

Concours Smart-City

Data is for Good



Emmanuel Letremble

Objectif du projet

Data is for Good

**L'optimisation des tournées pour
l'entretien des arbres de la ville.**

- L'arrosage
- Le suivi et diagnostic
- La taille (ou élagage)
- L'abattage
- Le plantage et replantage



Présentation du jeu de données

Data is for Good



Il s'agit d'un jeu de données issu d'opendata.paris qui liste **200 137** arbres relevant de la ville de Paris.

Ce jeu de données inclut les arbres :

- d'alignements (dans les rues)
- des espaces verts
- des équipements municipaux

Mais il n'inclut pas :

- les bosquets
- les ensembles forestiers
- les arbres du domaine privé

Présentation du jeu de données

Data is for Good



des méta-données

- id
- id_emplacement
- type_emplacement

des variables géographiques

- domanialite
- arrondissement
- complement_adresse
- numero
- lieu
- geo_point_2d_a
- geo_point_2d_b

Présentation du jeu de données

Data is for Good



des variables numériques

- **circonference_cm**
- **hauteur_m**

des variables nominales

- **libelle_francais**
- **genre**
- **espece**
- **variete**
- **stade_developpement**

une variable booléenne

- **remarquable**

Analyse de données

Data is for Good



Une **analyse univariée** des 18 colonnes, qui suggère :

- **la suppression de 6 colonnes**

(id, numero, type_emplacement, complement_adresse, id_emplacement, variete)

- **l'imputation de 5 colonnes**

(libelle_francais, genre, espece, stade_developpement, remarquable)

- **la suppression des valeurs aberrantes de 2 colonnes**

(circonference_cm, hauteur_m)

- **le formatage / découpage des valeurs de 2 colonnes**

(lieu + complement_adresse)

Analyse de données

Data is for Good

Une analyse multivariée orientée par des objectifs, qui suggère :

- **une corrélation importante entre certaines colonnes**

(la circonférence & la hauteur des arbres entre elles ou avec le stade de développement, le lieu & la domanialité, les variables décrivant le types d'arbres entre elles ...)

qui permettrait des imputations ou la mise en place de modèles prédictifs pour les tournées

- **une corrélation modérée entre certaines colonnes**

(les variables définissant le type d'arbre et la domanialité ou le lieu ...)

sur lesquelles on pourrait s'appuyer mais avec beaucoup de prudence.

- **une absence de corrélation entre certaines colonnes**

(visiblement les especes / genres / libelle_francais ne sont pas ou peu liées aux diverses valeurs géographiques)



Préparation des données

Data is for Good



Une **nettoyage** du jeu de données :

- **suppression des colonnes inutiles**
(id, numero, type_emplacement, complement_adresse, id_emplacement, variete)
- **suppression des valeurs aberrantes**
(circonference_cm, hauteur_m)
- **suppression des doublons**
(en général et sur la base des coordonnées gps)
- **l'imputation des valeurs**
(libelle_francais, genre, espece, stade_developpement, remarquable)
- **renommage des colonnes GPS**
(geo_point_2d_a et geo_point_2d_b deviennent latitude et longitude)

2+ jeux de données exploitables :

- **data_clean**
- **data_outliers**



Synthèse de l'analyse de données

Data is for Good



En analysant le jeu de donnée,
nous avons pu mettre en évidence
les **défauts** de certaines colonnes
(pas de valeurs ajoutée, valeurs aberrantes ou manquantes ...)
ou encore les **liaisons** plus ou moins
fortes qui existent entre celles-ci

C'est un **processus long**, mais
important qui nous permet de
préparer le jeu de données et
d'**identifier les variables clés** afin
de pouvoir produire des outils qui
puisse répondre à nos besoins.

Synthèse de l'analyse de données

Data is for Good



Nous avons utilisé les jeux de données `data_clean` et `data_outliers` pour présenter quelques exemples pouvant faciliter la planification des tournées...

- arrosages des jeunes arbres
- contrôle des valeurs aberrantes
- contrôle des grands arbres
- diversité pour le replantage

Mais les données peuvent être utilisées pour répondre à des questions bien plus précises en utilisant plus de variables.

Synthèse de l'analyse de données

Data is for Good



En plus de l'identification des arbres répondant à un objectif précis, il serait pertinent d'utiliser l'**algorithme A*** ou sa variante **Dijkstra** pour trouver le trajet le plus optimal.

Enfin, il serait plus facile d'organiser les tournées si le jeu de données pouvait être enrichi par quelques informations complémentaires :

- **dates utiles**
(arrosage, taille, contrôle, ...)

- **particularités**
(type de taille, proche d'un bâtiment, traitement particulier, ...)

Merci

Data is for Good

de m'avoir écouté,
évalué et conseillé

