Rapport Projet Big Data

Binôme

Mathys Bodennec – Nicolas Guermeur

Table des matières

Contexte du projet

Ce projet est la première partie du grand projet Big Data/IA/Web. Le but de ce grand projet est d'appliquer les compétences acquises dans divers modules (Big Data, Intelligence Artificielle, Développement Web) à une étude complète sur le patrimoine arboré de la ville de Saint-Quentin (Aisne). L'objectif principal est de concevoir et développer une application permettant de traiter et de visualiser les données des arbres, et d'identifier ceux nécessitant une attention particulière, comme les arbres à abattre.

Cahier des charges

- 1. Exploration des données
- 2. Visualisation des données des graphiques
- 3. Visualisation des données sur une carte
- 4. Prédiction de la variable « Age estimé »

Fonctionnalité 1 : Description et exploration des données :

- Description du jeu de données
- Statistiques descriptives univariées, bivariées
- Nettoyage des données
  - Valeurs manquantes, valeurs aberrantes
  - Doublons

Voici une description de chaque variable contenue dans le fichier Patrimoine_Arboré(RO).csv

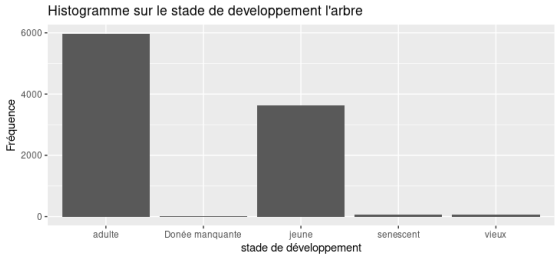
1. **X** : Coordonnée X en mètre de l'arbre suivant la norme ESPG3949 (suivant l'Est).
2. **Y** : Coordonnée Y en mètre de l'arbre suivant la norme ESPG3949 (suivant le Nord).
3. **OBJECTID** : Identifiant unique de l'objet dans la base de données.
4. **created_date** : Date de création de l'enregistrement.
5. **created_user** : Utilisateur ayant créé l'enregistrement.
6. **src_geo** : Source géographique de la donnée.
7. **clc_quartier** : Quartier de localisation de l'arbre.
8. **clc_secteur** : Secteur de localisation de l'arbre.
9. **id_arbre** : Identifiant unique de l'arbre.
10. **haut_tot** : Hauteur totale de l'arbre.
11. **haut_tronc** : Hauteur du tronc de l'arbre.
12. **tronc_diam** : Diamètre du tronc de l'arbre.
13. **fk_arb_etat** : État de l'arbre (par exemple, en place, à abattre).
14. **fk_stadedev** : Stade de développement de l'arbre.
15. **fk_port** : Port de l'arbre (sa forme générale).
16. **fk_pied** : Type de pied de l'arbre (par exemple, unique ou multiple).
17. **fk_situation** : Situation de l'arbre (par exemple, en bord de route, dans un parc).
18. **fk_revetement** : Type de revêtement autour de l'arbre.
19. **commentaire_environnement** : Commentaires sur l'environnement de l'arbre.
20. **dte_plantation** : Date de plantation de l'arbre.
21. **age_estim** : Âge estimé de l'arbre.
22. **fk_prec_estim** : Précision de l'estimation de l'âge de l'arbre.
23. **clc_nbr_diag** : Nombre de diagnostics effectués sur l'arbre.
24. **dte_abattage** : Date d'abattage de l'arbre.
25. **fk_nomtech** : Nom technique de l'arbre.
26. **last_edited_user** : Utilisateur ayant effectué la dernière modification.
27. **last_edited_date** : Date de la dernière modification.
28. **villeca** : Indicateur de localisation dans la ville.
29. **nomfrancais** : Nom français de l'arbre.
30. **nomlatin** : Nom latin de l'arbre.
31. **GlobalID** : Identifiant global unique de l'enregistrement.
32. **CreationDate** : Date de création (d'un format différent de `created_date`).
33. **Creator** : Créateur de l'enregistrement.
34. **EditDate** : Date de modification (d'un format différent de `last_edited_date`).
35. **Editor** : Utilisateur ayant modifié l'enregistrement.
36. **feuillage** : Type de feuillage de l'arbre.
37. **remarquable** : Indicateur si l'arbre est remarquable (oui ou non).

Extraction et Nettoyages des données

Les données ont été extraites du fichier CSV et nettoyées pour supprimer les enregistrements incomplets ou erronés. Les colonnes non pertinentes ont été éliminées pour simplifier l'analyse. Les valeurs manquantes ont été traitées et des doublons ont été supprimés.

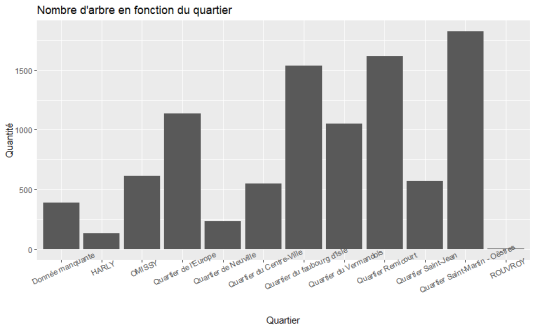
Certaines variables sont de type numérique, tandis que d'autres sont textuelles. Cependant, certaines variables textuelles peuvent être aisément converties en un autre type, tel qu'une date ou un booléen. De plus, de nombreuses variables présentent des informations manquantes (NA ou 0). Il y a également de nombreuses fautes de frappe ou un manque de cohérence (par exemple : Gricourt/Griourt, Orthophoto/orthophoto).

Fonctionnalité 2 : Visualisation des données sur des graphiques

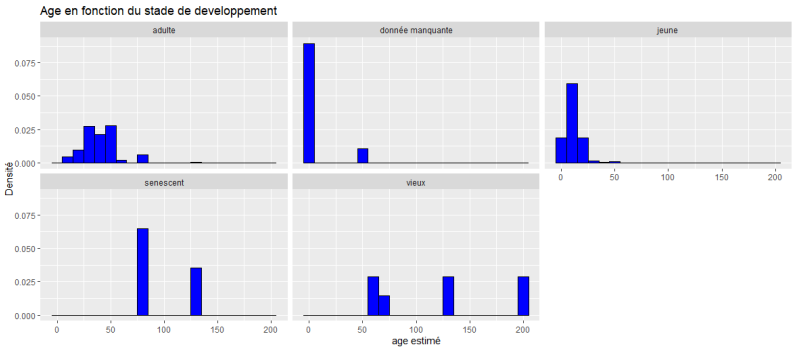


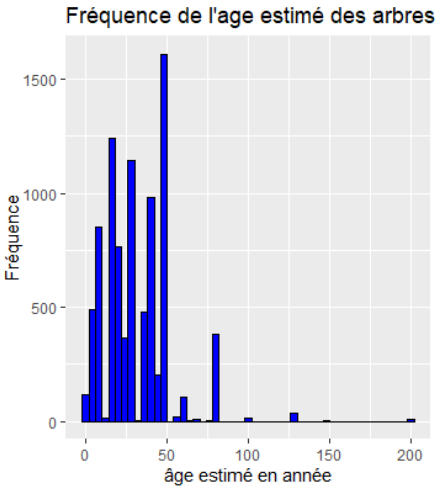
On peut voir qu'il y a très peu de données manquantes

Ainsi que peu de vieux arbres ou d'arbres sénescents

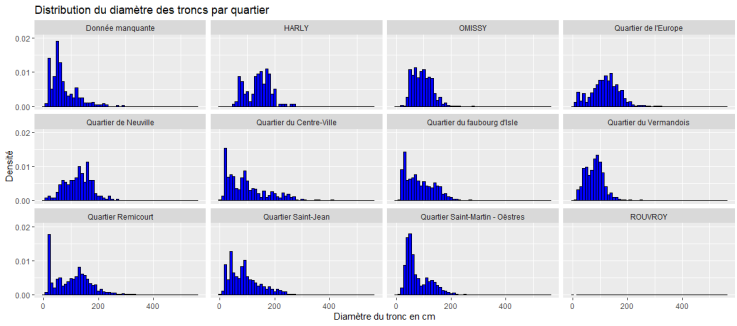


Le nombre d'arbres par quartier





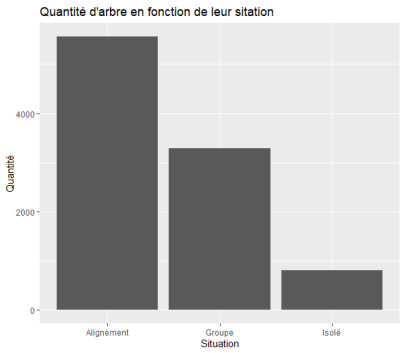
Représentation de l'Âge des arbres en fonction du stade de développement



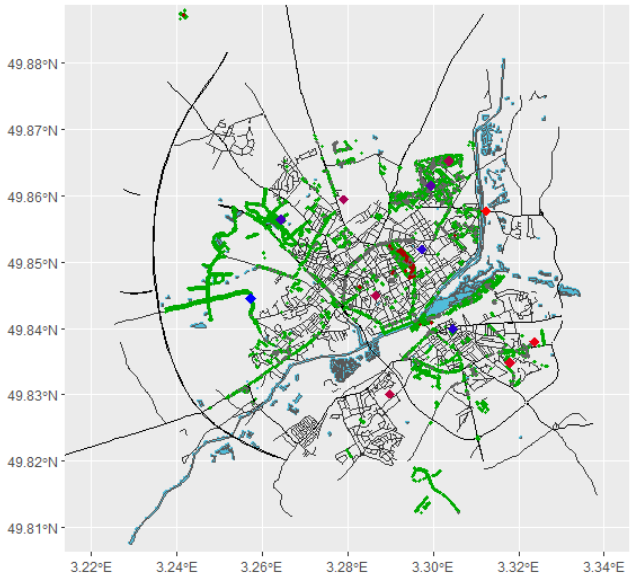
On peut voir la distribution du diamètre du troncs des arbres par quartier

Exemple : un pique à 0,10 pour un diamètre de 50 cm correspond à 10 % des arbres du quartier en question ont un diamètre de 50cm

On peut voir que très peu d'arbre sont isolés et que la majorité des arbres de ce projet sont dans un alignement



Fonctionnalité 3 : Visualisation des données sur une carte



Carte de Saint-Quentin (Aisne) et alentour :

- **Position des arbres :**
  - Points Vert : arbres en place
  - Points Gris : arbres qui sont abattu
  - Points Rouge : arbres remarquable (indépendamment de s'ils sont en place ou non
- **Répartition des arbres par quartier**
  - La couleur est bleue quand le quartier a beaucoup d'arbres
  - La couleur devient de plus en plus rouge quand il y a moins d'arbres

Fonctionnalité 4 : Etude des corrélations

Quels sont les liens entre les variables ?

La corrélation entre les variables les plus importantes :

	X	Y	OBJECTID	haut_tot	haut_tronc
X	1.000000000	0.07492028	0.35385877	0.1828903	0.08628804
Y	0.074920279	1.00000000	0.09867958	0.1352471	0.04989249
OBJECTID	0.353858770	0.09867958	1.00000000	-0.1471038	-0.11446146
haut_tot	0.182890281	0.13524712	-0.14710383	1.0000000	0.54321479
haut_tronc	0.086288037	0.04989249	-0.11446146	0.5432148	1.00000000
tronc_diam	0.004902468	-0.01999811	-0.29515575	0.6298171	0.42294979
age_estim	-0.106779967	-0.06711859	-0.46641570	0.5434735	0.50184086
fk_prec_estim	-0.061066110	-0.07148161	-0.50919374	0.4449263	0.42940243
clc_nbr_diag	-0.156300081	-0.13172844	-0.31704852	0.2463040	0.37817665

	tronc_diam	age_estim	fk_prec_estim	clc_nbr_diag
X	0.004902468	-0.10677997	-0.06106611	-0.1563001
Y	-0.019998108	-0.06711859	-0.07148161	-0.1317284
OBJECTID	-0.295155753	-0.46641570	-0.50919374	-0.3170485
haut_tot	0.629817061	0.54347349	0.44492633	0.2463040
haut_tronc	0.422949791	0.50184086	0.42940243	0.3781766
tronc_diam	1.000000000	0.78508748	0.62730878	0.2771263
age_estim	0.785087480	1.00000000	0.80271824	0.3523156
fk_prec_estim	0.627308781	0.80271824	1.00000000	0.3246315
clc_nbr_diag	0.277126311	0.35231562	0.32463147	1.0000000

Nous obtenons que les variables :

Age_estim & tronc_diam (0.78) ainsi que age_estim & fk_prec_estim(0.8) sont les plus corrélés.

Pour l'étude des relations entre variables qualitatives :

Nous avons utilisé le test de Chi-Carré, le test Chi2 est utilisé pour vérifier l'indépendance entre deux variables qualitatives. Le test de Chi2 sur R nous donne 3 informations :

p-value : Si le p-value < 0.05 l'hypothèse nulle est rejetée,

X-squared : la statistique de référence

df : degré de liberté de la loi chi2

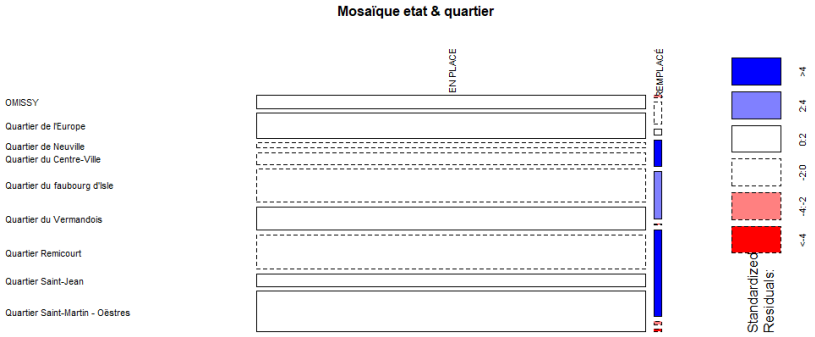
Fk_arb_etat & clc_quartier

```
        Pearson's Chi-squared test

data:  contingency_df
X-squared = 149.35, df = 8, p-value < 2.2e-16
```

Hypothèse nulle : fk_arb_etat & clc_quartier son indépendant

p-value < 0.05 donc l'hypothèse nulle est rejeté, on peut donc dire que fk_arb_etat & clc_quartier sont dépendants



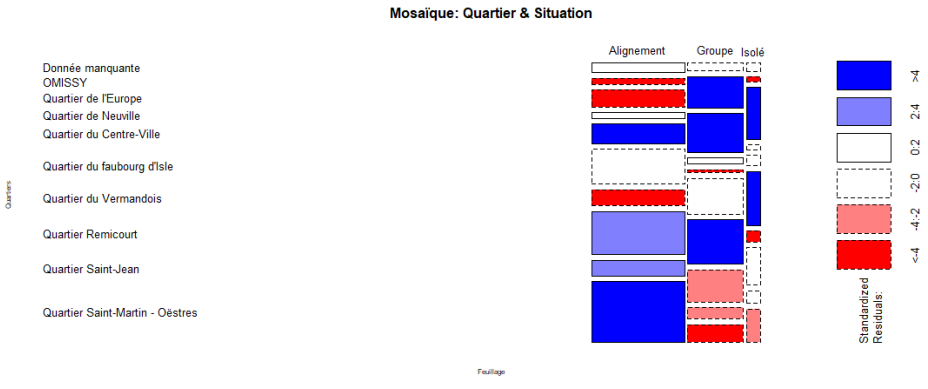
clc_quartier & fk_situation

```
Pearson's Chi-squared test

data:  tableau_croise
X-squared = 1602.4, df = 16, p-value < 2.2e-16
```

Hypothèse nulle : clc_quartier & fk_situation sont indépendant

p-value < 0.05 donc l'hypothèse nulle est rejeté, on peut donc dire que clc_quartier & fk_situation sont dépendants



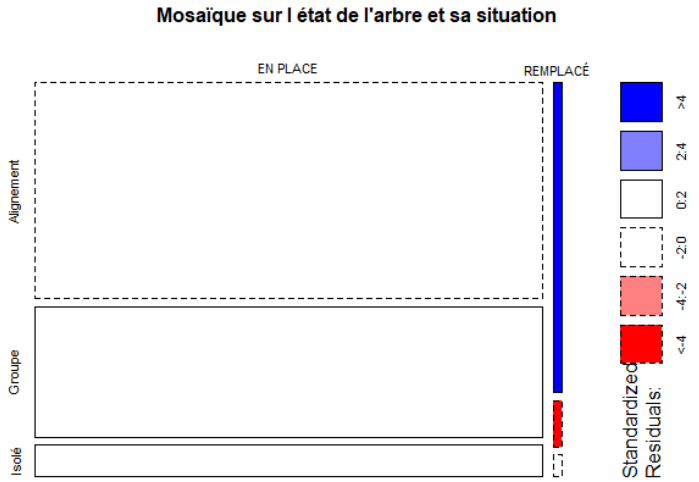
fk_arb_etat & situation

```
        Pearson's Chi-squared test

data:  contingency_df
X-squared = 44.634, df = 2, p-value = 2.032e-10
```

Hypothèse nulle : fk_arb_etat & fk_situation sont indépendant

p-value < 0.05 donc l'hypothèse nulle est rejeté, on peut donc dire que fk_arb_etat fk_situation sont dépendants



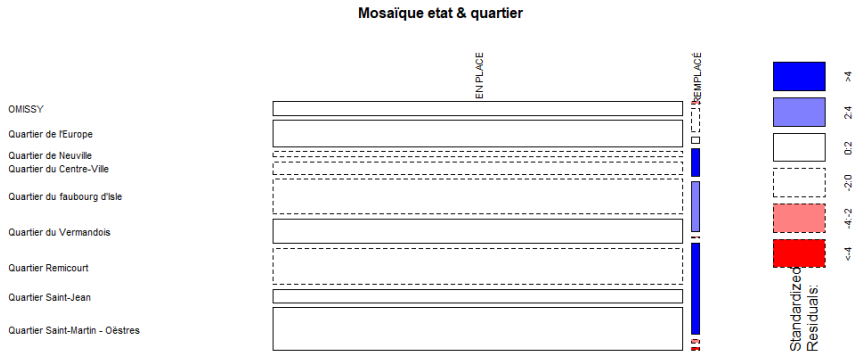
fk_arb_etat & remarquable

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_df
X-squared = 3.9683e-29, df = 1, p-value = 1
```

Hypothèse nulle : fk_arb_etat & remarquable sont indépendants

p-value > 0.05 donc hypothèse nulle validée

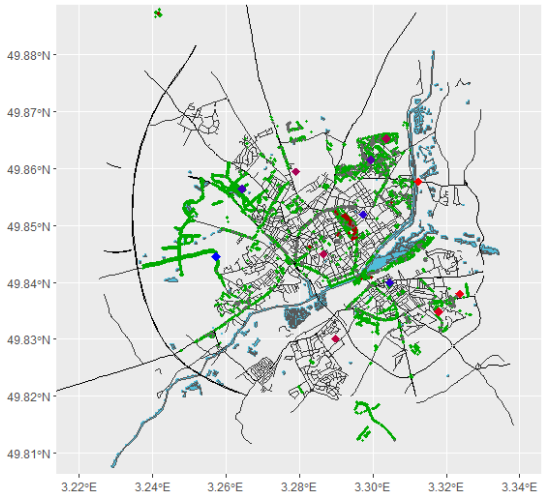


Fonctionnalité 5 : Etude des corrélations entre variables

• La ville a une politique urbaine qui consiste à planter des nouveaux arbres.

➢ Dans quelle zone faut-il les planter pour harmoniser le développement global de la ville ?

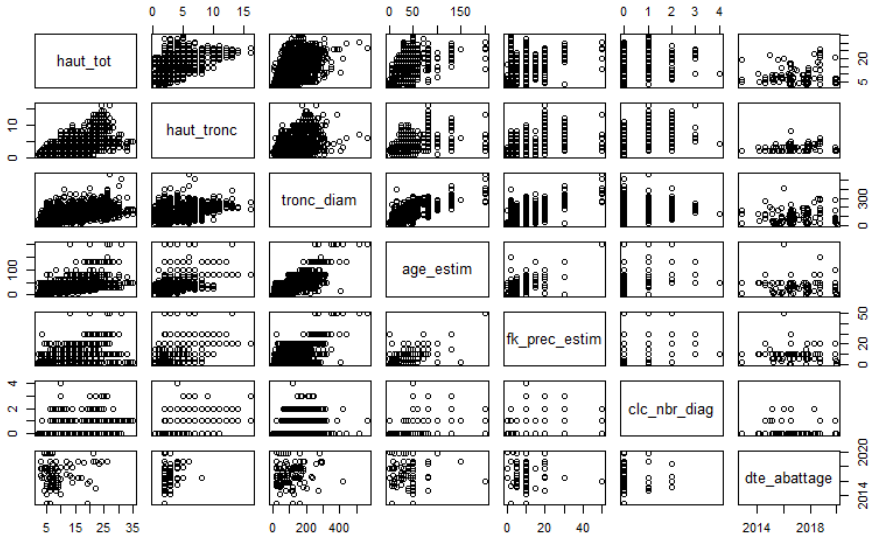
Il faut regarder sur la carte les losanges en rouge qui indiquent un manque relatif d'arbres par rapport aux autres quartiers



• On souhaite prédire la variable âge de l'arbre. Faire une étude de régression.

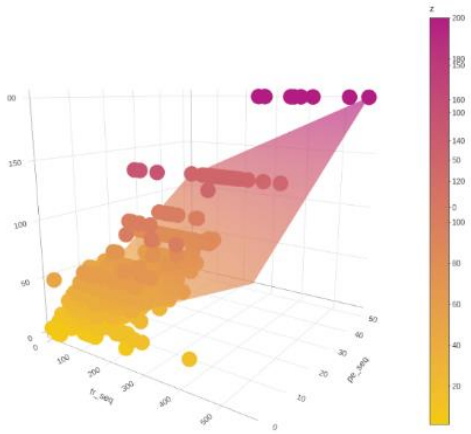
• On souhaite savoir quels sont les arbres à abattre. Faire une étude à l'aide de régression logistique

Étude de régression de la variable de l'âge de l'arbre



L'estimation de l'âge en fonction du diamètre du tronc

Et la précision de l'estimation de l'âge



```
Residual standard error: 9.334 on 8260 degrees of freedom
  (2 observations effacées parce que manquantes)
Multiple R-squared:  0.7913,    Adjusted R-squared:  0.7913
F-statistic: 1.566e+04 on 2 and 8260 DF,  p-value: < 2.2e-16
```

On peut voir que notre modèle sur la régression linéaire à une erreur résiduelle qui est faible (9.333), ça prouve que le modèle peut bien estimer l'âge de l'arbre

Regression logisitique

En raison du grand nombre de variables quantitatives à la base de ce modèle, la représentation graphique aurait été impossible sous forme d'image statique

On peut voir ici que notre modèle est largement moins bon que notre ancien modèle

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1626.3  on 8271  degrees of freedom
Residual deviance:  967.7  on 8031  degrees of freedom
AIC: 1449.7

Number of Fisher Scoring iterations: 20
```

Les arbres à abattre, le nombre à gauche étant l'ID de l'arbre

1936	0.6409588	TRUE
1952	0.6409588	TRUE
1956	0.6409588	TRUE
1970	0.6409588	TRUE
1975	0.6409588	TRUE
1980	0.6409588	TRUE
11182	0.8065934	TRUE
11184	0.8065934	TRUE

Fonctionnalité 6 : Export pour l'IA

- Exporter le fichier nettoyé en format csv pour une utilisation dans la partie Intelligence Artificielle.

Il s'agit tout simplement d'une seule ligne de code permettant de renvoyer la base de données après que le nettoyage ait été effectué

Sources :

https://www.statology.org

stackoverflow.com