Nom prénom : TAILLEBOIS Nicolas

Nom prénom : RIOU Margot

Nom prénom : AZEAU Julia

Classe : CIR

Module : Big Data

Responsable du module : Nesma SETTOUTI & Abir EL-HAJ

Titre du document : Rapport projet Big Data

Date : Vendredi 14 juin 2024

Nombre de mots :

- ☑ En adressant ce document à l'enseignant, je certifie que ce travail est le mien et que j'ai pris connaissance des règles relatives au référencement et au plagiat.

Sommaire

Introduction

Ce rapport va traiter de notre projet de Big Data à partir d'une base de données contenant les informations sur des arbres dans la ville de Saint-Quentin. Nous avions pour objectif différentes fonctionnalités que nous allons aborder et développer tout au long de ce rapport. Dans un premier temps nous nous sommes familiarisés avec la base de données en définissant chaque modalité puis en regardant à l'aide de différents graphique les liens entre ces dernières.

Nous avons pu remarquer que de nombreuses valeurs dans la base n'étaient pas exploitables, c'est pour cette raison que nous avons dû trier et nettoyer la base afin de pouvoir fournir des prédictions plus fiables.

Une fois toutes ces étapes suivies nous avons pu analyser les corrélations entre les variables qui nous paraissaient les plus pertinentes. Ces résultats nous ont ensuite permis de réaliser des régressions et des prédictions pour compléter les valeurs manquantes, notamment l'âge estimé des arbres.

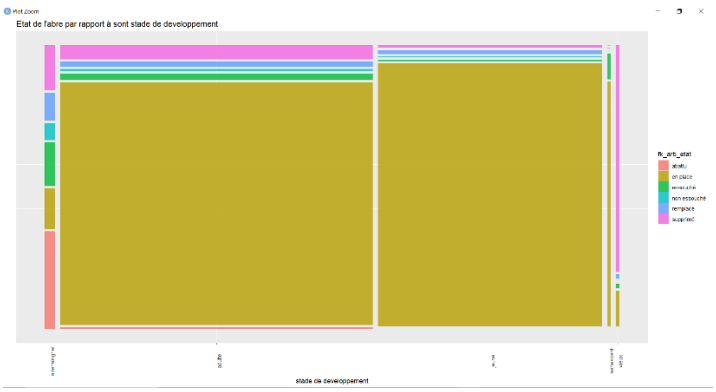
Fonctionnalité 1 : Exploration des données

Description du jeu de données :

Pour ce projet nous avons une base de données composée de 11421 exemples pour 37 variables qui sont :

- **X** : Coordonnées X de l'arbre (float)
- **Y** : Coordonnées Y de l'arbre (float)
- **OBJECTID** : ID de l'objet dans le tableau (int)
- **created_date** : Date de création de l'entrée dans le tableau (date)
- **created_user** : Nom de l'utilisateur ayant créé l'entrée dans le tableau (str)
- **src_geo** : Méthode d'images géographiques utilisée (str)
- **clc_quartier :** Quartier dans lequel est l'arbre (str)
- **clc_secteur :** Nom du secteur dans lequel est l'arbre (str)
- **id_arbre :** ID de l'arbre (int)
- **haut_tot :** hauteur totale de l'arbre (int)
- **haut_tronc :** hauteur du tronc de l'arbre (int)
- **tronc_diam :** Diamètre du tronc (int)
- **fk_arb_etat :** État de l'arbre (str)
- **fk_stadedev :** Stade de développement de l'arbre (str)
- **fk_port :** Aspect général de l'arbre (str)
- **fk_pied :** Terrain sur lequel pousse l'arbre (str)
- **fk_situation :** Situation de l'arbre (isolé, groupé, aligné, etc) (str)
- **fk_revetement :** Présence ou non d'un revêtement (bool)
- **commentaire_environnement :** Commentaires sur l'arbre (str)
- **dte_plantation :** Date de plantation (date)
- **age_estim :** Age estimé de l'arbre (int)
- **fk_prec_estim :** Tolérance de l'estimation de l'âge, pourcentage d'erreur (int)
- **clc_nbr_diag :** Nombre de diagnostiques fait sur l'arbre (int)

- **dte_abattage :** Date de l'abattage de l'arbre, s'il a été abattu (date)
- **fk_nomtech :** nom (str)
- **last_edited_user :** Dernier utilisateur à avoir modifié les informations de l'arbre (str)
- **last_edited_date :** Dernière date d'édition des informations de l'arbre (date)
- **villeca :** Info => Ville ou Casq
- **nomfrancais :** nom (str)
- **nomlatin :** nom (str)
- **GlobalID :** Id pour les informations de l'arbre (char)
- **CreationDate :** Date de création du global ID (date)
- **Creator :** Créateur du global ID (str)
- **EditDate :** Date de modification du global ID (date)
- **Editor :** Nom de la personne ayant modifié le global ID (str)
- **Feuillage :** Type de feuillage de l'arbre (str)
- **Remarquable :** Information si l'arbre est remarquable ou pas (bool)



Durant la visualisions graphique, nous avons eu une remise en question. Dans un premier temps nous avons fait le choix de différencier « supprimé » et « abattu ». Nous avons par la suite remarqué que les arbres « abattu » sont plus présents chez les adultes et les jeunes. Alors que « supprimé » concerne principalement les « vieux » et « sénescent ». On peut donc se demander si « abattu » n'est pas un acte de l'homme et « supprimé » serait un arbre mort de vieillesse et on l'aurait enlevé, donc sa disparition ne serais pas dû à l'homme. Finalement nous avons choisi de les considérer comme similaires.

Conversion du type de données :

Comme nous l'avons précisé dans la partie précédente chaque variable doit être d'un certain type or ce n'est pas toujours le cas. Il faut donc procéder à cette conversion manuellement.

On s'est rendu compte que toutes les dates d'une même colonne avaient une heure qui était toujours la même, on en a déduis que cette date n'est pas utile, on a donc converti toutes ces colonnes avec un as.Date().

On a fait le choix de transformer toutes les chaines de caractère en factor afin de faciliter l'affichage des graphiques et cartes.

Corriger l'encodage :

Pour avoir toutes les données en UTF-8 on a créé une boucle parcourant tous les caractères de la base de données pour les convertir en UTF-8.

Nettoyage des données :

Nous avons commencé par remplacer toutes les chaines de caractères vides par des NA puis nous avons supprimé toutes les lignes qui avaient au moins 13 valeurs vides car nous avons jugé que plus d'un tiers de données manquantes n'était pas assez précis pour pouvoir travailler avec ces données.

Nous avons remarqué sur les premières valeurs du tableau que les valeurs étaient les mêmes, on pensait donc pouvoir les fusionner. On a donc procédé à une vérification avec une boucle for qui nous a démontré que quelques valeurs étaient différentes ne permettant pas de fusionner ces colonnes sous risque de perdre des données pour plus tard.

Les deux colonnes fk_revetement et remarquable ont des valeurs qui peuvent être considérée comme booléennes. On transforme donc les Oui en True et les non par False. On doit le faire manuellement car en utilisant as.logical toutes les cases sont remplacées par NA.

Pour le taux d'erreur des estimations d'âge, nous avons choisi de remplacer les NA par des 0 pour éviter les erreurs plus tard dans le code.

Concernant la colonne avec les informations sur le type de feuillage, nous avons remarqué que les cases vide correspondaient aux arbres abattus ou supprimés. Pour éviter de futures erreurs due au fait d'avoir une case vide, nous avons choisi de les remplacer par la modalité « Non renseigné »

Pour les X et Y, nous avons vérifié s'il existait des valeurs manquantes, et nous avons supprimé les lignes où il y avait des NAs.

Nous avons aussi décidé de supprimer la colonne OBJECTID car les valeurs ne se suivaient pas, donc elles n'étaient pas exploitables, tout comme les lignes sans valeurs dans id_arbres

Pour la hauteur totale, la hauteur du tronc et son diamètre, nous avions commencé par remplacer les valeurs manquantes, s'il y en avait, par la valeur médiane affichée dans le summary de la base. Ensuite, pour plus de précision, nous avons effectué des régressions entre ces trois modalités en ajoutant aussi l'age estimé, le stade de développement de l'arbre et le type de feuillage de l'arbre. Malheureusement, cette méthode ne remplace pas la totalité des NA présents et la précision n'est pas exceptionnelle non plus (R-squared pour la hauteur totale : 0.4806, pour le diamètre du tronc : 0.6628 (cette valeur est mieux) et pour la hauteur du tronc : 0.322)

Pour les quartiers et les secteurs, nous avons recherché le quartier le plus proche des entrées n'ayant pas de quartier assigné en calculant la distance euclidienne de l'arbre avec chacun des arbres de la base pour garder la distance la plus courte et lui assigner le quartier correspondant. Pour les secteurs manquants, nous avons remplacé les NA par « Non renseigné »

Étant donné que les coordonnées X et Y de la base de données suivent le système de coordonnées français EPSG 3949 nous avons choisis de rajouter les coordonnées converties dans le système EPSG 4326 qui est le système de coordonnées que nous connaissons. Pour ce faire nous avons deux nouvelles variables à notre base de données x et y.

Pour les modalités fk_stadedev, fk_port, fk_pied et src_geo nous avons remarqué que certaines valeurs étaient les mêmes avec pour différence une majuscule, on a donc pour chaque exemple mis toute l'information en minuscule pour les rassembler ensemble. Pour éviter les erreurs d'apparition de NA à cause de nouveau nom de modalité dans le facteur. On modifie la variable en chaine de charactère avant puis à nouveau en facteur après modification. On a également modifié les valeurs nulles par « non renseigné », car on a vu que certaines n'avaient pas d'espace et d'autres en avaient un or il s'agit dans les deux cas d'une valeur inconnue.

Nous avons pu voir que les exemples des modalité fk_situation et fk_arb_etat avaient des variations au niveau de la casse, il nous a donc semblé pertinent de tout mettre en minuscule afin d'éviter de potentielles répétitions au sein de la base.

Nous avons fait le choix de laisser les valeurs NA pour les dates d'abattage et date de plantation car il est impossible de donner une date d'abattage si l'arbre est encore planté et la deuxième de peut être calculée à partir d'une estimation de l'âge à une date inconnue.

De même pour toutes les cases vides à propos des informations d'édition ou de création de l'arbre dans la table car on ne peut estimer les dates, on a donc choisi de les laisser telles quelle.

Pour finir, nous avons utilisé la fonction unique() pour supprimer les doublons dans la table.

Analyse exploratoire :

Pour cette partie nous utiliserons diverses fonctions, hist(), plot(), ggplot...

Les variables catégorielles choisies sur lesquelles nous voulons calculer les différentes fréquences :
Clc_quartier / Fk_arb_etat / Fk_stadedev / Fk_pied / Feuillage / Fk_situation

À première vue et avec nos connaissances ces variables s'influencent sûrement. La fonctionnalité 4 permet de confirmer nos choix.

Nous calculons les différentes fréquences pour chaque valeur, on peut utiliser deux fonctions :

- Prop.table(table (« la colonne qu'on souhaite »))
- Describe(data[« colonnes 1 », « colonnes 2 »)

« Describe » : Affiche dans le terminal différentes informations sur les variables demandé dont la fréquence. Les fréquences nous permettent d'obtenir une première vision sans graphique sur le type d'arbres le plus important, le quartier qui a le plus d'arbre...

L'analyse exploratoire nous a aussi permis de nous rendre compte qu'il fallait transformer les chaines de caractères en factor avec as.factor() pour pouvoir les utiliser dans divers graphiques.

Au final nous serons amenés à faire de l'analyse exploratoire tout au long du projet a fin de mieux comprendre certaine données, et d'actualiser nos connaissances sur la base de données en fonction du problèmes face à nous.

L'analyse exploratoire nous pousseras naturellement vers la fonctionnalité 2 avec la visualisation des données qui peut aussi faire partie de l'analyse exploratoire.

Fonctionnalité 2 : Visualisation des données sur des graphiques

Dimensions des arbres :

En premier lieux nous avons réaliser des graphiques simples avec des variables quantitatives pour visualiser par exemple : le nombres d'arbres selon la taille, le diamètre des arbres ...

Les arbres au sein des quartiers :

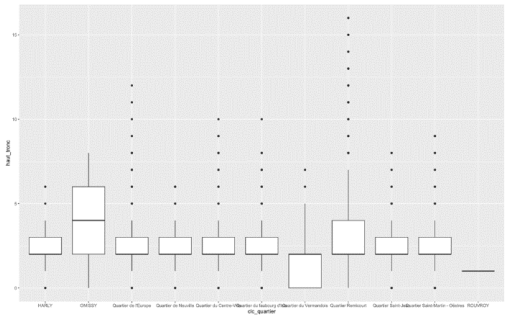


Figure : Hauteur des troncs par quartiers

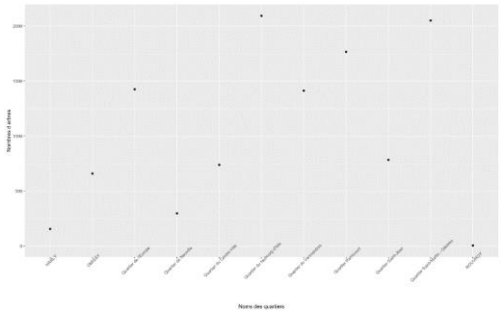


Figure : Nombres d'arbres par quartiers

Pour se faire une idée des positionnements des arbres sans carte nous avons réalisé des graphiques avec la variables clc_quartier.
On remarque qu'il n'y a que certains quartiers qui possédent des arbres dépassant les 10 m.
Le quartier Remicourt posséde tous les plus grands arbres.

Des chiffres sur les arbres :

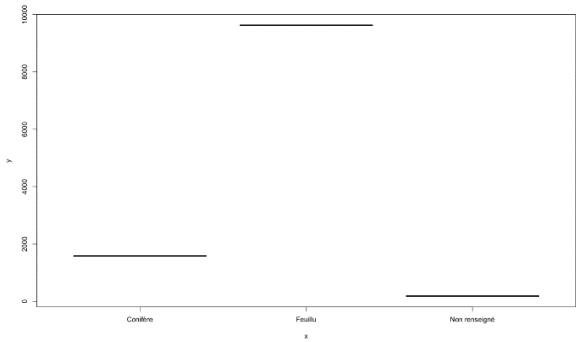
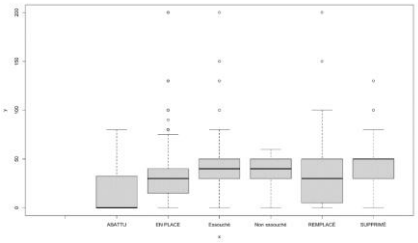


Figure : Le nombre de chaque type d'arbre

On a pu réaliser des graphiques pour mieux appréhender certaines particularités des arbres comme : le nombre de chaque type d'arbre, la quantité d'arbres en fonction de leurs situation...
Il y a en général beaucoup plus d'arbres feuillu que de conifère.
Et la plupart des arbres sont alignés surement le long des routes ou des chemins, les arbres isolés sont plus rares.

État de l'arbre et son évolution :



Nous remarquons que ni le stade de développement ni l'âge estimé n'influencent l'état de l'arbre.

Graphiques mosaïques :

Nous nous sommes attardés sur les graphiques mosaïques afin de pouvoir se familiariser avec les variables qualitatives et les visualiser ensemble, quelques exemples :

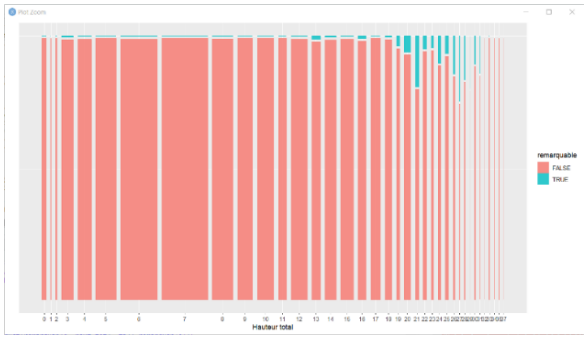


Figure : L'âge selon si l'arbre est remarquable

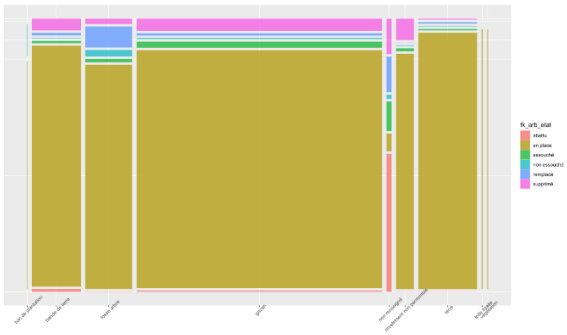
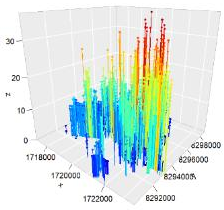
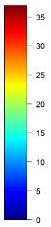


Figure : L'état des arbres en fonction du type de sol

Nous remarquons qu'en général plus un arbre est vieux plus il a des chances d'être inscrit comme remarquable. C'est une des possibilités de corrélation que nous avons pu visualiser avant de les étudier. Nous avons également noté que le type de sol n'influencent pas l'état de l'arbre.

Autres graphiques :

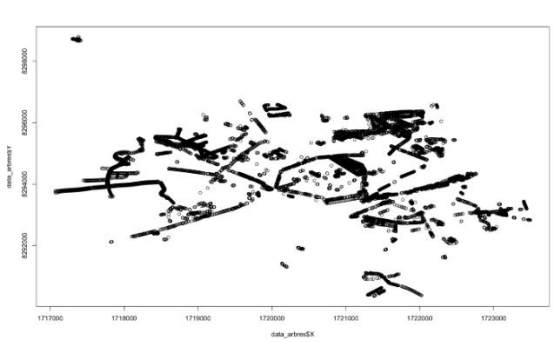


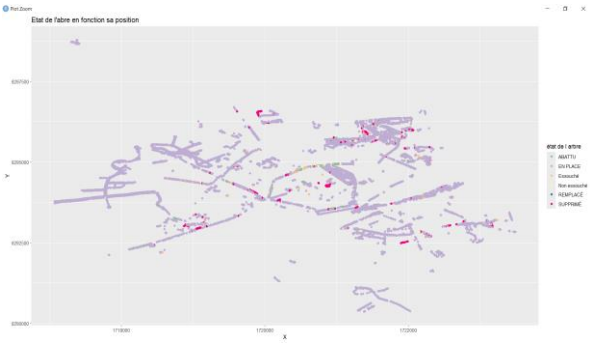


Nous avons également fait une tentative de graphique en 3D pour représenter les arbres en fonction de leur position et avec leur hauteur :
Au-delà de créer un rendu sympa on peut remarquer que les arbres les plus grands paressent se trouver pratiquement tous au même endroit.

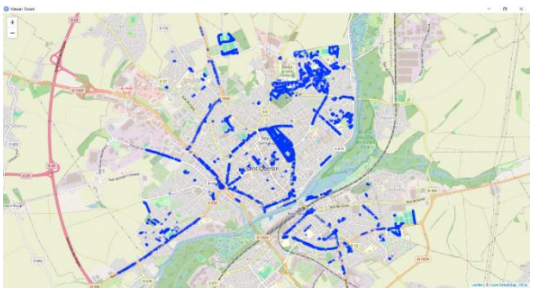
Fonctionnalité 3 : Visualisation des données sur une carte

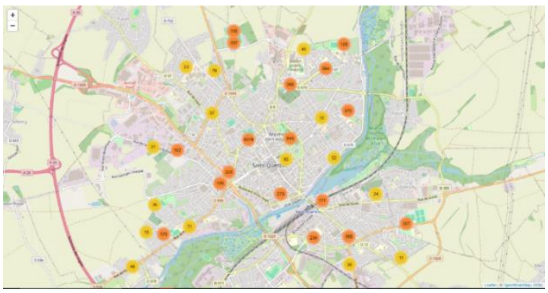
Nous avons choisi dans un premier temps de représenter la position des arbres en fonction de leurs coordonnées géographiques (X et Y). Puis on a ensuite représenté le même graphique en faisant varié la couleur de chaque arbre en fonction de son état :



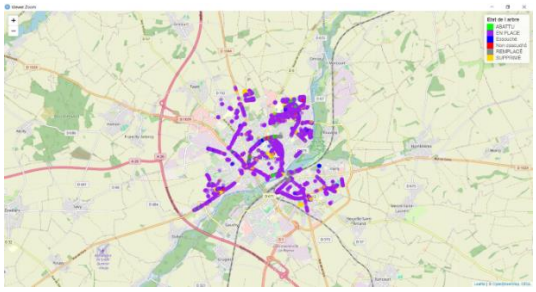


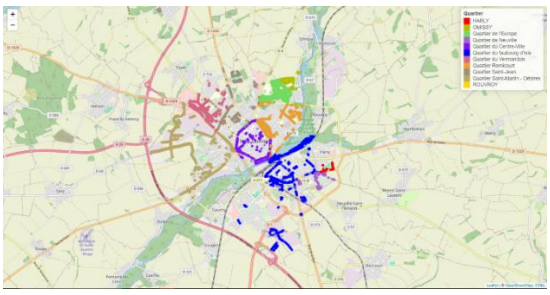
Nous avons également ajouté une carte interactive à l'aide de la librarie leaflet, dans un premier temps juste les arbres dans la ville. Et aussi avec une amélioration qui nous permets de mieux visualiser le nombre d'arbres :





Nous avons aussi une carte en fonction de l'état de l'arbre et une autre en fonction du quartier dans lequel ils sont situés :





Nous pouvons même comparer les deux cartes afin de se faire une idée rapide visuellement des états des arbres par quartier.

Fonctionnalité 4 : Étude des corrélations

En affichant les différents graphiques précédents on a remarqué que certaines étaient liés et d'autres non. Sachant que notre objectif était principalement sur les caractéristiques de

l'arbre, notamment son âge nous avons majoritairement cherché des corrélations sur ces modalités.

Pour calculer la corrélation entre deux de nos variables quantitatives qui suivent une loi normale on peut utiliser la fonction cor.test() avec la méthode Pearson :

Variable 1	Variable 2	Taux de corrélation
Hauteur totale de l'arbre	Âge estimé de l'arbre	0.6
Hauteur totale de l'arbre	Hauteur du tronc	0.526
Hauteur totale de l'arbre	Diamètre du tronc	0.695
Hauteur du tronc	Diamètre du tronc	0.394
Age estimé	Diamètre du tronc	0.76

Ce tableau rassemblant les résultats des différents tests que nous avons pu faire nous démontre bien que pour une grande partie des variables, notre instinct était le bon car le taux est entre 0.5 et 1 exprimant une corrélation entre les variables.

Nous avons également expliqué la variation d'une valeur par rapport à une autre en créant des modèles de régression, le résultat correspond au pourcentage expliquant la variation de la première variable par rapport à la deuxième :

Variable 1	Variable 2	Pourcentage de variation
Hauteur tronc	Coordonnées X et Y	11.24 %
Hauteur du tronc	Sol au pied de l'arbre	3.4 %
Hauteur de l'arbre	Sol au pied de l'arbre	10.8 %
Âge de l'arbre	Stade de développement de l'arbre	50 %

On a également testé la corrélation entre l'âge estimé de l'arbre et sa date de plantation. Nous avons été surpris de remarquer que le pourcentage de variation de l'âge par rapport à cette date est très faible dans un premier temps, environ 0.3 %. Ce pourcentage augmente nettement une fois la régression pour déterminer les âges manquants faite, il est désormais de 36 %.

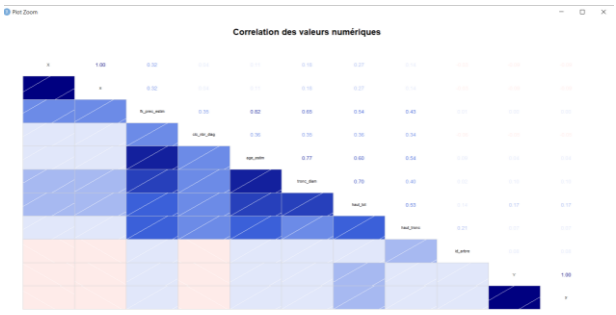
Pour vérifier tout cela, nous avons voulu créer un modèle de régression exprimant le diamètre du tronc en fonction de l'âge estimé, de la hauteur totale de l'arbre, le stade de développement de l'arbre et feuillage. Le modèle nous montrait que ces variables expliquent à 66% la variation du diamètre du tronc.

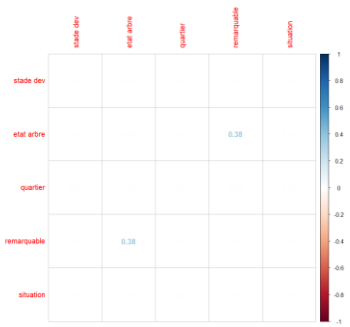
Pour la corrélation entre variables qualitatives on a pu utiliser chi 2 pour visualiser le p-value :
Si p-value > 5% (0.05) alors les variables sont indépendantes entres elles.
Si p-value est proche de zéro il est probable que les deux variables soient fortement corrélées.

Variable 1	Variable 2	p-value
Stade de développement	L'état de l'arbre	2.2e-16
Quartier	Remarquable	2.2e-16
Situation	Remarquable	4.109e-10
Stade de développement	Pied	2.2e-16
Quartier	Pied	2.2e-16

Il existe un lot de variables qualitatives ou les corrélations au sein de se lot sont importantes.

Exemples de matrices de corrélation :





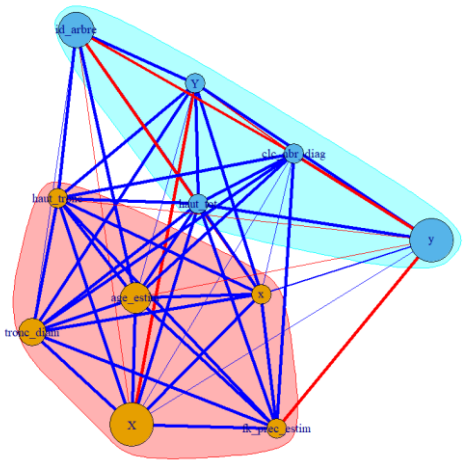
Matrice : selon le p-value

La première matrice exprime la corrélation entre les valeurs numériques grâce aux taux de corrélation, plus on est proche de 1 ou de – 1 plus il y a corrélation et plus on est proche de 0 plus les variables sont indépendantes.

La deuxième matrice est entre les valeurs qualitative, elle se base sur le p-value, plus c'est proche de 0 plus il y a corrélation.

Les deux matrices confirment donc les conclusions que nous avons pu réaliser précédemment avec nos tableaux.

Test avec la fonction corrgraph :



Nous avons découvert cette fonction permettant d'afficher graphiquement les corrélations entre les variables. Mais nous nous sommes rendu compte que malgré le fait qu'elle semble faciliter le travail, elle n'est pas vraiment fiable, en effet les corrélations entre les coordonnées et l'âge de l'arbre ne nous parais pas très logique. De même pour celle liant l'id de l'arbre à ces coordonnées ou dimension.

Fonctionnalité 5 : Prédiction de variables

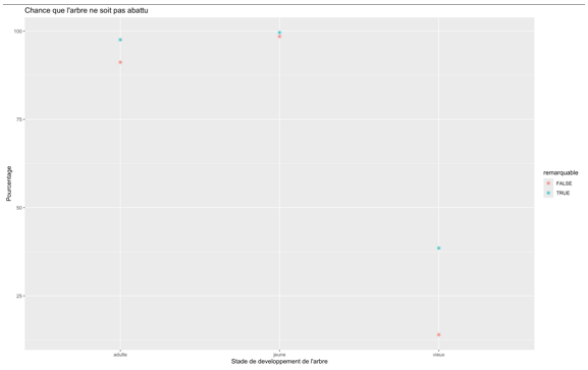
Prédiction de la zone où planter de nouveaux arbres :

Pour prédire le quartier dans lequel planter en priorité de nouveaux arbres, nous avons calculé la surface de chaque quartier puis sa densité d'arbre par hectare. Nous avons ensuite récupéré le quartier avec la plus faible densité d'arbres et en avons déduit que c'était celui là qui nécessitait le plus de la plantation de nouveaux arbres.

Prédiction de l’âge estimé :

Pour l’estimation de l’âge de l’arbre nous avons remarqué une corrélation entre la hauteur totale de l’arbre, le diamètre du tronc et le stade de développement de l’arbre. Nous avons donc fait le choix de faire notre model et notre prédiction sur ces modalités en ajoutant aussi le stade de développement de l’arbre (jeune, adulte, etc). On a pu obtenir une p-value de 2.2e-16 et un Adjusted R-squared de 0.6715, ce qui montre que la prédiction est assez précise.

Prédiction des arbres à abattre



Pour les arbres à abattre on a pris deux paramètres qui nous semblait important : le développement de l’arbre et s’il est remarquable. Qui nous a permis de réaliser un modèle de type binomial.
On a ensuite réalisé trois prédictions : Remarquable était à False pour les trois (on part du principe que s’il a le titre remarquable on n’y touche pas) et pour les 3 stade de développement (jeune, adulte et vieux).

On a donc réalisé un graphique du pourcentage de chance que l'arbre ne soit pas abattu.
On voit que le stade de développement joue un rôle important et que plus il est vieux (et non remarquable) plus ses chances d’être abattue/supprimer sont grandes.

Conclusion

Pour conclure ce rapport, cette semaine nous a permis de pleinement analyser la base de données afin de la nettoyer au mieux pour les rendre plus cohérentes et exploitables par un modèle d’entrainement de prédiction d’une IA.

Nous avons pu remarquer l’importance du traitement des données, même si nous ne l’avons pas fait dans ce projet avec un retour sur expérience il nous semble primordial pour les projets futurs de mieux automatiser les tâches répétitives, telles que les boucles permettant le nettoyage de variable similaires.

Nous avons fait face à des difficultés que nous avons réussis à surmontées grâce à notre travail d’équipe, même si nous sommes restés bloqué un certain temps avant de trouver la solution pour prédire les lieux où planter des arbres ou ceux qu’il faut abattre.

Cela nous a aussi permis de nous rapprocher du métier d'ingénieur et de nous rendre compte des différentes contraintes, notamment l'harmonisation dans les quartiers (type d'arbres, hauteur…). Avec plus de données, nous aurions pu prendre en compte, par exemple, les arbres qui dérangent ou non le voisinage, les câbles électriques à proximité et bien d'autres.

De plus, étant actuellement dans un contexte écologique important, la prise en compte de l'environnement, le respect des arbres (remarquables ou non) et la demande de certains quartiers ou villes de se revégétaliser, l'analyse de ce type de données devient de plus en plus essentielle.