1. CONTEXTE

RAPPEL DU CONTEXTE

l'ONG "**Data is for Good**" propose des challenges de Data Science en ligne sur des thématiques ayant trait au bien commun. Des associations et collectivités publiques sponsorisent ces challenges.

Rôle

Vous êtes fraîchement établi en tant qu'expert indépendant spécialisé en intelligence artificielle.
Vous participez régulièrement à des concours pour vous **faire la main sur de nouveaux sujets**.

Vous avez décidé de participer à un challenge proposé par la ville de Paris

Vos résultats contribueront à une optimisation des tournées pour l'entretien des arbres de la ville.

OBJECTIFS DE L'ANALYSE EXPLORATOIRE

contribuer à une optimisation des tournées pour l'entretien des arbres de la ville.

2. DEMARCHE METHODOLOGIQUE D'ANALYSE DE DONNEES

TABLE DES MATIERES

Paramétrage & Chargement du jeu de données

01

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import seaborn as sns
import pip
```

1. Installation et activation de l'environnement virtuel
2. Installation de Jupiter Notebook
3. Installation , importations et verifications des versions des librairies Python

```
data = pd.read_csv("p2-arbres-fr.csv", sep=";")
```

4. Chargement du jeux de données

5. Affichage des 5 premières lignes du jeux de données

```
data.head()
```

	id	type_emplacement	domanialite	arrondissement	complement_addresse	numero	lieu	id_emplacement	libelle_francais	genre	espece	variete	circonference_cm	hauteur_m	stade_developpement	remarquable	geo_point_2d_a	geo_point_2d_b
0	99874	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	19	Marronnier	Aesculus	hippocastanum	NaN	20	5	NaN	0.0	48.857620	2.320962
1	99875	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	20	If	Taxus	baccata	NaN	65	8	A	NaN	48.857656	2.321031
2	99876	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	21	If	Taxus	baccata	NaN	90	10	A	NaN	48.857705	2.321061
3	99877	Arbre	Jardin	PARIS 7E ARRDT	NaN	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	22	Erable	Acer	negundo	NaN	60	8	A	NaN	48.857722	2.321006
4	99878	Arbre	Jardin	PARIS 17E ARRDT	NaN	NaN	PARC CLICHY-BATIGNOLLES-MARTIN LUTHER KING	000G0037	Arbre à miel	Tetradium	daniellii	NaN	38	0	NaN	NaN	48.890435	2.315289

Exploration du jeu de données

02

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200137 entries, 0 to 200136
Data columns (total 18 columns):
 #   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
dtypes: float64(4), int64(3), object(11)
memory usage: 27.5+ MB
```

```
data.describe()
```

	id	numero	circonference_cm	hauteur_m	remarquable	geo_point_2d_a	geo_point_2d_b
count	2.001370e+05	0.0	200137.000000	200137.000000	137039.000000	200137.000000	200137.000000
mean	3.872027e+05	NaN	83.380479	13.110509	0.001343	48.854491	2.348208
std	5.456032e+05	NaN	673.190213	1971.217387	0.036618	0.030234	0.051220
min	9.987400e+04	NaN	0.000000	0.000000	0.000000	48.742290	2.210241
25%	1.559270e+05	NaN	30.000000	5.000000	0.000000	48.835021	2.307530
50%	2.210780e+05	NaN	70.000000	8.000000	0.000000	48.854162	2.351095
75%	2.741020e+05	NaN	115.000000	12.000000	0.000000	48.876447	2.386838
max	2.024745e+06	NaN	250255.000000	881818.000000	1.000000	48.911485	2.469759

```
print("Nombre de colonnes :", data.shape[1])
print("Nombre de lignes :", data.shape[0])
```

```
Nombre de colonnes : 18
Nombre de lignes : 200137
```

Data.info()

Fourni un résumé sur le contenu du DataFrame.. Nombre de lignes, colonnes, leur noms, le type de données (int, float) et l'estimation de la mémoire utilisée.

Data.describe()

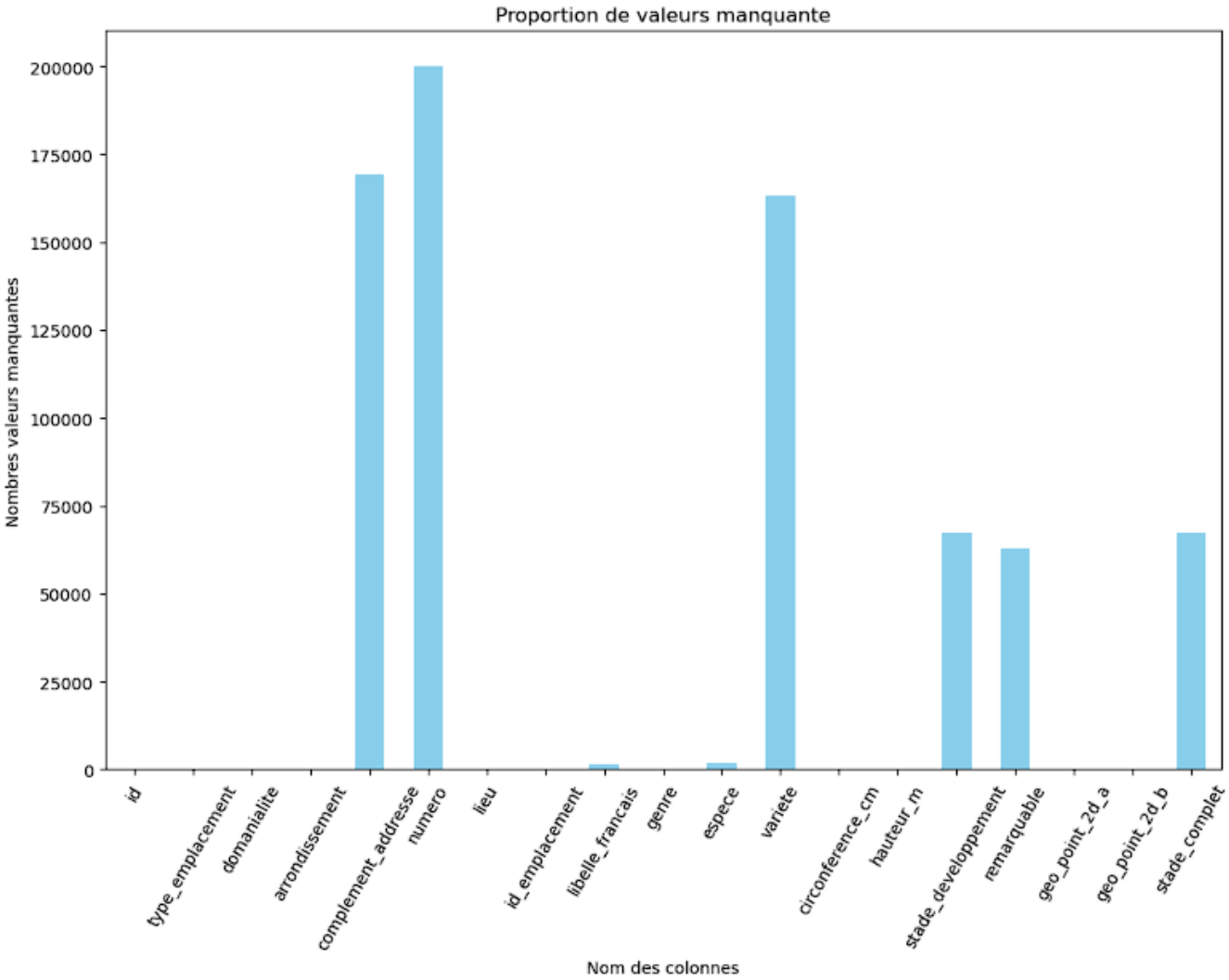
Calcul automatiquement plusieurs statistiques sur les colonnes numérique du dataset et permet de repérer d'éventuelles anomalies.

Data.shape

Permet d'obtenir les dimensions du jeu de données.

Nettoyage du jeu de données

03



Les valeurs manquantes

Nous constatons de nombreuses valeurs manquantes dans certaines colonnes

Les colonnes "numéro", "complement-addresse" et "variete" ne sont pas les colonnes les plus pertinentes pour notre objectif et en raison d'un manque de données à plus de 80%, nous les écarterons de notre analyse

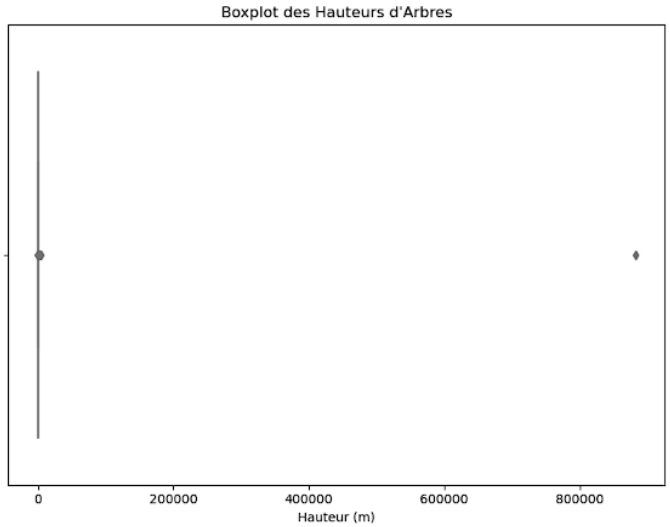
Les doublons

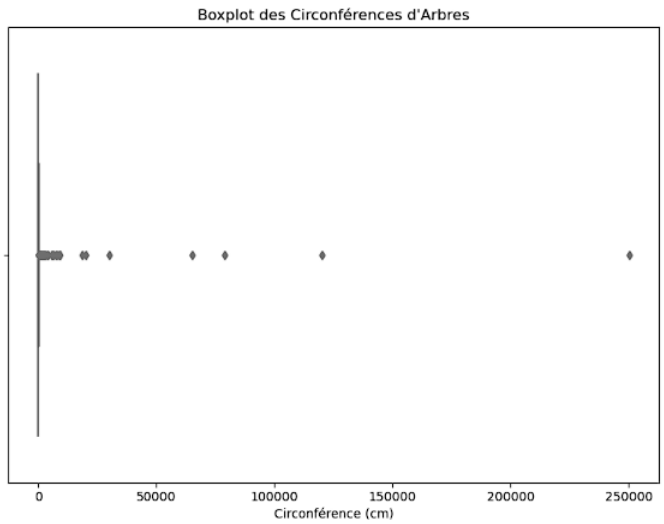
Nous constatons qu'il n'y a pas de doublons ce qui facilite la suite de notre analyse

```
Il manque 200137 valeurs, soit 100.0 % dans la colonne numero
Il manque 169235 valeurs, soit 84.56 % dans la colonne complement_addresse
Il manque 163360 valeurs, soit 81.62 % dans la colonne variete
Il manque 67205 valeurs, soit 33.58 % dans la colonne stade_developpement
Il manque 63098 valeurs, soit 31.53 % dans la colonne remarquable
```

```python
duplicates = data[data.duplicated()]

if len(duplicates) > 0:
    print("Notre jeu de données contient", len(duplicates), "doublon(s).")
else:
    print("Notre jeu de données ne contient pas de doublons.")
```



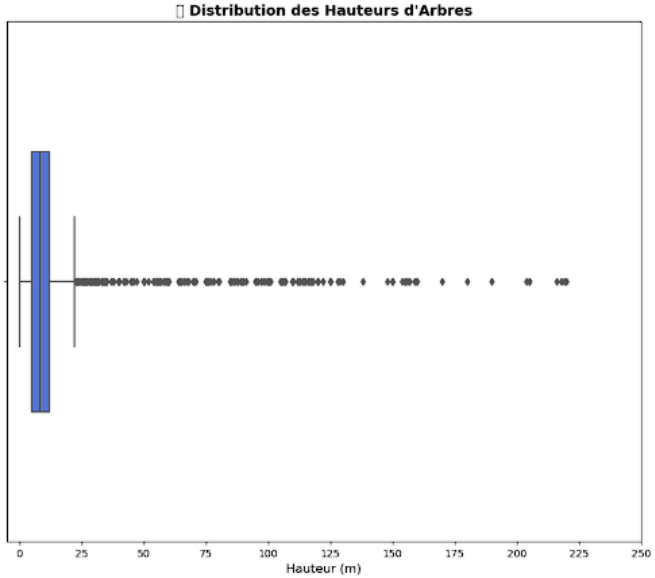


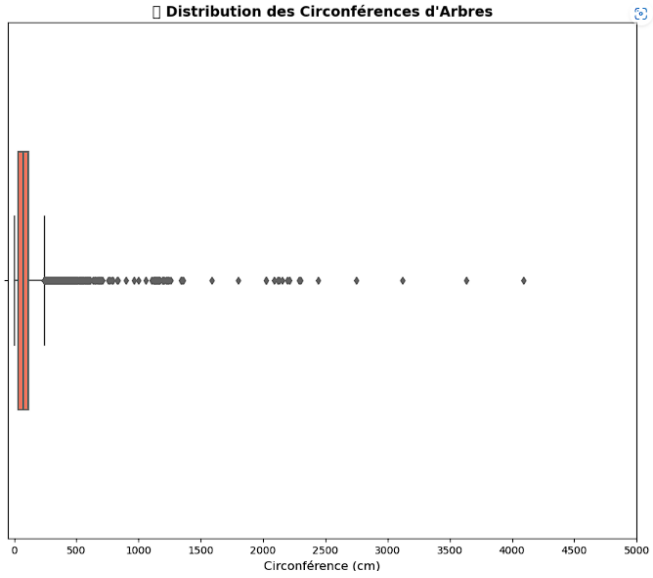
Gestion des outliers

A l'aide de ces boxplot, on est en mesure d'identifier les valeurs aberrantes.

Les graphiques du haut indiquent la hauteur d'un arbre allant au-delà de 800km (gauche) et la circonférence d'un arbre allant jusqu'à 2.5 km (droite)

Les graphiques du bas nous montrent que la majeur partie des arbres ont une hauteur entre 25m et 125m et une circonférence entre 2.5m et 7.5m





```
Pourcentage d'arbres avec hauteur = 0 : 19.60%
Pourcentage d'arbres avec circonférence = 0 : 12.92%
Pourcentage d'arbres avec hauteur = 0 ET circonférence = 0 : 12.74%
```

```
📊 Distribution des stades de développement pour les arbres avec hauteur = 0 et circonférence = 0 :
stade_developpement
Non spécifié    99.070625
J                0.478413
A                0.219599
JA               0.145092
M                0.086271
Name: proportion, dtype: float64
```

Les valeurs = 0

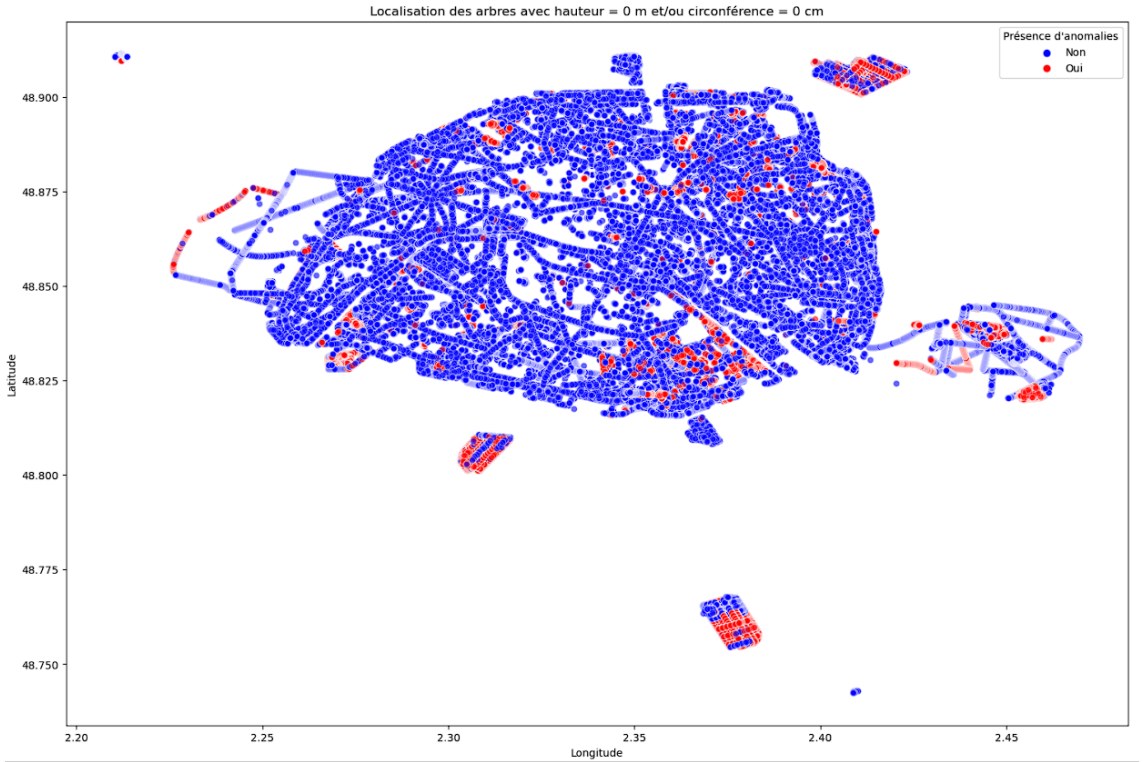
après avoir identifié les arbres dont la circonférence et/ou la hauteur = 0, on remarque quelques incohérences sur certains arbres adultes ou matures qui devraient avoir un certains volumes en raison de leur stade de développement mais la plupart sont bien des jeunes ce qui expliquerai donc un manque de précision.

plutôt que de les supprimer et oublier des arbres réellement existant ou de les remplacer par des estimations qui pourraient fausser l'analyse, je les conserve pour que les agents d'entretient confirme l'erreur ou non (shema ci-dessous).

Les hauteurs et circonférences excessives

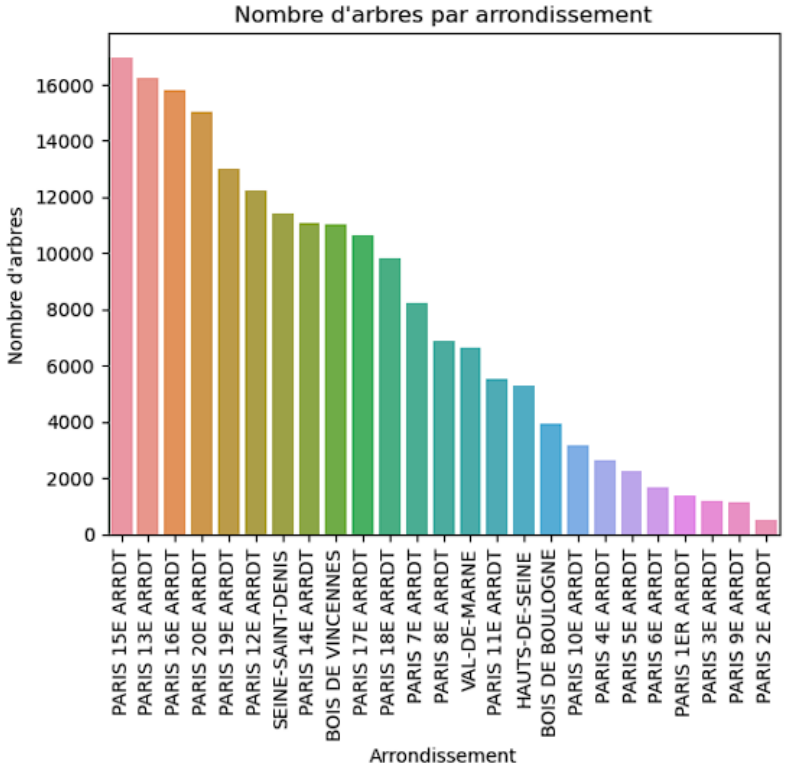
En réalisant une approche métier et en se renseignant sur les tailles des arbres de paris, j'ai utilisé la méthode interquatilles pour identifier et exclure les valeurs aberrantes

```
Nombre de valeurs aberrantes dans la colonne 'hauteur_m' : 3420
Nombre de valeurs aberrantes dans la colonne 'circonference_cm' : 3327
```



Analyse et représentations graphiques

04

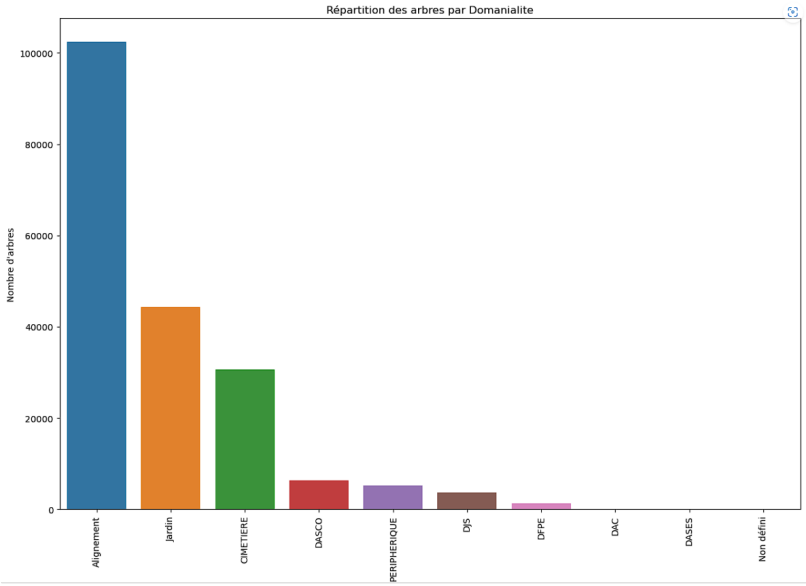


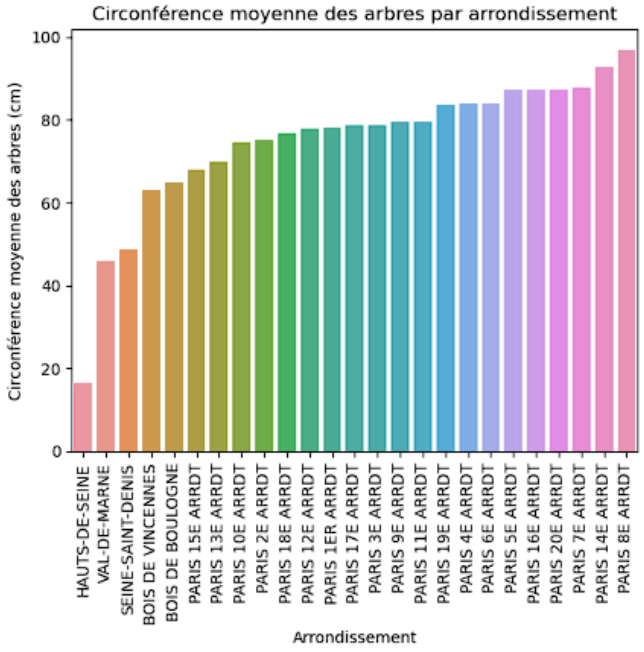
La répartitions des arbres dans Paris

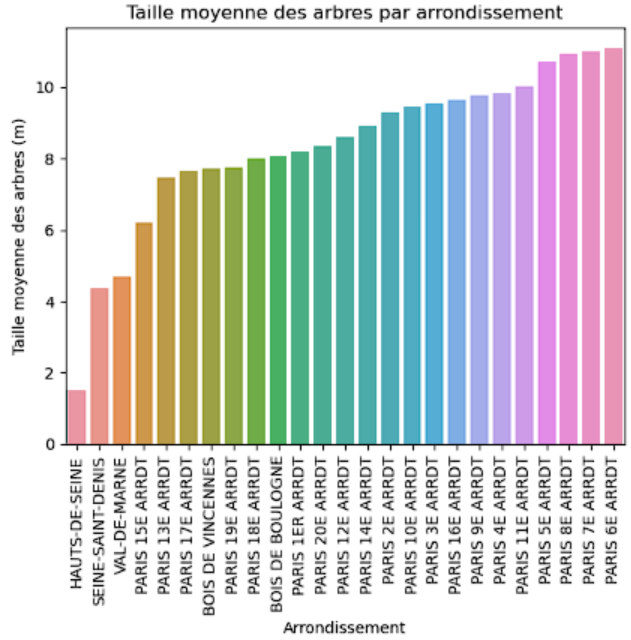
On observe que le 15ème arrondissement est le plus verdoyant contrairement au 2ème arrondissement qu'il est le moins boisé. (graphique du haut)

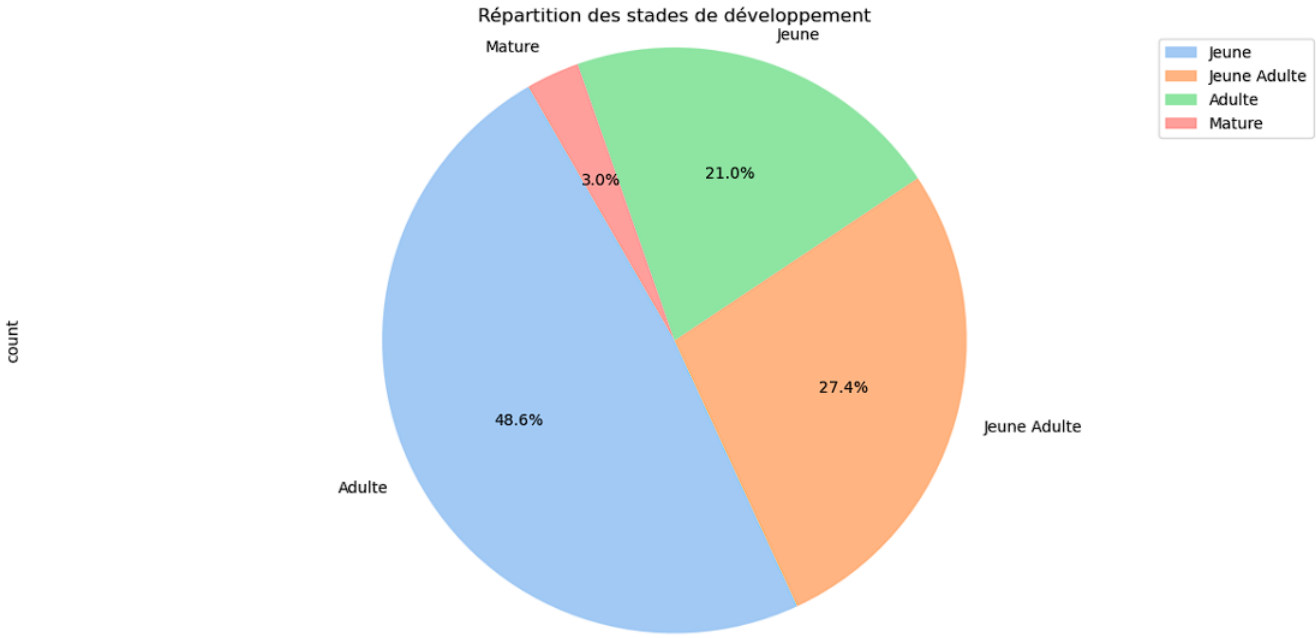
on remarque le la plus grande partie des arbres se trouvent dans l'alignement.

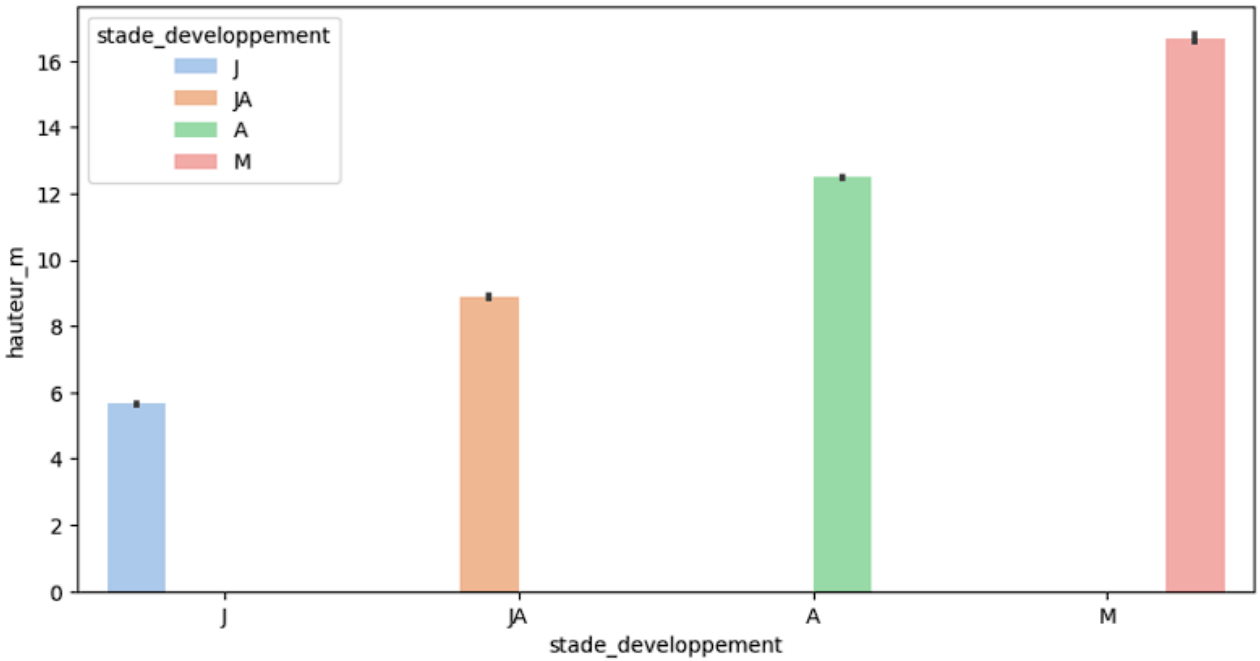
En ce qui concerne les tailles et circonferences des arbres les plus grands et volumineux se concentrent dans les arrondissements comme 6ème, 7ème, 8ème, 5ème et on note un retard pour les Haut-De-Seine, Seine-Saint-Denis, Val-De-Marne







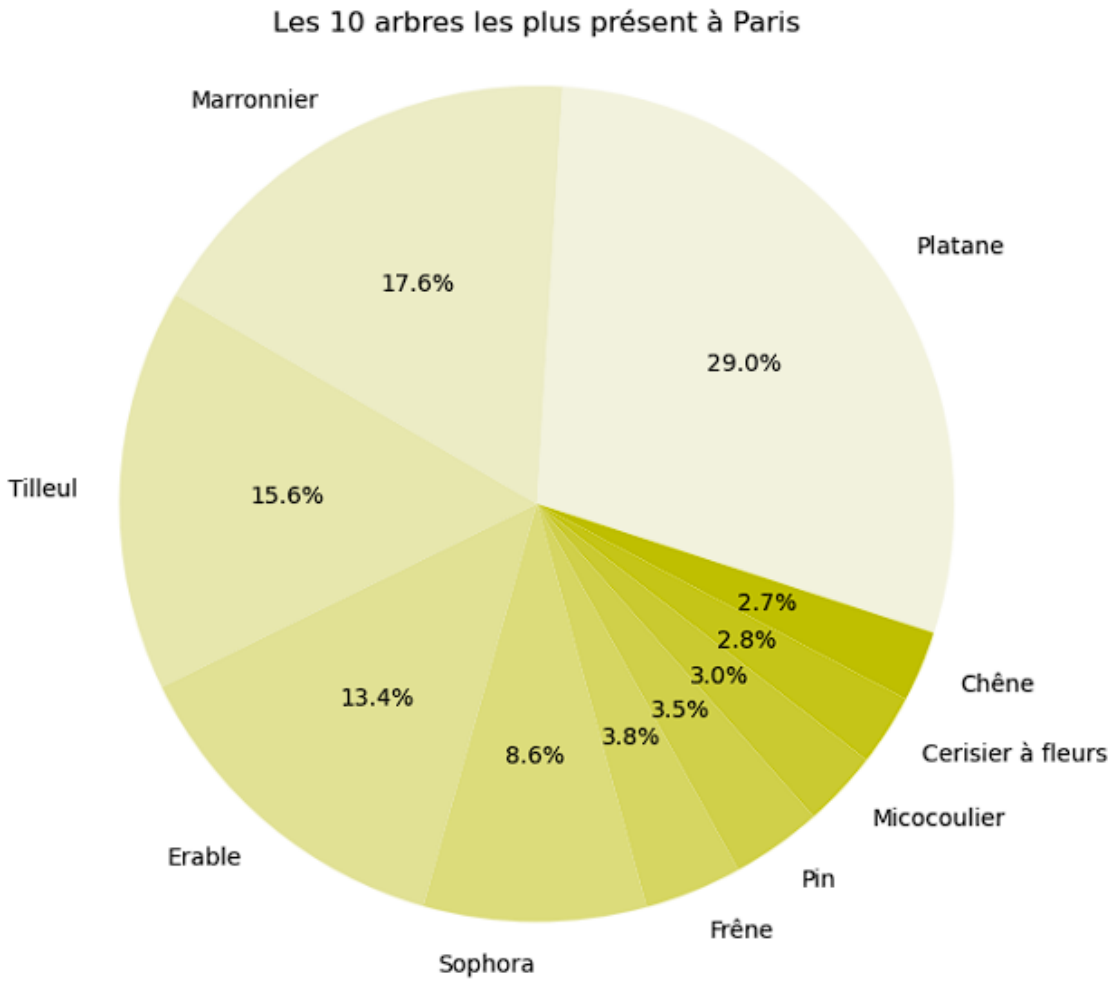




Les stades de développement

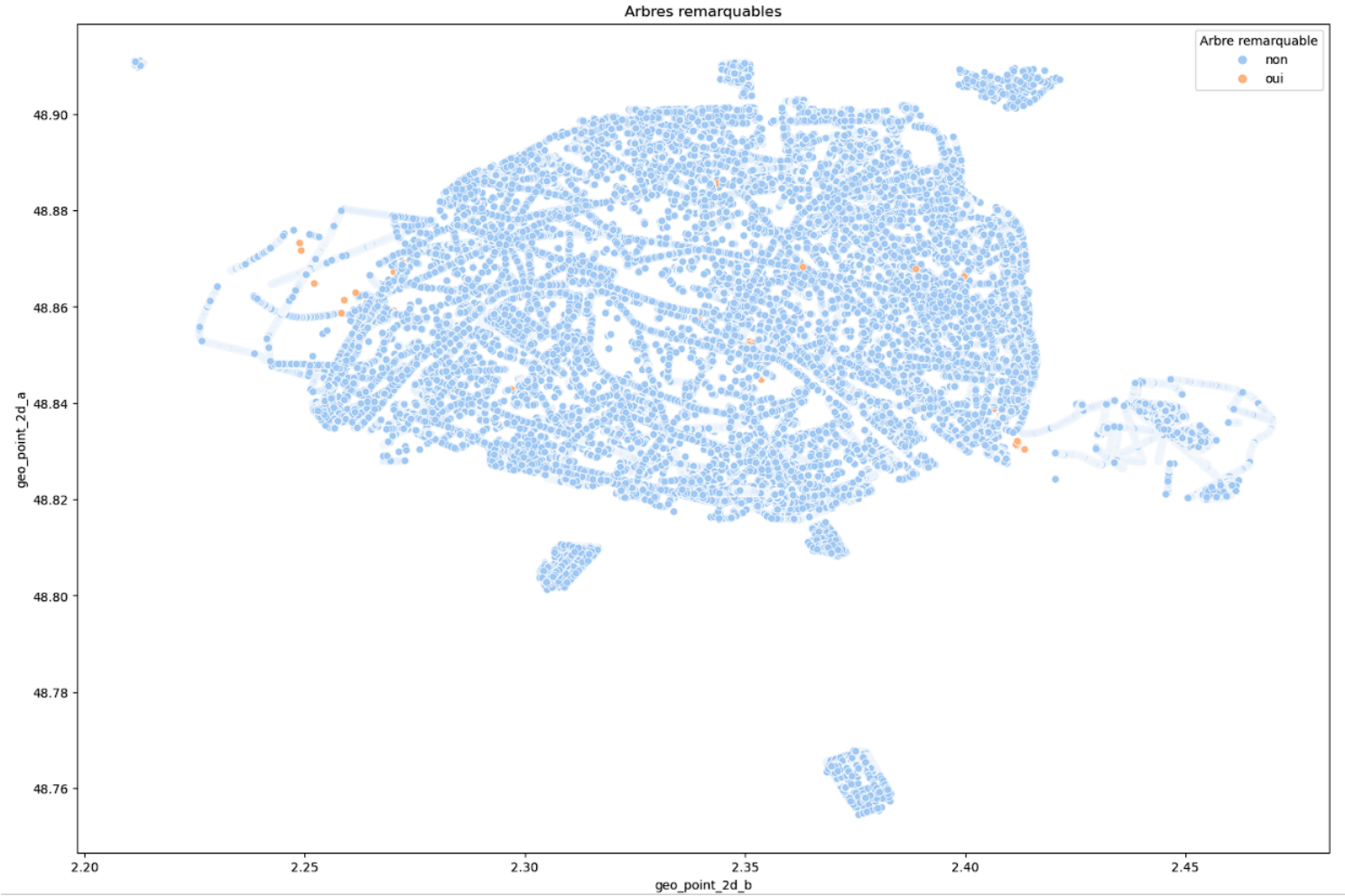
on observe grâce au grapique du haut la répartition des stades de développement entre les differents types de croissances des arbres. La majeure partie des arbres sont des adultes.

on constate grâce au graphique du bas les moyennes des hauteurs pour chaque stade de développement.



Les éspèces de Paris

on remarque un taux important de Platanes, suivi de Marroniers, de Tilleuls et d'Erables. Surement une selection bien choisi pour des raisons spécifiques dans l'entretient et des besoins de bien-être des Parisiens et/ou pour des raisons économiques.



Les arbres remarquables

La ville de Paris compte 78 arbres remarquables, ces arbres sont en moyenne plus grands (12.8 m) que les non remarquables (9.3m) et ont une circonférence beaucoup plus grande (163 cm) que les autres (83 cm)

on peux les observer en orange sur cette carte de Paris.

Conclusion

05

Cette analyse permet de connaitre la localisation des arbres, leurs stades de développement et toutes leurs caractéristiques ainsi que de connaitre les zones les plus verdoyantes ou les moins boisées.

on pourrait faire une analyse plus poussée si on avait plus d'éléments comme le nombre d'employés, le matériel utilisé et les centres de départ des agents

cette première analyse permet aux chefs d'équipe de prendre des décisions sur le matériel adapté aux arbres, la gestion des tournées des équipes, favoriser la diversité des arbres et la confirmations de certaines anomalies

il est possible d'approfondir d'aventage l'analyse au-dela de l'objectif demandé par les interlocuteurs.