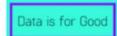
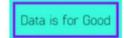
# **Concours Smart-City**





## Objectif du projet

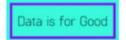




L'optimisation des tournées pour l'entretien des arbres de la ville.

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

### Présentation du jeu de données





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

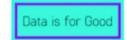
Ce jeu de données inclut les arbres :

- 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





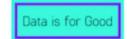
#### des méta-données

- id
- id\_emplacement
- type\_emplacement

### des variables géographiques

- o domanialite
- arrondissement
- o complement\_addresse
- numero
- lieu
- geo\_point\_2d\_a
- geo\_point\_2d\_b

## Présentation du jeu de données





### des variables numériques

- o circonference\_cm
- hauteur\_m

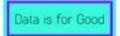
#### des variables nominales

- libelle\_francais
- genre
- espece
- variete
- stade\_developpement

#### une variable booléenne

remarquable

### **Analyse de données**





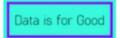
Une analyse univariée des 18 colonnes, qui suggère :

- la suppression de 6 colonnes (id, numero, type\_emplacement, complement\_addresse, 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\_addresse)

### Analyse de données





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

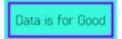
 une corrélation modérée entre certaines colonnes

(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



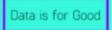


Une nettoyage du jeu de données :

- suppression des colonnes inutiles

   (id, numero, type\_emplacement, complement\_addresse, 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

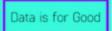




En analysant le jeu de donnée, nous avons pu mettre en évidence les défauts de certaines colonnes (pas de valeurs ajoutée, valeurs aberrantes ou manquantes ...) ou encore les liaisons plus ou moins

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



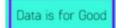


Nous avons utilisé les jeux de données data\_clean et data\_outliers pour présenter quelques exemples pouvant faciliter la planification des tournées...

- 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





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

