Analyse exploratoire univariée

Jeu de données arbres_paris

Par FEUYAN TCHOUO

Le problème

La ville

Depuis la COP21, le développement durable est une problématique d'actualité dans toutes les villes du monde. La ville de Paris dans cet élan s'engage à être une ville verte.

Le contexte

Dans le cadre du programme "Végétalisons la ville", Paris entend entretenir et développer son couvert végétal.

Définition du problème

En s'appuyant sur la base de données des arbres de la ville, comment optimiser les tournées pour l'entretien des arbres ?

Les défis

Défis 1

Défis 2

Défis 3

Consistance de la base de données

La base de données doit être bien structurée et consistante pour garantir de bons résultats

Extraction de l’information

Analyser les données et extraire de bonnes informations

Prise de décision

Sur la base des informations extraites, prendre les bonnes décisions.

Analyse exploratoire

Nous procéderons à une analyse exploratoire de la base de données fournie, afin d’aboutir à une prise de décision adéquate

Description de la base de données

Description du jeu de données

- 200137 arbres et 18 caractères;
- 13 variables qualitatives nominales, 3 variables qualitatives ordinales, 2 variables quantitatives continues;
- Valeurs des types entier, réel et objet (chaînes de caractères)

Analyse exploratoire

Analyse exploratoire - Nettoyage

Type de données

Tout semble correct au niveau du type de données de chaque variable.

```
arbres.dtypes
```

```
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
```

Analyse exploratoire - Nettoyage

Valeurs manquantes

18.5% des valeurs du jeu de données sont manquantes.

```
id                    0.000000
hauteur_m             0.000000
circonference_cm      0.000000
geo_point_2d_a        0.000000
id_emplacement        0.000000
lieu                  0.000000
geo_point_2d_b        0.000000
arrondissement        0.000000
type_emplacement      0.000000
domanialite           0.000500
genre                 0.007995
libelle_francais      0.747988
espece                0.875400
remarquable          31.527404
stade_developpement  33.579498
variete              81.624088
complement_addresse  84.559577
numero              100.000000
dtype: float64
```

Analyse exploratoire - Nettoyage

Valeurs manquantes

- Un seul arbre concerné par la valeur manquante pour “domanialite”;
- Tous les arbres du même arrondissement et du même lieu ont pour valeur “Jardin”.

```
arbres.loc[arbres["domanialite"].isnull(),:]
```

	id	type_emplacement	domanialite	arrondissement	complement_addresse	numero	lieu
197239	2020911	Arbre	NaN	PARIS 20E ARRDT	NaN	NaN	JARDINS D IMMEUBLES PORTE DE VINCENNES NORD / ...

Analyse exploratoire - Nettoyage

Valeurs manquantes

Les variables genre, libelle_francais et espece ne présentent pas de corrélations apparentes au niveau des valeurs manquantes

```
arbres.loc[arbres["genre"].isnull(),"genre"]="Non spécifié"
```

```
arbres.loc[arbres["libelle_francais"].isnull(),"libelle_francais"]="Non spécifié"
```

```
arbres.loc[arbres["espece"].isnull(),"espece"]="Non spécifié"
```

Analyse exploratoire - Nettoyage

Valeurs manquantes

- Les variables variete, complement_addresse et numero on plus de 75% de leurs données manquantes.

- Les variables stade_developpement et remarquable seront analysées d'avantage si besoin est.

```
arbres.drop(["variete", "complement_addresse", "numero"], inplace=True, axis=1)
```

```
arbres.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200137 entries, 0 to 200136
Data columns (total 15 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   id                   200137 non-null  int64
 1   type_emplacement     200137 non-null  object
 2   domanialite          200137 non-null  object
 3   arrondissement       200137 non-null  object
 4   lieu                 200137 non-null  object
 5   id_emplacement       200137 non-null  object
 6   libelle_francais     200137 non-null  object
 7   genre                200137 non-null  object
 8   espece               200137 non-null  object
 9   circonference_cm     200137 non-null  int64
 10  hauteur_m            200137 non-null  int64
 11  stade_developpement  132932 non-null  object
 12  remarquable          137039 non-null  float64
 13  geo_point_2d_a       200137 non-null  float64
 14  geo_point_2d_b       200137 non-null  float64
dtypes: float64(3), int64(3), object(9)
memory usage: 22.9+ MB
```

Analyse exploratoire - Nettoyage

Valeurs aberrantes

- L'écart type des variables circonference_cm et hauteur_m est très grand. L' écart entre la moyenne et la valeur maximale le démontre d'avantage.

```
arbres.describe()
```

	id	circonference_cm	hauteur_m	remarquable	geo_point_2d_a	geo_point_2d_b
count	2.001370e+05	200137.000000	200137.000000	137039.000000	200137.000000	200137.000000
mean	3.872027e+05	83.380479	13.110509	0.001343	48.854491	2.348208
std	5.456032e+05	673.190213	1971.217387	0.036618	0.030234	0.051220
min	9.987400e+04	0.000000	0.000000	0.000000	48.742290	2.210241
25%	1.559270e+05	30.000000	5.000000	0.000000	48.835021	2.307530
50%	2.210780e+05	70.000000	8.000000	0.000000	48.854162	2.351095
75%	2.741020e+05	115.000000	12.000000	0.000000	48.876447	2.386838
max	2.024745e+06	250255.000000	881818.000000	1.000000	48.911485	2.469759

Analyse exploratoire - Nettoyage

Doublons

```
arbres.loc[arbres[["domanialite", "arrondissement", "lieu", "genre", "espece", "circonference_cm", "hauteur_m", "geo_po
```

type_emplacement	domanialite	arrondissement	lieu	id_emplacement	libelle_francais	genre	espece	circonference_cm	hauteur_m	st
Arbre	Jardin	PARIS 12E ARRDT	JARDIN PARTAGE BEL-AIR	5	Ailante	Ailanthus	altissima	0	0	
Arbre	Jardin	PARIS 12E ARRDT	JARDIN PARTAGE BEL-AIR	6	Ailante	Ailanthus	altissima	0	0	
Arbre	Alignement	BOIS DE VINCENNES	ROUTE DAUPHINE	402029	Tilleul	Tilia	platyphyllos	0	0	
Arbre	Alignement	BOIS DE VINCENNES	ROUTE DAUPHINE	402030	Tilleul	Tilia	platyphyllos	0	0	

Doublons apparents, mais au final il s'agit très probablement de jeunes arbres très proches.

Analyse exploratoire - Répartitions

```
arbres.nunique()
```

```
id                    200137
type_emplacement           1
domanialite                9
arrondissement            25
lieu                    6921
id_emplacement         69040
libelle_francais         192
genre                    175
espece                   540
circonference_cm         531
hauteur_m                143
stade_developpement        4
remarquable                2
geo_point_2d_a        200107
geo_point_2d_b        200114
dtype: int64
```

```
arbres['domanialite'].value_counts(normalize=True)*100
```

```
Alignement      52.438580
Jardin          23.115666
CIMETIERE       15.952073
DASCO            3.208802
PERIPHERIQUE     2.661677
DJS              1.948665
DFPE             0.662046
DAC              0.010493
DASES            0.001999
Name: domanialite, dtype: float64
```

Analyse exploratoire - Répartitions

```
arbres['arrondissement'].value_counts(normalize=True)*100
```

```
PARIS 15E ARRDT      8.569630
PARIS 13E ARRDT      8.350280
PARIS 16E ARRDT      8.195886
PARIS 20E ARRDT      7.664750
PARIS 19E ARRDT      6.849808
PARIS 12E ARRDT      6.295687
SEINE-SAINT-DENIS    5.781040
BOIS DE VINCENNES    5.751061
PARIS 14E ARRDT      5.695599
PARIS 17E ARRDT      5.377317
PARIS 18E ARRDT      5.002074
PARIS 7E ARRDT       4.305551
VAL-DE-MARNE         3.787406
PARIS 8E ARRDT       3.620020
PARIS 11E ARRDT      2.827063
HAUTS-DE-SEINE       2.647187
BOIS DE BOULOGNE     1.987638
PARIS 10E ARRDT      1.691341
PARIS 4E ARRDT       1.369062
PARIS 5E ARRDT       1.183190
PARIS 6E ARRDT       0.881396
PARIS 1ER ARRDT      0.706016
PARIS 3E ARRDT       0.604086
PARIS 9E ARRDT       0.583101
PARIS 2E ARRDT       0.273812
Name: arrondissement, dtype: float64
```

```
arbres['hauteur_m'].value_counts(normalize=True)*100
```

```
0       19.596077
10      14.306200
5       13.163483
15       8.608103
8        6.809336
           ...
5155     0.000500
218      0.000500
91       0.000500
219      0.000500
255      0.000500
Name: hauteur_m, Length: 143, dtype: float64
```

```
arbres['circonference_cm'].value_counts(normalize=True)*100
```

```
0       12.924647
20       4.852176
70       3.387679
60       3.182320
80       3.100876
           ...
357      0.000500
485      0.000500
1125     0.000500
1205     0.000500
511      0.000500
Name: circonference_cm, Length: 531, dtype: float64
```

Analyse exploratoire - Répartitions

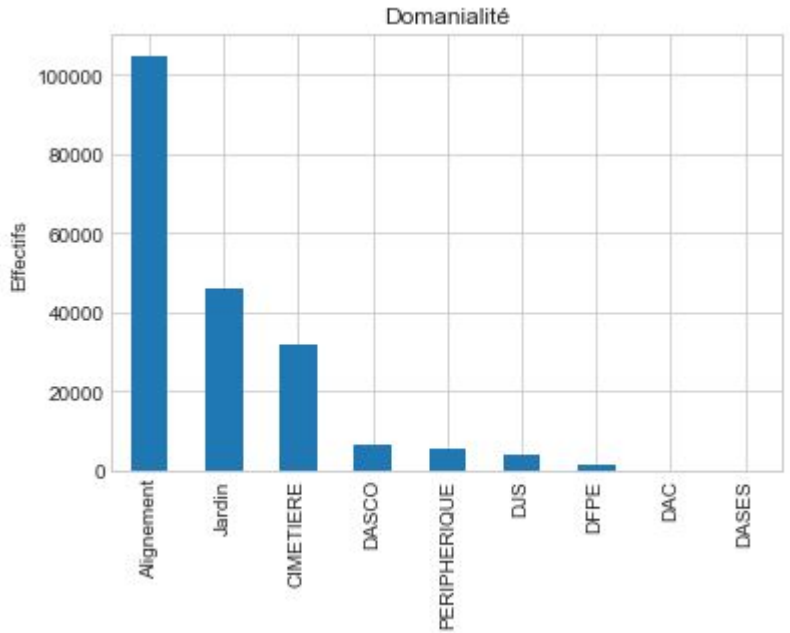
```
arbres['genre'].value_counts(normalize=True)*100
```

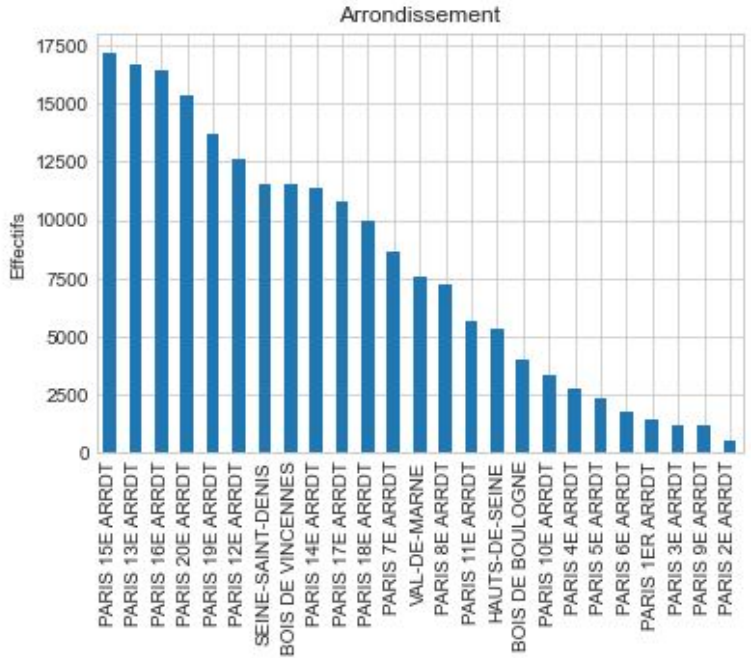
```
Platanus        21.280923
Aesculus        12.661827
Tilia           10.767624
Acer             9.229178
Sophora          5.910951
                   ...
Xanthoceras      0.000500
Ziziphus         0.000500
Sciadopitys      0.000500
Brachychiton     0.000500
Maackia          0.000500
Name: genre, Length: 175, dtype: float64
```

```
arbres['espece'].value_counts(normalize=True)*100
```

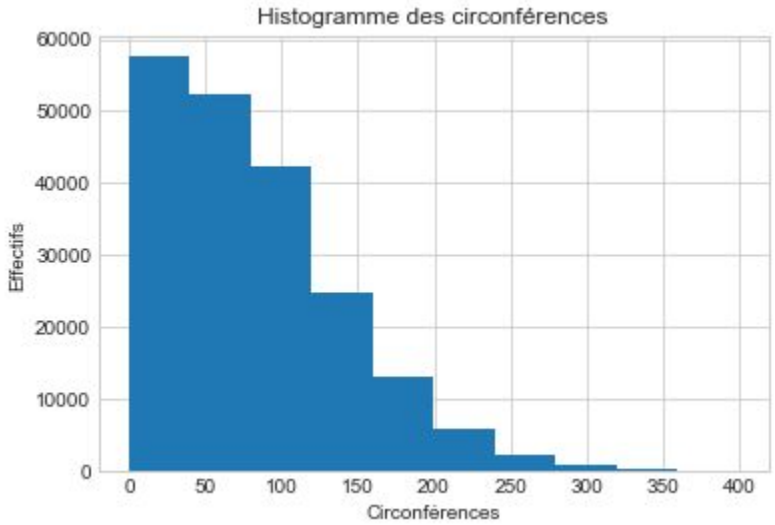
```
x hispanica      18.192038
hippocastanum    10.012641
japonica          5.906954
n. sp.            4.528398
tomentosa         4.477933
                    ...
muehlenbergii     0.000500
camphora          0.000500
verticillata      0.000500
koraiensis        0.000500
pekinensis        0.000500
Name: espece, Length: 540, dtype: float64
```

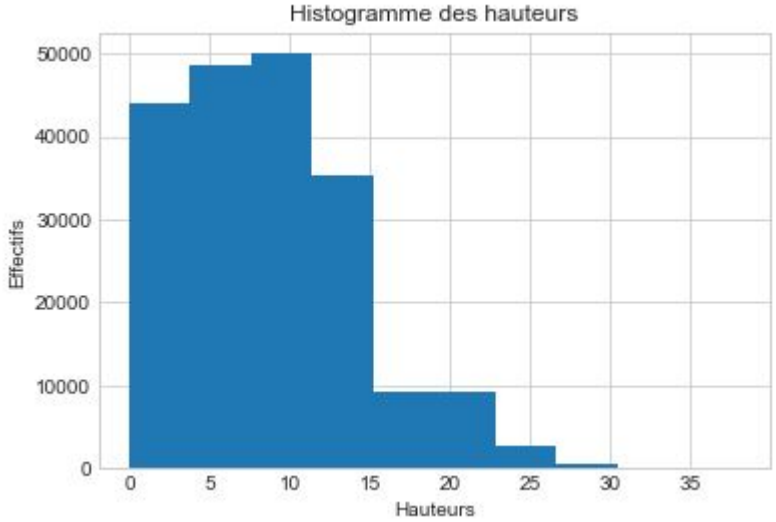
Analyse exploratoire - Distributions





Analyse exploratoire - Distributions





Analyse exploratoire - Dispersion et forme

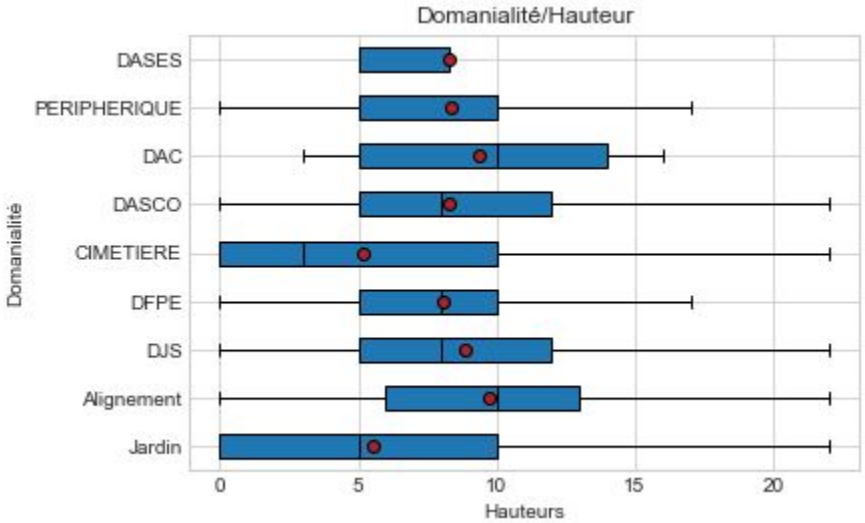
- Taux de dispersion
  - Hauteur: 15035.40%
  - Circonférence: 807.37%
- Asymétrie
  - Hauteur: 447.29
  - Circonférence: 298.16

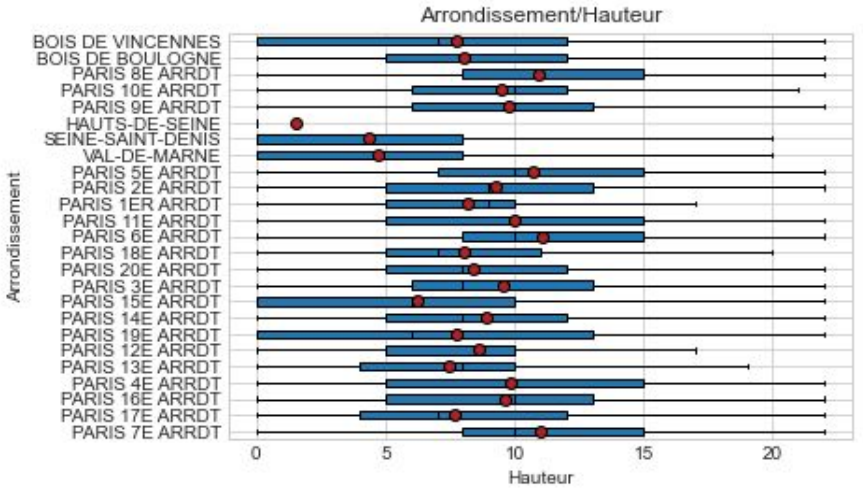
Données très dispersées, non asymétriques et étalées vers la droite.

Analyse exploratoire - Tendances centrales

- Hauteur
  - Moyenne: 7.93
  - Médiane: 8
  - Mode: 0
- Circonférence
  - Moyenne: 74.80
  - Médiane: 70
  - Mode: 0

Analyse exploratoire - Corrélations





Synthèse

Synthèse

- La plupart des arbres: 12e au 20e arrondissement;
- Les arbres les plus courts: les cimetières, les jardins, Haut-De-Seine, Seine-Saint-Denis, 15e, 19e arrondissement pour la plupart;
- Les arbres les plus hauts: DAC, Dasco, DJS, Alignement, 5e, 6e, 7e et le 8e arrondissement;
- Les jeunes arbres ont le plus fort effectif;
- La moitié des arbres: taille inférieure à 8 m et circonférence inférieure à 70 cm;
- Espèces les plus représentées: x hispanica, japonica, tomentosa;
- Genres les plus représentés: Platanus, Aesculus, Tilia, Acer.

Perspectives

Perspectives

- Analyses corrélations Arrondissement, Domanialité, Espèce, Genre, Hauteur;
- Distribution des espèces par arrondissement et domanialités
- Distribution des espèces par arrondissement et tailles