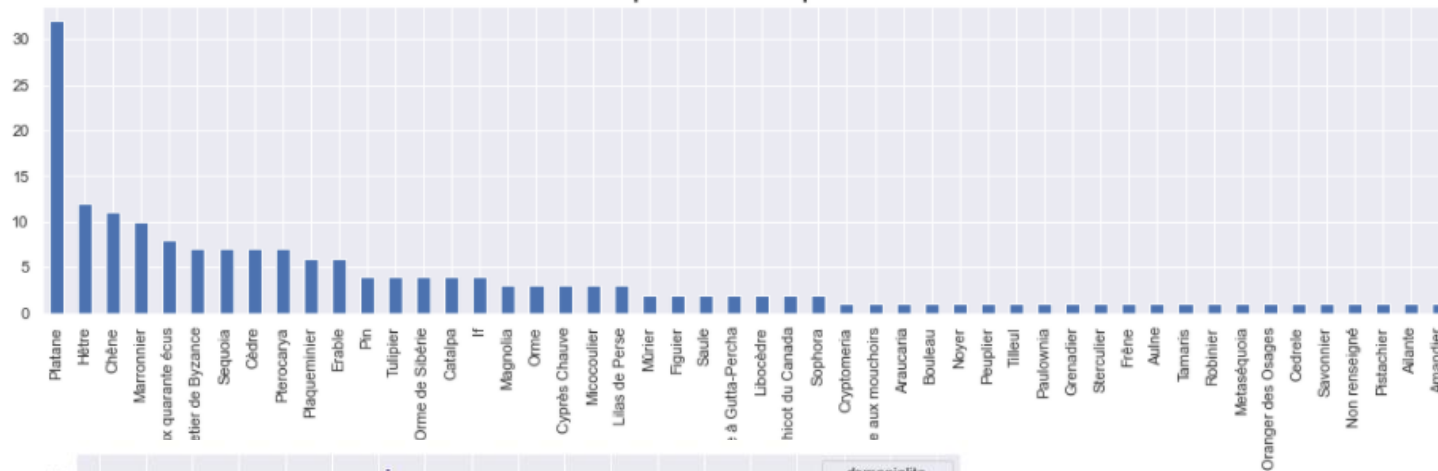


- Les platanes et les marronniers ont des troncs relativement imposants.
- La projection linéaire nous montre que, de taille égale, les platanes ont tendance à être plus hauts que les marronniers.

Synthèse de l'analyse de donnée : remarquable

184 de spécimens remarquables à Paris appartiennent à 50 essences d'arbres différentes.

spécimens remarquables

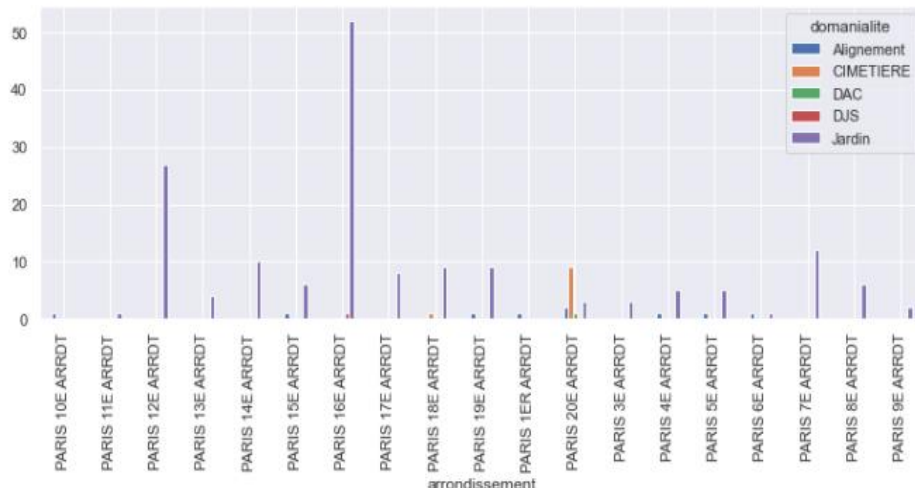


► Les quelques 184 spécimens remarquables répertoriés à Paris appartiennent à 50 essences d'arbres différentes.

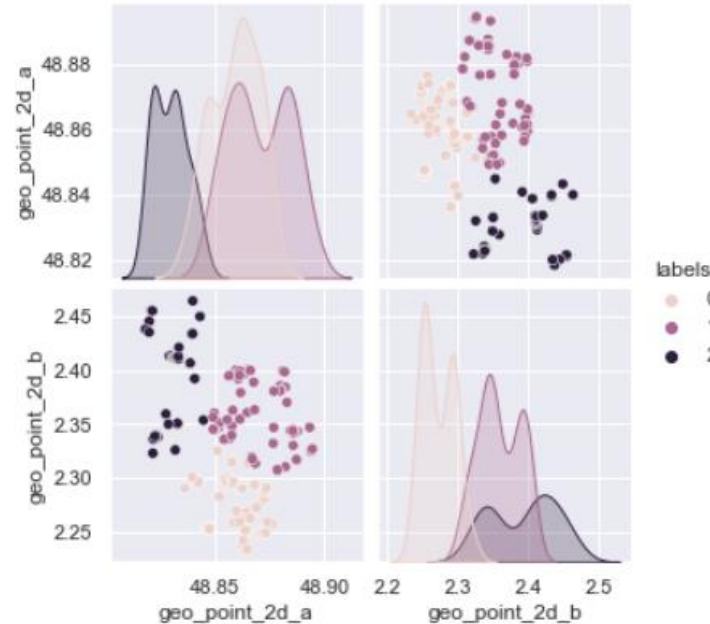
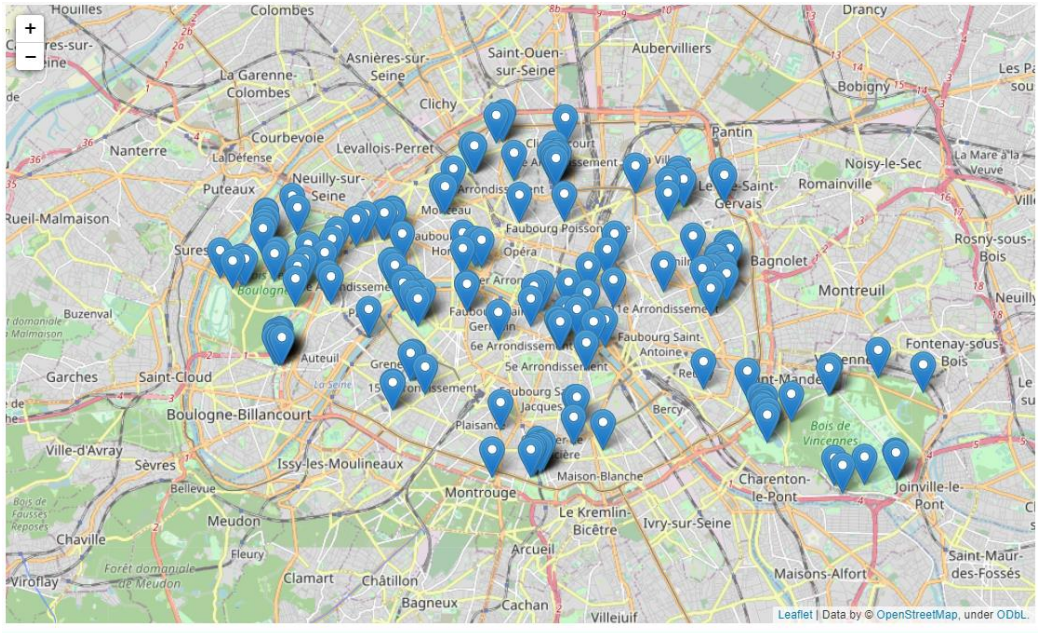
► Les platanes, hêtres et marronniers sont les plus représentés.

► Ils se trouvent majoritairement dans les jardins (163), mais aussi en voie publique (9), ou dans les cimetières (10).

► Le 16^{ème} arrondissement recense plus de sujets remarquables (53).



Synthèse de l'analyse de donnée : remarquable



- Les arbres remarquables doivent être surveillés de près, et son évolution est observée avec attention et chaque année. Kmeans permet de regrouper les sujets par apport à leur proximité géographique afin de faciliter l'organisation des tournées.