Participez à un concour sur la smartcity

AI Engineer - Projet 2

Ludovic Lafon

December 6, 2023

OpenClassrooms

Exploration des données

Traitement des données manquantes

Détection des outliers

Exploration des données

Nombre de données par colonnes

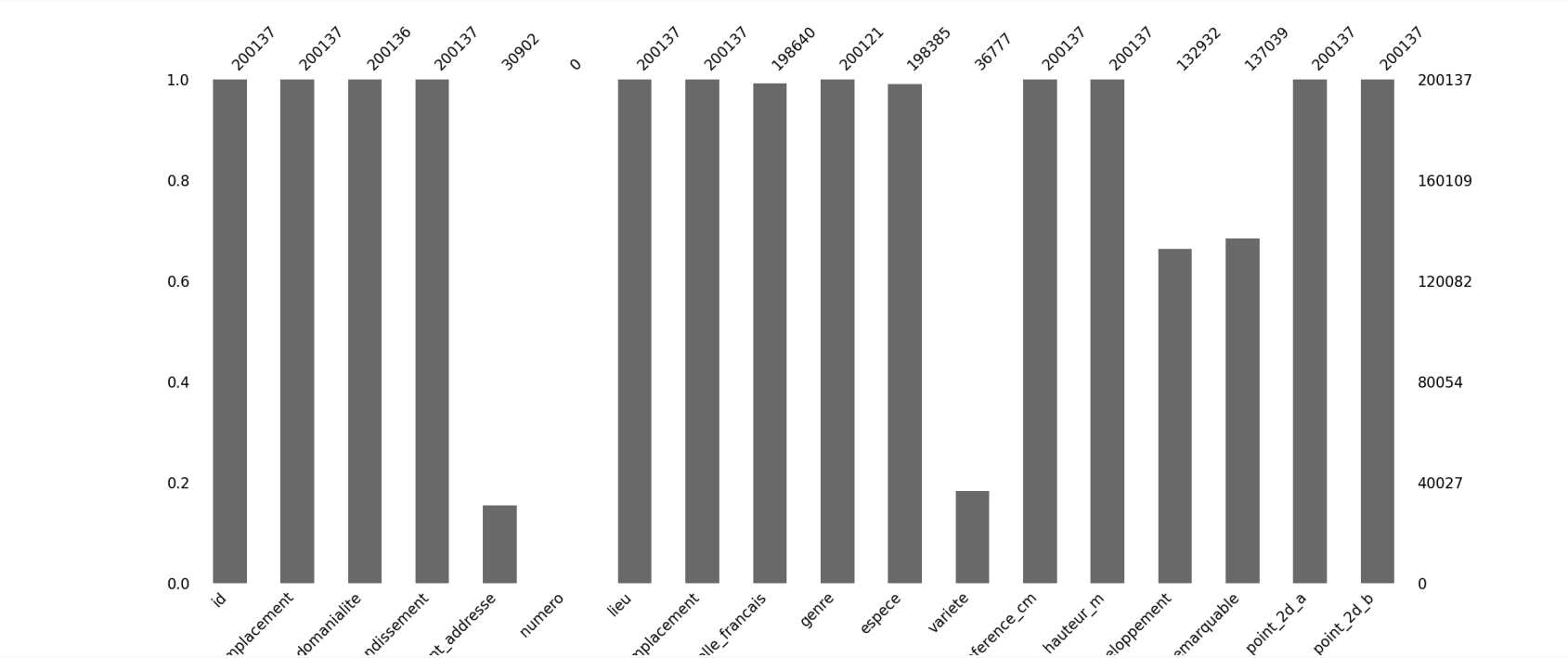

Figure 1: Nombre de données par colonnes

```
df.shape
✓ 0.0s                                                    Python
(200137, 18)
```

Figure 2: pandas.DataFrame.shape() - 200137 lignes, 18 colonnes

Données Manquantes

Données manquantes :

- domanialite
- complement addresse
- numero
- libelle francais
- genre
- espece
- variete
- stade devel-oppement
- remarquable

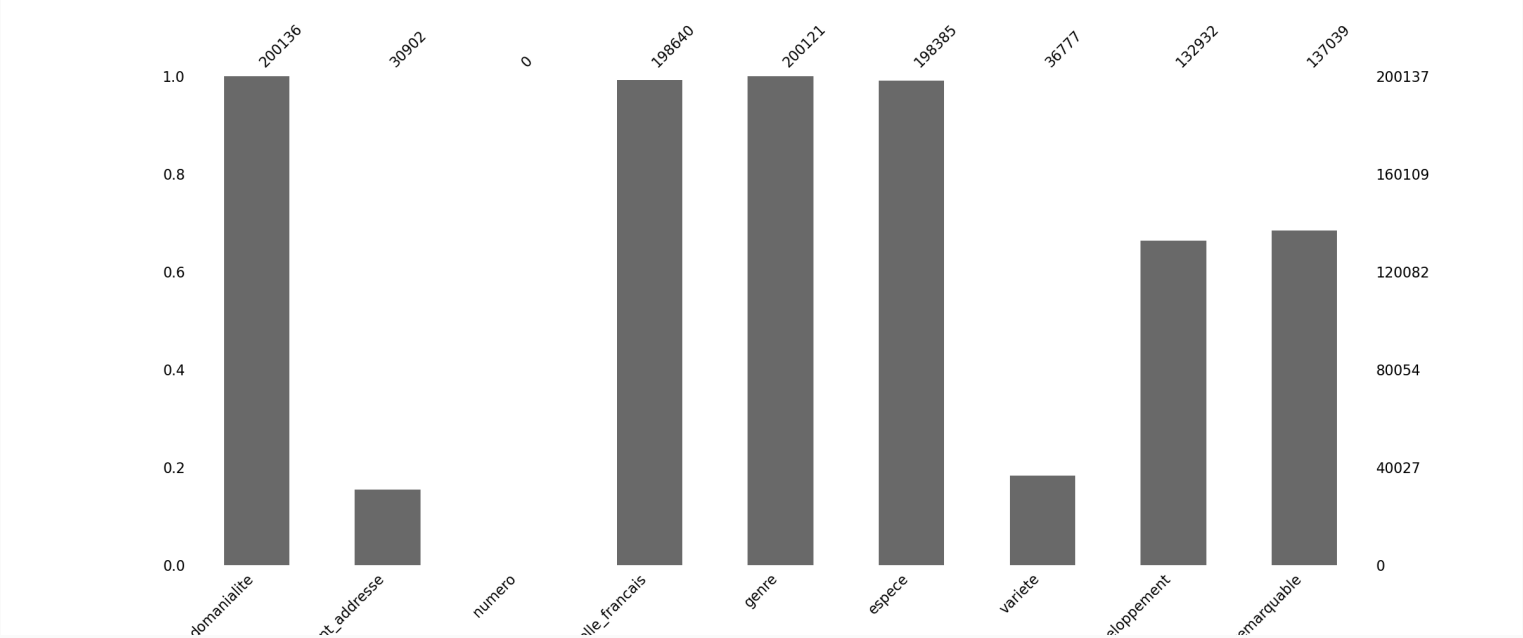

Figure 3: Nombre de données par colonnes pour les colonnes non intégralement remplies

Une colonne est vide !

- numero

Deux colonnes sont presque vide !

- complement addresse
- variete

Données Géographiques

- domanialite
- arrondissement
- complement addresse
- lieu
- geo point 2d a
- geo point 2d b

Données Géographiques

Colonnes gardée :

- domanialite
- arrondissement
- geo point 2d a
- geo point 2d b

Colonnes supprimées :

- complement addresse
- lieu

Traitement des données manquantes

Domanialité

- Une seule donnée manquantes

Domanialité

- Une seule donnée manquantes
- Le lieu de cette donnée est connu

Domanialité

- Une seule donnée manquantes
- Le lieu de cette donnée est connu
- L'ensemble des domanialités correspondant à ce lieu sont des Jardin

Domanialité

- Une seule donnée manquantes
- Le lieu de cette donnée est connu
- L'ensemble des domanialités correspondant à ce lieu sont des Jardin

Solution

- Remplacer la donnée manquante par Jardin

Libellé Français

```
sub_df = sub_df[sub_df['libelle_francais'].isna()]
sub_df
```

	id	libelle_francais	genre	espece
528	100589	NaN	Crataegus	japonica
1416	101521	NaN	Cladrastis	lawsoniana
2727	103209	NaN	Pinus	sylvestris
5282	106682	NaN	Sorbus	aria
5287	106687	NaN	Staphylea	colchica
...	...	...	...	...
200120	2024729	NaN	Magnolia	obovata
200121	2024730	NaN	Magnolia	delavayi subsp. potaninii
200128	2024737	NaN	Parrotia	n. sp.
200129	2024738	NaN	Non spécifié	sinensis
200132	2024741	NaN	Castanea	fargesii

1497 rows × 4 columns

Figure 4: DataFrame des données manquantes pour la colonne libelle francais

Libellé Français

- 1497 données manquantes
- Un libellé français correspond à un couple unique (genre, espece)
- Pour chaque donnée manquante, nous allons essayer de trouver un libellé français correspondant à un couple (genre, espece) parmis les donnée déjà présentes

```
sub_df = sub_df[sub_df['libelle_francais'].isna()]
sub_df
```

	id	libelle_francais	genre	espece
528	100589	NaN	Crataegus	japonica
1416	101521	NaN	Cladrastis	lawsoniana
2727	103209	NaN	Pinus	sylvestris
5282	106682	NaN	Sorbus	aria
5287	106687	NaN	Staphylea	colchica
...	...	...	...	...
200120	2024729	NaN	Magnolia	obovata
200121	2024730	NaN	Magnolia	delavayi subsp. potaninii
200128	2024737	NaN	Parrotia	n. sp.
200129	2024738	NaN	Non spécifié	sinensis
200132	2024741	NaN	Castanea	fargesii

1497 rows × 4 columns

Figure 5: DataFrame des données manquantes pour la colonne libelle francais

Libellé Français

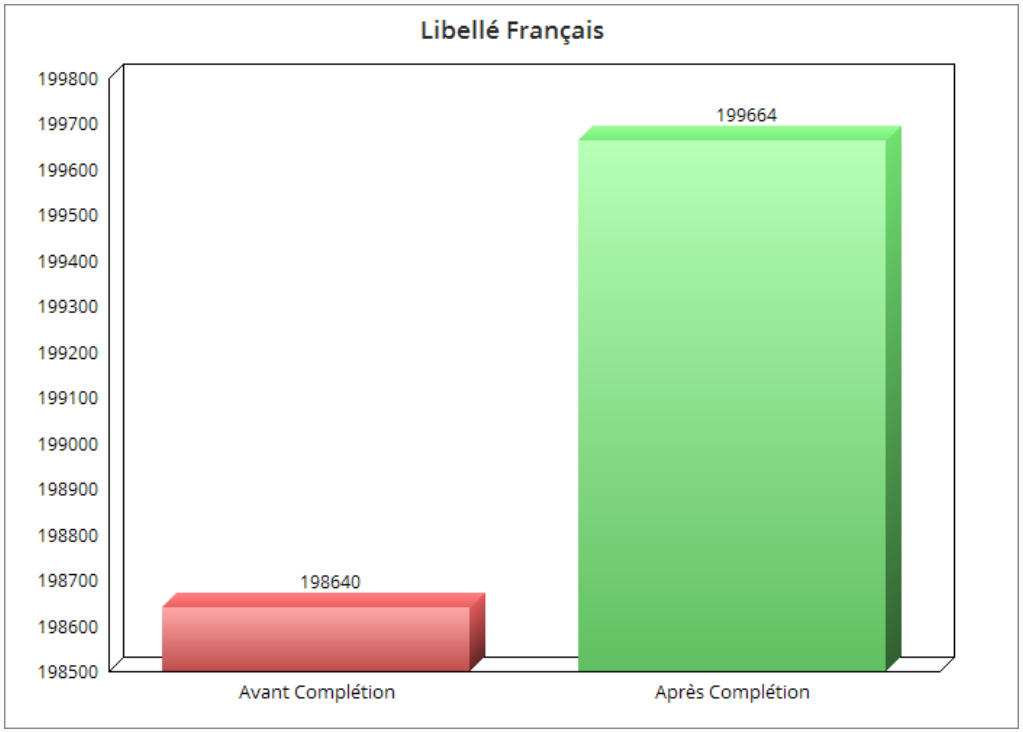

Figure 6: Résultat de la recherche de libellé français correspondant à un couple (genre, espece)

Résultat

- 1024 données manquantes ont pu être retrouvées

Détection des outliers

Données Géographiques

Nous avons les coordonnées GPS de chaque arbre :

- geo point 2d a
- geo point 2d b

Nous allons utiliser l'algorithme KNN pour détecter les outliers.

Données Géographiques - Paramétrage KNN

Données Géographiques - Résultats KNN

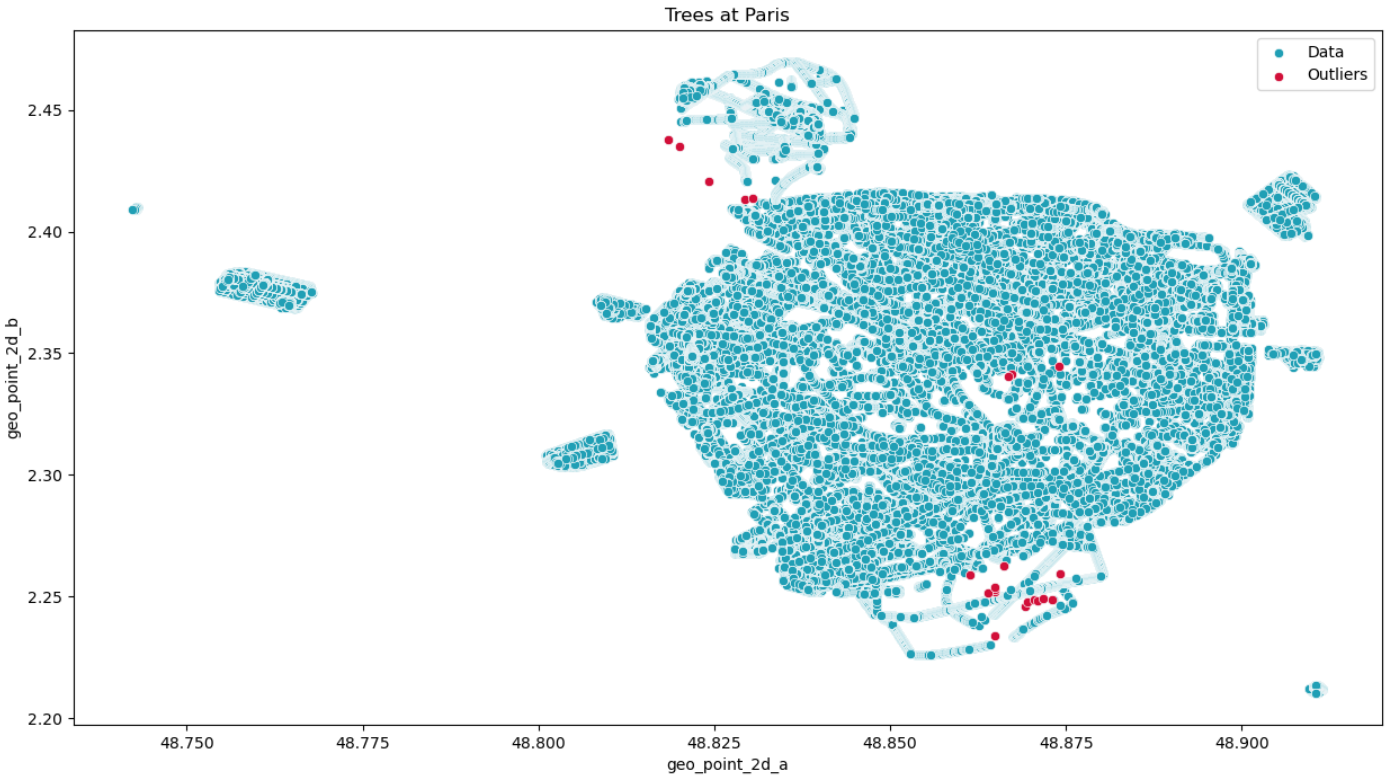

Figure 7: Résultat de la recherche des outliers avec KNN

Données numériques de l’arbre

Pour chaque arbre, nous avons les données suivantes :

- hauteur
- circonference

Nous allons utiliser la méthode des interquartiles pour détecter d’éventuels outliers.

Données numériques de l’arbre - IQR

Situation d’oringine :

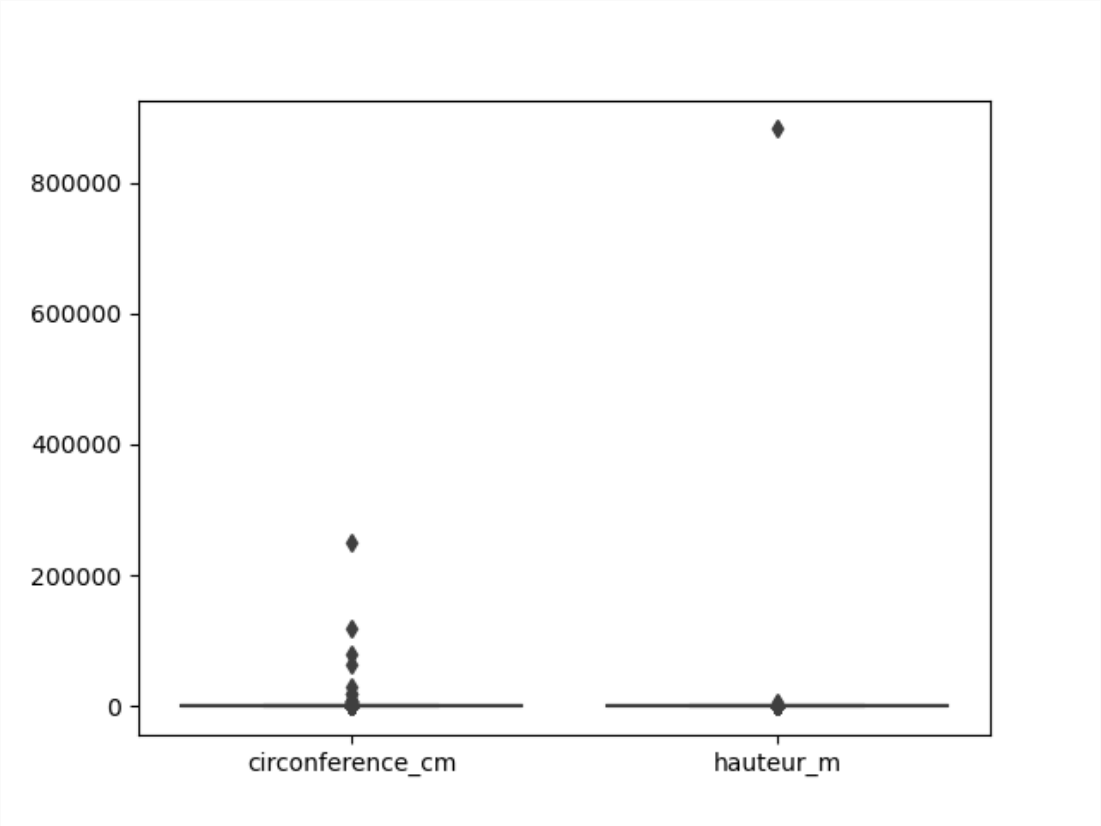

Figure 8: Boxplot des variables circonference et hauteur

Données numériques de l'arbre - IQR

- $IQR = Q3 - Q1$
- $Limite\ basse = Q1 - 1.5 \times IQR$
- $Limite\ haute = Q3 + 1.5 \times IQR$

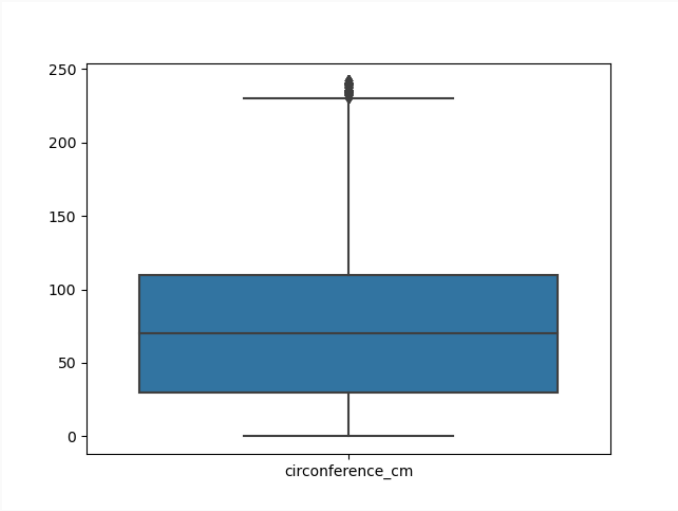

Figure 9: Boxplot de la variable circonference après application de la méthode des IQR

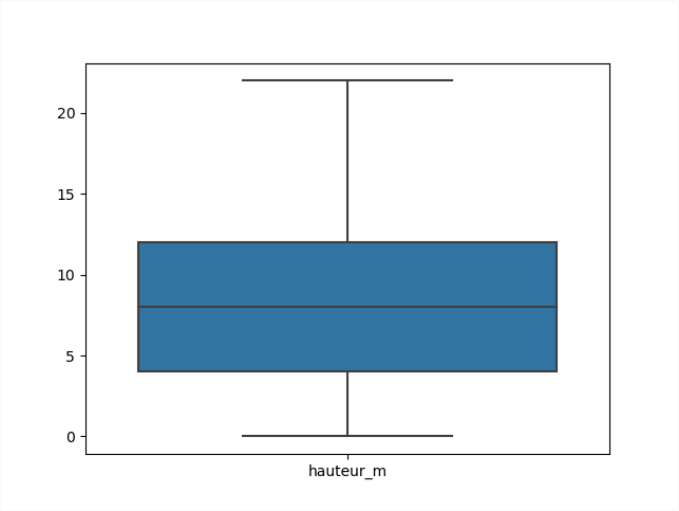

Figure 10: Boxplot de la variable hauteur après application de la méthode des IQR