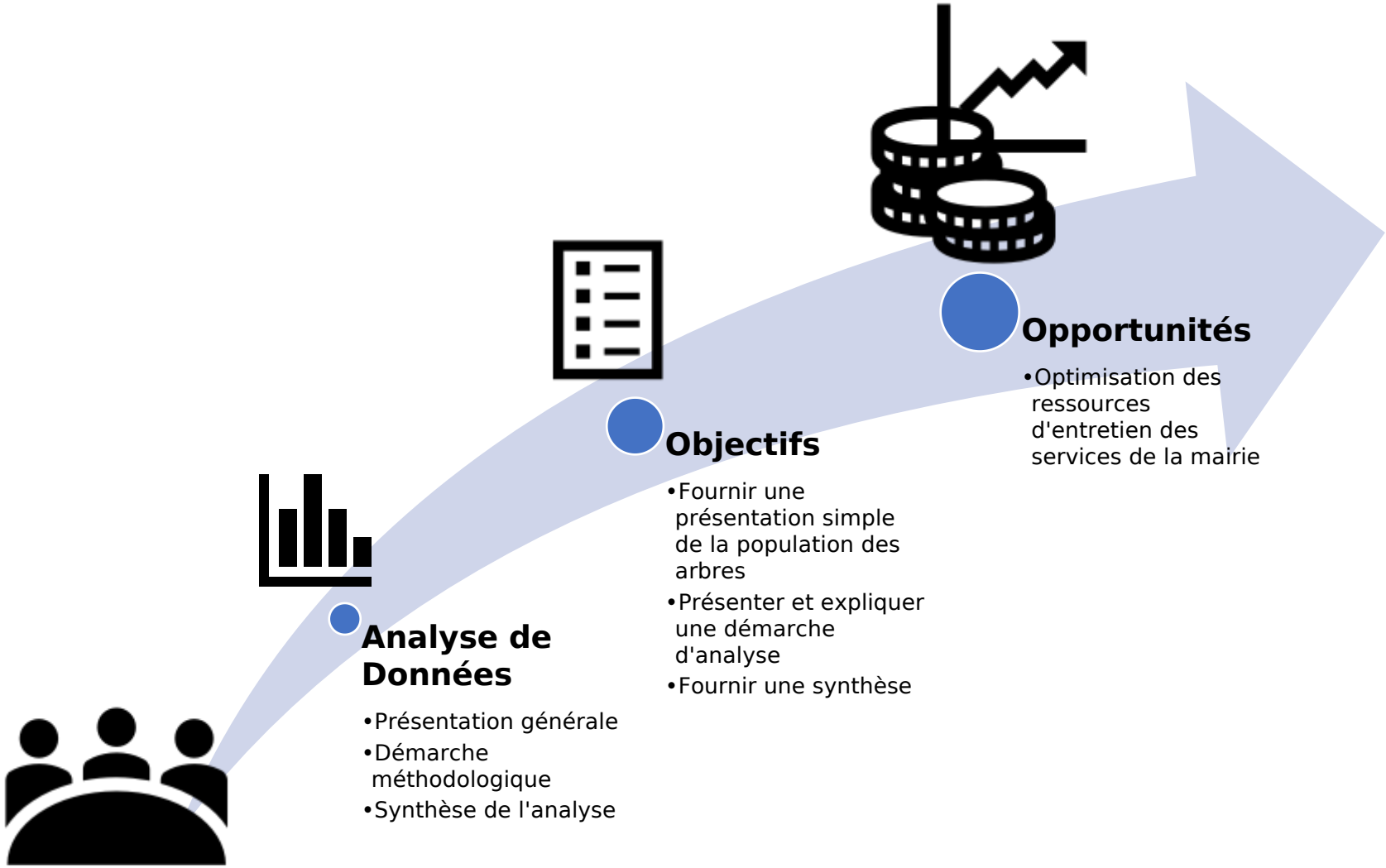
Projet
Végétalisons la ville

Analyse des données
sur les arbres de la ville
de Paris

Bailly DIOUNOU –
11/08/2020

Contexte & périmètre du projet



ONG "Data for Good" & Ville de Paris

Présentation générale du jeu de données

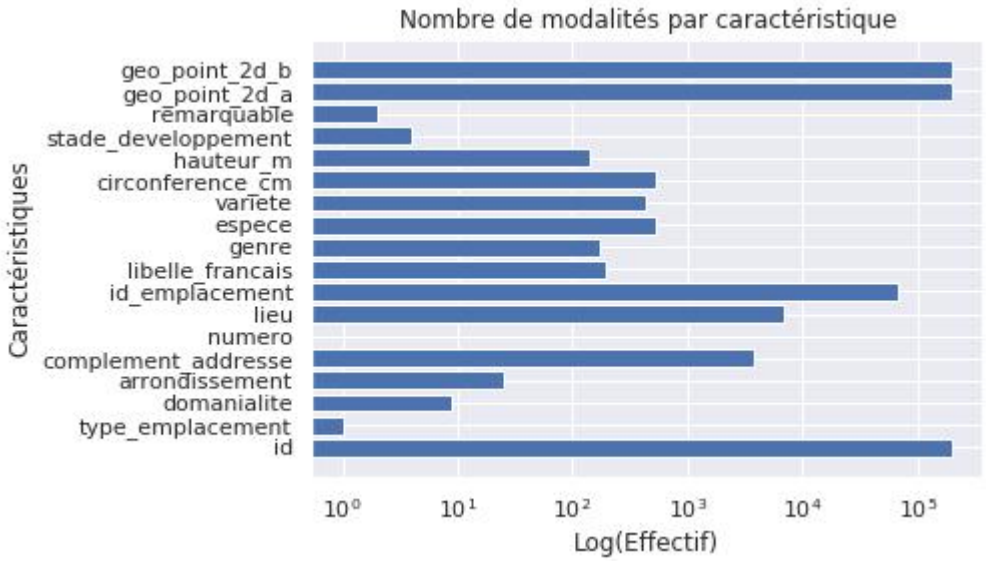
Données statistiques de base - Données brutes

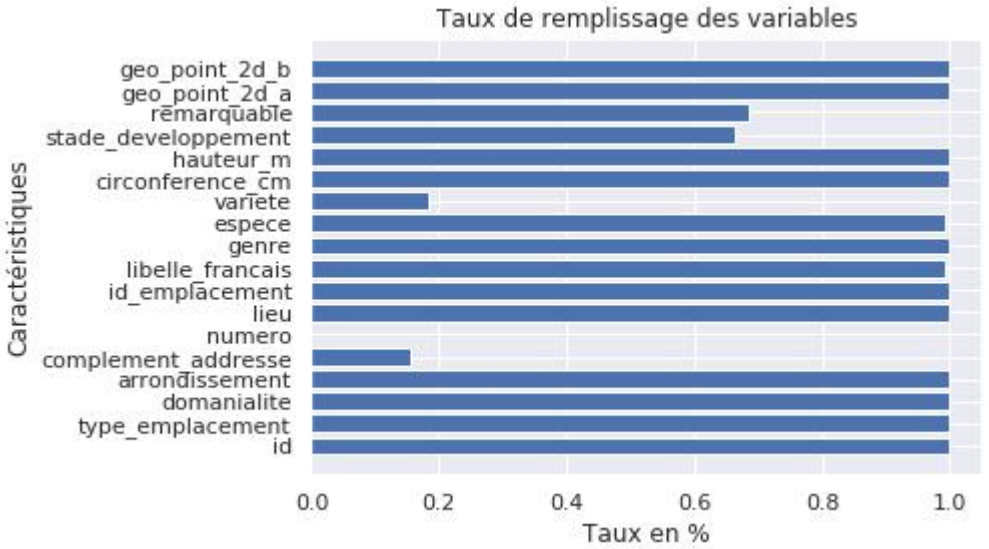
- Taille de la population **N = 200 137** individus
- Nombre de caractéristiques **18 + 1**
- Caractéristiques quantitatives **q = 3** (circ_cm, hauteur_m, remarquable toutes *continues*)
- Caractéristiques qualitatives **p = 16** - dont 4 *ordinales*.

Présentation générale du jeu de données

Indicateurs statistiques de base des variables numériques (quantitatives et qualitatives)							
	id	numero	circonference_cm	hauteur_m	remarquable	geo_point_2d_a	geo_point_2d_b
Count	200137	0	200137	200137	137039	200137	200137
Mean	387202,6823		83,38047937	13,1105093	0,001342683	48,85449079	2,348207896
Std	545603,2424		673,190213	1971,217387	0,036618171	0,030234145	0,05121964
Min	99874		0	0	0	48,7422903	2,210241229
25%	155927		30	5	0	48,83502074	2,307530321
50%	221078		70	8	0	48,85416209	2,351095166
75%	274102		115	12	0	48,87644714	2,386838072
Max	2024745		250255	881818	1	48,91148476	2,469759472

Présentation générale du jeu de données





Démarche méthodologique d'analyse de données

Nettoyage

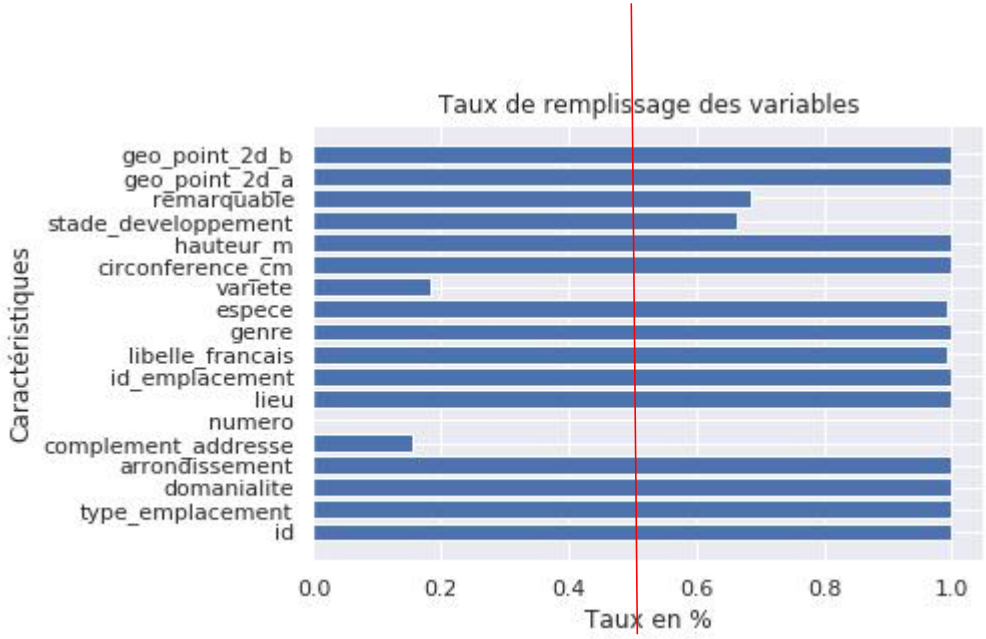
Imputation

Analyse Univariée

Démarche méthodologique

1. *Nettoyage*

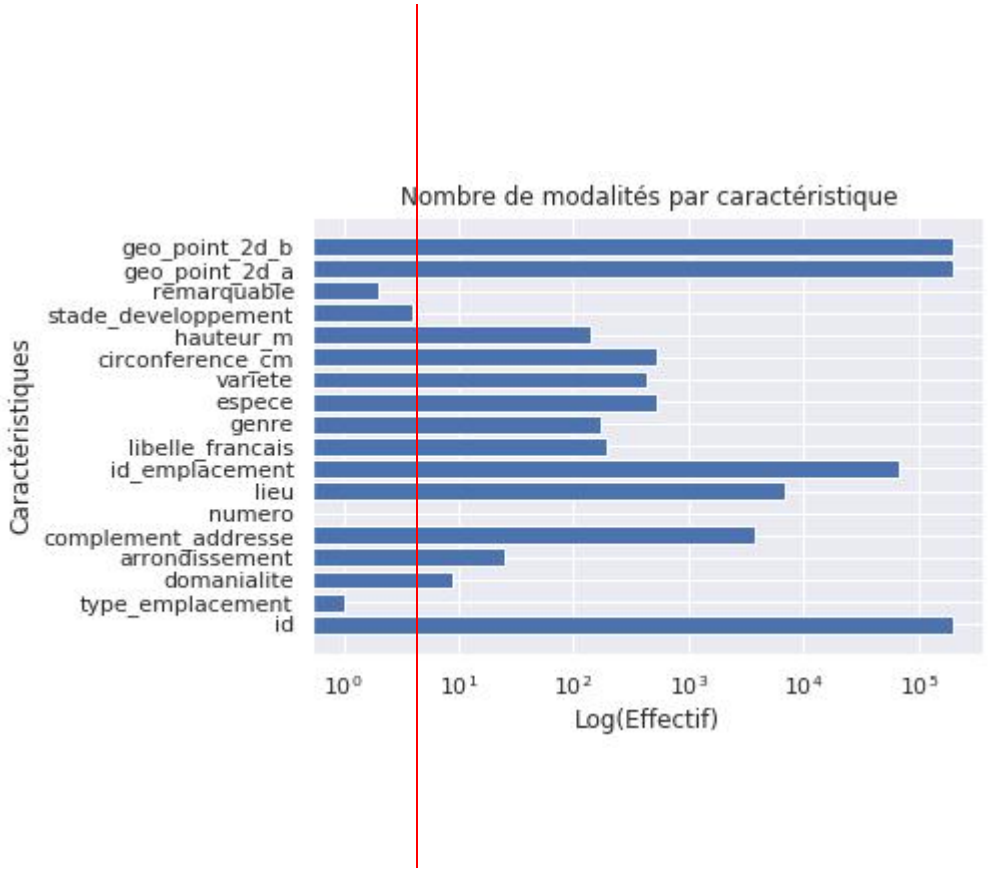
- **70% de plancher pour le taux de remplissage**
- Pertinence évaluée en fonction de qualité et nombre de modalités
- Filtre par identification des valeurs abbérhantes (quantiles)



Démarche méthodologique

1. *Nettoyage*

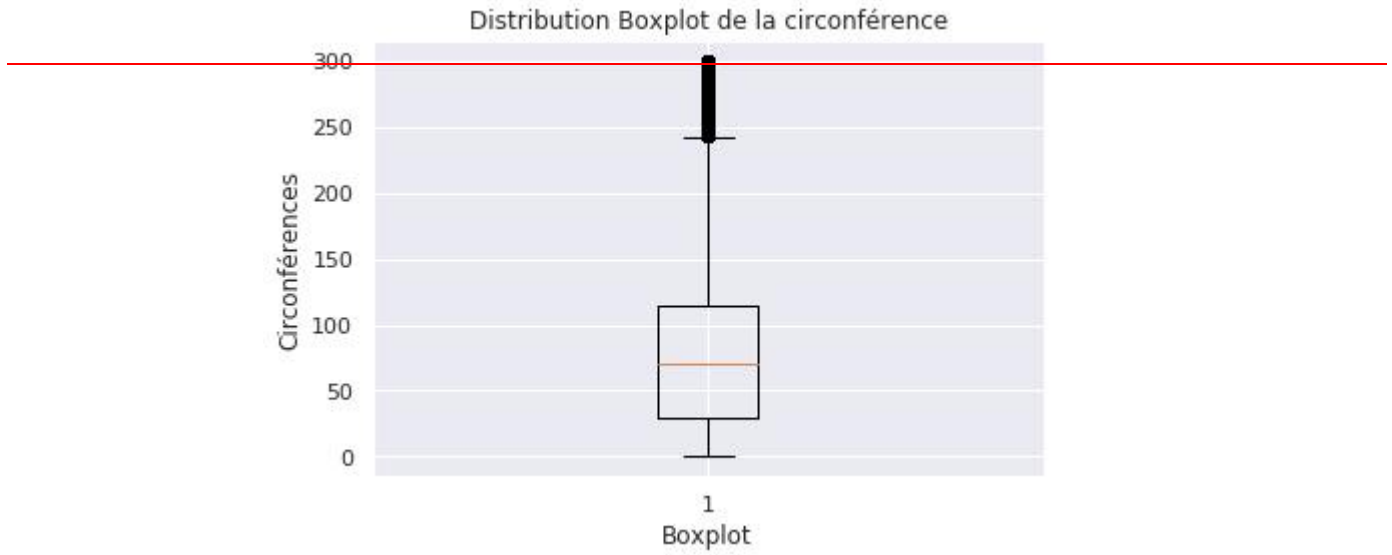
- 70% de plancher pour le taux de remplissage
- **Pertinence évaluée en fonction de qualité et nombre de modalités**
- Filtre par identification des valeurs abbérhantes (quantiles)



Démarche méthodologique

1. *Nettoyage*

- 70% de plancher pour le taux de remplissage
- Pertinence évaluée en fonction de qualité et nombre de modalités
- **Filtre par identification des valeurs abbérhantes (quantiles)**



Démarche méthodologique

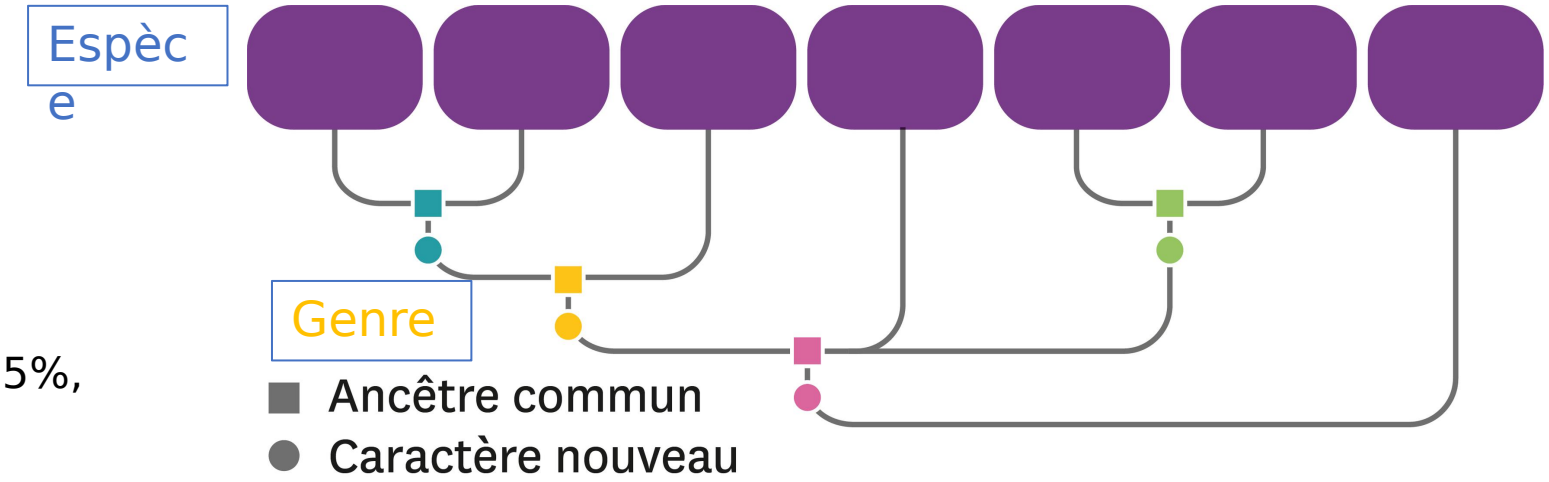
2. *Imputation*

- **Critère dendrologique**
- Imputation basée sur un apprentissage interne vs Imputation simple/itérative
- Gain -> libellé 0.69% / 0.75%, espèce: 0.31 / 0.88

Démarche méthodologique

2. *Imputation*

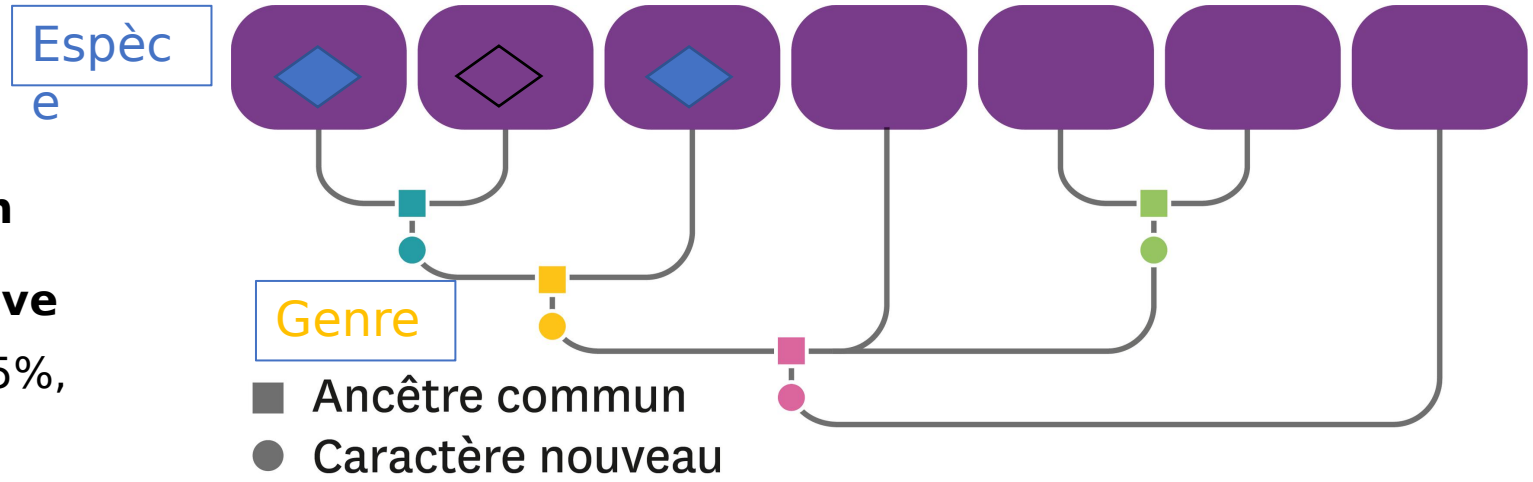
- **Critère dendrologique**
- Imputation basée sur un apprentissage interne vs Imputation simple/itérative
- Gain -> libellé 0.69% / 0.75%, espèce: 0.31 / 0.88



Démarche méthodologique

2. *Imputation*

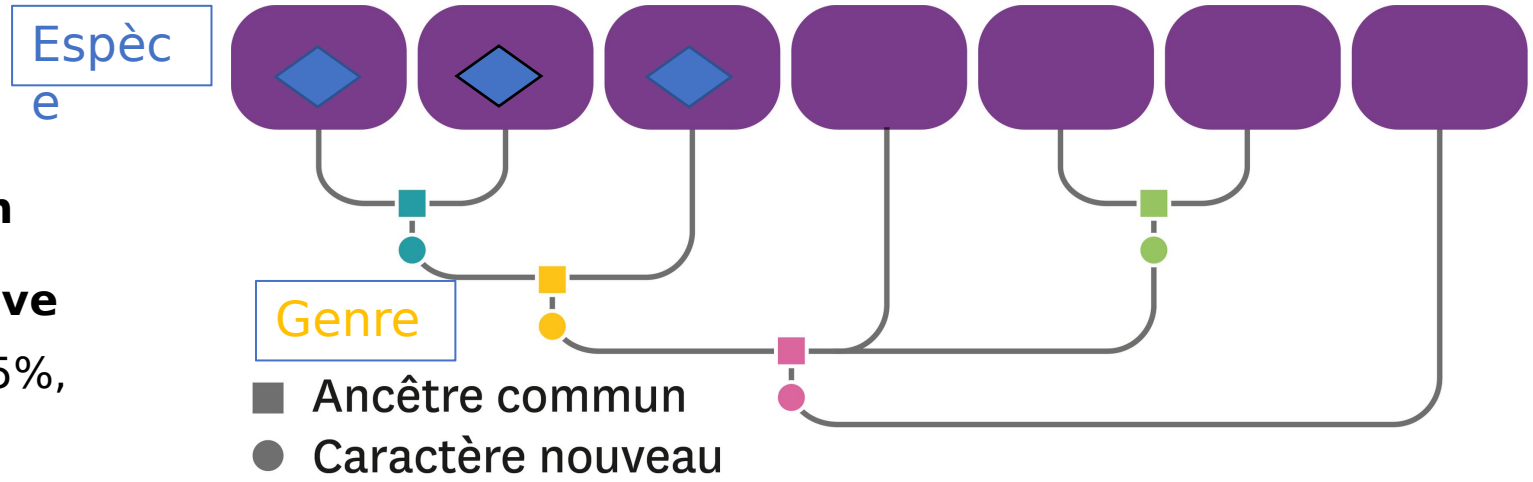
- Critère dendrologique
- **Imputation basée sur un apprentissage interne vs Imputation simple/itérative**
- Gain -> libellé 0.69% / 0.75%, espèce: 0.31 / 0.88



Démarche méthodologique

2. *Imputation*

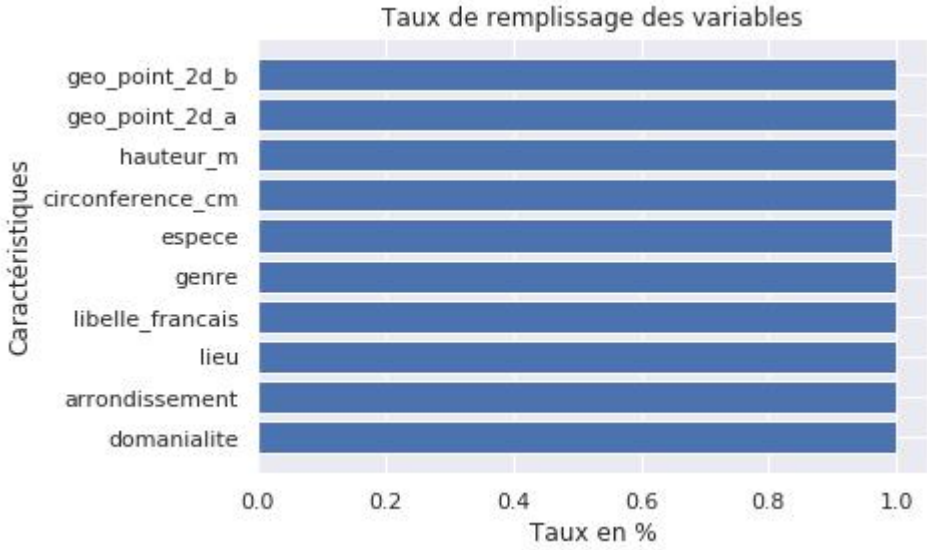
- Critère dendrologique
- **Imputation basée sur un apprentissage interne vs Imputation simple/itérative**
- Gain -> libellé 0.69% / 0.75%, espèce: 0.31 / 0.88



Démarche méthodologique

2. *Imputation*

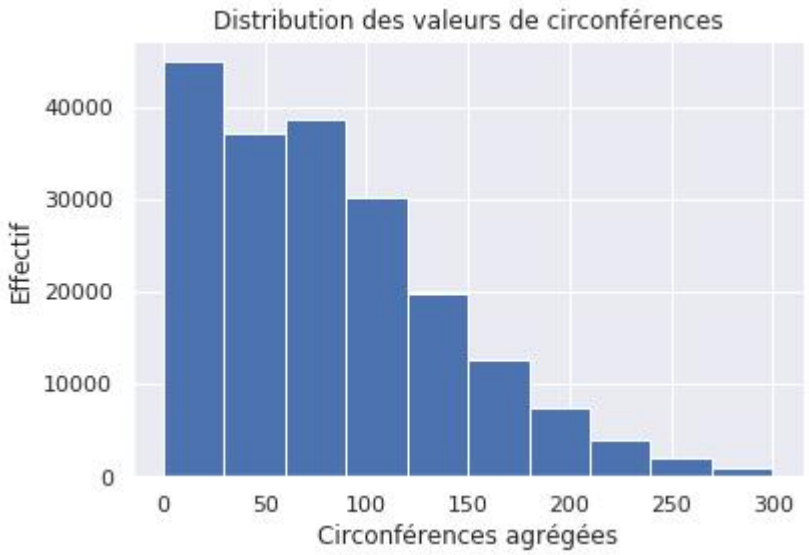
- Critère dendrologique
- Imputation basée sur un apprentissage interne vs Imputation simple/itérative
- **Gain**
  - > libellé: 0.69% / 0.75% = **92%**
  - > espèce: 0.31 / 0.88 = **35%**



Démarche méthodologique

3. *Analyse Univariée*

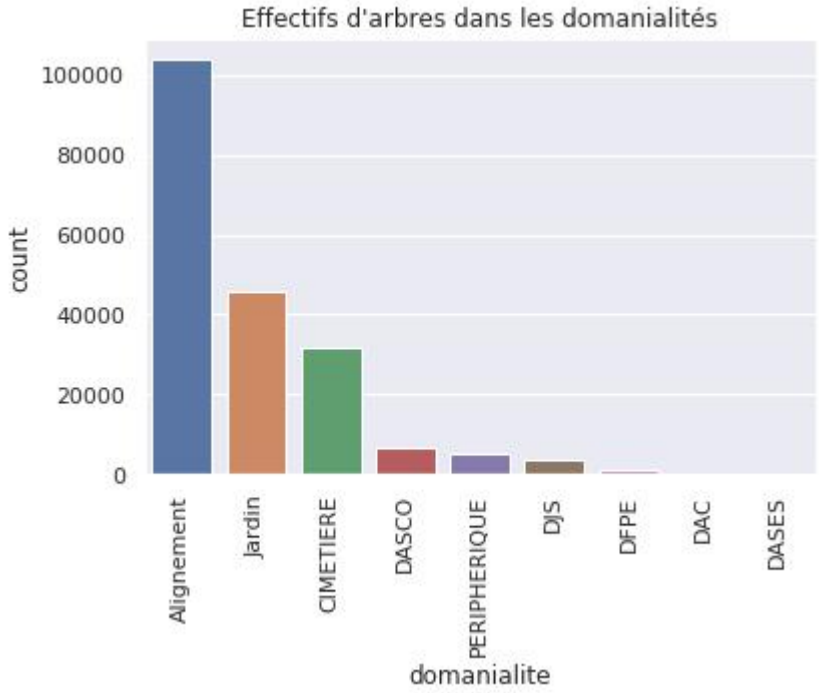
- **Circonférence**
  - Forme quasi gaussienne
- Domanialité
  - Trimodale



Démarche méthodologique

3. *Analyse Univariée*

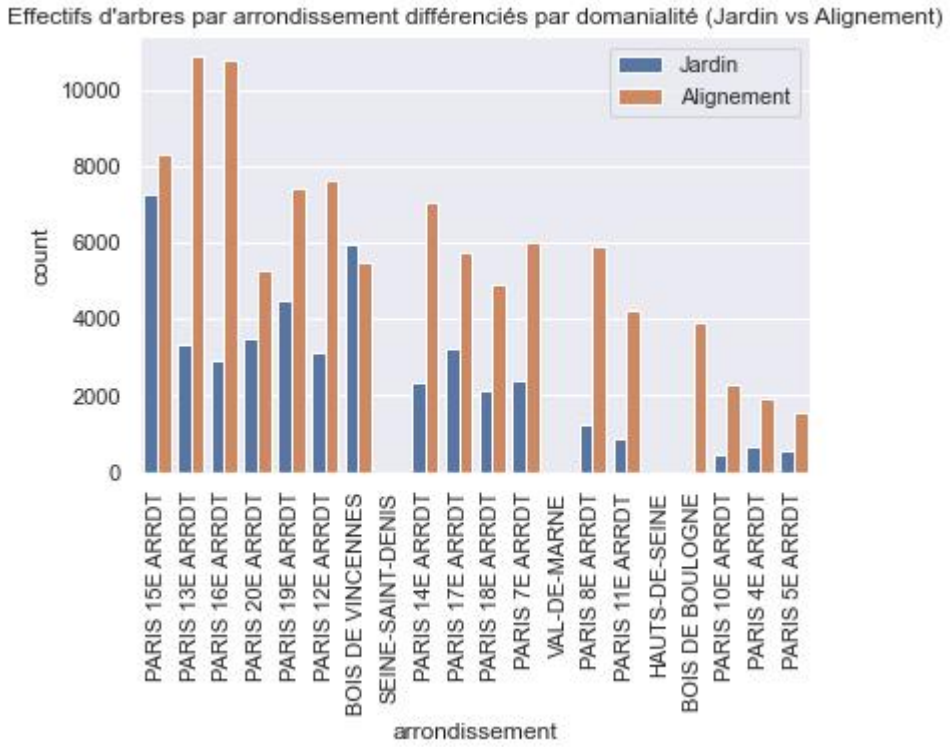
- Circonférence
  - Forme quasi gaussienne
- **Domanialité**
  - Trimodale



Démarche méthodologique

4. *Analyse bivariée & Stratégie*

Domanialité - Arrondissement

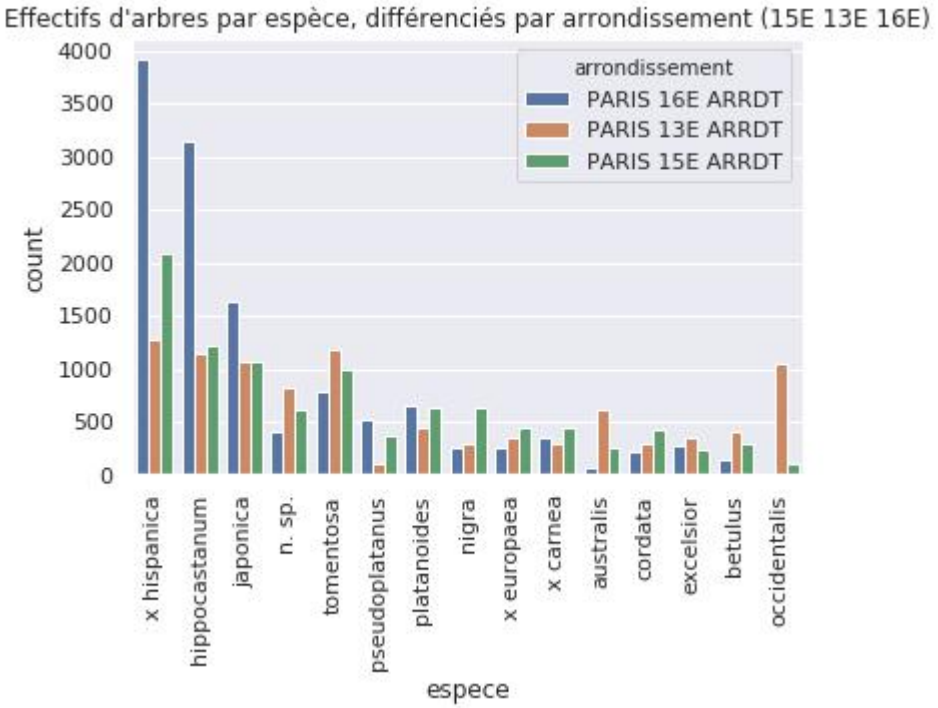


Densité d'arbres par arrondissement

Démarche méthodologique

4. *Analyse bivariée & Stratégie*

Arrondissement - Espèce



Densité d'arbres par arrondissement

Synthèse de l'analyse de données

Jeu de données des arbres de la ville de Paris

N = 200 137 individus, 19 caractéristiques (3 quantitatives)

Démarche méthodologique d'analyse

(1) Nettoyage

- Filtre "passe-haut" à taux de remplissage de 70%,
- Filtre de non-pertinence: critère du nombre et de la qualité des modalités
- Filtre sur valeurs aberrantes de dimensions : multiples des quantiles de leurs distributions

(2) Imputation

- Critère dendrologique sur relations genre-espèce et genre-libellé français. Valeurs statistique et dendrologique
- Gains sur valeurs manquantes: 92% sur libellé français et 35% sur espèce

⟹ **Bilan: Suppression: 13% de lignes, ~50% des colonnes**

(3) Analyse univariée et bivariée

- **Analyse univariée** sur la grande majorité des caractéristiques post-nettoyage.
  - Tracé des distributions,
  - Calcul des indicateurs statistique de base (moyenne, médiane, écart-type).
- **Analyse bivariée**
  - Croisement de la domanialité (2 plus fortes modalités: Jardin et Alignement), avec Arrondissement, Libellé français, Espèce et Circonférence.
  - Tracé des distributions différenciées

Merci pour votre attention

Temps de questions/réponses