Projet **Végétalisons la ville**

Analyse des données sur les arbres de la ville de Paris

Bailly DIOUNOU - 11/08/2020

Contexte & périmètre du projet



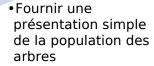


 Optimisation des ressources d'entretien des services de la mairie



Analyse de Données

- Présentation générale
- •Démarche méthodologique
- •Synthèse de l'analyse



Objectifs

- Présenter et expliquer une démarche d'analyse
- •Fournir une synthèse



ONG "Data for Good" & Ville de

Paric

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.

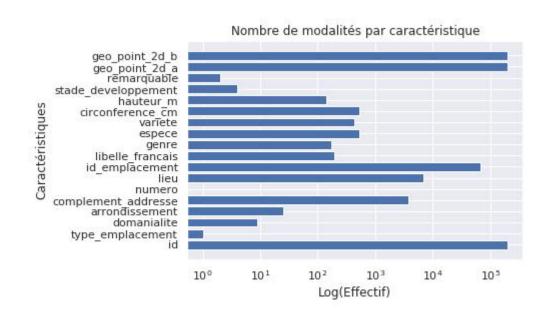
Présentation générale du jeu de données

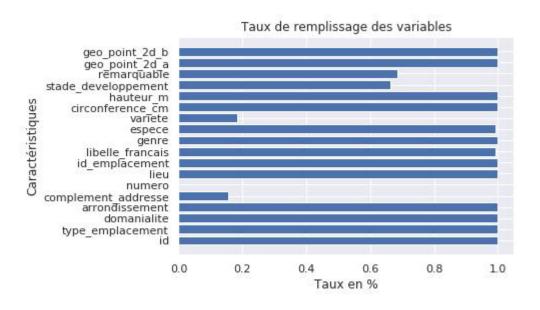
Indicateurs statistiques de base des variables numériques (quantitatives et qualitatives)

remeralish are point are point

		id	numero	circonferen	hauteur_m	remarquab		
		IG	Harriero	ce_cm	maaccar_m	le	2d_a	2d_b
	Count	200137	0	200137	200137	137039	200137	200137
	Mean	387202,68		83,380479	13,110509	0,0013426	48,854490	2,3482078
	Mean	23		37	3	83	79	96
	Std	545603,24		673,19021	1971,2173	0,0366181	0,0302341	0,0512196
	Stu	24		3	87	71	45	4
	Min	99874		0	0	0	48,742290	2,2102412
	141111	99074		U	U	U	3	29
	25%	155927		30	5	0	48,835020	2,3075303
	23/0	133927		50	J	O	74	21
!	50%	221078		70	8	0	48,854162	2,3510951
							09	66
	75%	274102		115	12	0	48,876447	2,3868380
	13/0	2/4102		113	12	U	14	72
	Max	2024745		250255	881818	1	48,911484	2,4697594
	MUA	<u> </u>		<u> </u>	001010		7.0	70

Présentation générale du jeu de données



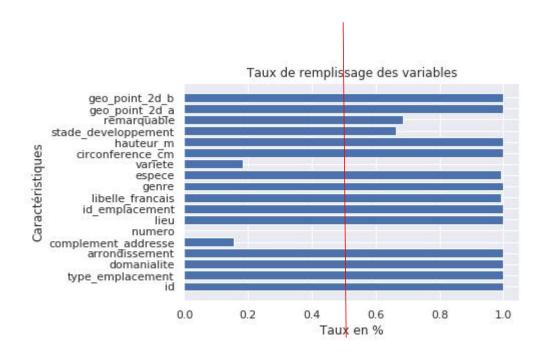


Démarche méthodologique d'analyse de données

Nettoyage Imputation Analyse Univariée

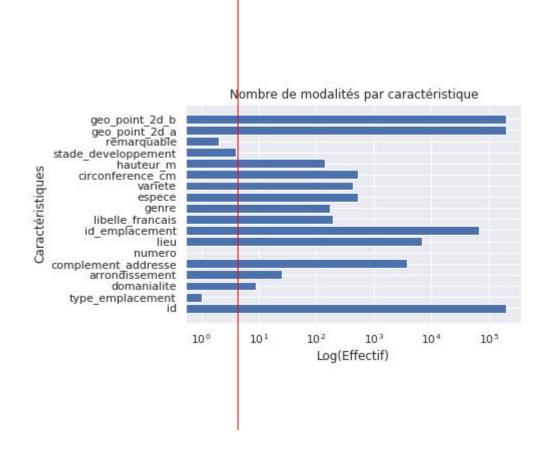
1. Nettoyage

- •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)



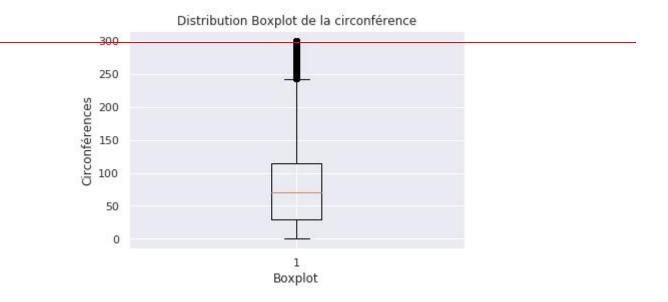
1. Nettoyage

- •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)



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)



2. Imputation

·Critère dendrologique

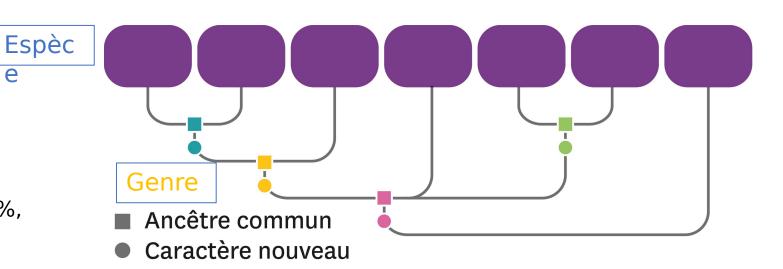
- Imputation basée sur un apprentissage interne vs Imputation simple/itérative
- •Gain -> libellé 0.69% / 0.75%, espèce: 0.31 / 0.88

```
Domaine Ex.: Bacteria
Règne Procaryotae
Phylum
Classe Schizomyctes
Ordre Micrococcales
Famille Micrococcaceae
Genre Staphylococcus
Espèce S. aureus
```

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

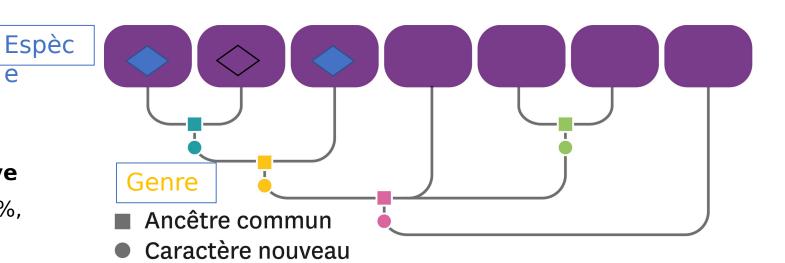


2. Imputation

Critère dendrologique

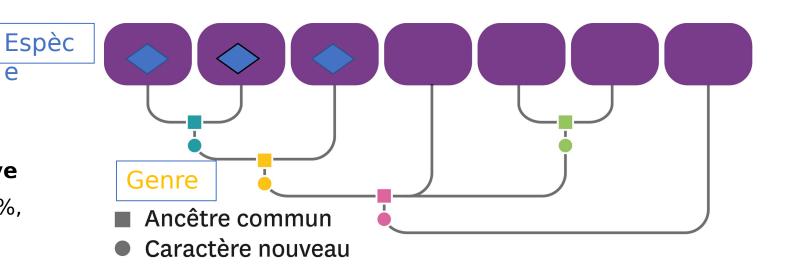
•Imputation basée sur un apprentissage interne vs Imputation simple/itérative

•Gain -> libellé 0.69% / 0.75%, espèce: 0.31 / 0.88



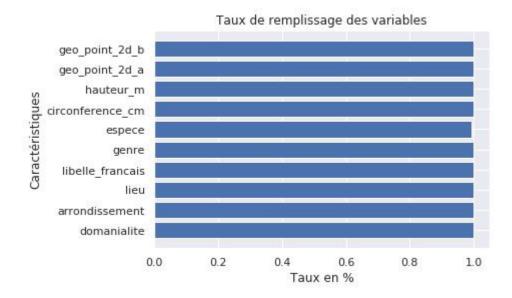
2. Imputation

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



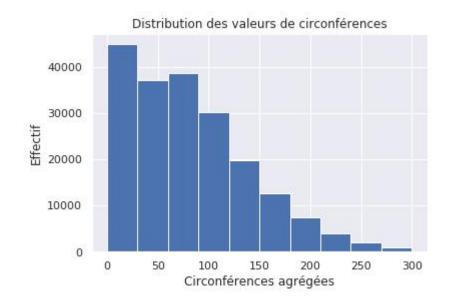
2. Imputation

- •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%



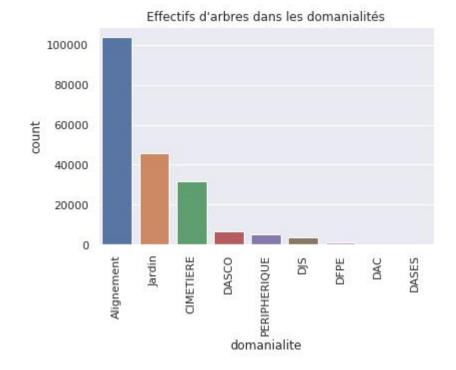
Démarche méthodologique 3. *Analyse Univariée*

- Circonférence
 - •Forme quasi gaussienne
- Domanialité
 - Trimodale



Démarche méthodologique 3. *Analyse Univariée*

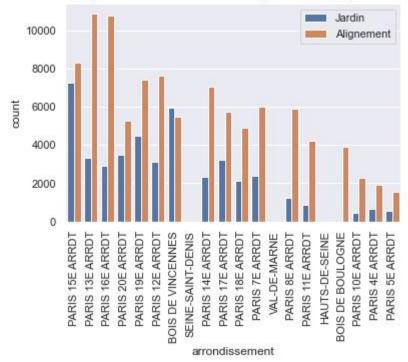
- Circonférence
 - •Forme quasi gaussienne
- Domanialité
 - Trimodale



4. Analyse bivariée & Stratégie

Domanialité - Arrondissement

Effectifs d'arbres par arrondissement différenciés par domanialité (Jardin vs Alignement)



Densité d'arbres par arrondissement



4. Analyse bivariée & Stratégie

Arrondissement - Espèce

Effectifs d'arbres par espèce, différenciés par arrondissement (15E 13E 16E) arrondissement ARIS 16E ARRDT 3500 ARIS 13E ARRD 3000 PARIS 15E ARRDT 2500 count 2000 1500 1000 500 australis betulus x hispanica hippocastanum aponica udoplatanus nigra cordata occidentalis platanoides enropaea espece

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)

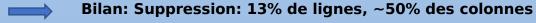
Démarche méthodologique d'analyse

(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



(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