PARTICIPER À UN CONCOURS SUR LA SMART CITY

Voahangy Joan ALEONARD – 02/09/2020

AGENDA DU JOUR

PRÉSENTATION

JEU DE DONNÉES

DÉMARCHE

ANALYSES

PRÉSENTATION DU PROJET

PARIS LA SMART CITY : « VÉGÉTALISONS LA VILLE »

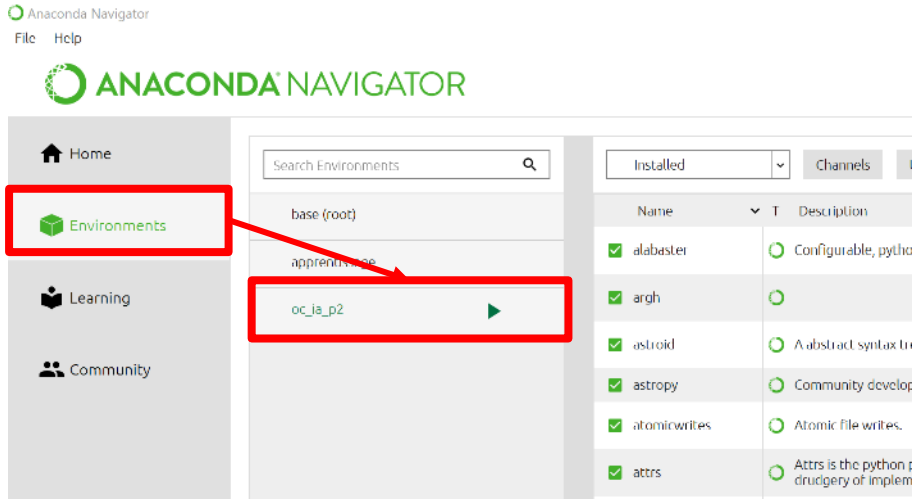
PARIS

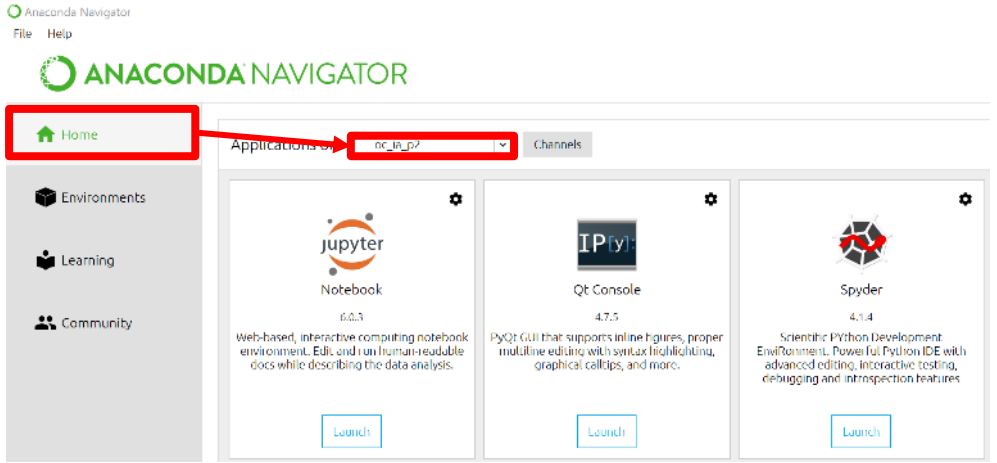
Projet	Participer au **challenge Data** sponsorisé par la ville de Paris.
Enjeu / Objectifs	**Aider Paris à devenir une smart-city** quant à l'entretien de ses arbres : • Réduire le temps de trajet des tournées d'entretien; • De fait, augmenter le nombre d'arbres entretenus.
Mission	Réaliser une analyse exploratoire et statistique des données sur lesdits arbres afin d'aider la ville à **optimiser ses tournées d'entretien**.
Approche	• Utiliser le jeu de données mis à disposition par la ville; • Explorer ces données à l'aide de Python et de ses librairies; • Faire appel à l'analyse exploratoire et l'analyse statistique univariée.
Résultat attendu	Livrer une analyse **complète, pertinente et présentable** dans un **Jupyter Notebook**.

ENVIRONNEMENT DE TRAVAIL FONCTIONNEL

L'environnement de développement est installé via la distribution Anaconda.

Un environnement virtuel - « oc_ia_p2 », a été créé pour l'isolement du projet et la gestion des dépendances, visible dans les menus « Environnements » et « Home » d'**Anaconda Navigator**





JEU DE DONNÉES

VUE D'ENSEMBLE

	Valeurs uniques	Valeurs non-nulles	Valeurs manquantes	% Manquants vs Total	Valeurs à zéro	% Zéro vs Total	Type Données
numero	1	0	200137	100.0	0	0.0	float64
complement_addresse	3796	30902	169235	84.6	0	0.0	object
variete	437	36777	163360	81.6	0	0.0	object
stade_developpement	5	132932	67205	33.6	0	0.0	object
remarquable	3	137039	63098	31.5	136855	68.4	float64
espece	540	198385	1752	0.9	0	0.0	object
libelle_francais	193	198640	1497	0.7	0	0.0	object
genre	176	200121	16	0.0	0	0.0	object
domanialite	10	200136	1	0.0	0	0.0	object
circonference_cm	531	200137	0	0.0	25867	12.9	int64
geo_point_2d_a	200107	200137	0	0.0	0	0.0	float64
hauteur_m	143	200137	0	0.0	39219	19.6	int64
id	200137	200137	0	0.0	0	0.0	int64
type_emplacement	1	200137	0	0.0	0	0.0	object
id_emplacement	69040	200137	0	0.0	0	0.0	object
lieu	6921	200137	0	0.0	0	0.0	object
arrondissement	25	200137	0	0.0	0	0.0	object
geo_point_2d_b	200114	200137	0	0.0	0	0.0	float64

Nombre de valeurs non nulles avec 0 valeurs manquantes

Notre DataFrame a :

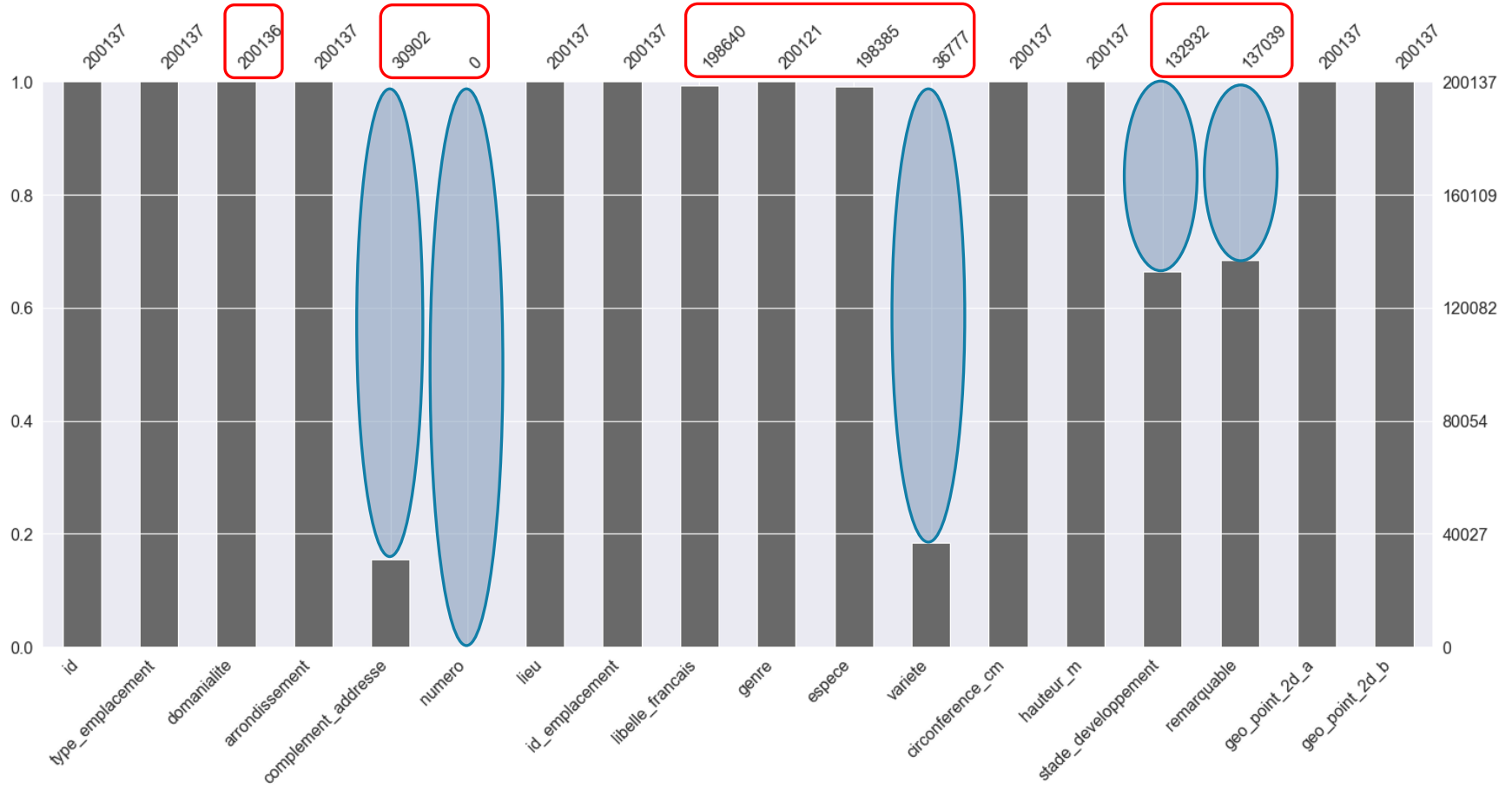
- **18 colonnes (ou variables)** ;
- **200.137 lignes** (y compris les en-têtes);
- on peut aussi en déduire qu'il y a **9 variables SANS valeurs manquantes**.
- **9 variables AVEC valeurs manquantes** ;
- **3 variables contenant des valeurs à zéro ;**
- **11 variables qualitatives** (décrites par la valeur *object*) **et 7 variables quantitatives** (décrites par les valeurs *float64* et *int64* ;

Il y a également 2 spécificités :

- La variable numero qui n'a aucune donnée (valeurs non-nulles = 0);
- La variable type_emplacement qui a la même donnée partout (Valeurs uniques = 1, Valeurs non-nulles = nombre total de lignes)

REPRÉSENTATION DES VALEURS MANQUANTES

La librairie **MissingNo** nous permet repérer **rapidement et visuellement** les données pour lesquelles le remplissage n'est pas complet.



Nombre total de lignes (*y compris les en-têtes*)

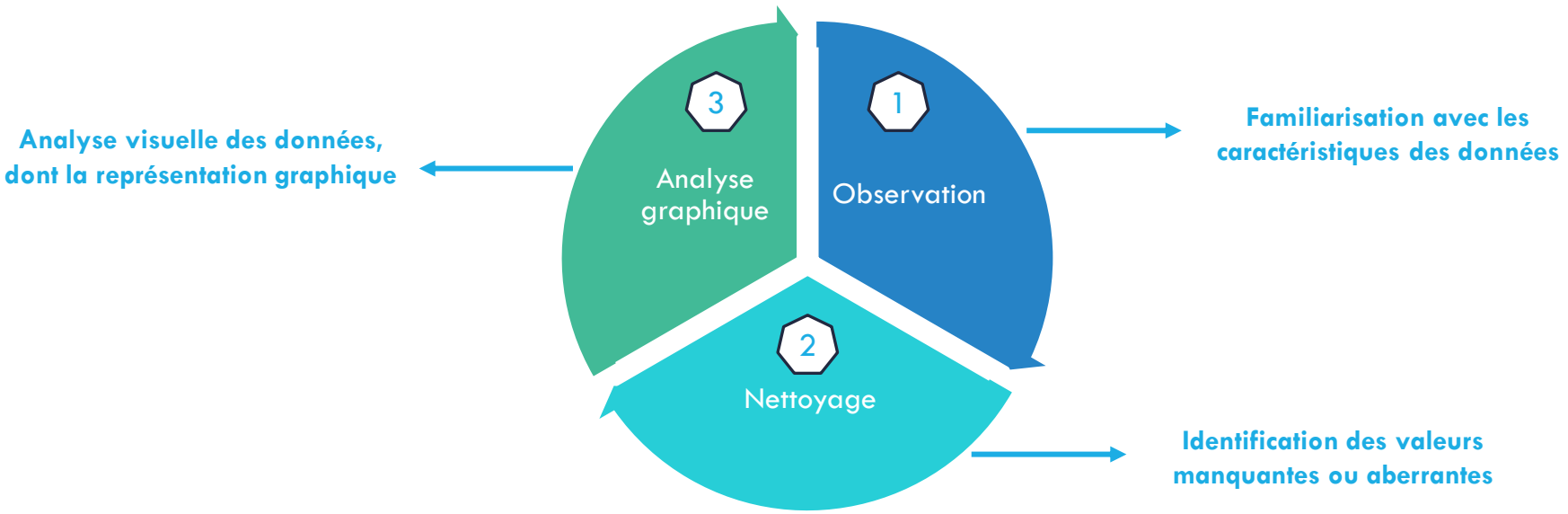
Repérage visuel des données manquantes

Comparaison du nombre de valeurs « non-nulles » par rapport au nombre total de lignes

DÉMARCHE MÉTHODOLOGIQUE

DÉMARCHE D'ANALYSE

L'analyse de données exploratoire est un travail **itératif** en **3 phases :**



Nous ferons appel exclusivement à la **statistique descriptive univariée** :

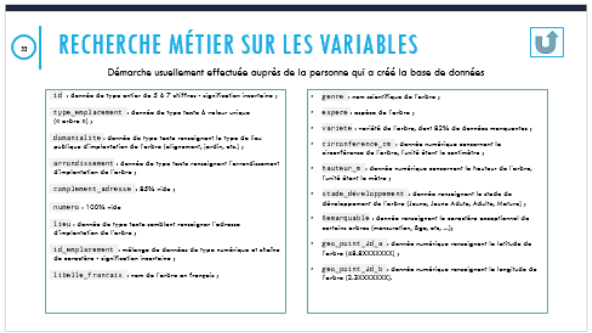
Les observations que nous ferons des variables se feront de manière **isolée, une par une, sans chercher à établir un lien** éventuel entre ces différentes variables

LES DONNÉES UTILES POUR L'ANALYSE

Rappel de l'objectif du challenge :

- Optimiser les tournées d'entretien des arbres de la ville de Paris.

Quelles données pour répondre à cet objectif ?



Une analyse qui permet de choisir les variables suivantes à des fins de :

Identification et spécificités

`libelle_francais`
`stade_developpement`
`remarquable`

Localisation

`domanialite`
`arrondissement`
`geo_point_2d_a`
`geo_point_2d_b`

Mensurations

`circonference_cm`
`Hauteur_m`

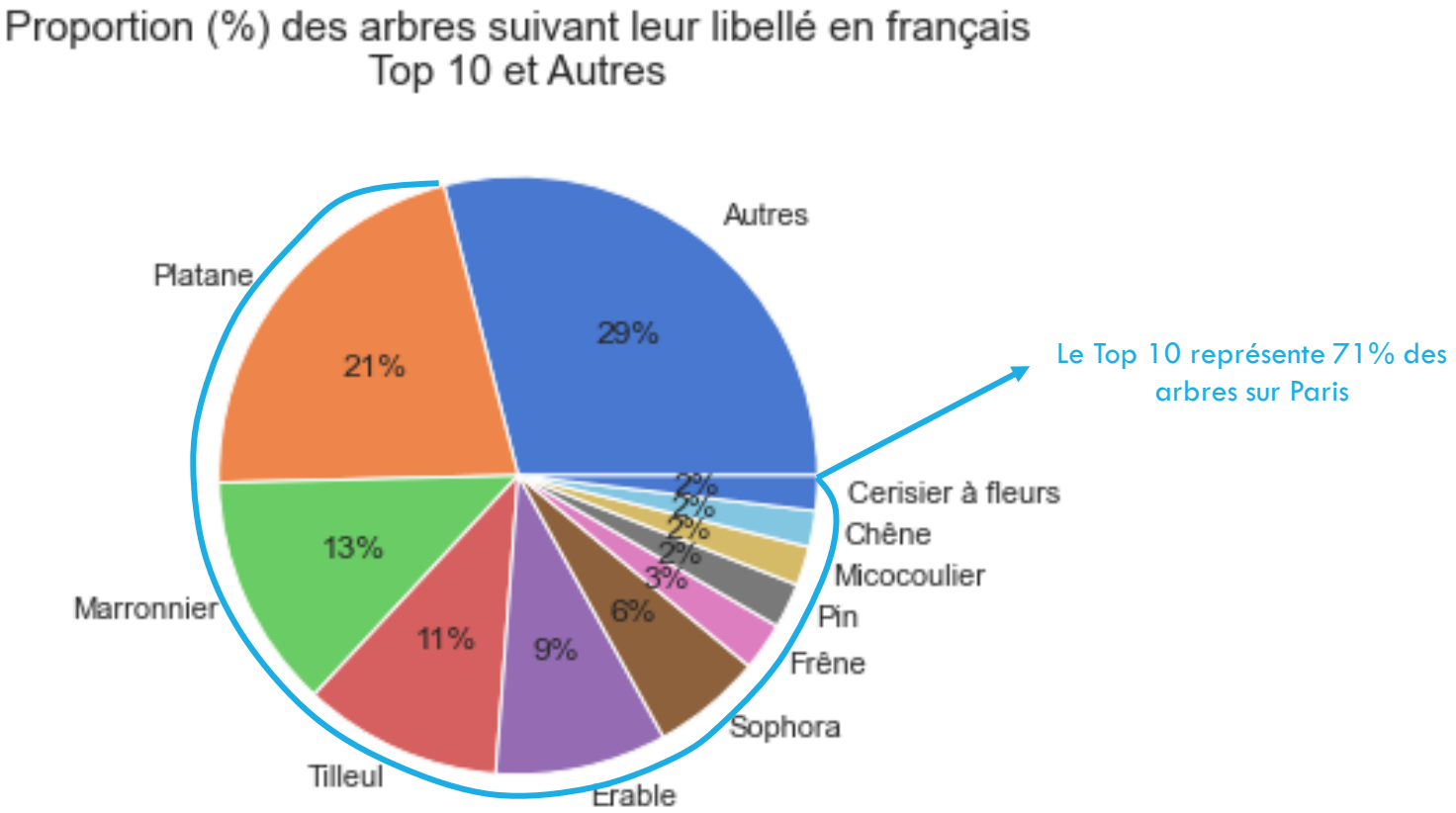
SOIT 9 VARIABLES ESSENTIELLES

	Valeurs uniques	Valeurs non-nulles	Valeurs manquantes	% Manquants vs Total	Valeurs à zéro	% Zéro vs Total	Type Données
STADE_DEVPT	4	132932	67205	33.6	0	0.0	object
REMARQUABLE	2	137039	63098	31.5	136855	68.4	float64
LIBELLE_FR	192	198640	1497	0.7	0	0.0	object
DOMANIALITE	9	200136	1	0.0	0	0.0	object
ARRONDISSEMENT	25	200137	0	0.0	0	0.0	object
LATITUDE	200107	200137	0	0.0	0	0.0	float64
LONGITUDE	200114	200137	0	0.0	0	0.0	float64
CIRCONF_CM	531	200137	0	0.0	25867	12.9	int64
HAUTEUR_M	143	200137	0	0.0	39219	19.6	int64

Dont 4 variables avec valeurs manquantes

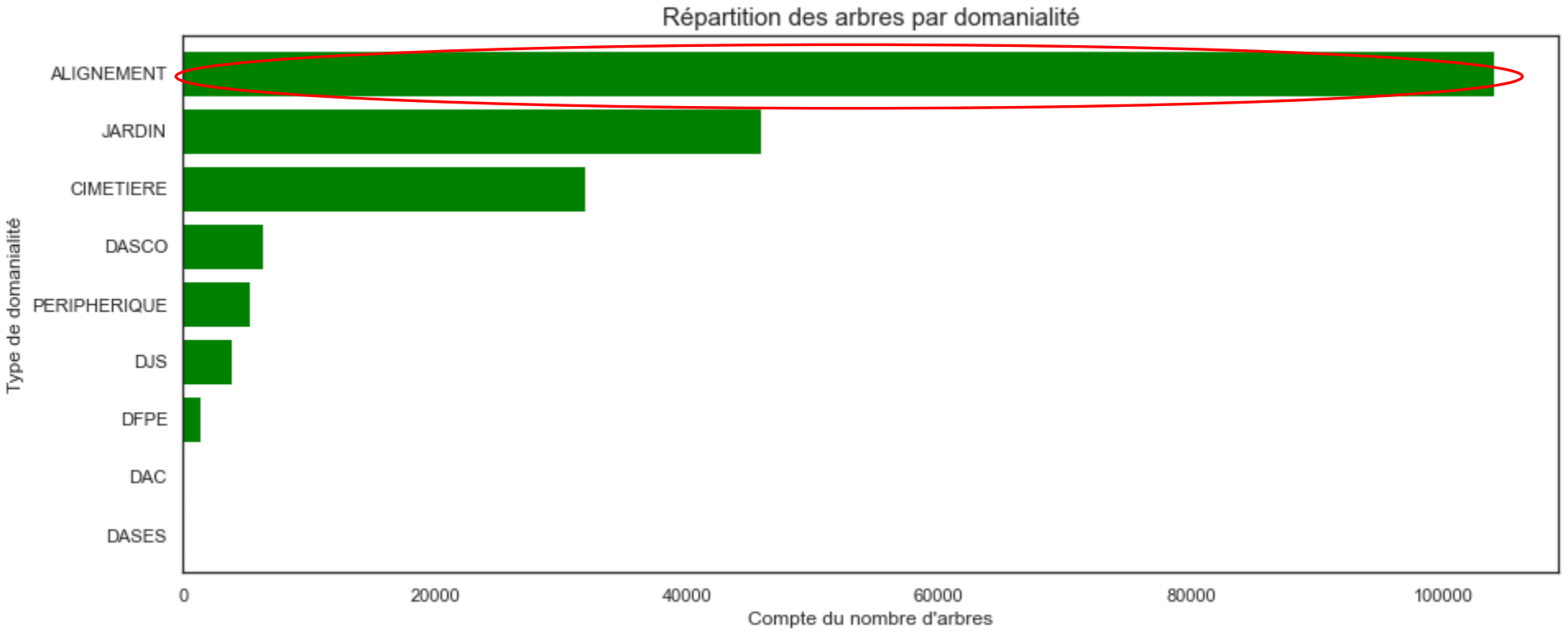
ANALYSE : VARIABLES QUALITATIVES

198.640 arbres hors valeurs manquantes, représentés majoritairement par le Platane



Le Top 10 représente 71% des arbres sur Paris

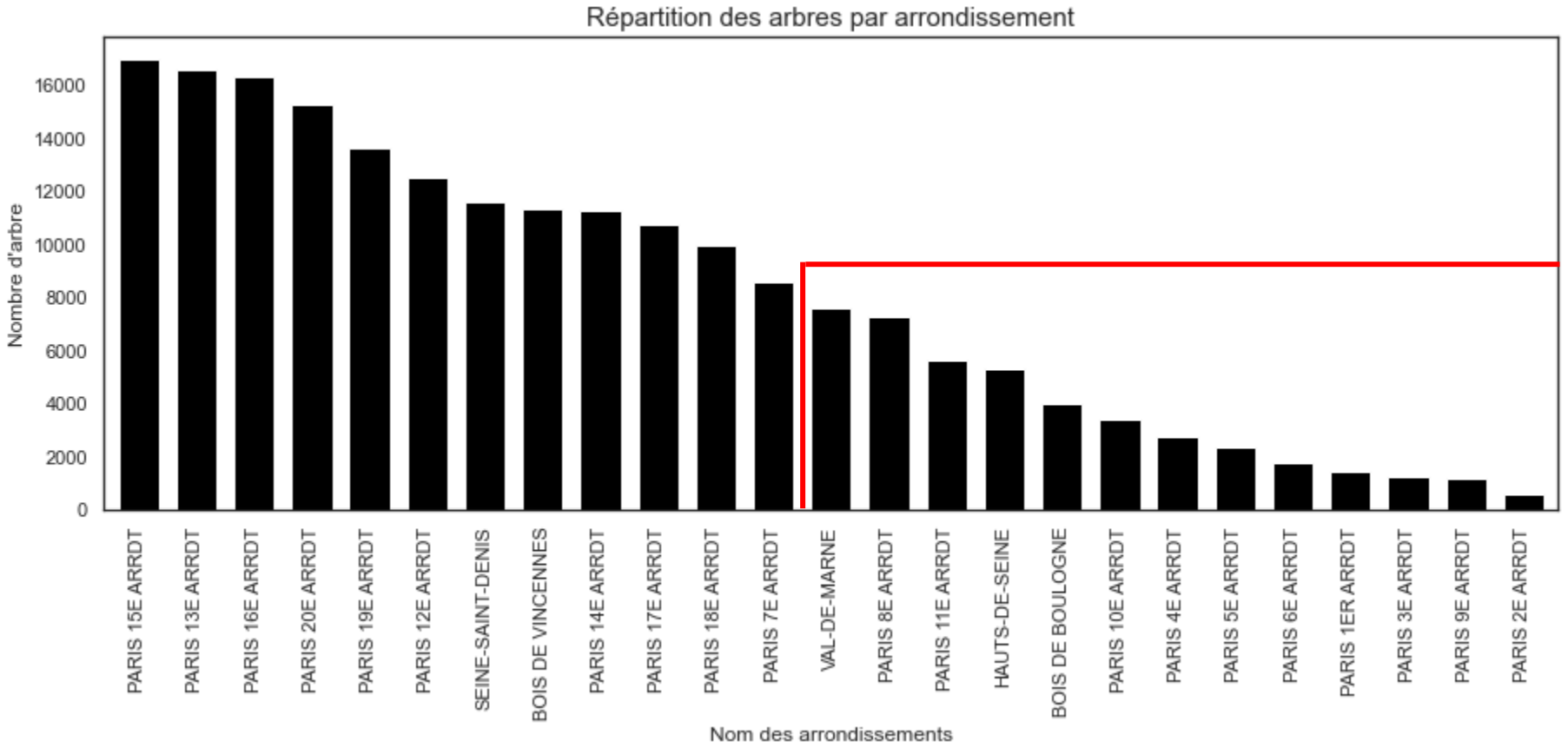
Plus de la moitié des arbres sont des arbres d'alignement



Les arbres d'alignement sont ceux qu'on retrouve le long des voies publiques (les rues).

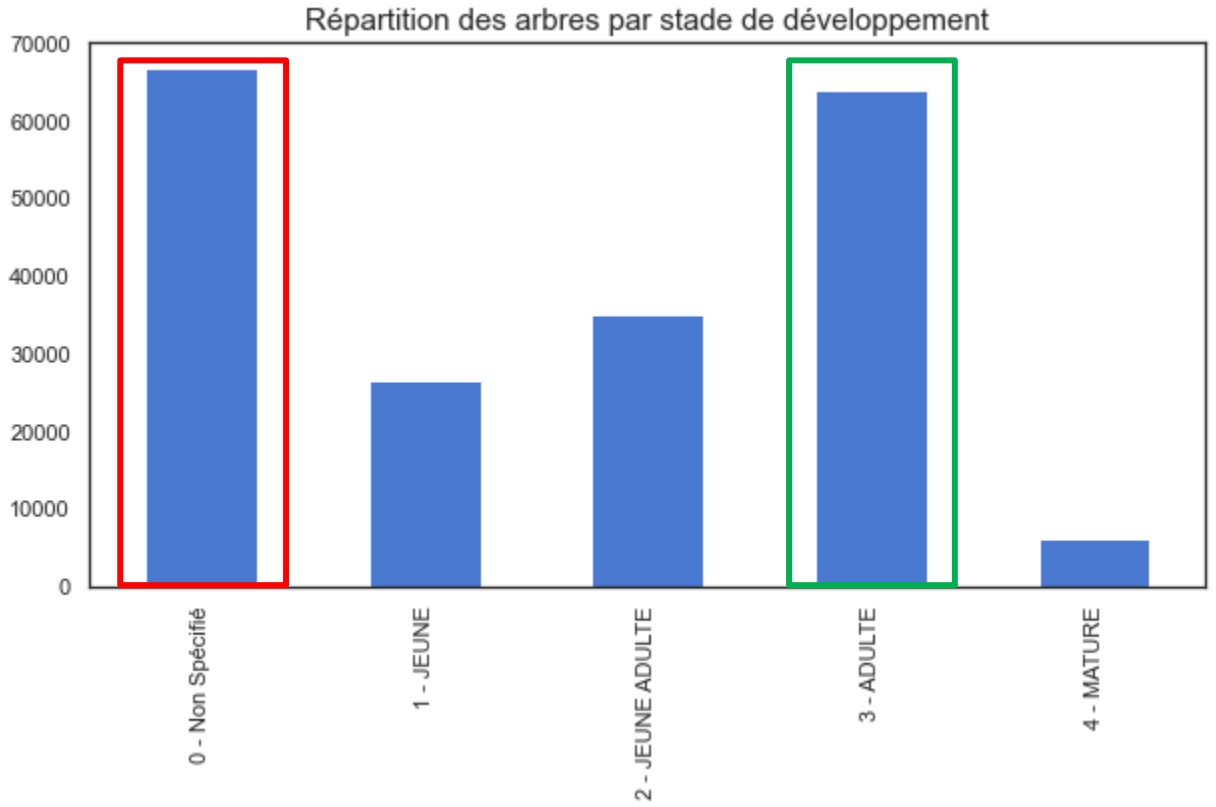
On peut noter qu'il y a plus d'arbres dans les rues que dans les jardins.

50% des arrondissements sont moins végétalisés que d'autres



La répartition des arbres n'est pas la même sur les 25 arrondissements : il y en a où la végétalisation ne semble pas une priorité à ce jour.

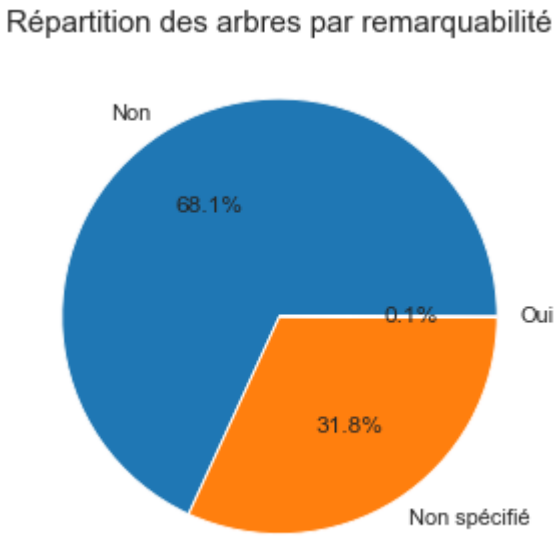
Une majorité d'arbres 'adulte'

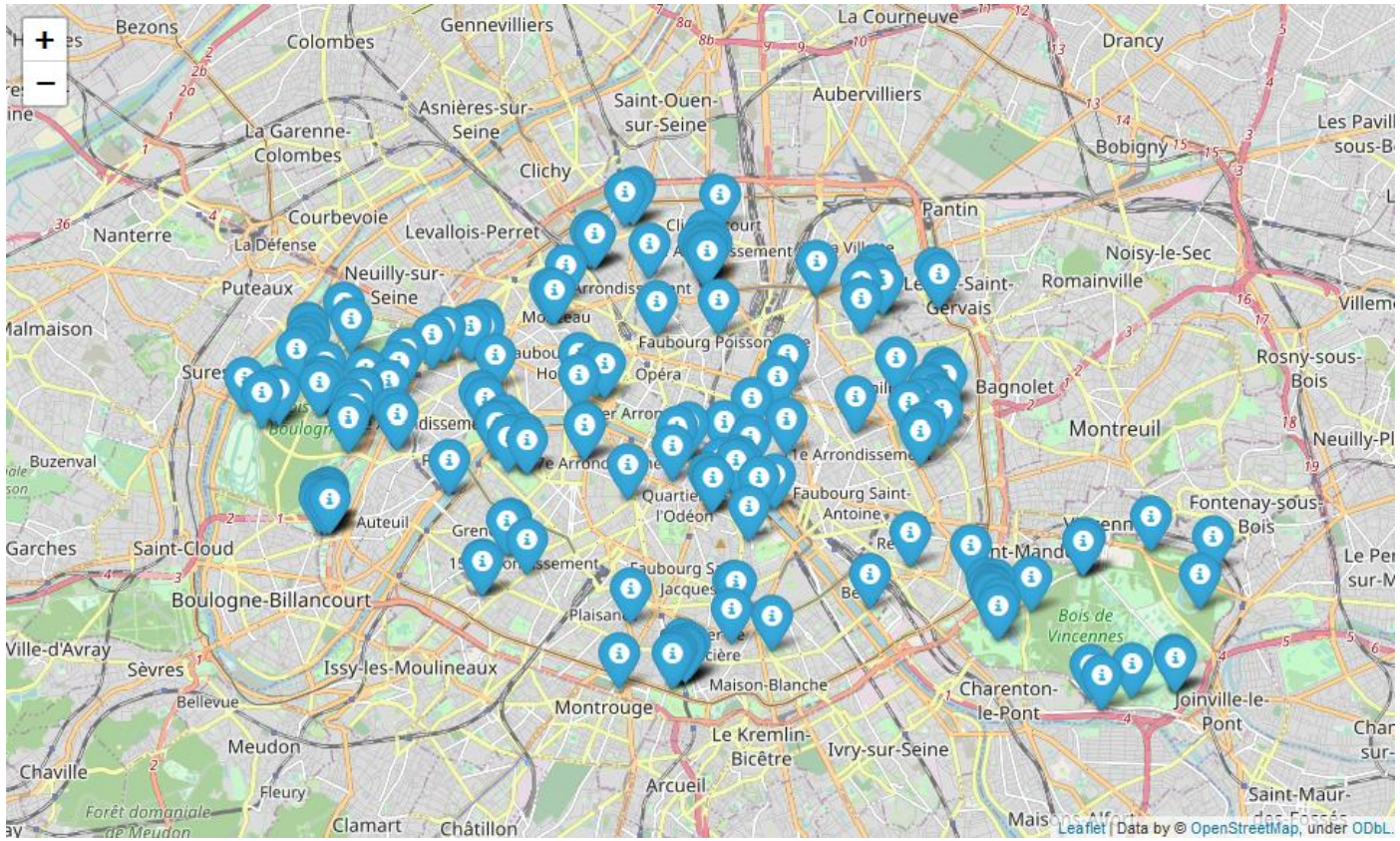


La majorité des arbres dont le stade de développement est identifié sont **adultes**.

Pour près de 70.000 individus, le stade de développement n'est pas connu.

183 arbres remarquables, assez dispersés





Utilisation des coordonnées géographiques (Latitude, Longitude) avec les cartes Folium

Ces arbres remarquables nécessitent une attention particulière.

Il est important de pouvoir les localiser et les surveiller régulièrement.

ANALYSE : VARIABLES QUANTITATIVES

Description statistique des données quantitatives

	CIRCONF_CM	HAUTEUR_M
count	198639.00	198639.00
mean	83.56	13.16
std	675.69	1978.64
min	0.00	0.00
25%	30.00	5.00
50%	70.00	8.00
75%	115.00	12.00
max	250255.00	881818.00

Avec une circonférence ou une hauteur à « 0 », l'arbre n'existe pas

Une circonférence de 2,5km ou une hauteur de 881km semble exagérée

Nous avons ici affaire à des **valeurs aberrantes,** qu'il faut « nettoyer » : corriger ou éliminer.

Focus sur le nettoyage de données

Le **nettoyage de donnée** est l'opération de détection et de correction d'erreurs présentes dans les données : c'est une étape très importante avant l'analyse ou la modélisation des données.

Objectif : améliorer la cohérence, la fiabilité et la valeur des données, afin de prendre des décisions avisées et de définir des stratégies efficaces.

On distingue 2 catégories de nettoyage :

Nettoyage métier

Objectif :
Supprimer les valeurs aberrantes grâce à la connaissance venue de l'extérieur.

Sur Internet pour notre cas :
Circonférence max = 800 centimètres
Hauteur max = 30 mètres

Nettoyage statistique

Objectif :
Repérer les valeurs atypiques/rares grâce aux méthodes statistiques.

Ce nettoyage est toujours appliqué après le nettoyage métier.

Regardons nos résultats lorsqu'on applique les différents nettoyages.

Résultat du nettoyage

Nettoyage métier

	CIRCONF_CM	HAUTEUR_M
count	158555.00	158555.00
mean	92.63	10.38
std	58.76	5.12
min	1.00	1.00
25%	50.00	6.00
50%	80.00	10.00
75%	125.00	14.00
max	790.00	30.00

Nettoyage statistique

	CIRCONF_CM	HAUTEUR_M
count	154664.00	154664.0
mean	88.03	10.1
std	50.87	4.8
min	1.00	1.0
25%	48.00	6.0
50%	80.00	10.0
75%	120.00	14.0
max	237.00	26.0

Le **nettoyage métier** nous renseigne sur **l'étendue des valeurs** présentes dans notre jeu de données, même si elles sont atypiques.

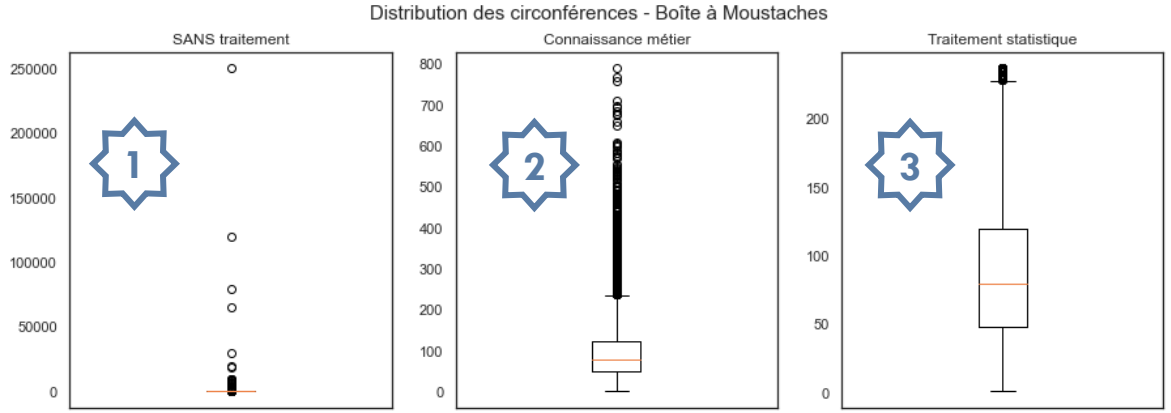
Le **nettoyage statistique** nous renseigne **sur les données qui sont considérés comme rares :** les circonférence au-delà de 237cm et les hauteurs au-delà de 26m.

Représentons graphiquement nos données :

les valeurs sans traitement vs les valeurs avec les nettoyages.

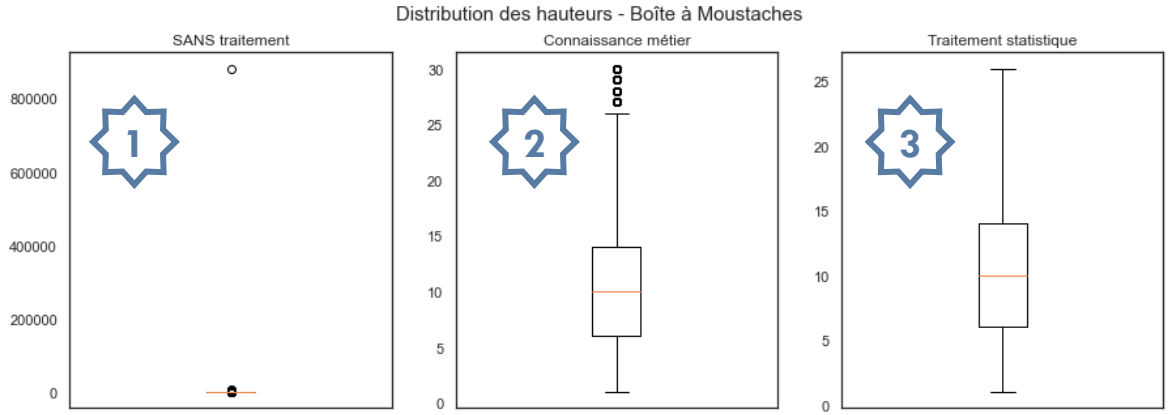
La boîte à moustache (boxplot)

La **boîte à moustaches** permet de représenter, pour un jeu de données, les indicateurs de position (la médiane au centre du rectangle ; le 1er quartile en bas et le 3e quartile en haut), et les *outliers* (valeurs atypiques).



Pour les circonférences :

- Le **1er graphique** (SANS traitement) montre un écrasement de la boîte à moustache dû à la valeur aberrante de 2,5km ;
- Le **2e graphique** (traitement métier) affiche la distribution hors valeurs aberrantes (>800cm) ;
- Le **3e graphique** (traitement statistique) exclut les valeurs rares (>237cm).



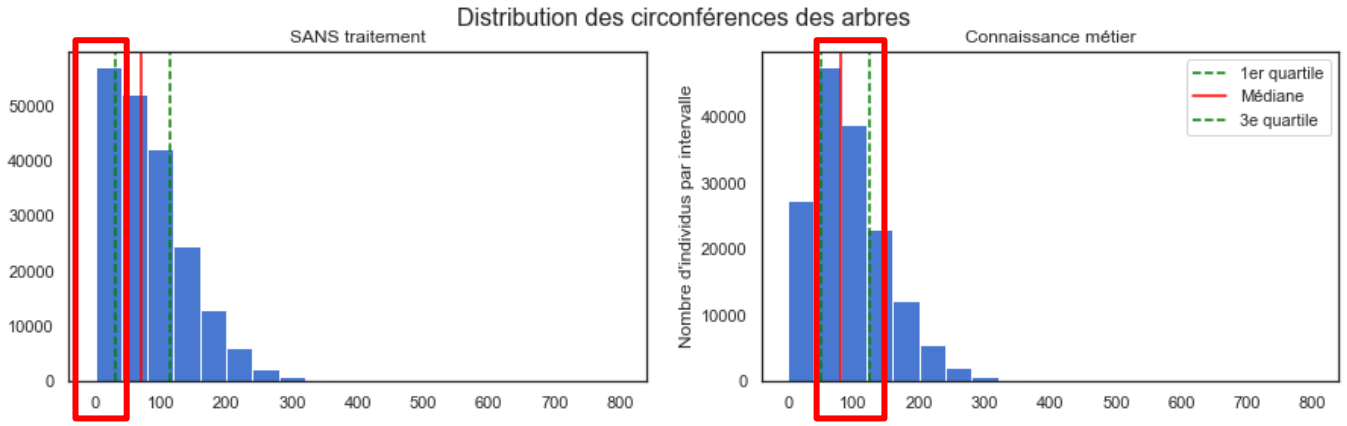
Pour les hauteurs :

- Le **1er graphique** (SANS traitement) montre un écrasement de la boîte à moustache dû à la valeur aberrante de 881km ;
- Le **2e graphique** (traitement métier) affiche la distribution hors valeurs aberrantes (>30m) ;
- Le **3e graphique** (traitement statistique) exclut les valeurs rares (>26m).

Néanmoins, **l'objectif ici n'est pas d'exclure ces arbres atypiques** mais bien d'aider la ville de Paris à les identifier pour mieux les entretenir : nous continuerons nos analyses **uniquement avec le nettoyage métier.**

Histogramme basé sur les données ‘métier’

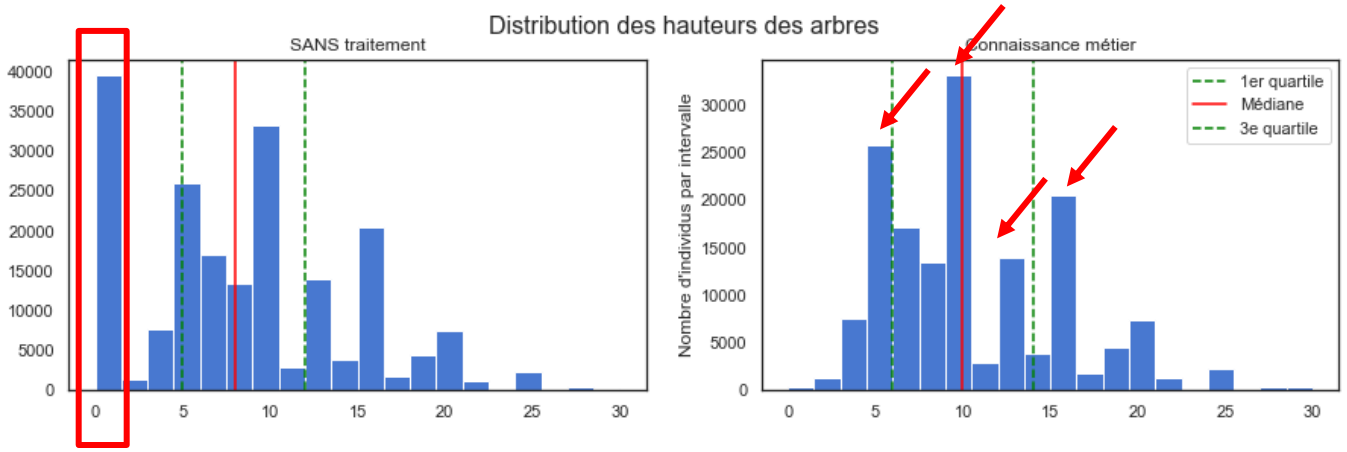
L’**histogramme** regroupe les observations en **classes** (intervalles) dont on représente la fréquence (ici, le nombre d’arbres).



L’histogramme des circonférences indique une distribution plutôt uniforme (resserrée autour de la médiane), avec ou sans valeur aberrante.

Sur le graphique sans traitement, la concentration de données à 0 biaise la dispersion.

Sur le graphique corrigé des valeurs aberrantes, la majorité des arbres affichent une circonférence entre 50 et 125cm. Les valeurs autour de 100 sont plus fréquentes que les valeurs plus élevées (le graphique se trouvant principalement sur la gauche).



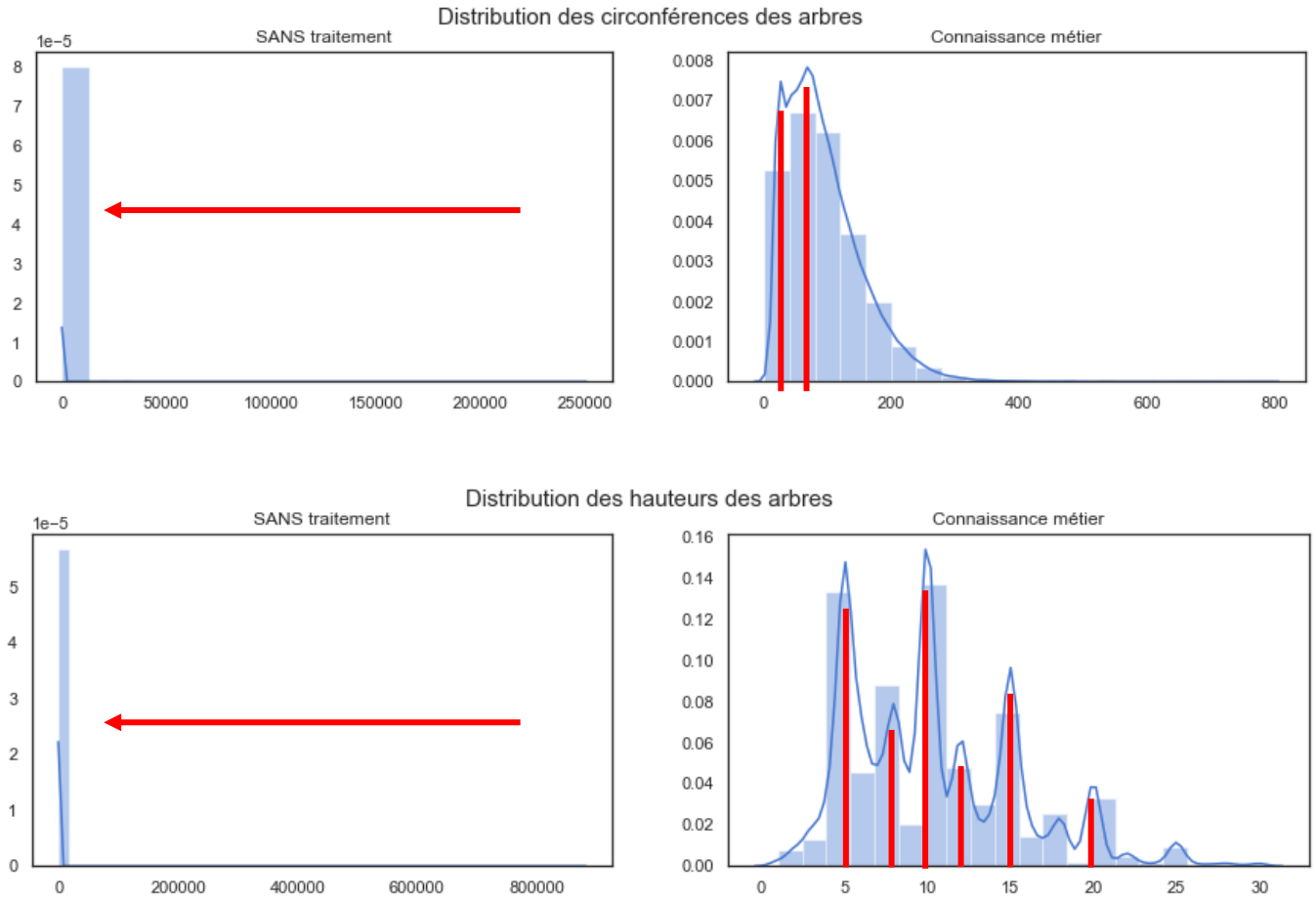
L’histogramme des hauteurs indique une distribution assez disparate.

Sur le graphique sans traitement, l’erreur due aux valeurs aberrantes est toujours explicite.

Les hauteurs comportent des pics et des creux. Il est difficile d’y apporter une explication claire sans analyse croisée avec d’autres variables telles que le type d’arbres ou le stade de développement.

Histogramme avec estimation de densité

L'**histogramme avec estimateur de densité** ajoute une courbe traçant la probabilité d'apparition des observations.



Sans surprise, les graphiques sans traitement affichent un écrasement des données sur la gauche due aux valeurs aberrantes.

Les courbes de densité montrent la fréquence de chaque intervalle de valeurs.

Plus l'intervalle est fréquent, plus il y a de pic élevé.

Les points de données sur l'axe des x et se trouvant au bout de chaque ligne rouge représentent les circonférences ou hauteurs qui apparaissent le plus fréquemment dans l'intervalle.

SYNTHÈSE

SYNTHÈSE SUR LE PATRIMOINE ARBORÉ

Identification & spécificités

~158.600 arbres identifiés

~169 types d'arbres différents

Top 10 = 74% des types d'arbres

Top Individu : Platane (25%)

39,5% d'arbres d'âge adulte

171 arbres remarquables

Localisation

62% d'arbres en alignement

9,4% dans le 16e arrondissement

Mensurations

Hauteur max : 30 m

Circonférence max: 800 cm

Par la connaissance de toutes ses caractéristiques, la Ville de Paris pourra monter une stratégie d'optimisation des tournées d'entretien de ses arbres.

CONCLUSION

Par la connaissance de toutes les caractéristiques décrites dans notre synthèse, la ***Direction des Espaces Verts et de l'Environnement*** de la ville de Paris pourra monter une stratégie d'optimisation des tournées :

- par le **choix du type d'entretien** (élagage ; abattement des arbres dépérissants ou dangereux ; remplacement des arbres vieillissants ou malades);
- la **détermination des fréquences des entretiens et l'allocation de ressources spécifiques** en fonction du type d'arbres, de son stade de développement et de sa qualité d'arbre "remarquable" ou non ;
- par **la création de routine de tournées** grâce à leur localisation ;
- etc.

Tout dépendra évidemment **des objectifs et des priorités définis** par la ville dans son programme **"Végétalisons la ville".**

Une stratégie à réfléchir serait d'essayer de végétaliser un peu plus les arrondissements les moins fournis en arbres.

QUESTIONS / RÉPONSES

ANNEXES

LIBRAIRIES PYTHON POUR L'ANALYSE DE DONNÉES

- **Pandas** : extension permettant la manipulation et l'analyse de données sur des tableaux numériques ou des séries temporelles.

- **NumPy** (Numerical Python) : extension permettant de manipuler et d'opérer des fonctions mathématiques sur des matrices ou des tableaux multidimensionnels.

- **Matplotlib** : extension permettant de tracer et de visualiser des données sous forme de graphiques.

- **Seaborn** : surcouche de Matplotlib, apportant des améliorations de visualisation.

D'autres librairies seront également utilisées dans cette analyse de manière ponctuelle :

- **MissingNo** : extension de visualisation des données manquantes.
- **Folium** : extension permettant la manipulation de cartes géographiques.

ET JUPYTER NOTEBOOK

Le **Jupyter Notebook** est un cahier électronique permettant de rassembler dans un même document du texte, des images et du code informatique exécutable et manipulable dans un navigateur web.

Il est utilisé pour exposer et partager notre analyse, notamment ses 3 grandes parties :

VILLE DE PARIS

Challenge Data Paris : "Végétalisons la ville"

Depuis l'année 2015, la ville de Paris communique largement sur son plan stratégique "**Ville intelligente et durable**", qui contribuera à faire de la ville de Paris **une smart city** d'ici 2050. Un des 3 volets de ce plan concerne la "**Ville Ingénieuse**", c'est-à-dire, une ville qui réinterroge le fonctionnement des réseaux, des aménagements et des flux urbains afin d'optimiser et d'économiser les ressources.

Le programme **Végétalisons la ville** fait partie du volet "**Ville ingénieuse**", et est géré par la **DEVE** - la *Direction des Espaces Verts Et de l'Environnement*. Dans ce cadre, la ville de Paris sponsorise un challenge Data ayant pour objectif d'aider la ville à **optimiser les tournées d'entretien de ses arbres**.

PARTIE 1 - Présentation générale du jeu de données

Le **jeu de données** portant sur le patrimoine arboré de la ville de Paris est disponible sur ce lien : opendata.paris.fr.

Dans cette 1ère partie, nous allons **établir le profil** du jeu de données : c'est un processus qui nous aide à avoir un aperçu des données à notre disposition.

PARTIE 2 - Démarche méthodologique d'analyse

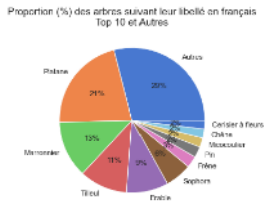
La démarche d'**analyse de données exploratoire** (ou analyse descriptive) est un travail itératif en 3 phases :

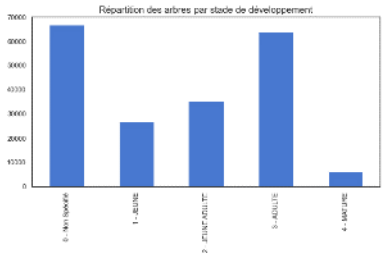
- Observation de la donnée afin de se familiariser avec ses caractéristiques ;
- Nettoyage de la donnée suite à l'identification de valeurs manquantes ou aberrantes ;
- Analyse de la donnée, incluant la visualisation sur un graphique.

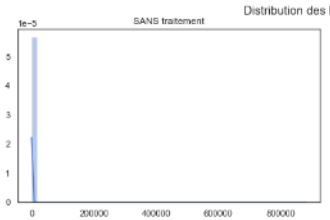
PARTIE 3 - Synthèse de l'analyse

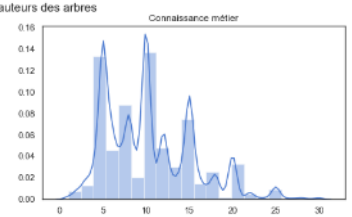
Maintenant que nous avons bien exploré notre jeu de données, nous pouvons en faire une synthèse fiable.

… avec une alternance de textes explicatifs, de codes et ses résultats en graphiques et tableaux









	CIRCONF_CM	HAUTEUR_M
count	158555.00	158555.00
mean	92.63	10.38
std	58.76	5.12
min	1.00	1.00
25%	50.00	6.00
50%	80.00	10.00
75%	125.00	14.00
max	790.00	30.00

RECHERCHE MÉTIER SUR LES VARIABLES

Démarche usuellement effectuée auprès de la personne qui a créé la base de données

`id` : donnée de type entier de 5 à 7 chiffres - signification incertaine ;

`type_emplacement` : donnée de type texte à valeur unique (« arbre ») ;

`domanialite` : donnée de type texte renseignant le type de lieu publique d'implantation de l'arbre (alignement, jardin, etc.) ;

`arrondissement` : donnée de type texte renseignant l'arrondissement d'implantation de l'arbre ;

`complement_adresse` : 85% vide ;

`numero` : 100% vide

`lieu` : donnée de type texte semblant renseigner l'adresse d'implantation de l'arbre ;

`id_emplacement` : mélange de données de type numérique et chaîne de caractère - signification incertaine ;

`libelle_francais` : nom de l'arbre en français ;

- `genre` : nom scientifique de l'arbre ;
- `espece` : espèce de l'arbre ;
- `variete` : variété de l'arbre, dont 82% de données manquantes ;
- `circonference_cm` : donnée numérique concernant la circonférence de l'arbre, l'unité étant le centimètre ;
- `hauteur_m` : donnée numérique concernant la hauteur de l'arbre, l'unité étant le mètre ;
- `stade_développement` : donnée renseignant le stade de développement de l'arbre (Jeune, Jeune Adute, Adulte, Mature) ;
- `Remarquable` : donnée renseignant le caractère exceptionnel de certains arbres (mensuration, âge, etc, ...);
- `geo_point_2d_a` : donnée numérique renseignant la latitude de l'arbre (48.8XXXXXXX) ;
- `geo_point_2d_b` : donnée numérique renseignant la longitude de l'arbre (2.3XXXXXXX).

Ce document a été produit dans le cadre de la soutenance du projet n°2 du parcours Ingénieur IA d'OpenClassrooms :
« Participez à un concours sur la Smart City »

Mentor : Thierno DIOP
Evaluateur : Olivier CHOTIN

OPENCLASSROOMS
En partenariat avec
CentraleSupélec Microsoft