Data is for Good : végétalisons la ville

Parcours Ingénieur IA

Katrin-Misel Ponomarjova, P2, 09/06/2022

Contexte : Végétalisons la ville

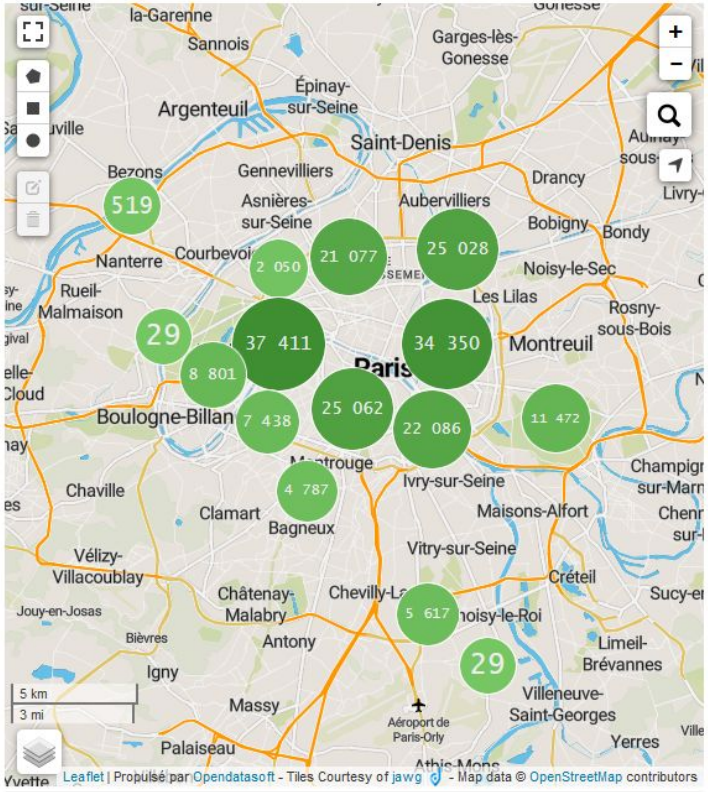
Challenge Data dans le cadre du programme “Végétalisons la ville”

Contribution à une optimisation des tournées pour l’entretien des arbres de la ville

Jeu de données des arbres de la ville de Paris sur opendata.paris.fr

Organisation de la présentation :

1. Présentation générale du jeu de données
2. Démarche méthodologique
3. Synthèse de l’analyse de données

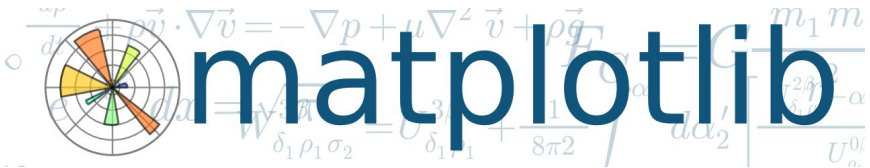


Ressources

In [2]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import missingno
```

- numpy
- pandas
- seaborn
- pyplot
- missingno



Présentation générale du jeu de données

Présentation générale du jeu de données

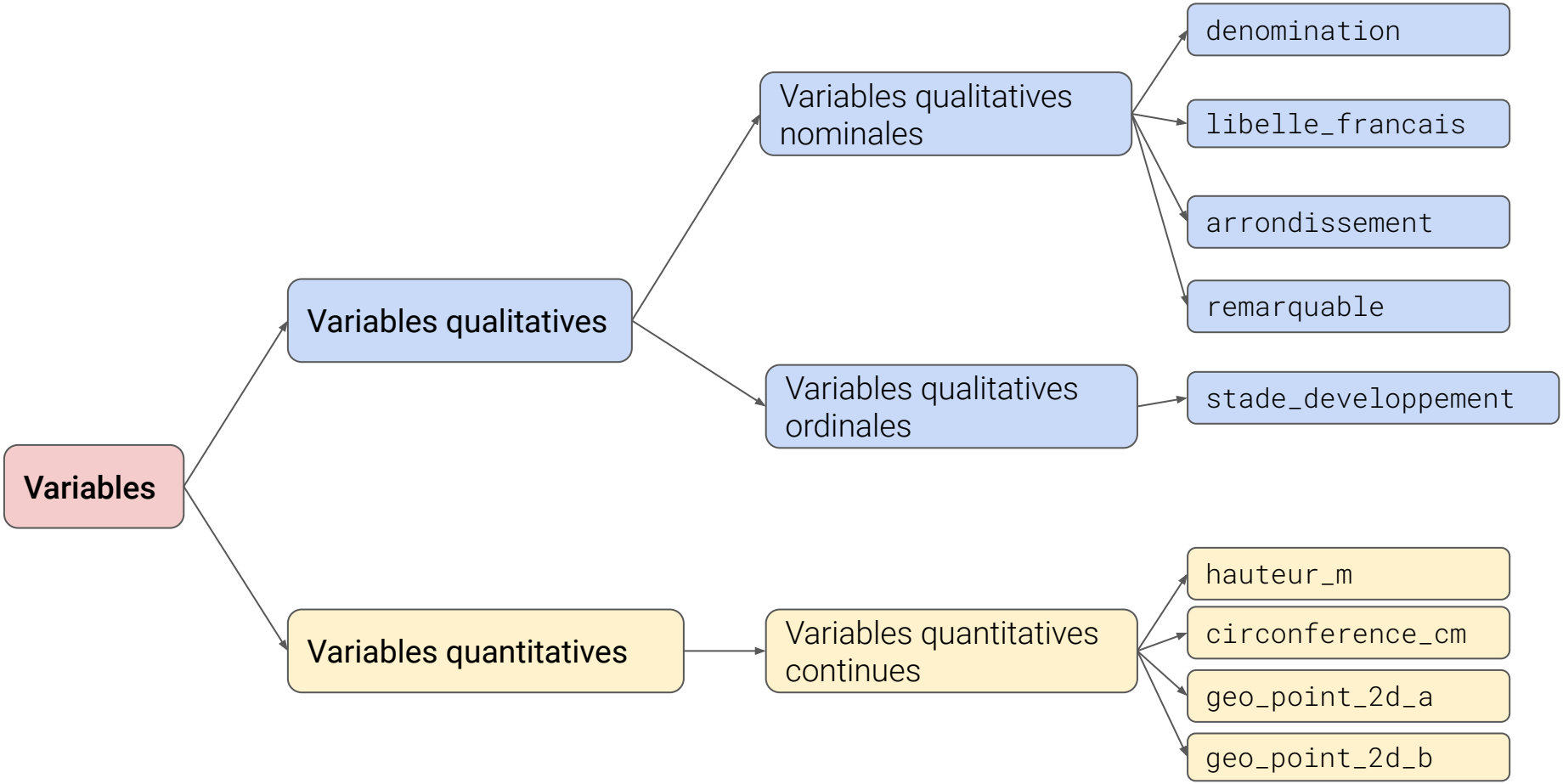
```
In [7]: df.dtypes

Out[7]: id                      int64
        type_emplacement       object
        domanialite            object
        arrondissement         object
        complement_addresse    object
        numero                float64
        lieu                   object
        id_emplacement         object
        libelle_francais       object
        genre                  object
        espece                 object
        variete                object
        circonference_cm        int64
        hauteur_m               int64
        stade_developpement    object
        remarquable           float64
        geo_point_2d_a        float64
        geo_point_2d_b        float64
        dtype: object
```

- 200136 individus
- 18 variables

Variables à exploiter

Nous n’allons pas utiliser toutes les variables dans l’analyse. Ceux qui nous seront utiles sont les suivantes :

- domanialite
- arrondissement
- libelle_francais
- circonference_cm
- hauteur_m
- stade_developpement
- remarquable
- geo_point_2d_a
- geo_point_2d_b


Variables
Variables qualitatives
Variables quantitatives
Variables qualitatives nominales
Variables qualitatives ordinales
Variables quantitatives continues
denomination
libelle_francais
arrondissement
remarquable
stade_developpement
hauteur_m
circonference_cm
geo_point_2d_a
geo_point_2d_b

Démarche méthodologique d’analyse de données

Démarche :

1. Identifier les valeurs aberrantes
2. Nettoyer les valeurs aberrantes par imputation
3. Identifier les valeurs manquantes
4. Nettoyer les valeurs manquantes par imputation ou suppression
5. Supprimer les doublons

Valeurs aberrantes/atypiques

- Détection des valeurs aberrantes avec l’ écart inter-quartile :
  - 1.70% de nos circonférences sont des outliers
  - 1.95% de nos hauteurs sont des outliers.
- Importance de l’hauteur / la circonférence
- Nettoyage

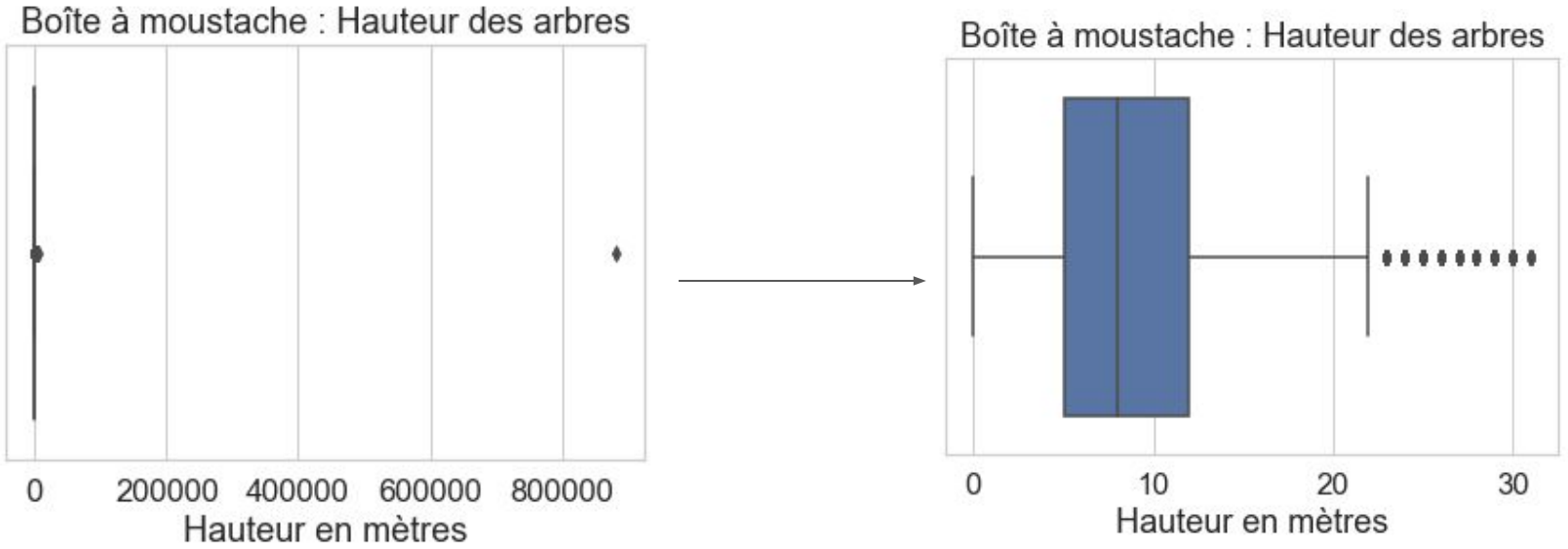
LE PLATANE D’ORIENT DU PARC MONCEAU

Le plus gros arbre de Paris est à admirer dans l’ouest du parc Monceau, au cœur du 8ème arrondissement. Ce platane d’Orient est reconnaissable entre mille à son énorme **tronc noueux qui mesure près de 8 mètres de circonférence**, ses branches bien étalées, ainsi que sa **hauteur de 30 mètres**. Il a été planté ici en 1814, ce qui fait de lui un bicentenaire, mais n’est pas pour autant le plus vieux de Paris, le doyen des platanes se trouvant lui au Jardin des Plantes.

Parc Monceau, 75008

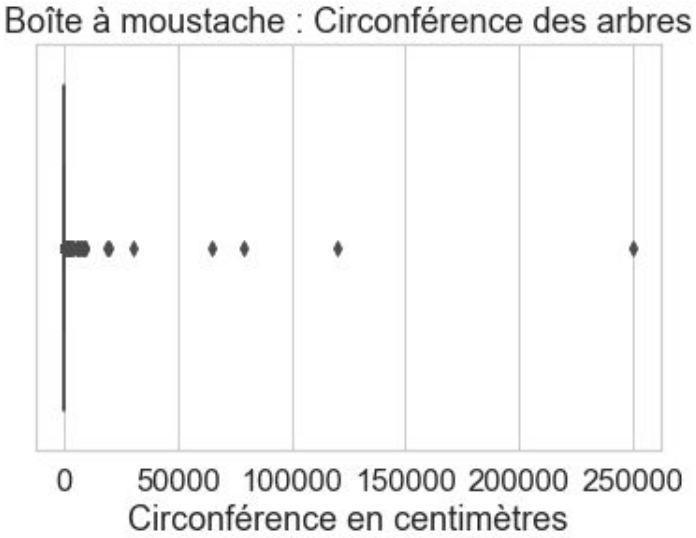
Valeurs aberrantes/atypiques

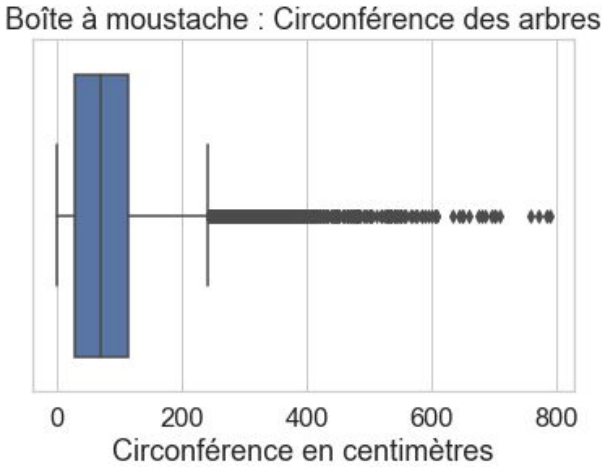
- `hauteur_m` : 558 valeurs > 30m



Valeurs aberrantes/atypiques

- `circonference_cm :` 77 valeurs > 800m





Valeurs aberrantes/atypiques

- 25501 arbres avec une hauteur de 0m et une circonférence de 0cm
- Idée : utiliser le stade de développement pour estimer l'hauteur et la circonférence

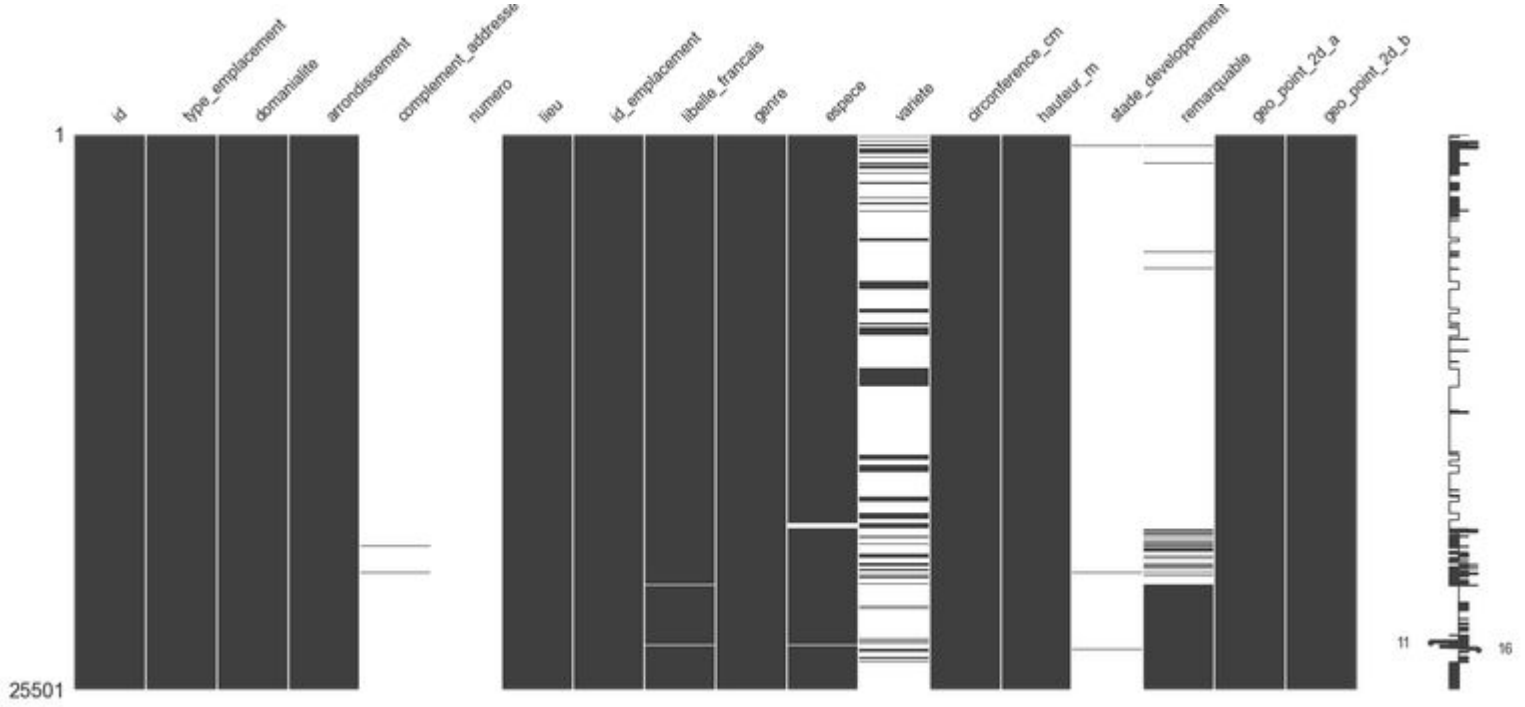
LE FIGARO

Paris : plus de 20.000 arbres plantés pendant l'hiver, selon la mairie

Par Le Figaro avec AFP

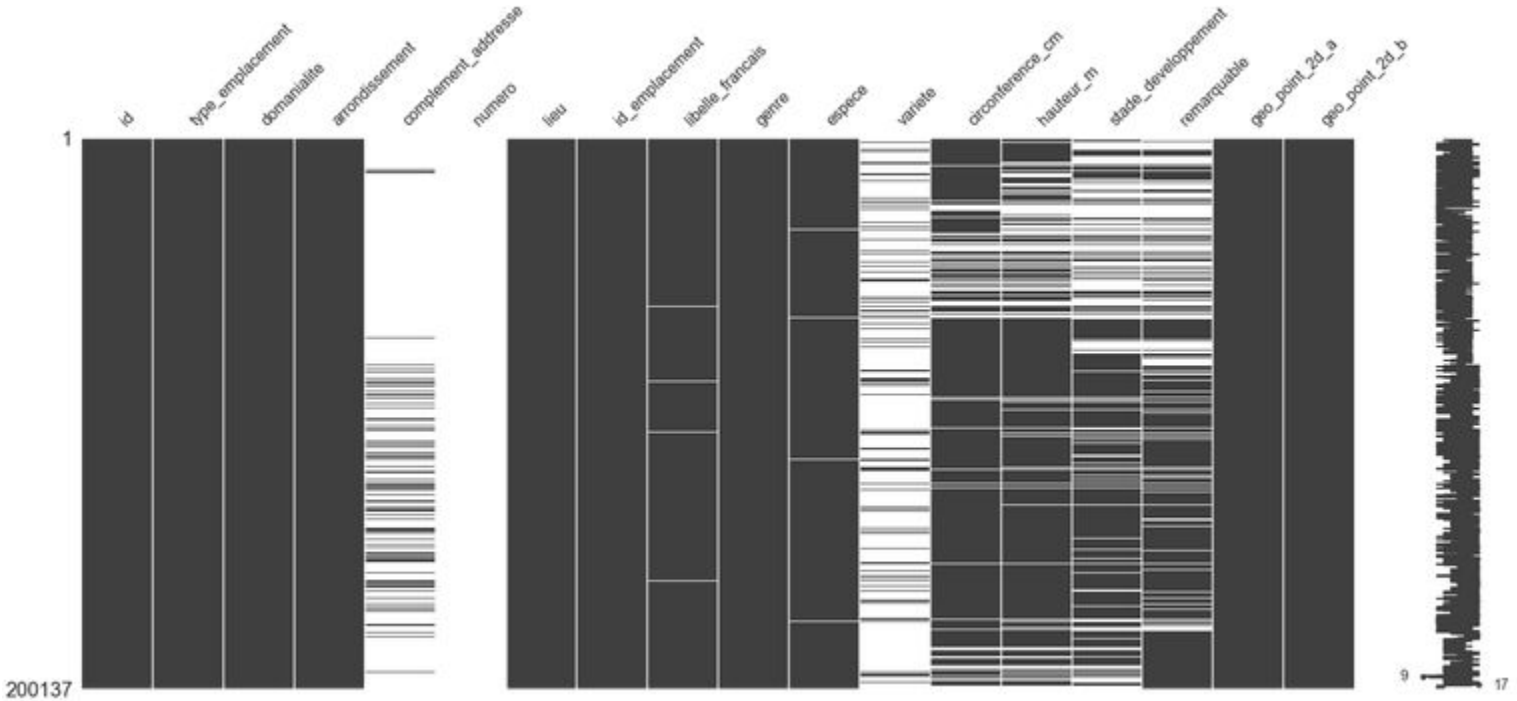
Publié le 21/04/2022 à 20:39, mis à jour le 21/04/2022 à 20:39

missingno.bar pour le dataset de hauteur_m=0, circonference_cm=0



Solution : Manque de données : transformer en NaN

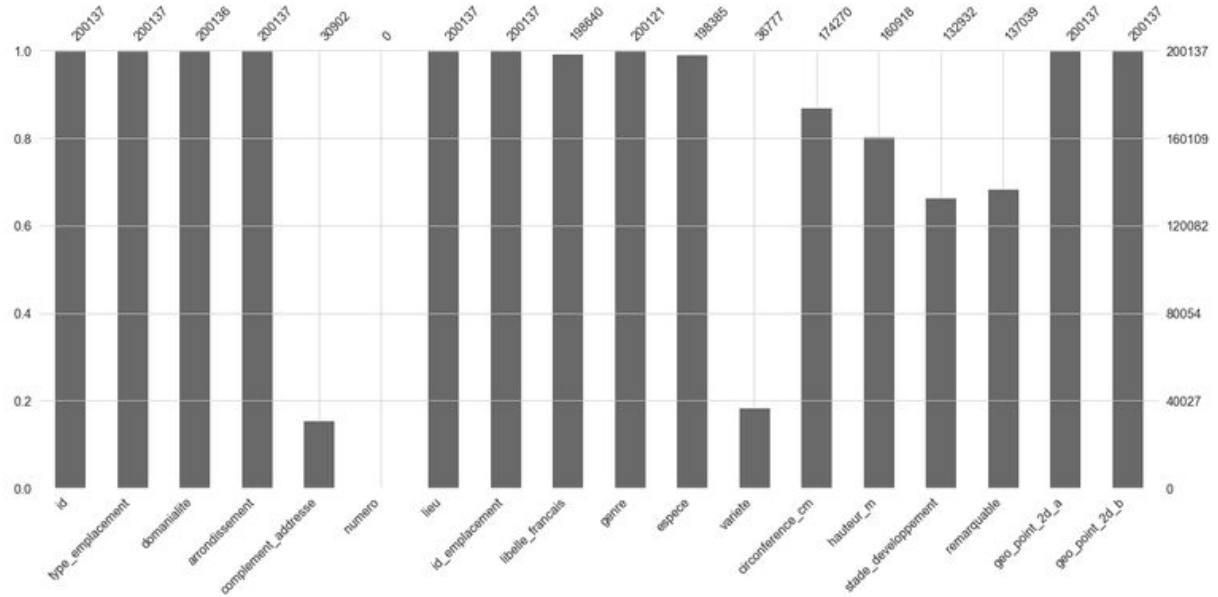
Valeurs manquantes



Valeurs manquantes

```
df.isna().sum()
```

```
id                          0
type_emplacement            0
domanialite                 1
arrondissement              0
complement_addresse    169235
numero                 200137
lieu                        0
id_emplacement              0
libelle_francais         1497
genre                      16
espece                   1752
variete                163360
circonference_cm        25867
hauteur_m               39219
stade_developpement     67205
remarquable             63098
geo_point_2d_a              0
geo_point_2d_b              0
dtype: int64
```



Nettoyage du jeu de données

- Elimination de la colonne “numero” car elle ne contient pas de valeurs
- Nettoyage de la variable “remarquable”
- Nettoyage de la variable “stade_developpement”

Doublons

- Il y a 22 arbres qui partagent des coordonnées géographiques avec un autre arbre.

- La plupart de ces arbres se trouvent dans le Bois de Vincennes, qui est un quartier très boisé.

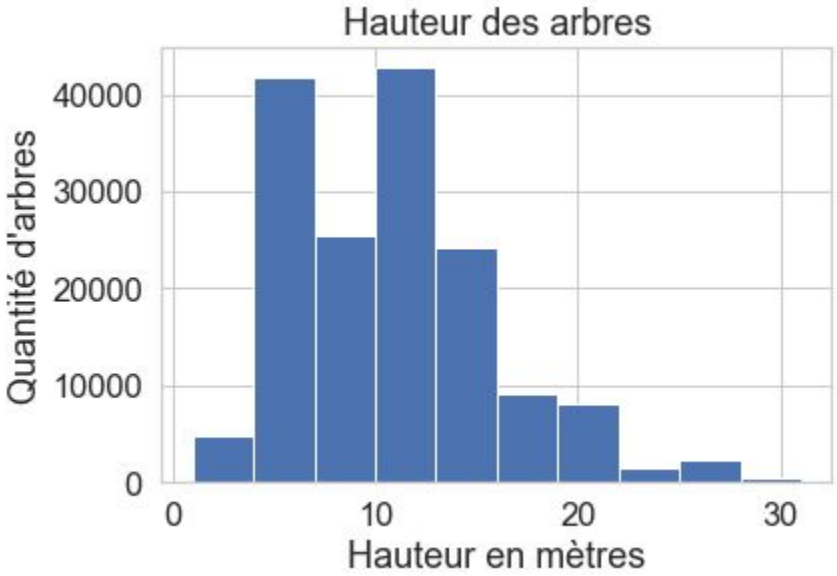
Variables quantitatives

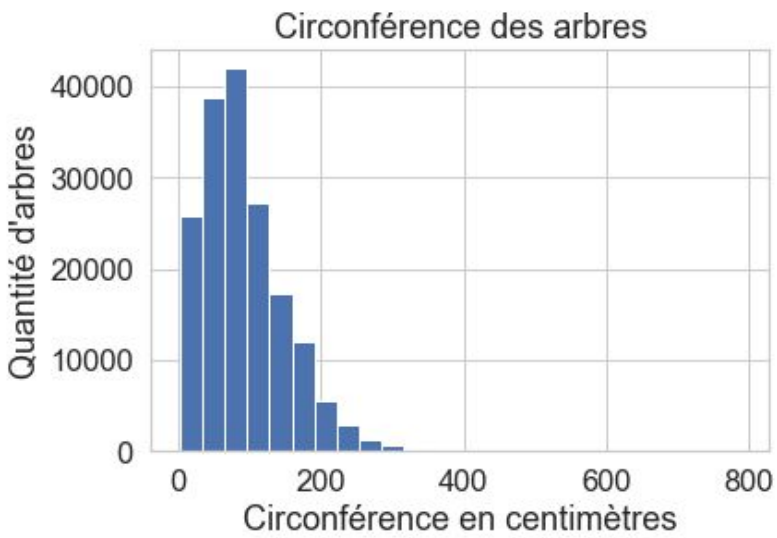
	hauteur_m	circonference_cm
count	160918.000000	174270.000000
mean	10.347761	91.470913
std	5.114521	58.956751
min	1.000000	1.000000
25%	6.000000	45.000000
50%	10.000000	80.000000
75%	14.000000	122.000000
max	31.000000	790.000000

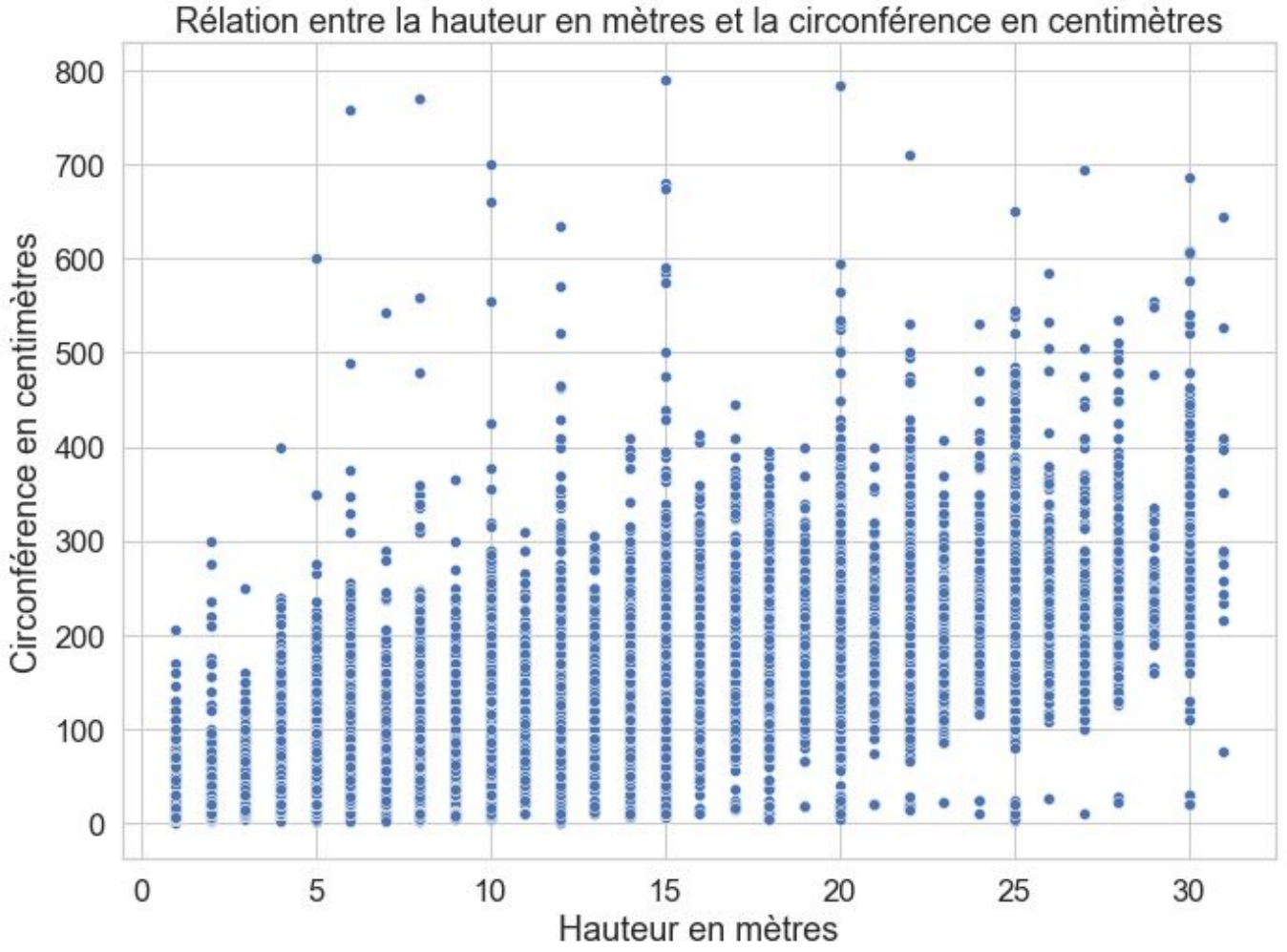
L’utilité des variables quantitatives

- Organiser les tournées d'entretien en fonction de la taille de l'arbre
- Les dimensions de l'arbre peuvent également être utiles en combinaison avec le stade de développement d'un arbre.
- Les variables quantitatives sont très utiles si on les analyse ensemble avec les variables qualitatives dans une analyse bivariée.
- Les coordonnées géographiques sont évidemment essentielles pour planifier les tournées dans la ville de Paris.

Variables quantitatives continues (hauteur, circonférence et coordonées géographiques)






Rélation entre la hauteur en mètres et la circonférence en centimètres
Circonférence en centimètres
Hauteur en mètres
0
100
200
300
400
500
600
700
800
0
5
10
15
20
25
30

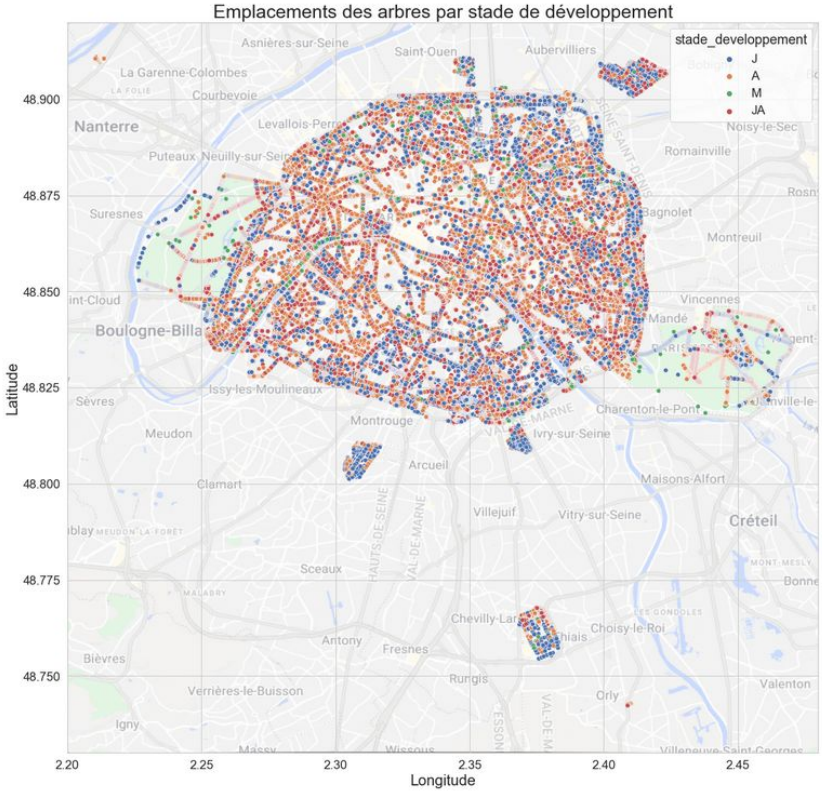
Variables qualitatives

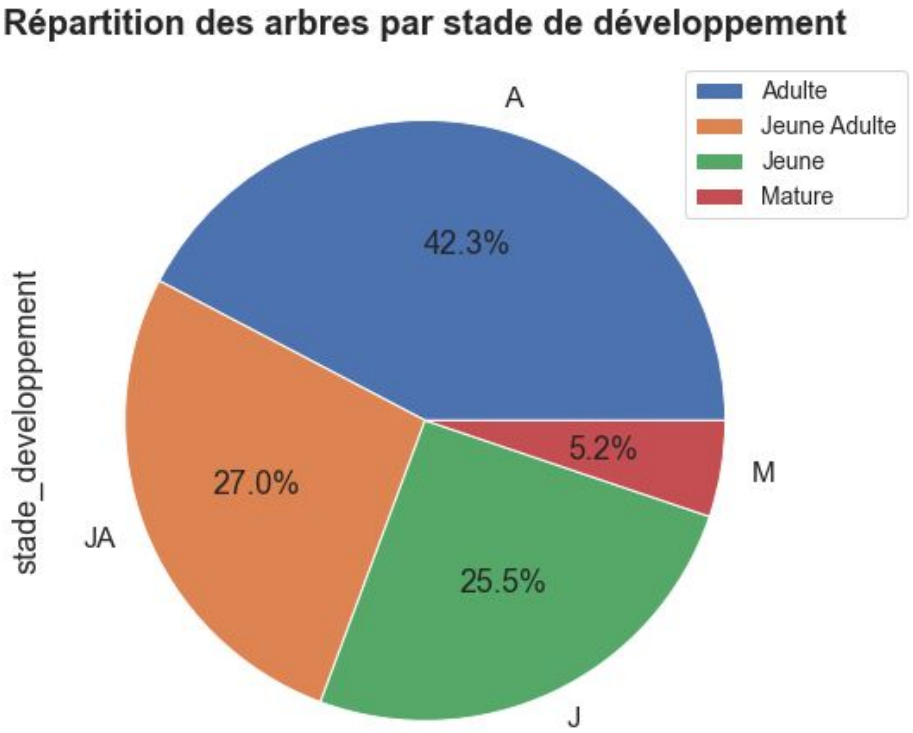
	domanialite	arrondissement	libelle_francais	stade_developpement	remarquable
count	200136	200137	198640	163185	200137
unique	9	25	192	4	2
top	Alignement	PARIS 15E ARRDT	Platane	A	False
freq	104949	17151	42508	69035	199953

L’utilité des variables qualitatives

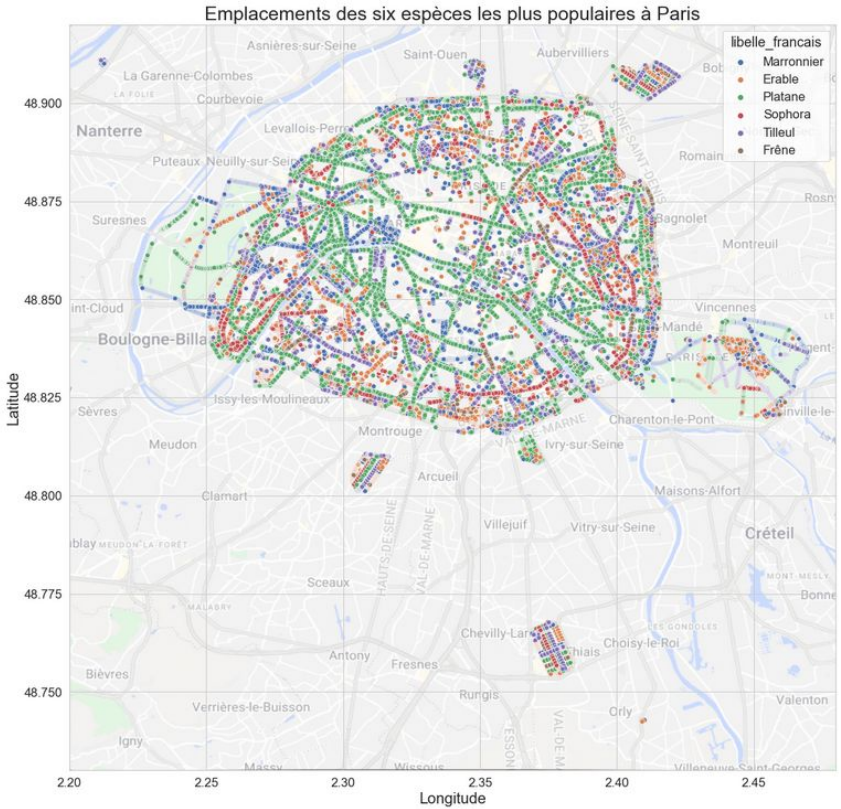
- L'espèce : un entretien spécifique ?
- Le stade de développement d'un arbre peut aider à planifier l'arrosage ou les soins particuliers.
- L'arrondissement nous aide à comprendre quels quartiers ont plus d'arbres, et quels quartiers peuvent peut-être bénéficier des prochaines plantations afin de les rendre plus verts.
- La domanialité (jardin, alignement, etc.) est également utile pour catégoriser les arbres selon les soins. Peut-être qu'un arbre de jardin a moins besoin d'élagage que les arbres d'alignement.

Variable qualitative ordinale : stade de développement

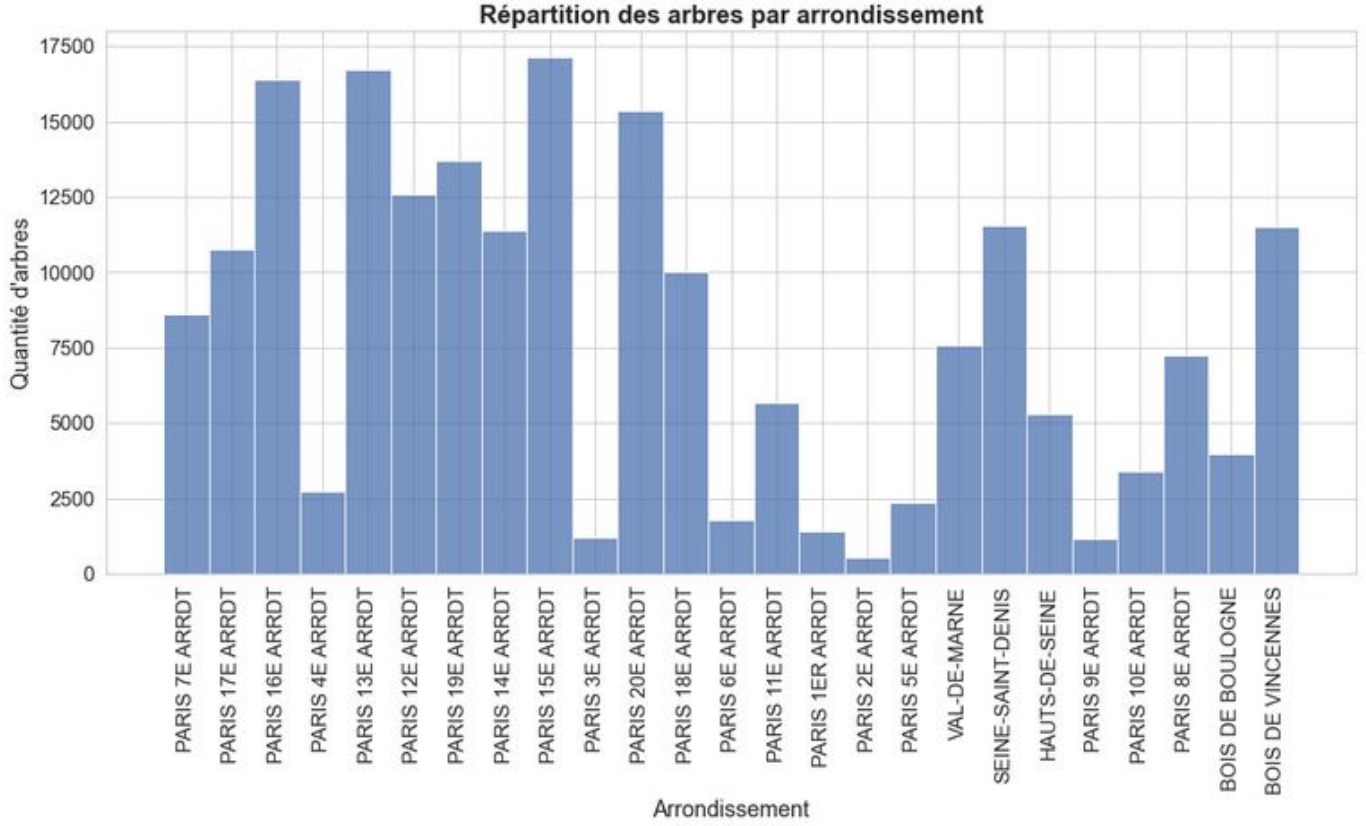




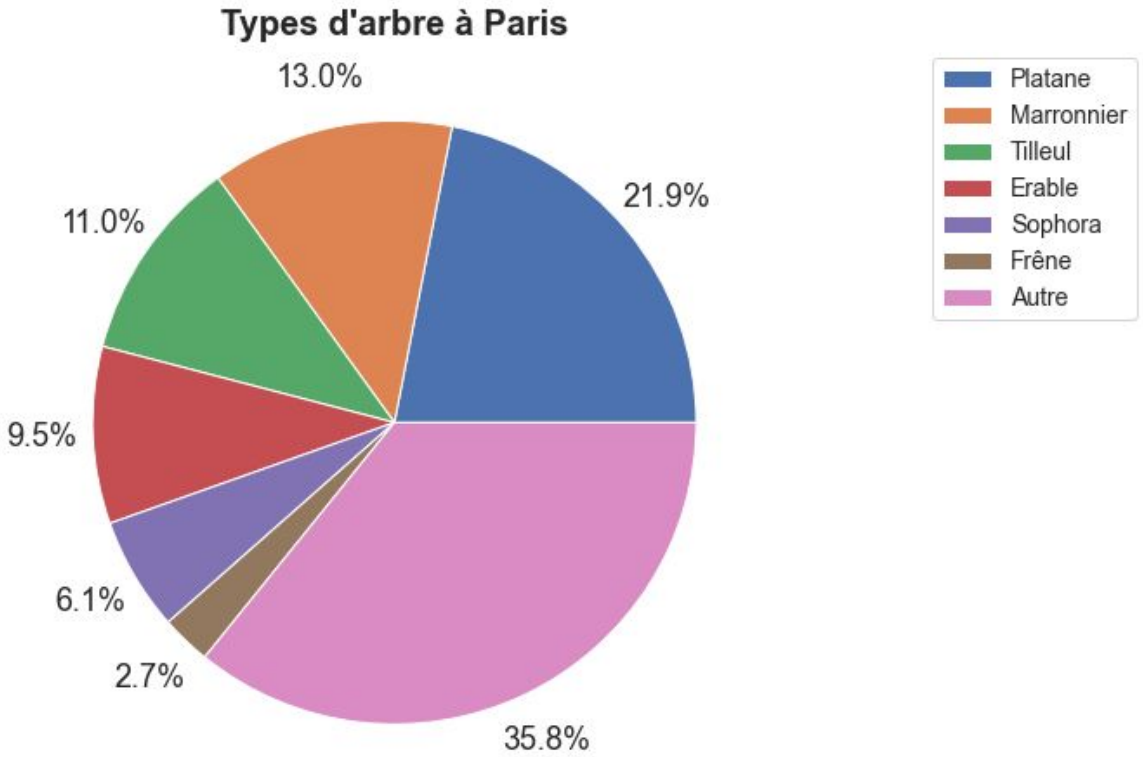
Variables qualitatives nominales : arrondissement



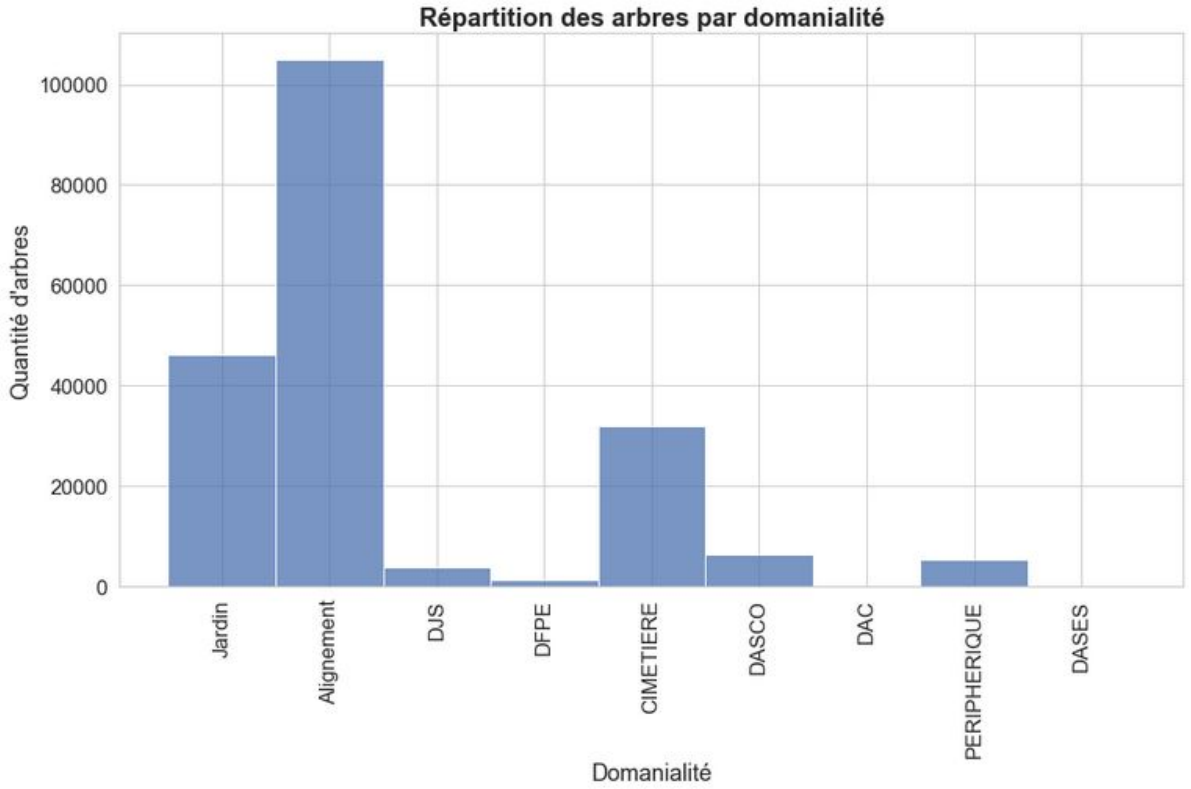
Variables qualitatives nominales : arrondissement



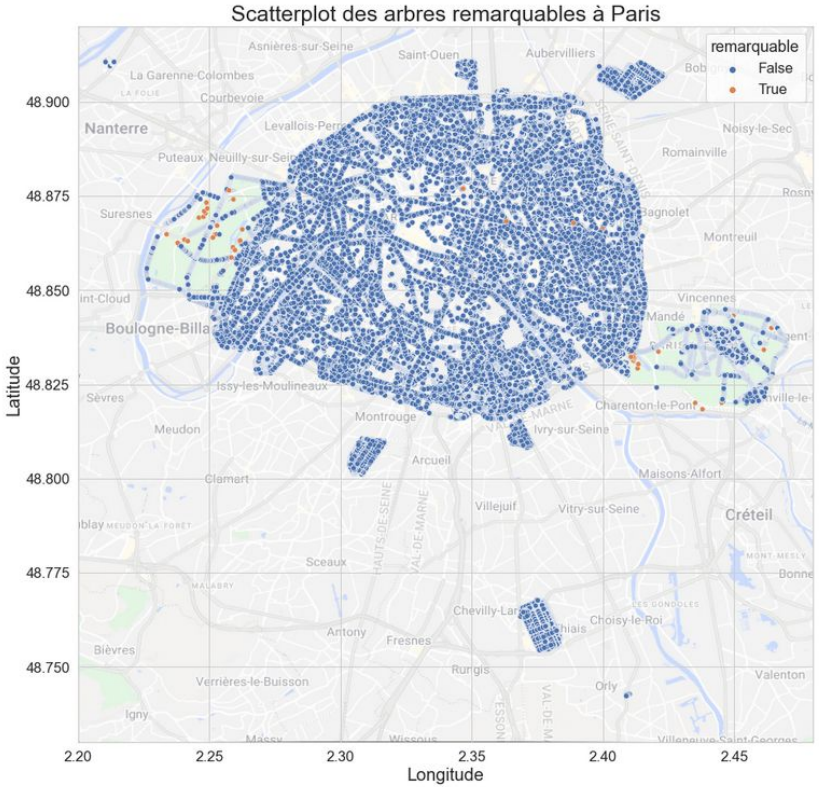
Variables qualitatives nominales : libellé français



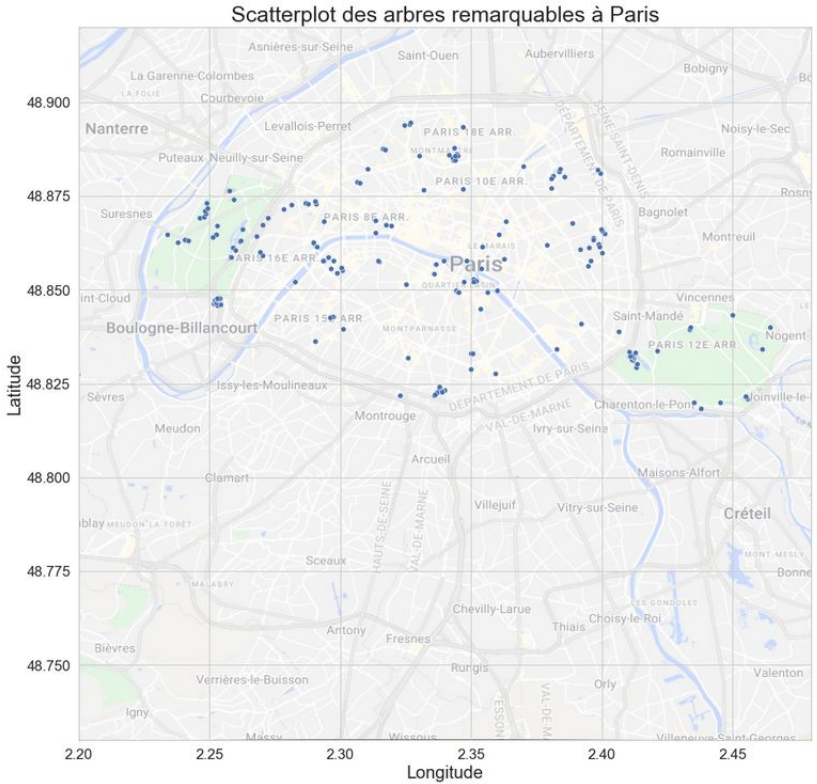
Variables qualitatives nominales : type de l’emplacement



Variable qualitatives nominales : arbres remarquables



Variable qualitatives nominales : arbres remarquables



Synthèse de l’analyse de données

La qualité du jeu de données

- Valeurs manquantes

- Valeurs atypiques non expliquées

Qualité et recommandations

- Rigueur dans la saisie des données
- Dès plantation, fournir des mesures
- Domanialités “DJS”, “DFPE”, “DASCO”, etc.