

Projet 02

Participez à un concours sur la Smart City

Analyse exploratoire de données



David Scanu | Septembre 2024



Parcours **AI Engineer**







Sommaire

1. Contexte
2. Démarche méthodologique d'analyse de données
3. Présentation générale du jeu de données
4. Analyse des données
5. Synthèse et recommandations

Contexte

- **Rôle**
 - **Expert indépendant** spécialisé en intelligence artificielle
- **Challenge de l'ONG Data is for Good**
 - **Analyse exploratoire** avec un jeu de données portant sur les arbres de la ville de Paris, dans le cadre du programme "Végétons la ville".
- **Objectif de l'analyse exploratoire**
 - **Optimiser les tournées d'entretien** des arbres de la ville de Paris.

Démarche méthodologique d'analyse de données

-  Installation environnement et bibliothèques
-  Importation du jeu de données
-  Visualisation générale du jeu de données
-  Nettoyages des données
-  Analyse univariée et bivariée
-  Synthèse et recommandations

Installation environnement et importation

- Activation d'un **environnement virtuel** dédié au projet.
- **Installation et importation des bibliothèques** Python
 - Pandas, Numpy, Matplotlib, Seaborn, Plotly
- **Importation du jeu de données**  **PARIS | DATA**
 - [Jeu de données "Arbres"](#) fourni par la ville de Paris
 - Géo-localise une partie des arbres relevant de la Ville de Paris
 - Mensurations, classification des arbres, arbres remarquables

```
david@David-PC:~$ python3 -m venv .venv
david@David-PC:~$ source .venv/bin/activate
david@David-PC:~$ pip install pandas numpy matplotlib
seaborn plotly==5.24.0 --no-cache-dir --quiet
```

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import plotly.express as px
```

```
data_filepath = "/data/p2-arbres-fr.csv"
data = pd.read_csv(data_url, delimiter=";")
```

Visualisation générale du jeu de données



```
data.head(10)
```

	id	type_emplacement	domanialite	arrondissement	complement_adresse	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 ARRD		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 ARRD		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 ARRD		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 ARRD		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 ARRD		NaN	PARC CLICHY-BATIGNOLLES-MARTIN LUTHER KING	000G0037	Arbre à miel	Tetradium	daniellii	NaN	38	0		NaN	48.890435	2.315289	
5	99879	Arbre	Jardin	PARIS 17E ARRD		NaN	PARC CLICHY-BATIGNOLLES-MARTIN LUTHER KING	000G0036	Arbre à miel	Tetradium	daniellii	NaN	38	0		NaN	48.890470	2.315228	
6	99880	Arbre	Jardin	PARIS 17E ARRD		NaN	PARC CLICHY-BATIGNOLLES-MARTIN LUTHER KING	000G0035	Arbre à miel	Tetradium	daniellii	NaN	37	0		NaN	48.890504	2.315168	
7	99881	Arbre	Jardin	PARIS 16E ARRD		NaN	SQUARE ALEXANDRE ET RENE PARODI / 1 PLACE DE L...	35	Platane	Platanus	x hispanica	NaN	260	17		NaN	48.876722	2.280222	
8	99882	Arbre	Jardin	PARIS 16E ARRD		NaN	JARDIN DE L'AVENUE FOCH / 10 AVENUE FOCH	802008	Sophora	Sophora	japonica	NaN	145	14	A	0.0	48.871990	2.275814	
9	99883	Arbre	Jardin	PARIS 16E ARRD		NaN	JARDIN DE L'AVENUE FOCH / 10 AVENUE FOCH	802009	Sophora	Sophora	japonica	NaN	135	10	A	0.0	48.872046	2.275752	



```
data.shape
```

Le DataFrame de notre jeu de données contient :

- 200137 lignes
- 18 colonnes.

Type de données



```
data.info()
```

• Variables qualitatives

- id (int)
- type_emplacement (str)
- domanialite (str)
- arrondissement (str)
- complement_adresse (str)
- numéro (float)
- lieu (str)
- id_emplacement (str)
- libelle_francais (str)
- genre (str)
- espece (str)
- variete (str)
- stade_developpement (str)
- remarquable (float → bool)

• Variables quantitatives

- circonference_cm (float)
- hauteur_m (int)
- geo_point_2d_a (float)
- geo_point_2d_b (float)

Bien qu'elles soient des valeurs numériques, les variables suivantes sont qualitatives :

- La variable **id** est une variable **qualitative nominale**.
- La variable **remarquable** est une variable **qualitative ordinale**, car elle ne contient que deux valeurs 0 et 1 (valeurs booléennes).

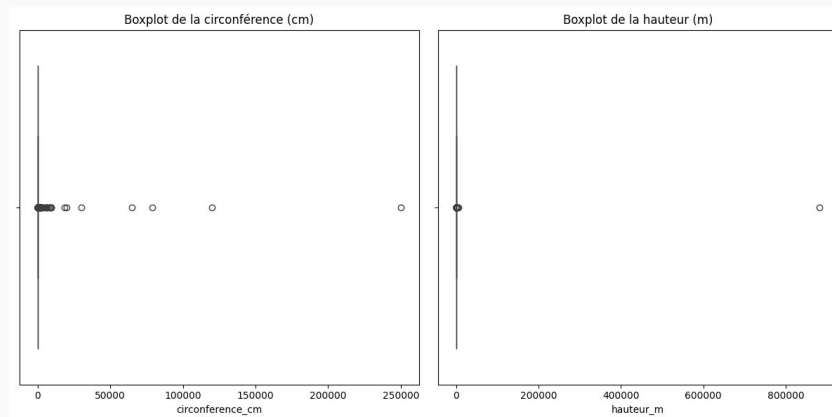
Analyse univariée préliminaire



```
data.describe()
```

	circonference_cm	hauteur_m	geo_point_2d_a	geo_point_2d_b
count	200137.000000	200137.000000	200137.000000	200137.000000
mean	83.380479	13.110509	48.854491	2.348208
std	673.190213	1971.217387	0.030234	0.051220
min	0.000000	0.000000	48.742290	2.210241
25%	30.000000	5.000000	48.835021	2.307530
50%	70.000000	8.000000	48.854162	2.351095
75%	115.000000	12.000000	48.876447	2.386838
max	250255.000000	881818.000000	48.911485	2.469759

Nous découvrons des **valeurs aberrantes** (minimum, maximum) qui peuvent fausser notre analyse univariée (moyenne). L'écart entre la médiane et la moyenne est aussi un indicateur de la présence de valeurs aberrantes ou atypiques.



Circonférence en cm

- Moyenne : 83.38
- Médiane : 70
- Min : 0
- Max : 250255

Hauteur en m

- Moyenne : 13.11
- Médiane : 8
- Min : 0
- Max : 881818



Nettoyages des données

Objectif : Améliorer la distribution des données

- Détection des doublons
- Suppression de variables inutiles
- Traitement des valeurs manquantes (NaNs)
- Traitement des valeurs aberrantes
- Traitement des valeurs atypiques

Doublons



```
# Vérification de doublons pour les coordonnées de localisation
data.loc[data[['geo_point_2d_a', 'geo_point_2d_b', 'id_emplacement']].duplicated(keep=False)]
```

	id	type_emplacement	domanialite	arrondissement	complement_adresse	numero	lieu	id_emplacement	libelle_francais	genre	espece	variete	circonference_cm	hauteur_m	stade_developpement	remarquable	geo_point_2d_a	geo_point_2d_b
189133	2011522	Arbre	Jardin	BOIS DE VINCENNES	NaN	NaN	PARC FLORAL DE PARIS / ROUTE DE LA PYRAMIDE	190042	Peuplier	Populus	nigra	NaN	0	0	NaN	0.0	48.836416	2.446277
189134	2011523	Arbre	Jardin	BOIS DE VINCENNES	NaN	NaN	PARC FLORAL DE PARIS / ROUTE DE LA PYRAMIDE	190042	Peuplier	Populus	nigra	Italica	100	22	A	0.0	48.836416	2.446277



```
# Suppression de la ligne index=189133
data.loc[data[['geo_point_2d_a', 'geo_point_2d_b', 'id_emplacement']].duplicated(keep=False)]
```

Suppression de variables inutiles

type_emplacement ne contient qu'une seule modalité **Arbre**.
Elle ne nous apporte aucune information et peut donc être supprimée.



```
data.drop('type_emplacement', axis=1, inplace=True)
```

Traitement des valeurs manquantes



```
# Afficher les valeurs manquantes pour chaque colonne  
print(data.isna().sum())
```

id	0
domanialite	1
arrondissement	0
complement_adresse	169234
numero	200136
lieu	0
id_emplacement	0
libelle_francais	1497
genre	16
espece	1752
variete	163359
circonference_cm	0
hauteur_m	0
stade_developpement	67204
remarquable	63098
geo_point_2d_a	0
geo_point_2d_b	0

Valeurs manquantes laissées

- Variable **complement_adresse** :
 - Optionnel
 - Sans impact sur notre analyse future
- Variables **domanialite, genre, espece, libelle_francais** :
 - NaNs peu nombreux
 - Sans impact sur notre analyse future
- Variable **variete** :
 - Caractères considérés comme mineurs.

Valeurs manquantes supprimées

- Variable **numero** :
 - toujours vide.
 - Nous **supprimons cette colonne**



```
data.drop(columns=['numero'], inplace=True)
```



Recommandation : Effectuer une tournée de relevé pour obtenir les informations et remplir ces champs.

Traitement des valeurs manquantes

Variable stade_developpement

- Nous décidons de **remplacer les NaNs par la modalité I** (pour inconnu).
- Puis de **remplacer les initiales par des noms**.

```
# Remplacer Nan par une nouvelle modalité : 'I'
data['stade_developpement'] = data['stade_developpement'].fillna(value='I')

# Remplacer les initiales par des noms
stade_developpement_modalites = {
    'J' : 'Jeune',
    'JA' : 'Jeune Adulte',
    'A' : 'Adulte',
    'M' : 'Mature',
    'I' : 'Inconnu'
}

data['stade_developpement'].replace(to_replace=stade_developpement_modalites, inplace=True)
```

```
data["stade_developpement"].value_counts()
```

• I = Inconnu	67204
• A = Adulte	64438
• JA = Jeune Adulte	35444
• J = Jeune	26937
• M = Mature	6113

Variable remarquable (Booléen)

- Les valeurs uniques sont 0 ou 1, il s'agit d'une **variable qualitative booléenne**.
- Nous considérons qu'une valeur vide (NaN) est 'non-remarquable'
- **Action :**
 - **Remplacement des Nans par 0** (soit False)
 - **Conversion de cette variable au bon 'dtype'**

```
# Remplacer les NaNs par 0
data['remarquable'] = data['remarquable'].fillna(value=0.)

# Conversion en booléen
data['remarquable'] = data['remarquable'].astype('bool')
```

Valeurs aberrantes



```
data[['hauteur_m', 'circonference_cm', 'geo_point_2d_a',  
'geo_point_2d_b']].describe()
```

	hauteur_m	circonference_cm	geo_point_2d_a	geo_point_2d_b
count	200136.000000	200136.000000	200136.000000	200136.000000
mean	13.110575	83.380896	48.854491	2.348207
std	1971.222311	673.191869	0.030234	0.051219
min	0.000000	0.000000	48.742290	2.210241
25%	5.000000	30.000000	48.835021	2.307530
50%	8.000000	70.000000	48.854163	2.351095
75%	12.000000	115.000000	48.876447	2.386834
max	881818.000000	250255.000000	48.911485	2.469759

Taille de l'arbre (hauteur_m, circonference_cm)



Informations sur les plus grands arbres de Paris

Le plus grand : le Séquoia des Buttes-Chaumont : D'une circonférence de 4,70 m et d'une **hauteur de plus de 35 mètres**

Le plus gros : platane d'Orient du Parc Monceau : son tronc mesure en effet **7 mètres de circonférence** pour une hauteur de 31 mètres environ !

Source :

<https://www.unjourdeplusaparis.com/paris-vert/arbres-remarquables-paris>

Valeurs qui dépassent les valeurs limites décrites ci-contre :

- Arbres de hauteur de 0 m : 39218
- Arbres de hauteur supérieure à 35 m : 509
- Arbres de circonférence 0 cm : 25866
- Arbres de circonférence supérieure à 700 cm : 82

Action : **Suppression** des individus présentant ces valeurs pour effectuer par la suite notre analyse univariée/bivariée.

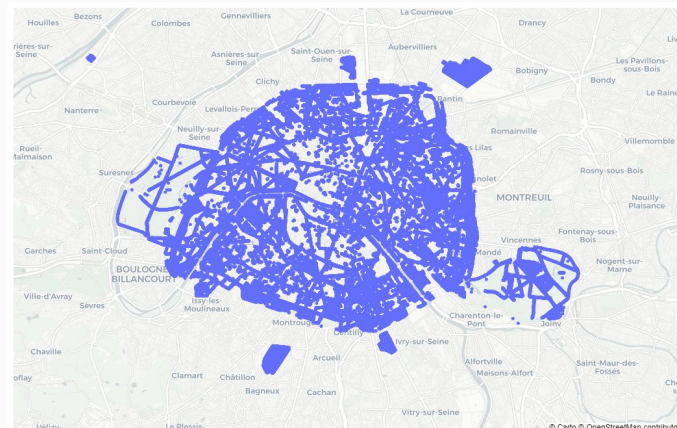
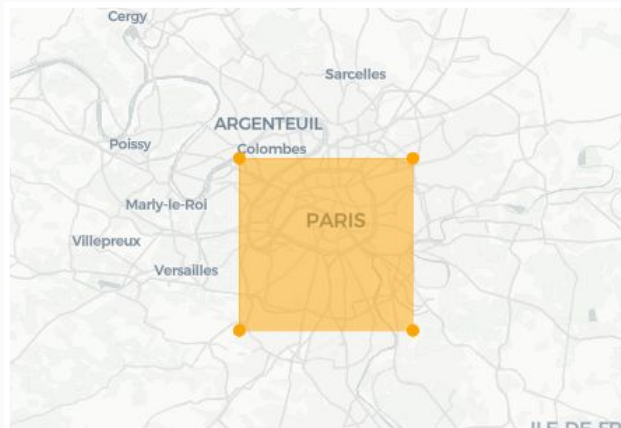
Géolocalisation

Latitude (geo_point_2d_a)

- min : 48.742290
- max : 48.911485

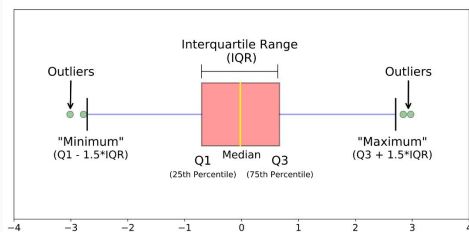
Longitude (geo_point_2d_b)

- min : 2.210241
- max : 2.469759



Les individus sont bien compris dans la zone de Paris et la région parisienne.

Valeurs atypiques



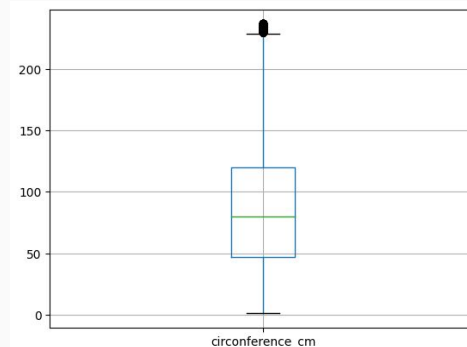
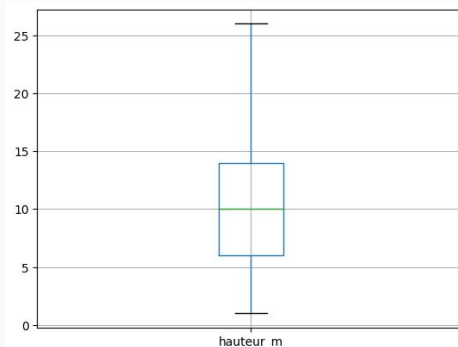
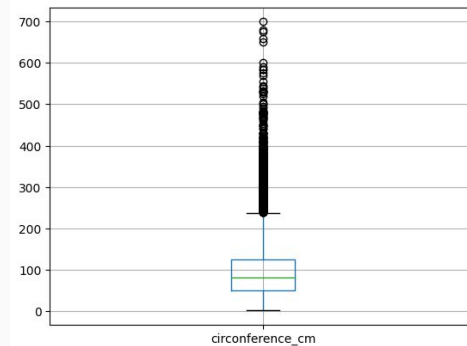
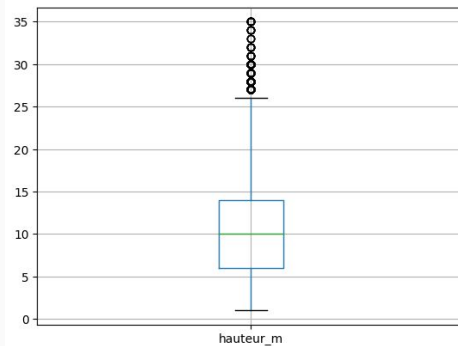
Détection des valeurs atypiques en utilisant la méthode interquartile (IQR)

Hauteur

- Plage interquartile (IQR) : 8.0
- Limite basse : -6.0
- Nombre de lignes : 0
- Limite haute : 26.0
- Nombre de lignes : 731

Circonférence

- Plage interquartile (IQR) : 75.0
- Limite basse : -62.5
- Nombre de lignes : 0
- Limite haute : 237.5
- Nombre de lignes : 3227



Action : Remplacement des lignes qui dépassent la limite haute, par la valeur de la limite haute. (Nous pourrions également les supprimer)

Distribution empirique et analyse univariée

D'après des **professionnels des espaces verts**, les variables suivantes sont les plus importantes à prendre en considération :

- **Espèce** : entretien spécifique (taille, fertilisation, etc.)
- **Stade de développement** : besoins en eau et en soins
- **Hauteur et circonférence** : complexité des travaux d'élagage et sécurité des opérations
- **Domanialité** : emplacement de l'arbre dans l'espace public
- **Arrondissement** : répartition des budgets (mairies d'arrondissement)
- **Géolocalisation** : itinéraire optimale pour la tournée d'entretien

Distribution par arrondissements

Objectif métier : La distribution par arrondissement peut être utile pour la **répartition budgétaire** des mairies d'arrondissement.

Interprétation

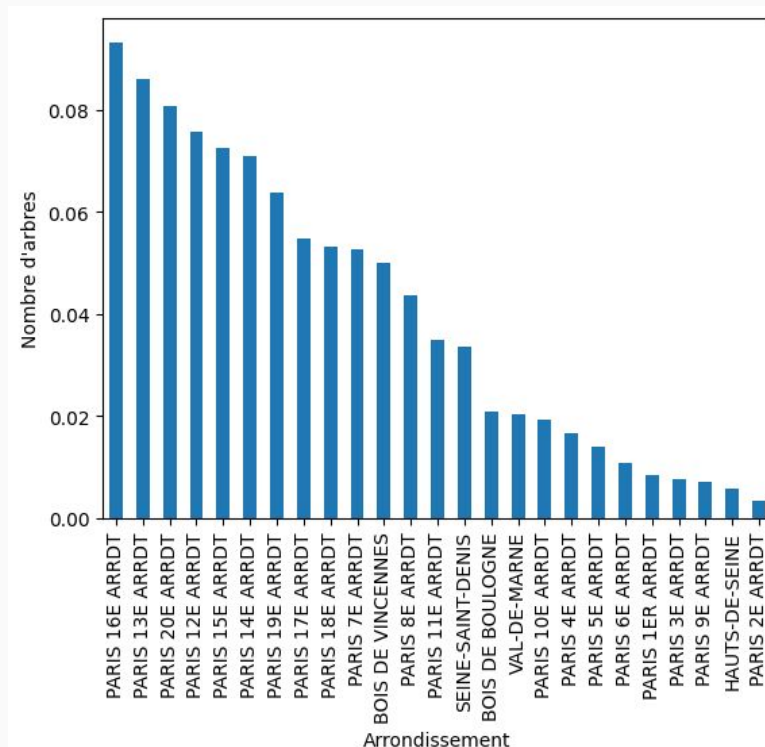
Les arrondissements comprenant le plus grand nombre d'arbres sont :

- Paris 16ème
- Paris 13ème
- Paris 20ème

Les arrondissements comprenant le moins d'arbres sont:

- Paris 2ème
- Les Hauts-de-Seine
- Paris 9ème

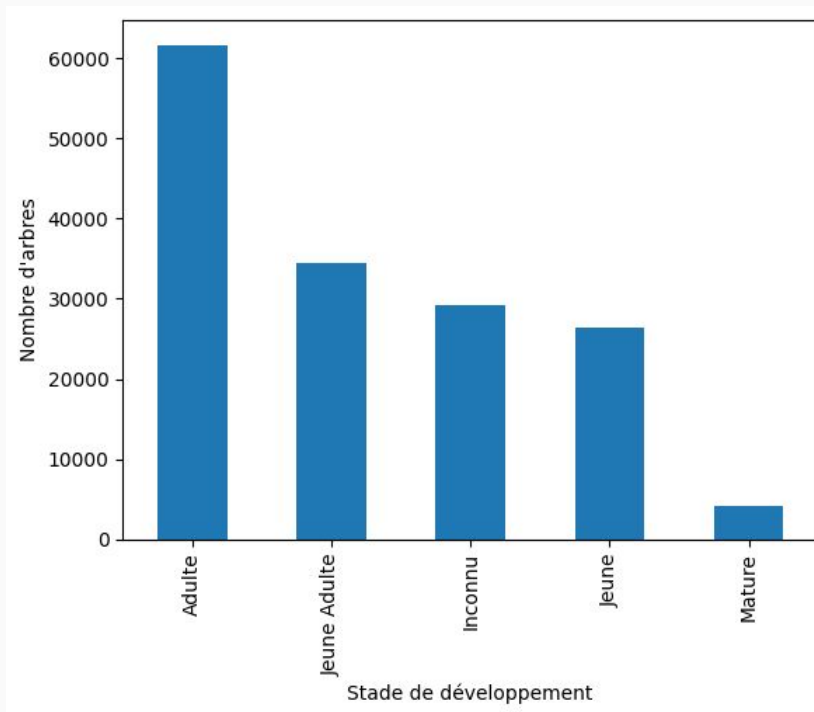
Les ressources budgétaires et matérielles devront être réparties en prenant en considération le nombre d'arbres par arrondissement.



Distribution par stade de développement

Objectif métier : Le stade de développement peut avoir une incidence sur la quantité d'eau à apporter.

Interprétation : Les arbres adultes sont les plus nombreux, suivis des jeunes adultes.

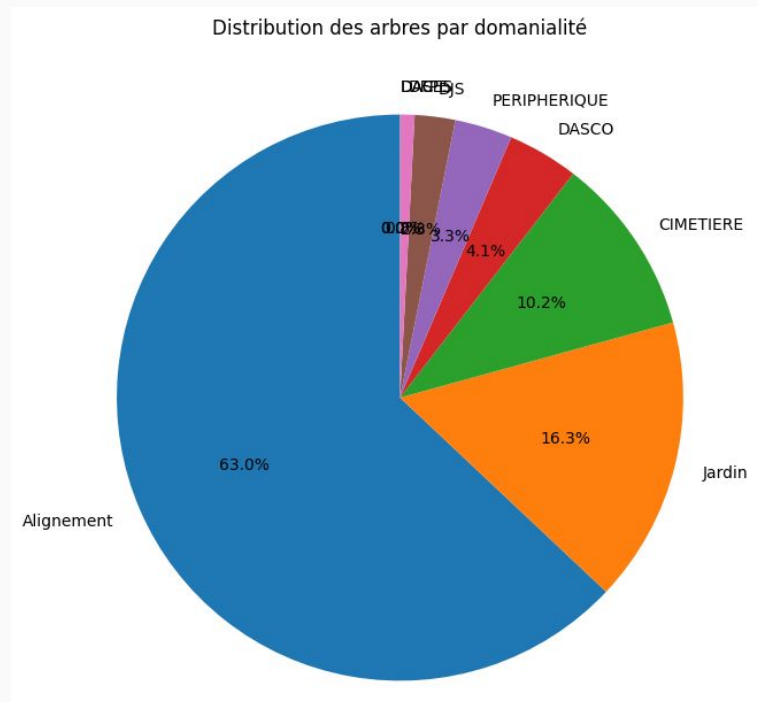


Distribution par domanialité

Objectif métier : La domanialité, cad l'emplacement des arbres dans l'espace public, a un impact sur le type de matériel nécessaire à l'entretien des arbres.

Interprétation : La domanialité "Alignement" est largement majoritaire avec (63%), suivi de "Jardin" (16,3%) et "Cimetière" (10,2%).

Action : L'approvisionnement en matériel et outils devra refléter cette distribution.



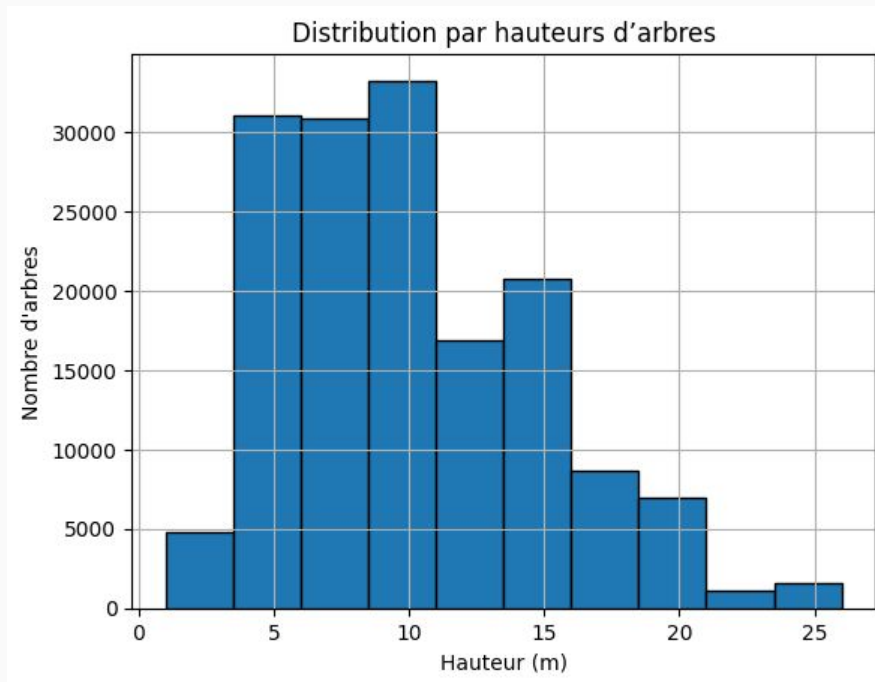
Distribution par hauteurs d'arbres



```
data_clean['hauteur_m'].describe()
```

Indicateurs statistiques

- **Total** : 156 013
- **Moyenne** : 10,08
- **Ecart-type** : 4,8
- **Minimum** : 1
- **Premier quartile (25%)** : 6
- **Médiane (50%)** : 10
- **Troisième quartile (75%)** : 14
- **Maximum** : 26



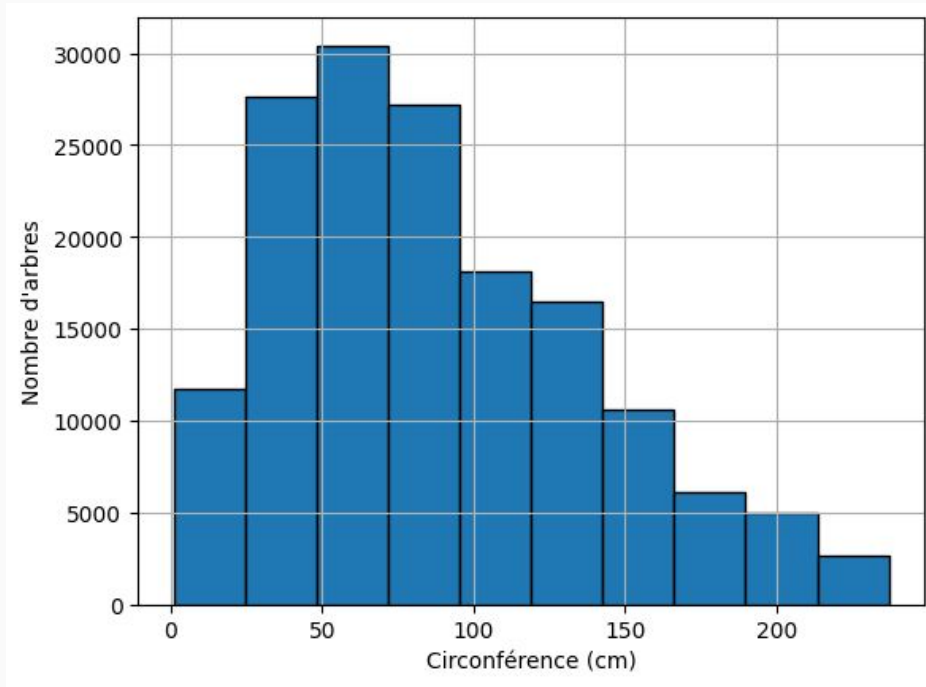
Distribution par circonférence des arbres



```
data_clean['circonference_cm'].describe()
```

Indicateurs statistiques

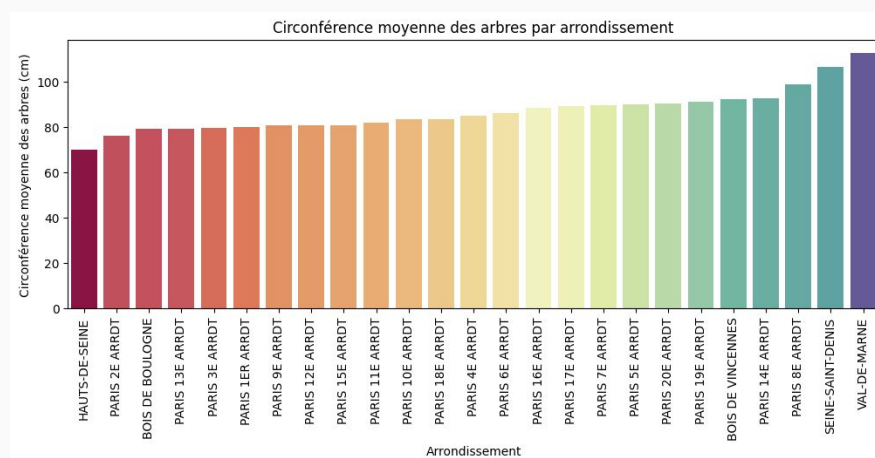
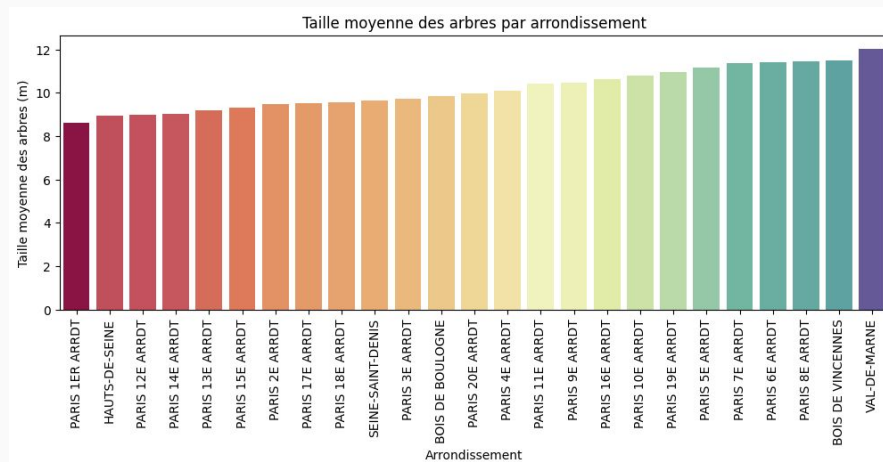
- **Total** : 156 013
- **Moyenne** : 87,81
- **Écart-type** : 50,86
- **Minimum** : 1
- **Premier quartile (25%)** : 47
- **Médiane (50%)** : 80
- **Troisième quartile (75%)** : 120
- **Maximum** : 237



Analyse bivariée

Etude de la relation entre deux variables

Hauteur et circonférence moyenne des arbres par arrondissement

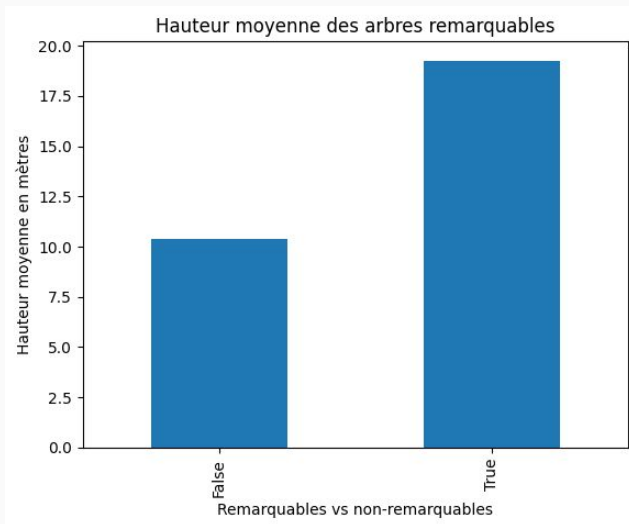


Hauteur et circonférence moyenne des arbres remarquables

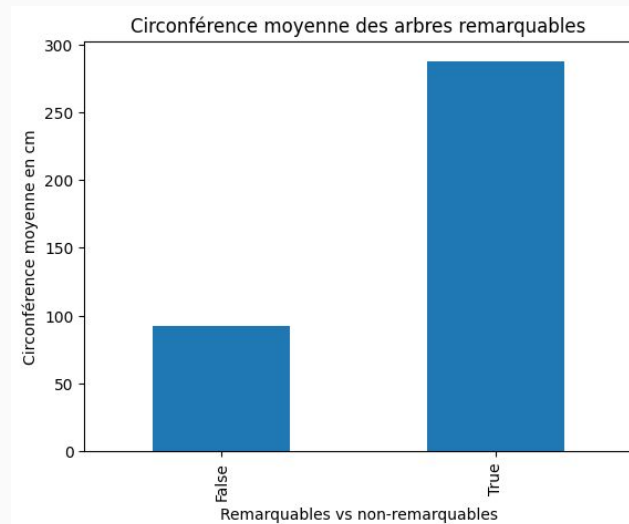
Interprétation

Les arbres remarquables sont **plus haut et gros**.

Il est possible qu'ils demande un entretien particulier qui peut faire l'objet d'une **tournée d'entretien spéciale**.



	count	mean	std	min	25%	50%	75%	max
remarquable								
False	159792.0	10.353278	5.122160	1.0	6.0	10.0	14.0	35.0
True	179.0	19.245810	7.680847	3.0	13.5	20.0	26.0	35.0



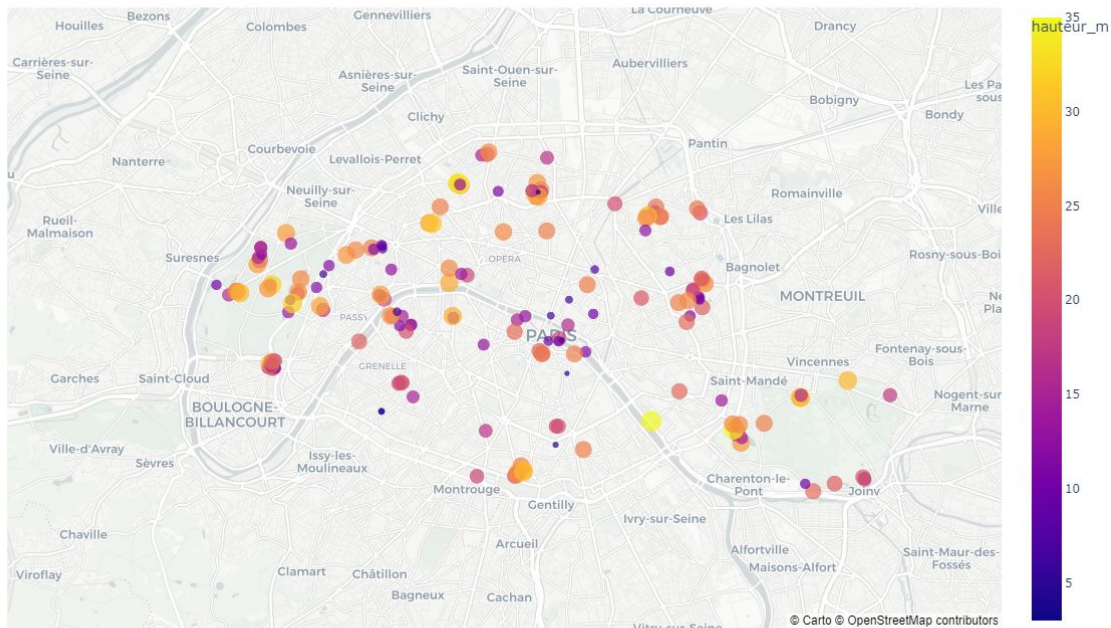
	count	mean	std	min	25%	50%	75%	max
remarquable								
False	159792.0	92.225343	58.312070	1.0	50.0	80.0	125.0	700.0
True	179.0	287.837989	142.117104	30.0	174.0	250.0	384.0	695.0

Emplacement des arbres remarquables

Interprétation

Les arbres remarquables sont **plus haut et gros**.

Il est possible qu'ils demande un entretien particulier qui peut faire l'objet d'une **tournée d'entretien spéciale**.



Corrélation entre circonférence et hauteur

Deux variables quantitatives.

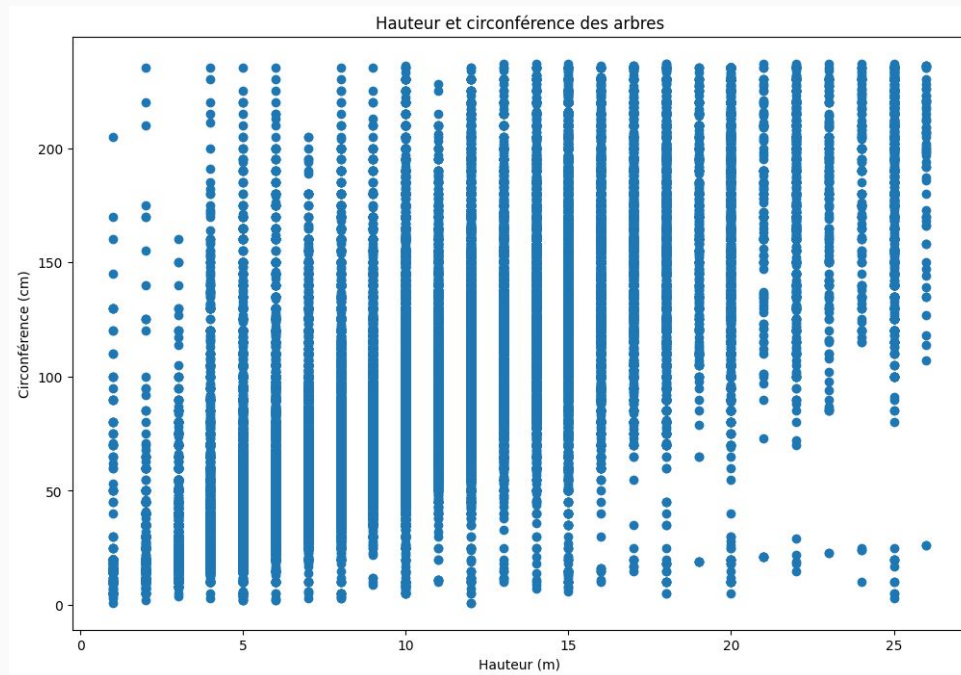
Méthode utilisées

- Diagramme de dispersion
- Coefficient de corrélation de Pearson : **0,80**



```
# Coefficient de corrélation de Pearson  
data_clean[['circonference_cm', 'hauteur_m']].corr()
```

Interprétation : Nous observons une forte corrélation entre la hauteur et la circonférence.



Hauteur moyenne par libellé français (les 20 plus fréquents)

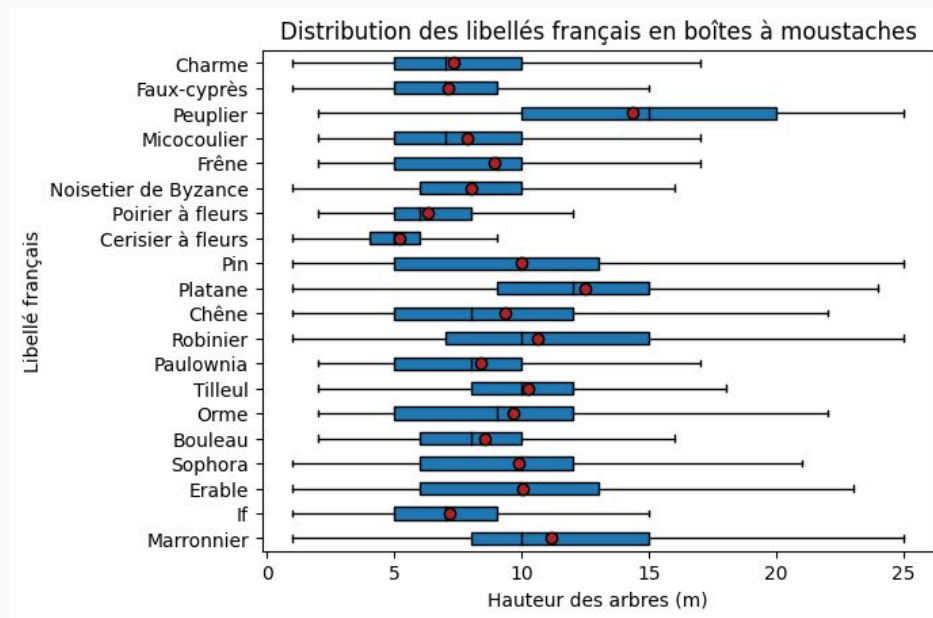
Une variable qualitative et une variable quantitative.

Méthode utilisées

- Analyse de la variance (ANOVA)
- Représentation en **boîtes à moustache**
- Calcul du **rapport de corrélation** (eta carré) : **0,22**

Utilité éventuelle : Étudier la corrélation entre le libellé français et la hauteur de l'arbre pourrait nous permettre d'**extrapoler une hauteur pour les arbres dont les hauteurs sont aberrantes/erronées**.

Interprétation : Ce résultat montre une légère association entre le libellé et la hauteur des arbres.



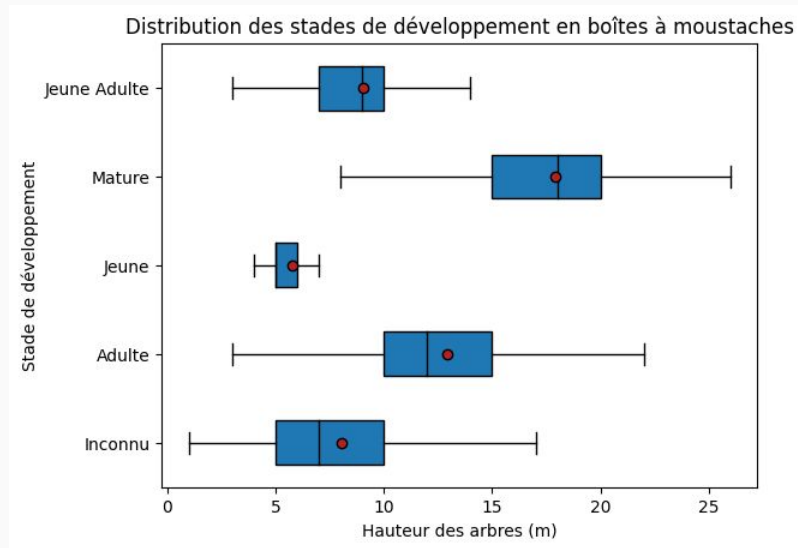
Corrélation entre stade de développement et hauteur

Une variable qualitative et une variable quantitative.

Rapport de corrélation entre le stade de développement et la hauteur (eta carré) : **0.39**

Interprétation : La hauteur et le stade du développement sont corrélés (moyennement).

Attention : la catégorie **inconnu** brouille la qualité de nos résultats d'analyse. La corrélation est **probablement plus grande**.



Synthèse et recommandations

Enjeux liés à la qualité des données

- **Qualité des données** : Eviter les données manquantes ou erronées.
- **Équilibre des données** : Une répartition inégale des données pour ces différentes caractéristiques peut biaiser les résultats de l'optimisation des tournées.

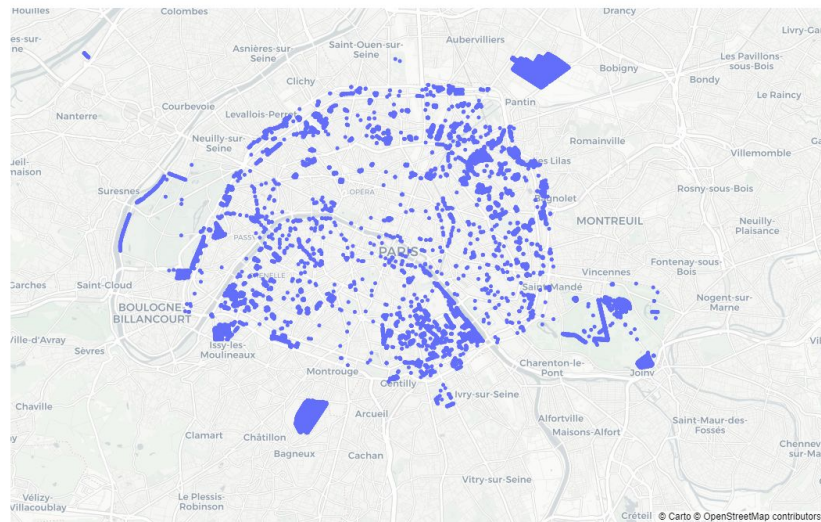
Recommandations

1. Réaliser une tournée de mesures

- **genre** (valeurs manquantes)
- **espece** (valeurs manquantes)
- **libelle_francais** (valeurs manquantes)
- **domanialité** (valeurs manquantes)
- **hauteur_m** (valeurs aberrantes: 0 et > 35 m)
- **circonference_cm** (valeurs aberrantes: 0 et > 700 cm)

2. Ajouter la variable "remarquable" aux variables importantes

Les arbres remarquables étant en moyenne plus hauts et gros, il est possible qu'ils demandent un **entretien particulier** qui peut faire l'objet d'une tournée d'entretien spéciale, qui mobilise une **équipe dédiée**.



Développement de modèles d'optimisation de tournée d'entretien

Ajustement des modèles d'optimisation : Les modèles d'optimisation des tournées devront être adaptés pour prendre en compte les spécificités de chaque espèce d'arbre et les contraintes liées à la hauteur, au stade de développement, à la circonférence et à la domanialité, ainsi que l'arrondissement pour la composition des équipes et la répartition des budgets.

Intégration de la dimension géographique : La géolocalisation doit être intégré dans l'algorithme d'optimisation pour minimiser les distances parcourues.

Problème du voyageur de commerce (TSP) : Notre problème se rapproche du problème algorithmique qui consiste à trouver le chemin le plus court entre un ensemble de points et de lieux à visiter.


Choix de l'algorithme optimal :

- Algorithmes
- Machine learning
- Deep learning



Merci

Ressources et documents

- 
- [Présentation](#)
 - [Notebook Colab](#)
 - [Dépôt GitHub](#)