Végétalisons la ville

CHALLENGE DATA - ENTRETIEN DES ARBRES DE LA VILLE DE PARIS

Sommaire

1) Contexte, objectifs, et contraintes du challenge

Participation, en tant que Data Scientist indépendant, au challenge Smart City proposé par la ville de Paris organisé par l'ONG Data is for Good dans le cadre du programme "Végétalisons la ville".

Objectifs

Réalisation d'une analyse exploratoire sur un jeu de données portant sur les arbres de la ville de Paris, pour améliorer les tournées d'entretien de ces arbres.

Livrables

- Un support de présentation en 3 parties :
  - Présentation générale du jeu de données
  - Démarche méthodologique d'analyse et de nettoyage des données
  - Synthèse de l'analyse de données et visualisation des données
- Un Notebook Jupyter, à destination d'un public non technique, documenté pour expliciter les différents traitements, calculs ou graphiques effectués en utilisant les fonctionnalités d'édition de texte de Jupyter.

2) Mise en place de l'environnement de travail

- Utilisation des librairies :
  - Pandas et Numpy pour les calculs et le traitement des données ;
  - Matplotlib et Seaborn pour la visualisation des données.

- Installation de Folium pour visualiser les données géospatiales sous forme de cartes ;
- Utilisation d'un Notebook Jupyter pour écrire le code Python
- Mise en place d'un environnement virtuel dédié au projet

3) Analyse des données

A) Présentation générale
B) Analyses Univariées
C) Nettoyage des données
D) Visualisation des données
E) Analyses et propositions

3) Analyse des données

A) Présentation générale des données

Ce jeu de données contient :

200 137 lignes ;

18 colonnes dont 7 variables quantitatives et 11 variables qualitatives.

```
<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
```

3) Analyse des données

A) Présentation générale des données

Caractéristiques des arbres :

emplacement, identification (espèce, variété, ...), et état (circonférence, hauteur, stade de développement, ...).

Plus de détails dans le Notebook Jupyter.

```
# Afficher les premières lignes du jeu de données pour vérifier le chargement
data.head()
```

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

3) Analyse des données

B) Analyses Univariées des variables

But : comprendre chaque variable et son contexte afin de pouvoir interpréter correctement les résultats de l'analyse.

- méthodes statistiques pour calculer les caractéristiques principales de chaque variable (médiane, moyenne, écart-type,...)
- graphiques appropriés pour chaque type de variables, notamment : diagramme en bar ou pie chart pour les variables qualitatives, courbes de densité ou boîtes à moustache pour les variables quantitatives

3) Analyse des données

B) Analyses Univariées des variables

id : identifiant, valeurs uniques
numéro : toutes les valeurs sont manquantes
circonférence/hauteur : écart-type élevé, présence de valeurs aberrantes / variables intéressantes pour l'analyse
remarquable : 2 valeurs possibles (0 ou 1) -> variable catégorielle, identifie les arbres avec des particularités
geo point : écart-type très faible, absence de valeurs aberrantes / donne la localisation exacte des arbres

Indicateurs statistiques pour les 7 variables quantitatives:

	id	numero	circonference_cm	hauteur_m	remarquable	geo_point_2d_a	geo_point_2d_
count	2.001370e+05	0.0	200137.000000	200137.000000	137039.000000	200137.000000	200137.00000
mean	3.872027e+05	NaN	83.380479	13.110509	0.001343	48.854491	2.34820
std	5.456032e+05	NaN	673.190213	1971.217387	0.036618	0.030234	0.05122
min	9.987400e+04	NaN	0.000000	0.000000	0.000000	48.742290	2.21024
25%	1.559270e+05	NaN	30.000000	5.000000	0.000000	48.835021	2.30753
50%	2.210780e+05	NaN	70.000000	8.000000	0.000000	48.854162	2.35109
75%	2.741020e+05	NaN	115.000000	12.000000	0.000000	48.876447	2.38683
max	2.024745e+06	NaN	250255.000000	881818.000000	1.000000	48.911485	2.46975

3) Analyse des données

B) Analyses Univariées des variables

11 variables qualitatives : 6 variables liées à l'emplacement et 5 variables liées à l'espèce

- type_emplacement : valeur constante (arbre) / inutile dans cette analyse ;
- domanialité (type de lieu où se trouve l'arbre), arrondissement, complement_adresse, lieu, id_emplacement (cardinalité très élevée) décrivent les lieux de localisation des arbres ;
- libelle_francais : nom français du type d'arbre ;
- genre, espece, variété : différentes nominations supplémentaires pour les arbres / a priori ils font doublons avec libelle_francais
- stade_developpement : jeune / Jeune Adulte / Adulte / Mature -> décrit âge de l'arbre, pourrait influencer le type de traitement utilisé

[5]:

ıplacement	domanialite	arrondissement	complement_addresse	lieu	id_emplacement	libelle_francais	genre	espece	variete	stade_developpement
Arbre	Jardin	PARIS 7E ARRDT	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	19	Marronnier	Aesculus	hippocastanum	NaN	NaN
Arbre	Jardin	PARIS 7E ARRDT	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	20	If	Taxus	baccata	NaN	A
Arbre	Jardin	PARIS 7E ARRDT	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	21	If	Taxus	baccata	NaN	A
Arbre	Jardin	PARIS 7E ARRDT	NaN	MAIRIE DU 7E 116 RUE DE GRENELLE PARIS 7E	22	Erable	Acer	negundo	NaN	A
Arbre	Jardin	PARIS 17E ARRDT	NaN	PARC CLICHY-BATIGNOLLES-MARTIN LUTHER KING	000G0037	Arbre à miel	Tetradium	daniellii	NaN	NaN

3) Analyse des données

C) Nettoyage des données

Objectif : jeu de données nettoyé, sans les valeurs aberrantes identifiées et amélioration de la distribution des données après traitement.

2 étapes communes : la détection et le traitement des valeurs aberrantes rencontrées.

A) variables inutiles

Logique métier -> **détection** rapide des variables qui n'apporteront aucune information significative à l'analyse : id (identifiant), numéro (valeurs manquantes), type_emplacement (valeur constante : arbre), complément_adresse (cardinalité très élevée et pourcentage valeurs manquantes élevé), id_emplacement (inutile). **Traitement** -> suppression

3) Analyse des données

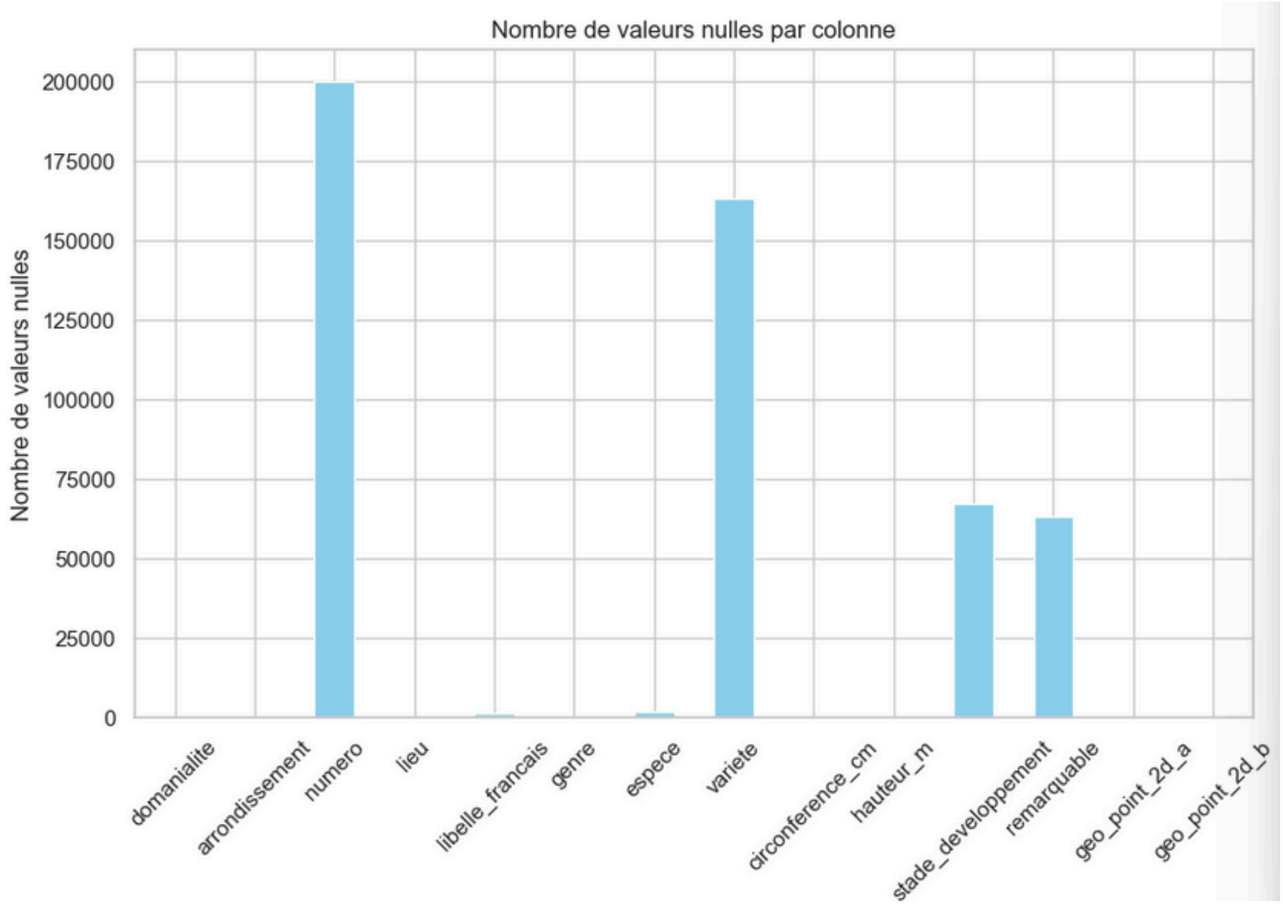
C) Nettoyage des données

B) Variables nulles

Les valeurs nulles peuvent affecter la qualité des résultats.

Calcul du nombre et du taux de valeurs nulles par colonnes.
Numéro et variété : taux 0 trop important –> suppression

Remarquable : informations importantes mais taux 0 important –> Suppression inadaptée donc remplacement par "Non".



3) Analyse des données

C) Nettoyage des données

D) valeurs atypiques (outliers)

Détection par une logique métier :

- La circonférence la plus élevée pour un arbre en France = 13 m ;
- L'arbre le plus haut de France a une hauteur de 66 m.

Traitement : remplacement des variables (circonférence supérieure à 14 m et hauteur supérieure à 67 m) par les médianes de groupe

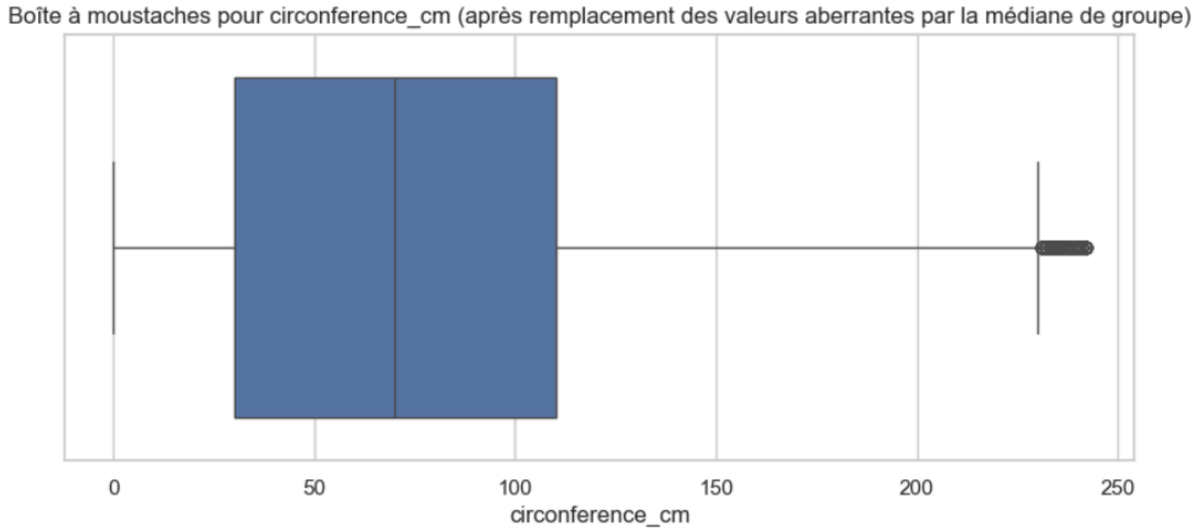
	circonference_cm	hauteur_m
count	200136.000000	200137.000000
mean	79.863028	8.357772
std	64.863088	6.301715
min	0.000000	0.000000
25%	30.000000	5.000000
50%	70.000000	8.000000
75%	115.000000	12.000000
max	1360.000000	67.000000

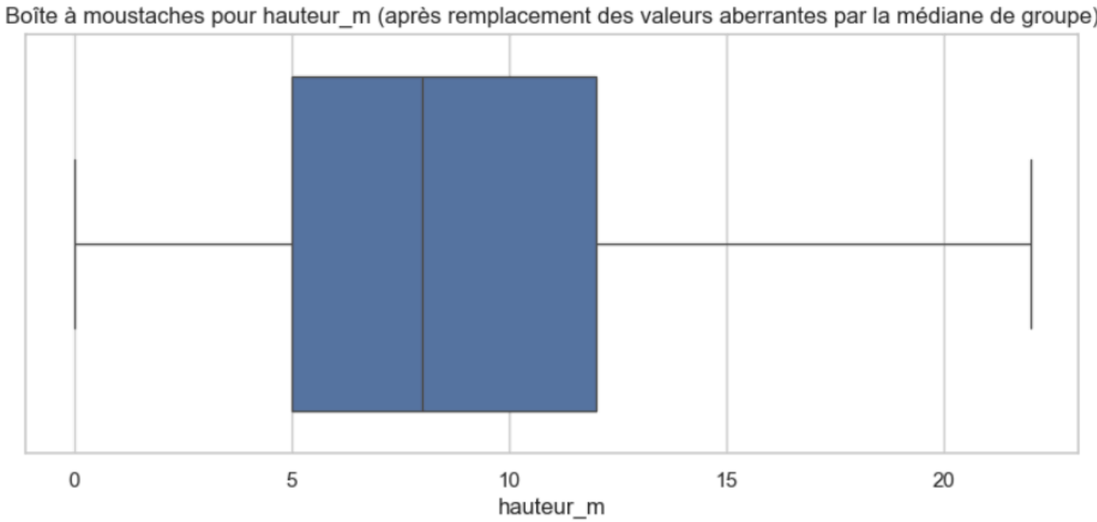
3) Analyse des données

C) Nettoyage des données

Détection par une approche technique :

détection –> Utilisation de méthodes statistiques : calcul des bornes inférieure et supérieure en utilisant les quantiles (Q1 et Q3) et l'écart interquartile (IQR). Les valeurs aberrantes = valeurs inférieures à la borne inférieure (Q1 – 1.5 * IQR) ou supérieures à la borne supérieure (Q3 + 1.5 * IQR). –> Traitement : Remplacement par des médianes de groupe



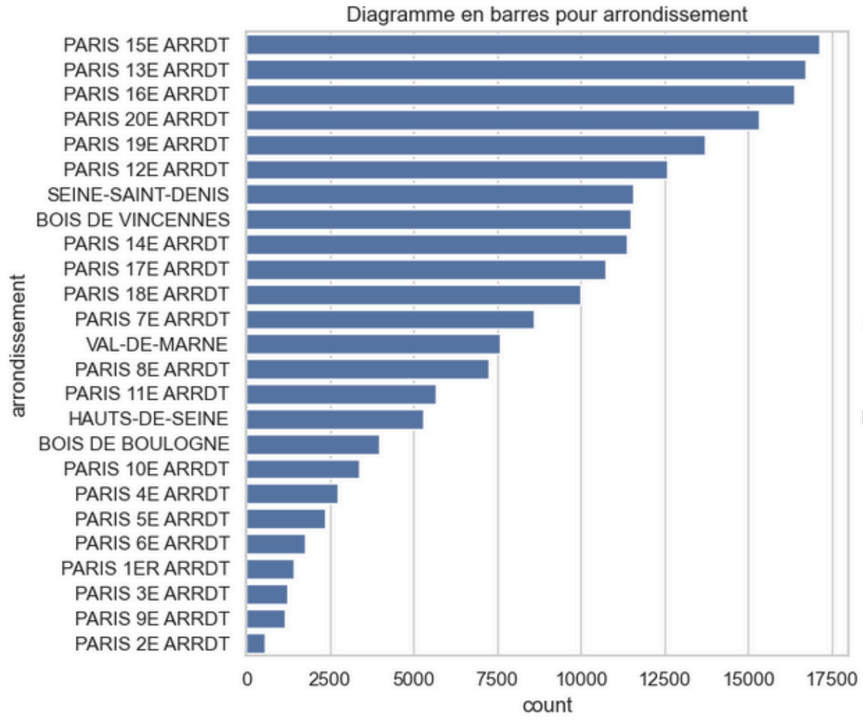


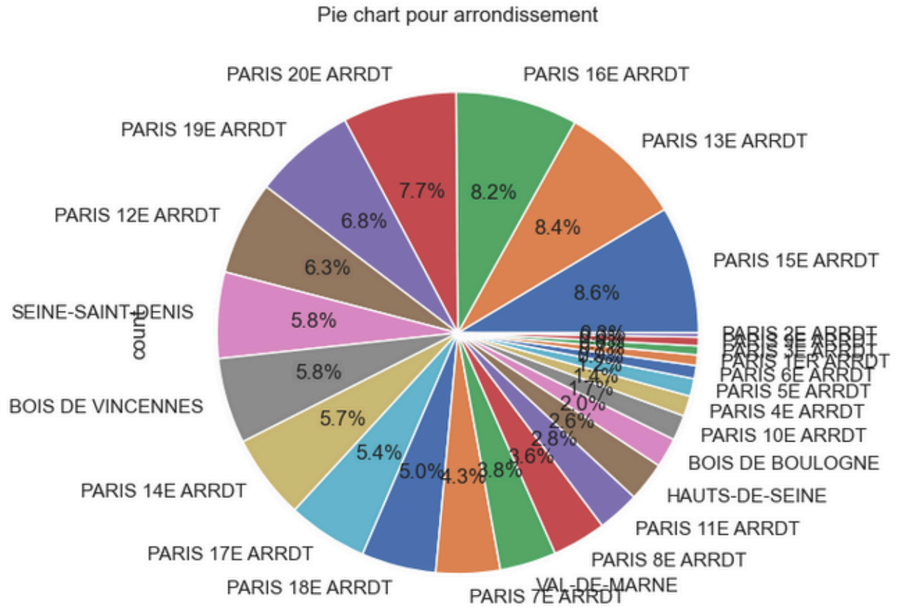
3) Analyse des données

D) Visualisation des données

2 types de variables : celles qui concernent la localisation des arbres, et celles qui informent sur le type des arbres.

Exemple concernant la localisation : Arrondissement. On peut en déduire que la répartition géographique des arbres n'est pas équitable. Quelque soit les traitements retenus, il faudra répartir les arrondissements de manière équitable entre les équipes des agents d'entretiens.

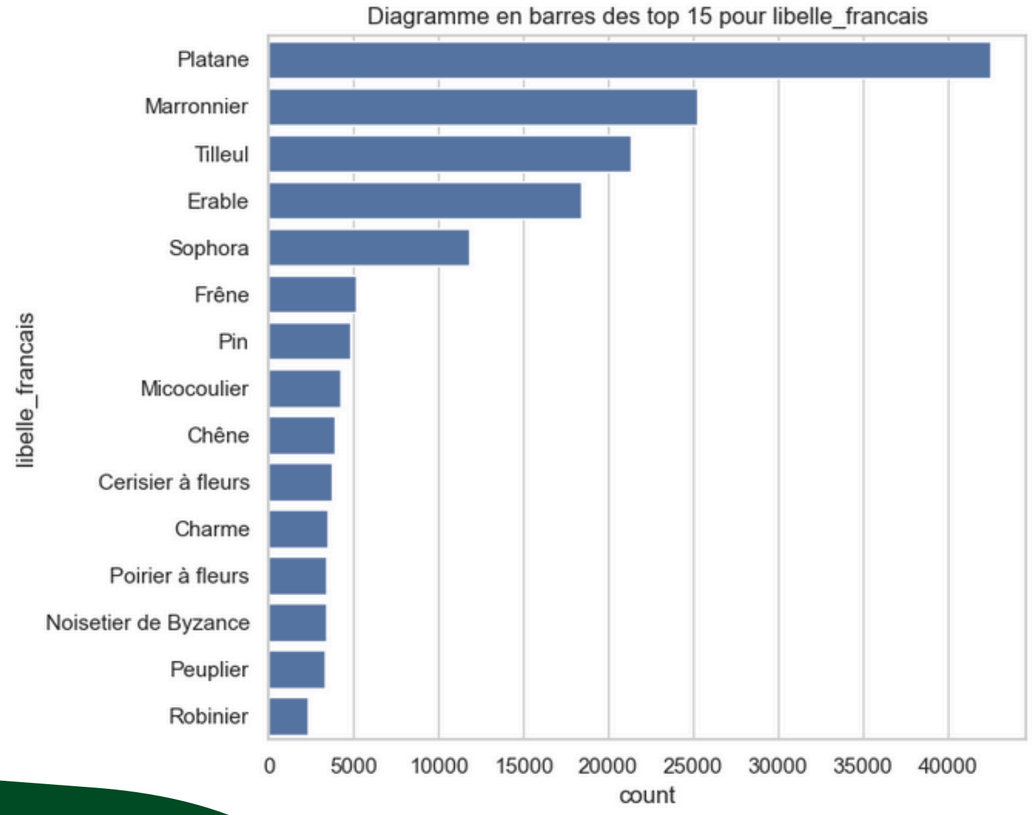


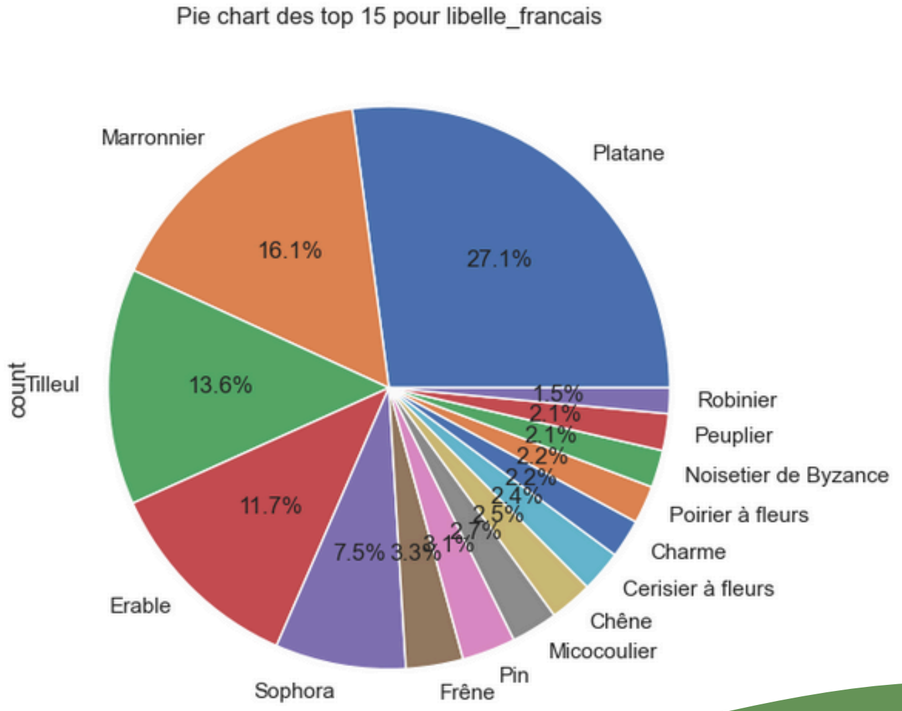


3) Analyse des données

D) Visualisation des données

Exemple de variables concernant le type d'arbres : libelle_francais.
Pour cette étude, nous nous concentrerons sur les 15 libellés les plus répandus. Il s'agit d'une variable importante car les traitements seront différents selon le type d'arbre.





3) Analyse des données

E) Analyses et conclusions

Pour optimiser la tournée d'une équipe d'agents d'entretien d'arbres, nous allons suivre plusieurs étapes :

1. Analyser les besoins d'entretien des différents types d'arbres : chaque type d'arbre peut avoir des besoins d'entretien spécifiques en fonction de son âge, sa hauteur et sa circonférence.
2. Catégoriser les arbres selon leurs besoins d'entretien : diviser les arbres en groupes basés sur leurs caractéristiques communes et leurs besoins.
3. Planifier la tournée des équipes d'entretien : répartir les travaux d'entretien entre plusieurs groupes d'agents, en tenant compte de la proximité géographique des arbres et de la charge de travail.

3) Analyse des données

E) Analyses et conclusions

Étape 1 : Analyser les besoins d'entretien des différents types d'arbres 1/2

Voici une liste des types d'entretien courants en fonction des caractéristiques des arbres :

- Élagage : pour contrôler la taille et la forme de l'arbre.
- Traitement phytosanitaire : pour prévenir ou traiter les maladies et parasites.
- Arrosage : pour les jeunes arbres ou en période de sécheresse.
- Fertilisation : pour assurer une croissance saine.
- Enlèvement des branches mortes : pour la sécurité et la santé de l'arbre.

3) Analyse des données

E) Analyses et conclusions

Étape 1 : Analyser les besoins d'entretien des différents types d'arbres 2/2

Variable	Pourquoi ?	Comment ?
Circonférence	La circonférence donne une indication de la taille et de la robustesse de l'arbre. Les arbres avec une grande circonférence peuvent nécessiter des équipements spécifiques pour l'élagage.	Utiliser la circonférence pour planifier le type d'équipement nécessaire et le temps estimé pour l'entretien.
Hauteur	La hauteur de l'arbre est importante pour évaluer les risques potentiels (branches tombantes, proximité des lignes électriques) et pour planifier les méthodes d'élagage.	Classifier les arbres par hauteur pour assigner les agents avec les compétences et les équipements appropriés (échelles, nacelles).
Stade de développement	Le stade de développement (jeune, jeune adulte, adulte, mature) indique les besoins spécifiques en soins et en traitement.	Identifier le stade de développement pour adapter les interventions.

3) Analyse des données

E) Analyses et conclusions

Étape 2 : Catégoriser les arbres selon leurs besoins d'entretien

1. Arbres nécessitant un élagage fréquent : platane, marronnier, tilleul, érable, sophora, frêne, micocoulier, chêne, cerisier à fleurs.
2. Arbres nécessitant des traitements phytosanitaires fréquents : platane (maladie du chancre coloré), marronnier (mineuse du marronnier), peuplier (rouille), robinier (acarien du robinier).
3. Arbres nécessitant un arrosage fréquent (jeunes arbres ou arbres en zones sèches) : pin, poirier à fleurs, noisetier de Byzance.
4. Arbres nécessitant une fertilisation régulière : charme, chêne, cerisier à fleurs, noisetier de Byzance.

3) Analyse des données

E) Analyses et conclusions

Étape 3 : Planifier la tournée des équipes d'entretien

Pour répartir les travaux d'entretien, nous allons diviser les arbres en groupes basés sur le type d'entretien approprié.

Groupes	Types d'entretien	Types d'arbres	Circonférence	Hauteur	Stade de développement
Groupe 1	Élagage fréquent : Élagage pour contrôler la taille et la forme de l'arbre, enlever les branches mortes ou dangereuses.	Platane, Marronnier, Tilleul, Érable, Sophora, Frêne, Micocoulier, Chêne, Cerisier à fleurs	51-100 cm	6-15 m	Jeune adulte, adulte, mature
Groupe 2	Traitements phytosanitaires fréquents : Traitement contre les maladies et les parasites, prévention et intervention.	Platane, Marronnier, Peuplier, Robinier	51-100 cm	0-15 m	Jeune adulte, adulte, mature
Groupe 3	Arrosage fréquent : Arrosage régulier pour les jeunes arbres nécessitant plus d'eau.	Pin, Poirier à fleurs, Noisetier de Byzance	0-50 cm	0-5 m	Jeune
Groupe 4	Fertilisation régulière : Fertilisation pour assurer une croissance saine et vigoureuse.	Charme, Chêne, Cerisier à fleurs, Noisetier de Byzance	0-100 cm	0-15 m	Jeune, jeune adulte, adulte
Groupe 5	Surveillance et soins des arbres matures : Surveillance régulière pour détecter et traiter les maladies, élagage des branches mortes, évaluation de la stabilité structurelle. + traitement des arbres remarquables	Chêne, Platane, Marronnier, Micocoulier	100+ cm	15+ m	Mature

3) Analyse des données

E) Analyses et conclusions

Étape 3 : Planifier la tournée des équipes d'entretien

Division de Paris en plusieurs secteurs géographiques auxquels attribuer une équipe, en veillant à avoir :

- une répartition géographique logique (arrondissements proches) ;
- une quantité d'arbres égale de sorte que chaque équipe ait une charge de travail équilibrée en fonction des arbres présents.

Groupes	Centre/Est	Rive Gauche	Nord/Nord-Est	Ouest
Arrondissements	Paris 1, Paris 2, Paris 3, Paris 4, Paris 8, Paris 9, Paris 10, Paris 11, Paris 12, Bois de Vincennes, Val de Marne	Paris 5, Paris 6, Paris 7, Paris 13, Paris 14, Paris 15	Paris 18, Paris 19, Seine Saint-Denis, Paris 20	Paris 16, Paris 17, Hauts de Seine, Bois de Boulogne

Conclusion générale

Analyse à compléter avec des données concernant les tournées actuelles, les modalités d'organisation précises de ces tournées, avec l'ensemble des types d'arbres, ...

Merci pour votre écoute