Participez à un concours sur la Smart City

Projet 2:

➔ Effectuer une analyse statistique univariée
➔ Utiliser des librairies python pour réaliser une analyse de données exploratoire

Formation: Ingénieur IA

Marius BARTCUS

Sommaire

Context du travail

Environnement du travail

Présentation générale du jeu de données

Démarche méthodologique d’analyse de données

Synthèse de l’analyse de données

Context du travail

- ➔ **Aidez Paris à devenir une smart-city !**
- ➔ ONG “**Data is for Good**” propose des challenges de Data Science en ligne sur des thématiques ayant trait au bien commun.
- ➔ **Contribué à une optimisation des tournées pour l’entretien des arbres de la ville**. (moins de tournées égale moins de trajets, et plus d’arbres entretenus.)

Environnement du travail

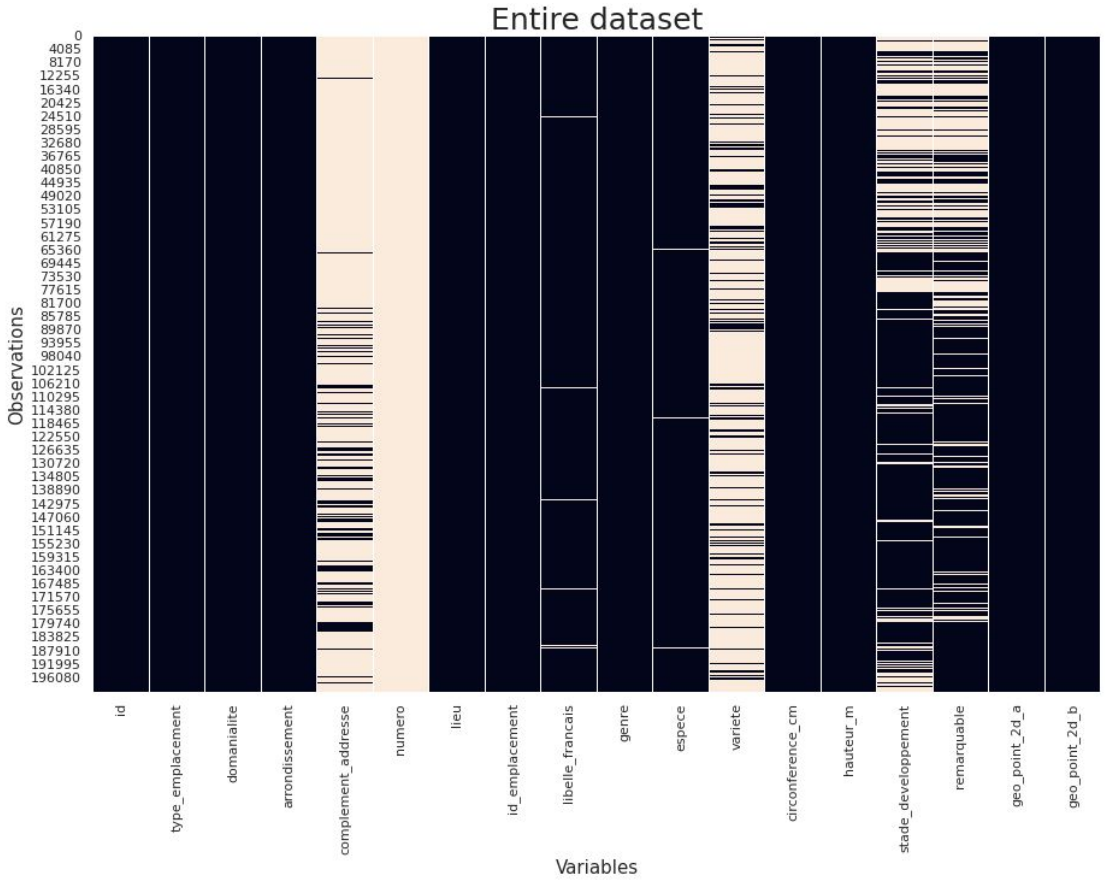
- ➔ Linux
- ➔ Anaconda
  - ◆ python
  - ◆ jupyter
  - ◆ numpy
  - ◆ pandas
  - ◆ matplotlib
  - ◆ seaborn
  - ◆ folium

Présentation générale du jeu de données

- **Description des variables**: The variable description are also available on: Arbre-Paris Data
  - Identification de donnée
    - id : id de l'arbre
    - id_emplacement : identifiant de l'emplacement
  - Description de l'arbre
    - domanialite : type de lieu auquel appartient l'arbre
    - type_emplacement : le type d'emplacement
    - libelle_francais : nom commun (vernaculaire) de l'espèce de l'arbre
    - genre : genre de l'arbre
    - espece : espèce de l'arbre
    - variete : variété de l'arbre
    - circonference_cm : circonférence en centimètres de l'arbre
    - hauteur_m : taille en mètres de l'arbre
    - stade_developpement : stade de développement de l'arbre
    - remarquable : si l'arbre est "remarquable" ou non (0, 1 ou NaN)
  - Emplacement de l'arbre
    - arrondissement : arrondissement de Paris où est situé l'arbre
    - complement_addresse : complement d'adress
    - numero : numéro de l'adress
    - lieu : adresse de l'arbre
    - geo_point_2d_a : latitude de la position de l'arbre
    - geo_point_2d_b : longitude de la position de l'arbre

- **Lignes et colonnes**: 200137, 18
- **Types des variables**: qualitative: 13 quantitative: 5
  - **quantitatives**
    - **discrètes** : id, circonference_cm, hauteur_m
    - **continues** : geo_point_2d_a, geo_point_2d_b
  - **qualitatives**
    - **nominales** : type_emplacement, domanialite, arrondissement, complement_addresse, numero, lieu, id_emplacement, libelle_francais, genre, espece, variete
    - **ordinales** : stade_developpement, remarquable

Présentation générale du jeu de données



Démarche méthodologique d'analyse de données

Première analyse de donnée

décrire les données
calculé les indicateurs statistiques
comparez les ordres de grandeur
vérifier les valeurs manquants
voir les valeurs aberrantes
vérifier les doublons

Nettoyage des données

elimination des valeurs aberrantes
elimination des colonnes inutiles
résoudre valeurs manquants
résoudre valeurs lieu: split '/'

Analyse des données

- ➔ **variable libelle_francais**
  - ◆ top 5
- ➔ **variable arrondissement**
  - ◆ Combien des arbres?
  - ◆ Combien des arbres remarquables?
- ➔ **variable lieu et split_lieu**
  - ◆ l'intérêt de création d'un nouvelle variable?
  - ◆ genre des arbres majoritairement planté au lieu le plus vert de Paris?
- ➔ **variable stade_developpement**
  - ◆ Les variables de grandeur (hauteur_m et circunference_cm) par rapport au stade_developpement d'un arbre, aussi classifié par le variable remarquable.
- ➔ **variable domanialite**
  - ◆ Sur quelle domanialite on a des arbres remarquables?
  - ◆ Quest que on peux dire de les taille des arbres (hauteur_m et circonference_cm) par rapport au domanialite?
  - ◆ Dans quelle domanialite se trouve plus des arbres remarquable?

Afficher les arbres remarquables dans le Jardin

Synthèse de l'analyse de données

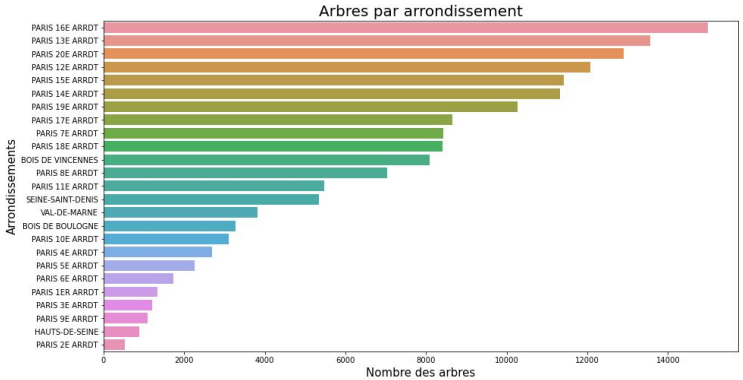
```
df.describe()
```

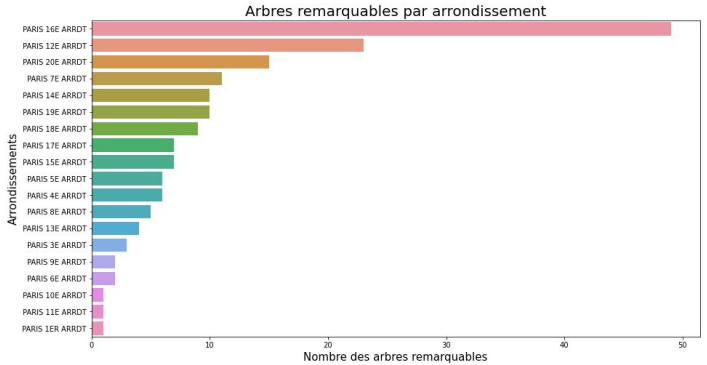
	id	numero	circonference_cm	hauteur_m	remarquable	geo_point_2d_a	geo_point_2d_b
count	2.001370e+05	0.0	200137.000000	200137.000000	137039.000000	200137.000000	200137.000000
mean	3.872027e+05	NaN	83.380479	13.110509	0.001343	48.854491	2.348208
std	5.456032e+05	NaN	673.190213	1971.217387	0.036618	0.030234	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

```
df.describe(include=[object])
```

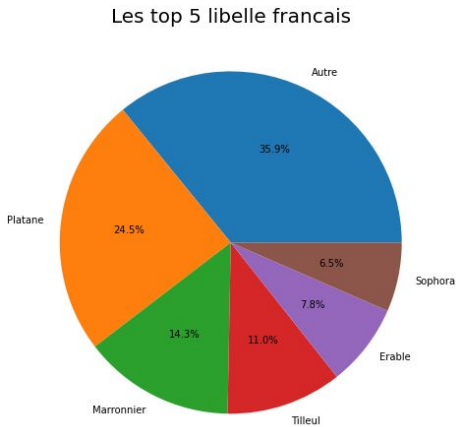
	type_emplacement	domanialite	arrondissement	complement_addresse	lieu	id_emplacement	libelle_francais	genre	espece	variete	sta
count	200137	200136	200137	30902	200137	200137	198640	200121	198385	36777	
unique	1	9	25	3795	6921	69040	192	175	539	436	
top	Arbre	Alignement	PARIS 15E ARRDT	SN°	PARC FLORAL DE PARIS / ROUTE DE LA PYRAMIDE	101001	Platane	Platanus	x hispanica	Baumannii'	
freq	200137	104949	17151	557	2995	1324	42508	42591	36409	4538	

Synthèse de l’analyse de données





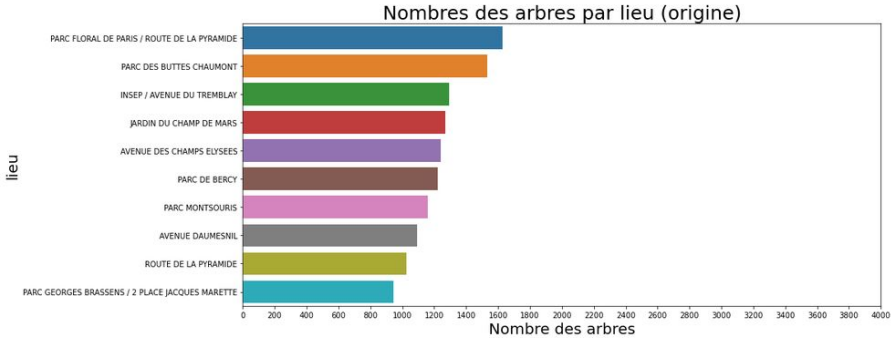
- La zone le plus vert de Paris c'est dans 16EME, 13EME, 20EME, 12EME arrondissements. Il faut avoir plus des personelle pour traiter ces arbre que dans les arrondissements le moins verte comme on voit dans 9EME, 2EME arrondissements et haute de seine.
- On trouve beaucoup des arbre remarquables dans 16EME, 12EME, 20EME, l'arrondissements. Les personelle qui font la route sur ces arrondissement doivent être formé pour une traitement spécifique des arbres remarquables.
- On trouve très peux d'arbre remarquables dans 10EME, 1ERE, 11EME arrondissement.

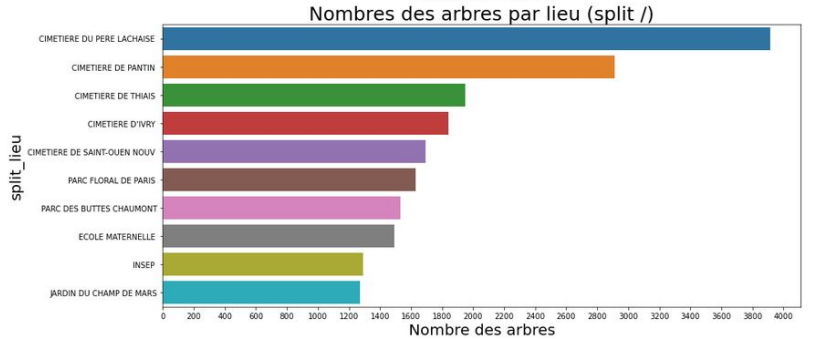


- Les top 5 libelle français sont Platane (24.5%), Marronier (14.3%), Tilleuil (11%), Erable (7.8%) et Sophora(6.5%). Tout les autre arbre sont de 35.9%.

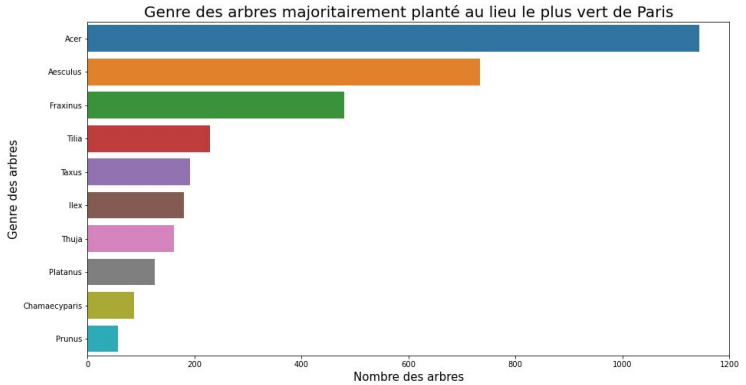
Synthèse de l’analyse de données

- En splittant le variable lieu on trouve les lieu avec plus d'importance.
- On remarque que les cimetières sont les plus vert à Paris.

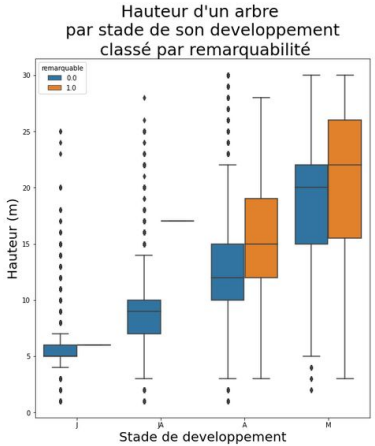


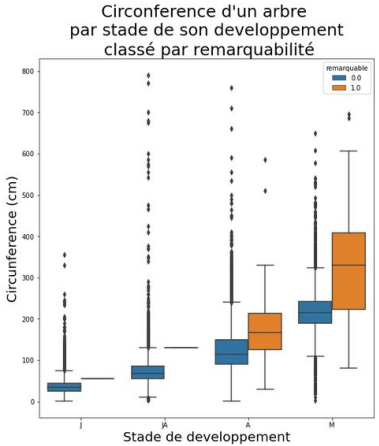


- Les genre des arbres les majoritairement planté à Paris c'est Acer (plus de 1000 arbres), Aesculus (plus de 700 arbres) et Fraxinus (plus de 400 arbres). Les autres genre des arbres sont autour de 200 arbres et moin.



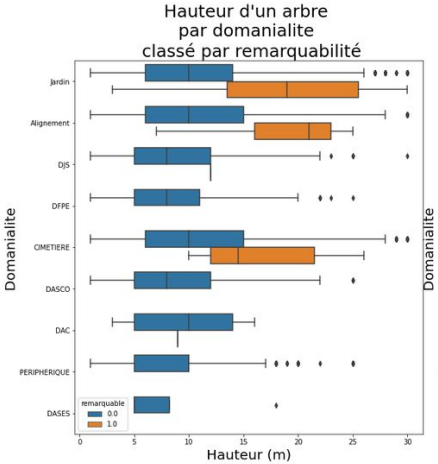
Synthèse de l’analyse de données

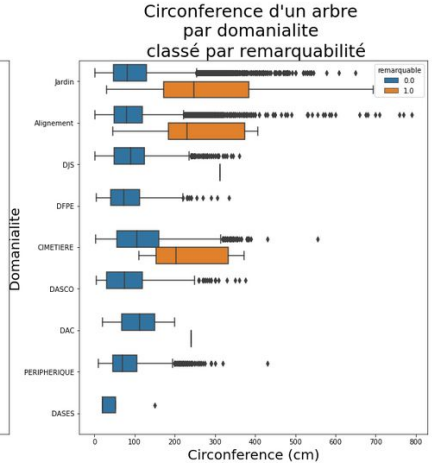




- Le plus le `hauteur_m` ou Le `circunference_cm` est grande, le `stade_developpement` change de Jeune à Jeune Adulte, Adulte et jusqu'à Mature.
- On observe des valeures aberrantes de `hauteur_m` et `circunference_cm` pour chaque `stade_developpement`. Il faut peut être re-mesurer ces arbre, ou changé le `stade_developpement`.
- On a 1 arbre remarquable Jeune et 1 Jeune Adulte, par contre on trouve beaucoup des arbre remarquable Adulte et Mature.

- les arbres remarquables sont plus grands en hauteur et en circonférence que ce qui ne sont pas remarquable.
- les arbres remarquables se trouve dans des Jardin, Alignment, et Cimetière.
- on a une arbre remarquable à DJS et une arbre remarquable à DAC. Il faut vérifier ces deux arbres.





Arbres remarquable dans le Jardin