ISEN
BREST

Rapport de Projet A3
Partie BigData

Du 3 juin au 7 juin
Avec Yann MOALIC

PORCHER Jeanne
ROUILLE Alban
POURCHASSE Lana

Sommaire

Introduction

Objectif principal du projet :

Concevoir et développer une application d'étude du patrimoine arboré de la ville de Saint-Quentin.

Objectifs détaillés de la partie Big Data :

1. Extraction des données : à partir du fichier de données brutes Patrimoine_Arbore.csv.
2. Visualisation d'un grand volume de données : création de graphiques et de cartes.
3. Nettoyage des données : suppression des données incomplètes et erronées, formatage textuel.
4. Application de modèles statistiques : utilisation d'études de corrélations, de régressions linéaires, régressions logistiques afin d'analyser les données.

Pour démarrer ce projet, nous avons commencé par nous répartir les tâches à partir d'un diagramme de Gantt (figure 1).

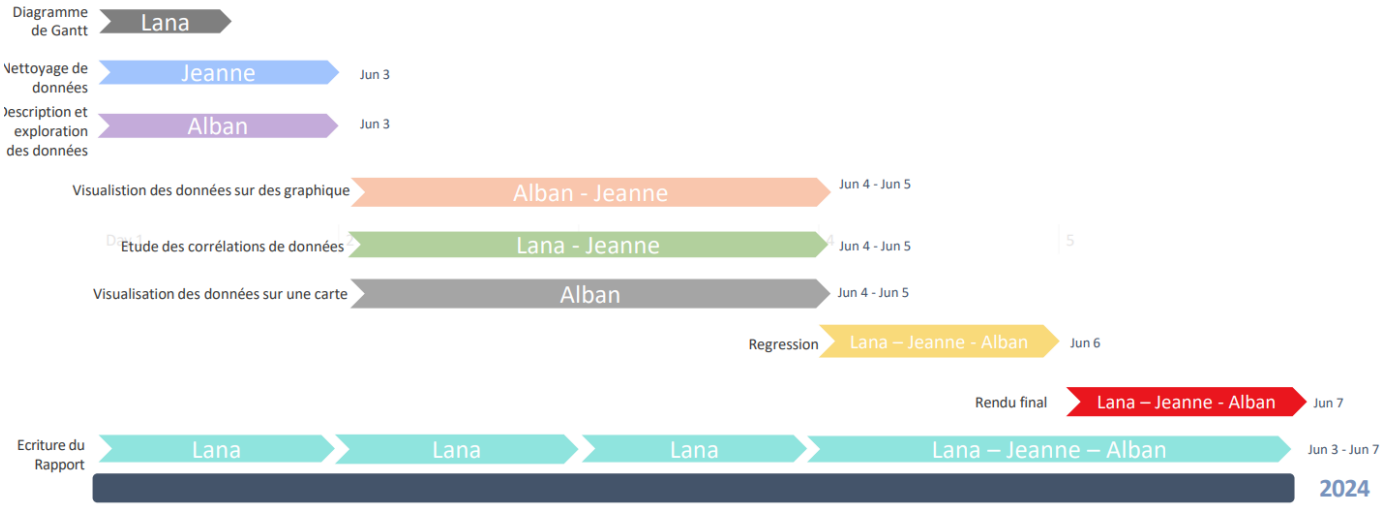

Figure 1: Diagramme de Gantt

Traitement des données

Description du jeu de données

Ce jeu de données provient du patrimoine arboré de la ville de Saint-Quentin (Aisne). Il s'agit d'un fichier contenant l'ensemble des arbres de la ville de Saint-Quentin, accompagnés de leur position, caractéristiques (taille, diamètre, texture, ...), et détails d'édition.

Exploration des données

- Position en RGF93-CC49 de l'arbre :

  X : variable quantitative

  Y : variable quantitative

- Identifiant de l'objet dans la base de données :

  OBJECTID : variable quantitative, identifiant local de la table

  GlobalID : variable quantitative, identifiant globale de l'arbre dans la base de données

  Id_arbre : variable quantitative, identifiant du type d'arbre

- Informations sur l'arbre :

  Hauteur totale de l'arbre en mètres : variable quantitative (haut_tot)

  Hauteur du tronc de l'arbre en mètres : variable quantitative (haut_tronc)

  Diamètre du tronc de l'arbre en centimètres : variable quantitative (tronc_diam)

  État de présence de l'arbre : variable qualitative (fk_arb_etat)

  État de développement de l'arbre : variable qualitative (fk_stadedev)

  Type de coupe et d'entretien de l'arbre : variable qualitative (fk_port)

  Composition du pied de l'arbre : variable qualitative (fk_pied)

  Situation de l'arbre (isolé ou aligné) : variable qualitative (fk_situation)

  Présence de revêtement sur l'arbre : variable qualitative (fk_revetement)

  Type du feuillage de l'arbre : variable qualitative (feuillage)

  Remarquable (oui/non) : variable qualitative (remarquable)

- Informations sur la variété de l'arbre :

  Nom français de la variété de l'arbre : variable qualitative (nomfrancais)

  Nom latin de la variété de l'arbre : variable qualitative (nomlatin)

- Informations sur la base de données et les modifications :

  Date de création de l'objet dans la base de données : variable quantitative (CreationDate)

  Créateur de l'objet dans la base de données : variable quantitative (Creator)

  Date de la dernière modification de l'objet : variable qualitative (EditDate)

  Personne ayant modifié l'objet : variable quantitative (Editor)

Nettoyage des données :

Après l'exploration des données, nous avons décidé de supprimer certaines données qui ne seront pas utiles pour l'exploitation statistique. Voici la liste des données que nous n'avons pas conservée :

Données supprimées :

Date de création de l'entrée dans la base de données (created_date)

Créateur de l'entrée dans la base de données (created_user)

Commentaire d'environnement autour de l'arbre (commentaire_environnement)

Date d'abattage de l'arbre (dte_abattage)

Nom technique de l'arbre (fk_nomtech)

Dernière personne ayant effectué une modification (last_edited_user)

Date de la dernière modification (last_edited_date)

Nom français de la variété de l'arbre (nomfrancais)

Nom latin de la variété de l'arbre (nomlatin)

Identifiant global de l'arbre (GlobalID)

Date de création de l'objet dans la base de données (CreationDate)

Créateur de l'objet dans la base de données (Creator)

Date de la dernière modification de l'objet (EditDate)

Personne ayant modifié l'objet (Editor)

Suppression des valeurs manquante et nettoyage des valeurs aberrantes :

Nous avons procédé au nettoyage des données en plusieurs étapes clés :

1- Traitement des Valeurs Manquantes : Remplacement des valeurs vides par `NA`, suppression des doublons, et imputation des valeurs manquantes par la médiane pour les colonnes numériques. Que nous avons déterminé avec la figure 2

2- Normalisation des Données Textuelles : Conversion des textes en minuscule et remplacement des espaces par des underscores.

3- Suppression des lignes incomplètes : Élimination des lignes contenant plus d'une valeur manquante.

4- Gestion des valeurs aberrantes : Application de la Winsorisation pour limiter l'impact des valeurs extrêmes. Nous avons pu détecter les valeurs aberrantes avec la figure 3

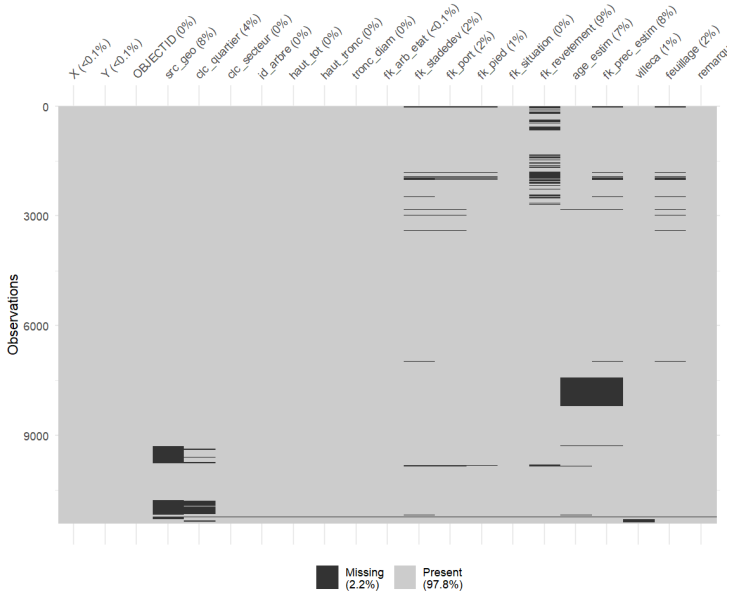

Figure 2 : graphique vis_miss

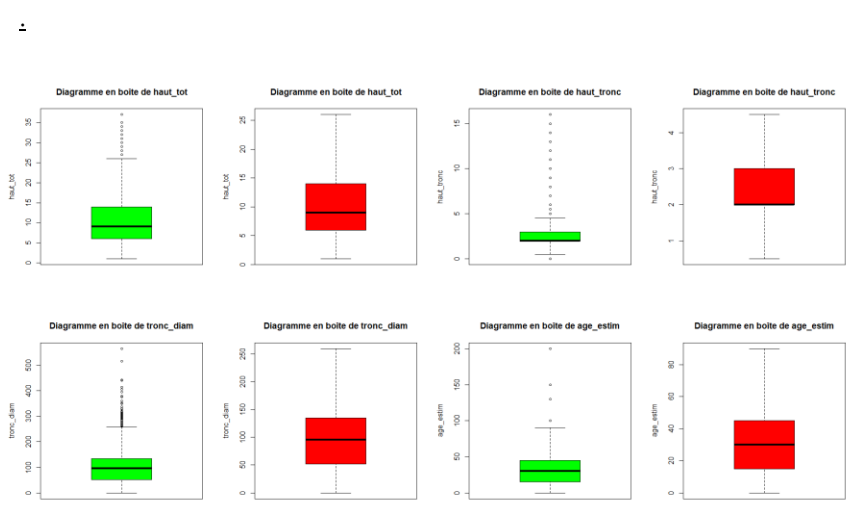

Figure 3 : Boxplot pour mettre en avant les valeurs aberrantes, en vert les valeurs avant traitement et en rouge après

Visualisation des données sur des graphiques

Répartition des arbres :

Les graphiques ci-dessous présentent une analyse visuelle de la répartition des arbres selon trois critères principaux : leur situation, leur localisation par quartier et leur stade de développement. Ces visualisations permettent de mieux comprendre la distribution et les caractéristiques des arbres dans une zone urbaine donnée.

Répartition des arbres selon leur situation :

La figure 4 illustre la répartition des arbres en fonction de leur situation. Les catégories analysées comprennent les arbres en alignement, en groupe, et isolés. On observe que la majorité des arbres sont alignés (5753), il s'agit d'arbres plantés sur les bords de rues. Suivis par ceux en groupe (3505), qui correspondrait à des arbres présents dans des parcs ou espaces verts. Les arbres isolés sont ceux les moins nombreux avec seulement 887 et correspondrait à des plantations spécifiques d'arbres.

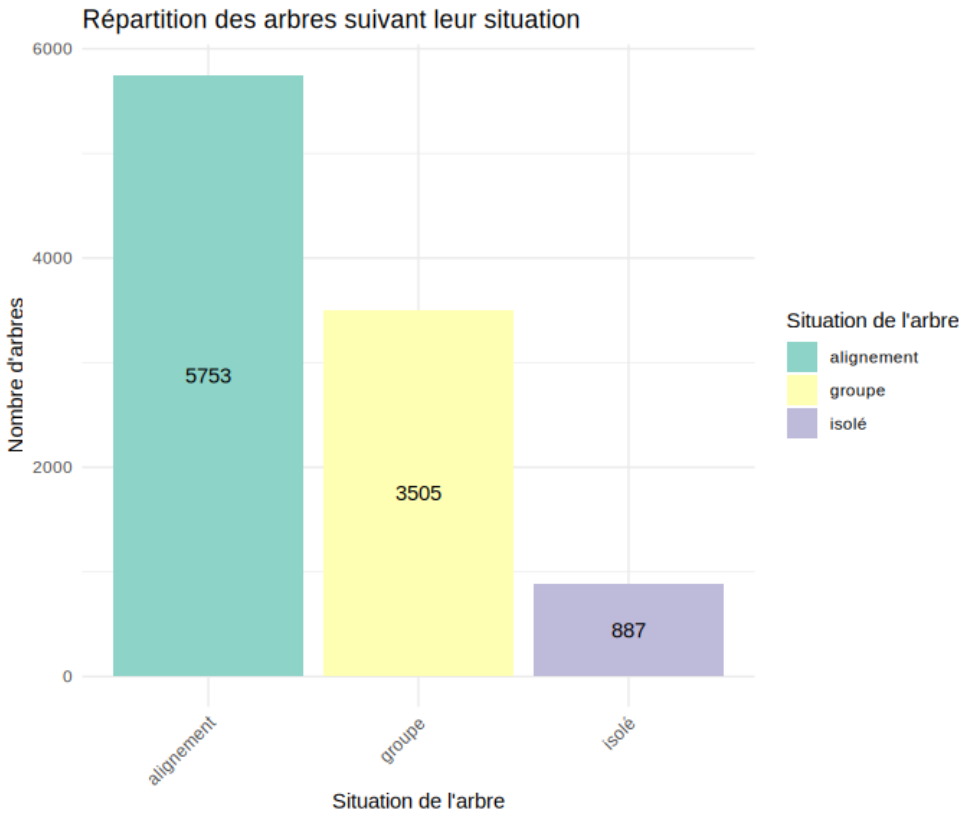

Figure 4 : Répartition des arbres suivant leur situation

Répartition des arbres selon leur quartier :

La figure 5 montre la répartition des arbres par quartier. Les quartiers incluent Harly, Omissy, Quartier de l'Europe, Quartier de Neuville, Quartier du Centre-ville, Quartier du Faubourg d'Isle, Quartier du Vermandois, Quartier Remicourt, Quartier Saint-Jean, Quartier Saint-Martin - Oestres, et Rouvroy. Les quartiers de Saint-Martin - Oestres et de Remicourt se distinguent par un nombre particulièrement élevé d'arbres, correspondant à des quartiers contenant des parcs ou de nombreux axes principaux arborés. Tandis que d'autres quartiers comme Rouvroy, Neuville et Harly, dénués de parcs ou de grandes rues contiennent beaucoup moins d'arbres.

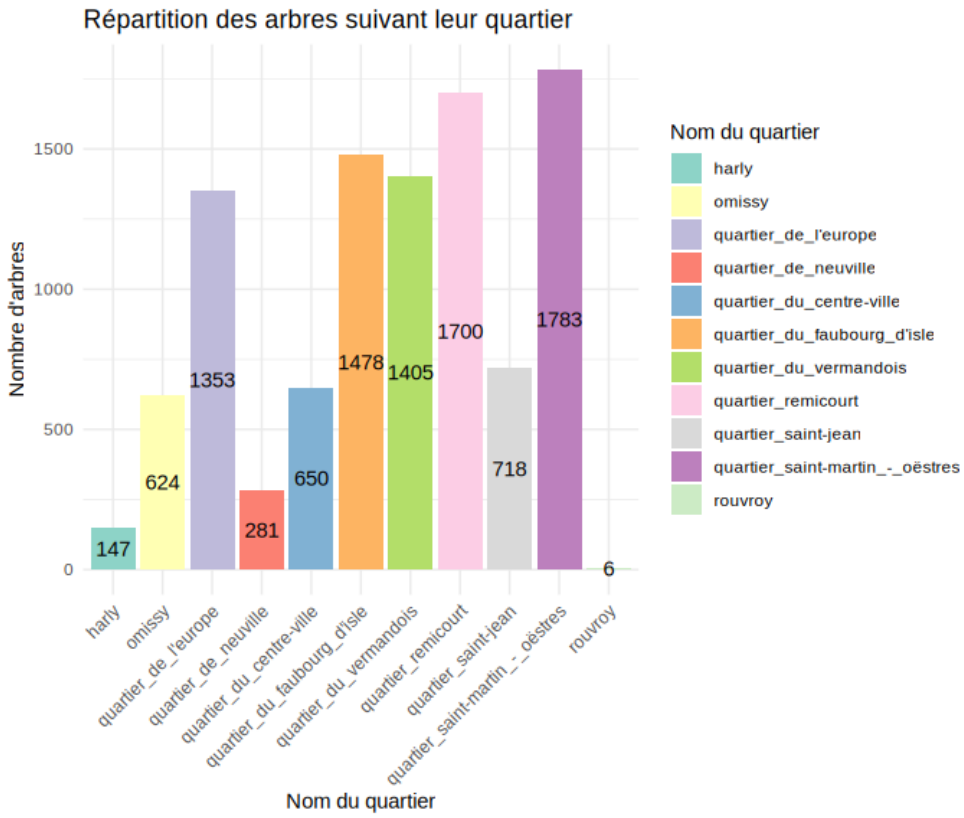

Figure 5 : Répartition des arbres suivant leur quartier

Répartition des arbres selon leur stade de développement :

La figure 6 représente la répartition des arbres selon leur stade de développement : adulte, jeune, sénescent, et vieux. Les arbres adultes dominent largement, il s'agit d'arbres matures et en bonne santé, suivis par les arbres jeunes, ce qui montre une volonté de la commune et de l'agglomération de développer leur politique d'urbanisme en plantant de nouveaux arbres ou en remplaçant les arbres en mauvaise santé. Les arbres sénescents et vieux sont en très faible proportion, ce qui montre que la politique d'urbanisme de la ville consiste à un renouvellement des arbres avant qu'ils atteignent leur fin de vie et à maintenir la couverture arborée de la ville.

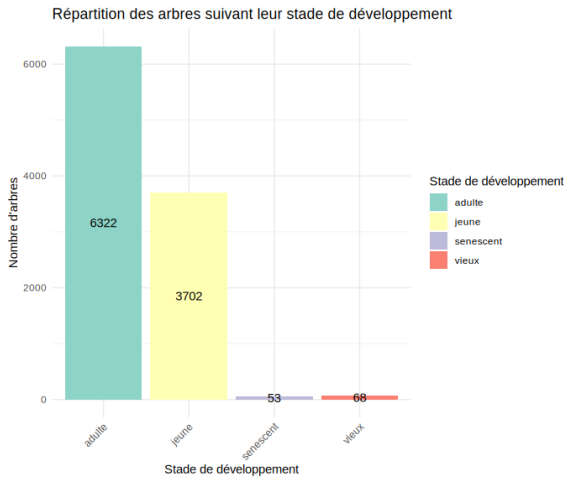

Figure 6 : Répartition des arbres suivant leur stade de développement

Ces hypothèses sont confirmées par la figure 7, présentant la répartition des arbres suivant leurs états. Nous pouvons voir qu'il y a 171 arbres qui ont été remplacés et 416 supprimés, ce qui correspond probablement à des arbres qui était sénescents ou vieux.
La figure 7 illustre la répartition des arbres en fonction de la hauteur total de l'arbre. On observe que les données suivent une loi normal centré sur 7.

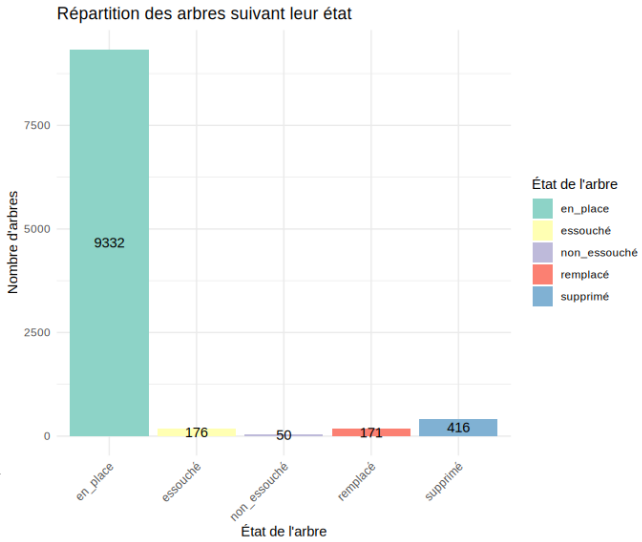

Figure 7 : répartition des arbres en fonction de la hauteur total de l'arbre

Ces hypothèses sont confirmées par la figure 7, présentant la répartition des arbres suivant leurs états. Nous pouvons voir qu'il y a 171 arbres qui ont été remplacés et 416 supprimés, ce qui correspond probablement à des arbres qui était sénescents ou vieux.

Visualisation des données sur une carte

Carte des arbres répertoriés :

Pour visualiser les arbres répertoriés, nous avons décidé de créer une carte interactive, comme illustré dans la figure 8 (voir code R). Cette carte permet de visualiser la position réelle de chaque arbre dans la ville. Les points noirs sur la carte représentent les arbres ordinaires, tandis que les points jaunes indiquent les arbres remarquables, offrant ainsi une distinction visuelle claire entre les deux catégories.

Grâce à cette carte interactive nous pouvons voir que les arbres remarquables se situent dans les différents parcs du centre-ville, avec une majorité au sein du parc des champs Élysées.

En outre, des pictogrammes d'arbres sont utilisés pour représenter les différents quartiers. Lorsque l'on clique sur ces pictogrammes avec la souris, une info-bulle apparaît, affichant le nombre total d'arbres présents dans le quartier correspondant. Cette fonctionnalité interactive aide, non seulement à comprendre la distribution spatiale des arbres, mais aussi à obtenir des informations instantanées sur la densité arboricole de chaque quartier.

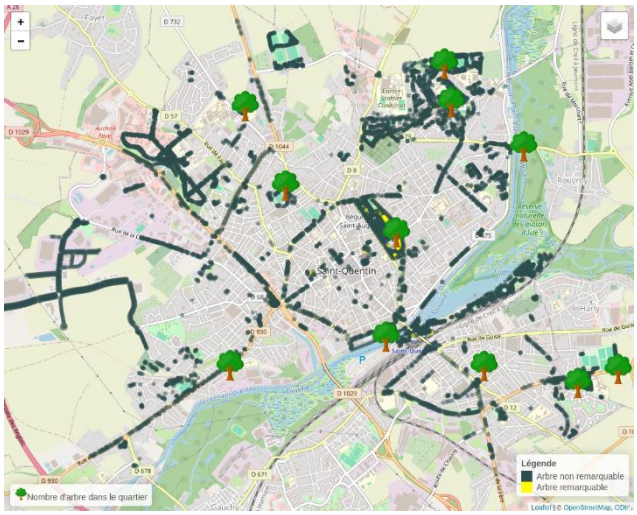

Figure 8 : Carte représentant les arbres répertoriés

Cette représentation graphique permet de mieux visualiser la répartition des arbres dans les différents quartiers et secteurs géographiques. De plus, elle permet d'identifier les zones moins arborées dans lesquels il serait possible de planter de nouveaux arbres afin de respecter la politique urbaine de la ville.

Etude des corrélations entre variables

Liens entre les variables

Le graphique est une matrice de corrélation illustrant les relations entre différentes variables associées aux arbres. Chaque point représente la corrélation entre une paire de variables, avec les couleurs et la taille indiquant la force et la direction de cette corrélation. Les corrélations positives sont représentées par des points bleus, les négatives par des points rouges. On remarque également que les corrélations semblent être proportionnelles. Par exemple, `haut_tot` et `haut_tronc` montrent que les arbres plus grands ont aussi des troncs plus hauts. Les points situés au niveau de la diagonale principale montrent la corrélation d'une variable avec elle-même qui est toujours parfaite c'est-à-dire égale à 1.

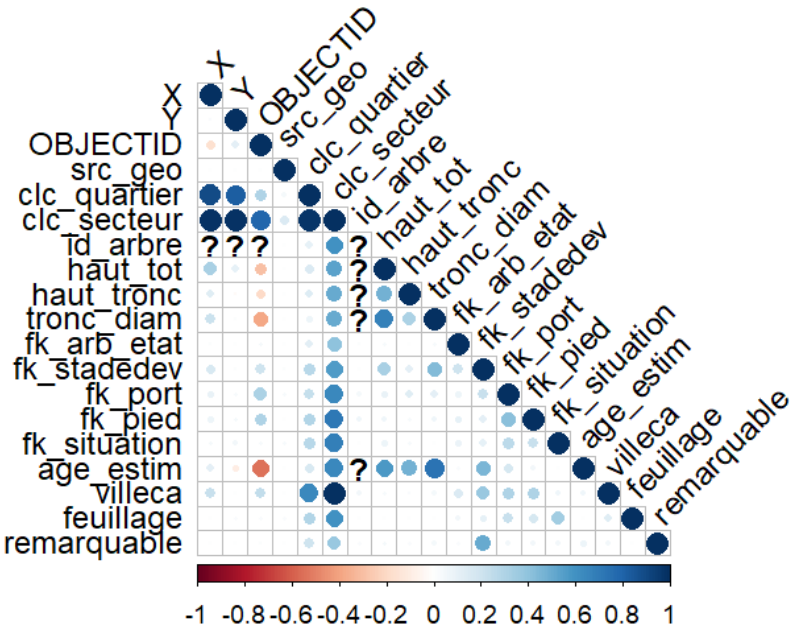

Figure 9 : Matrice de corrélation

Afin de calculer ces corrélations, nous avons utilisé trois méthodes selon le type de chaque paire :

- Paire quantitative : coefficient de Pearson

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2 \sum\left(y-\bar{y}\right)^2}}$$

- Paire qualitative : V de Cramer

$$\text{V de Cramer} = \sqrt{\frac{\text{Khi}^2}{n \times (K-1)}}$$

- Paire qualitative/quantitative : rapport de corrélation

$$\eta^2(y,x) = \frac{SCE_{inter}}{SCE_{totale}}$$

Afin de mieux appréhender les relations entre les différentes variables, nous avons choisi de réaliser des tableaux croisés et des tests d'indépendance du khi2. Les tableaux croisés offrent une représentation visuelle de l'interaction entre deux variables catégorielles, mettant en évidence des patterns et des tendances dissimulées dans les données.

Dépendance entre les variables par le khi2

On utilise le test d'indépendance du khi2 afin de vérifier si les associations présentes dans les tableaux croisés sont statistiquement concrètes. Autrement dit, cette épreuve nous permet de déterminer si les fluctuations d'une variable sont liées aux fluctuations d'une autre variable ou si elles sont le fruit du hasard.

En mettant en œuvre ces techniques, nous pouvons repérer des liens significatifs entre les variables, ce qui est crucial pour formuler des recommandations éclairées et prendre des décisions.

Le graphique est une Heatmap des p-values des tests du khi deux, montrant les relations de dépendance entre les variables catégorielles. Pour l'échelle de couleurs nous avons choisi du bleu foncé pour les p-value proche de 0 c'est à dire qu'il y a une forte dépendance entre les variables. Le rouge correspond au p-value proche de 1, marqueur d'indépendance. Enfin, le blanc correspond à la dépendance d'une variable avec elle-même. Ce graphique permet d'identifier rapidement les relations significatives entre variables catégorielles et aide à comprendre les influences mutuelles et les indépendances entre variables.

.

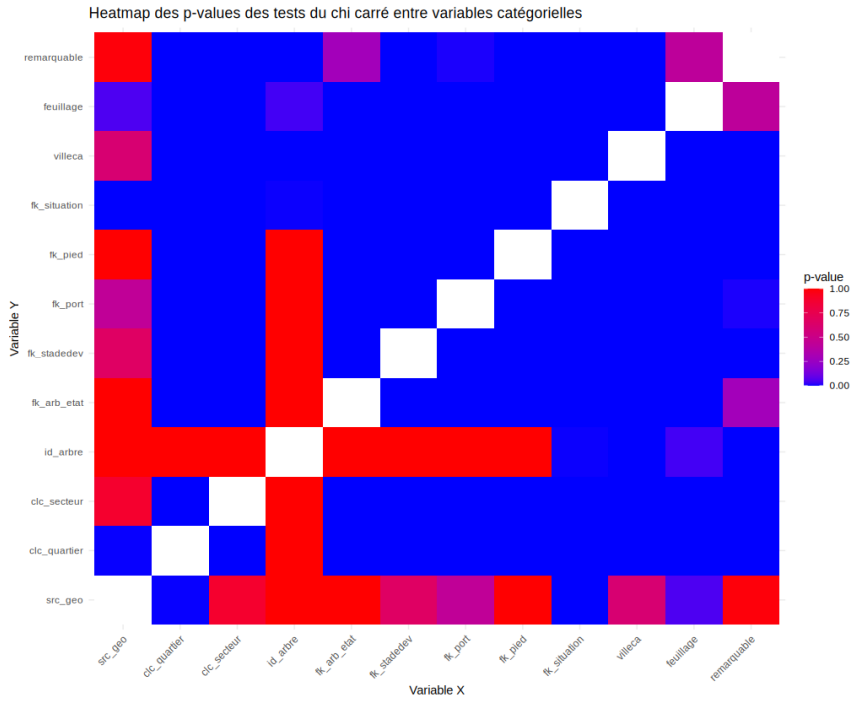

Figure 10 : Heatmap des p-values du test du khi2

Tableaux croisés

Afin de mieux appréhender les relations entre les différentes variables, nous avons choisi de réaliser des tableaux croisés et des tests d'indépendance du chi2 entre elles. On utilise le test d'indépendance du chi2 afin de vérifier si les associations présentes dans les tableaux croisés sont statistiquement concrètes. Autrement dit, cette épreuve nous permet de déterminer si les fluctuations d'une variable sont liées aux fluctuations d'une autre variable ou si elles sont le fruit du hasard. En mettant en œuvre ces techniques, nous pouvons repérer des liens significatifs entre les variables, ce qui est crucial pour formuler des recommandations éclairées et prendre des décisions.

Mosaic plot

Chaque tableau croisé nous a permis de créer un diagramme en mosaïque associé ou une Heatmap interactive (cf. clc_secteur) dont le but est d'observer deux à deux les variables qualitatives corrélées.

Ce choix nous permet de visualiser efficacement les relations entre les différentes variables catégorielles. Un Mosaicplot représente les données sous forme de mosaïques, où la taille de chaque tuile est proportionnelle à la fréquence des combinaisons de catégories. Cette représentation graphique est particulièrement utile pour détecter des patterns et des interactions complexes entre les variables, facilitant ainsi l'interprétation des résultats des tests d'indépendance du khi2.

Sur la figure 11, on voit que la ville et l'agglomération ont des stratégies de plantation similaires dans les quartiers. "rouvroy" et "quartier_saint-martin_-_oëstres" ont des proportions de plantation équivalentes, tandis que "quartier_du_faubourg_d'isle" et "quartier_remicourt" montrent plus de diversité dans les efforts. Cette répartition équitable contribue à l'égalité environnementale.

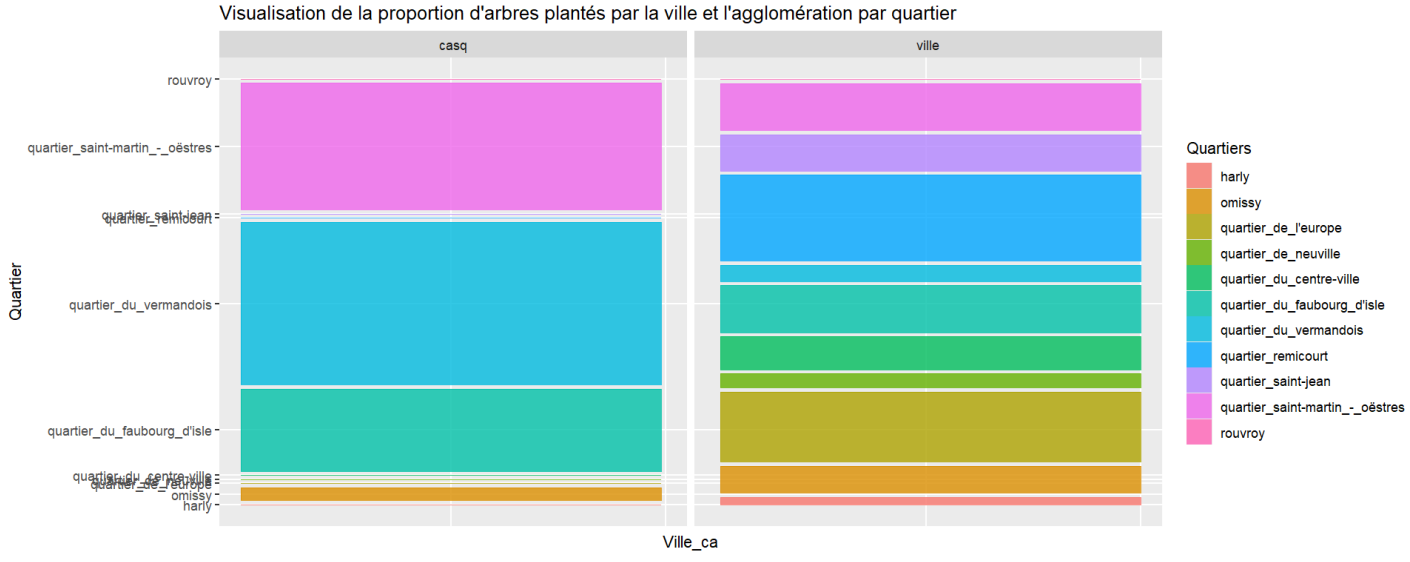

Figure 11 :Visualisation de la proportion d'arbres plantés par la ville et l'agglomération par quartier

Sur la figure 12, on observe que les arbres remarquables sont principalement adultes et sénescents, tandis que les non remarquables sont majoritairement adultes. Cela indique que les arbres adultes et sénescents sont plus souvent remarquables en raison de leur taille et âge. La gestion devrait se concentrer sur la conservation de ces arbres et encourager la croissance des jeunes vers un état remarquable.

Figure 12 : Visualisation de la proportion d'arbres remarquables par rapport

D'après la figure 13, on remarque que les quartiers "rouvroy" et "quartier_saint-martin_-_oëstres" ont principalement des arbres adultes et jeunes, tandis que "quartier_saint-jean" et "quartier_remicourt" ont une répartition équilibrée entre adultes et sénescents. "quartier_du_faubourg_d'isle" et "quartier_du_vermandois" sont dominés par des arbres adultes et sénescents. Les quartiers "harly" et "omissy" ont principalement des arbres adultes. Les quartiers avec de nombreux arbres jeunes et adultes verront leur feuillage croître, tandis que ceux avec beaucoup de sénescents nécessiteront des interventions.

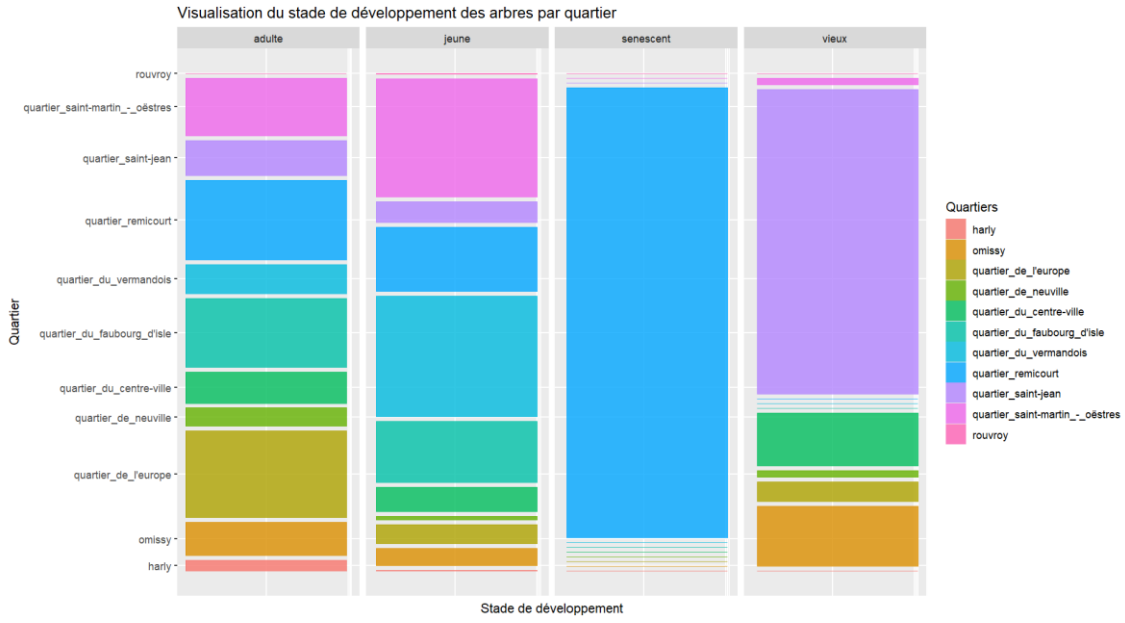

Figure 13 : Visualisation du stade de développement des arbres par quartier

Prédiction et étude de régression

Prédiction de l'âge de l'arbre :

Pour prédire l'âge de l'arbre, nous avons fait une régression linéaire avec pour paramètre les valeurs qui sont corrélées à l'âge estimé. C'est-à-dire le quartier, le secteur le stade de développement de l'arbre, comment les arbres sont taillés (fk_port), le type de sol (fk_pied) , s'il est aligné aux autres en groupe ou seul (fk_situation), la hauteur du tronc, l'objectid, la hauteur totale et le diamètre du tronc.

Sur la figure 14, nous pouvons observer une droite bleue qui correspond à l'âge estimé par rapport à la hauteur de l'arbre, on observe que plus l'arbre est grand plus l'âge estimé est élevé. Malgré tout, nous pouvons quand même observer des points qui s'éloignent particulièrement de cette droite ceux-ci est dû au fait que les arbres sont entretenus et donc tailler, leur taille totale ne reflète donc pas leurs âges, pour palier à ce type de cas nous avons aussi fait une prédiction par rapport au diamètre du tronc. (figure 15)

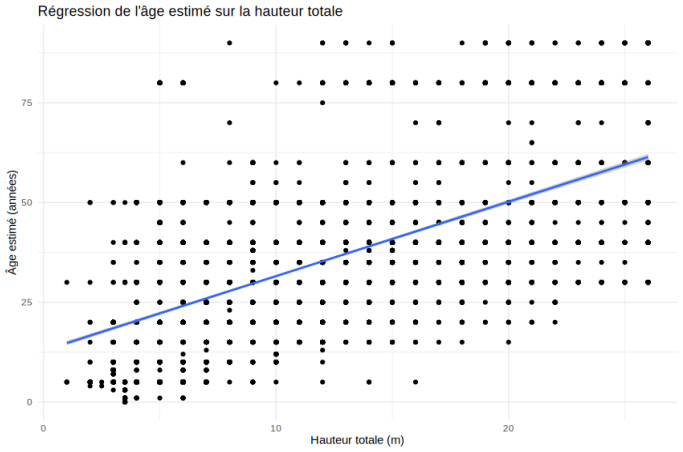

Figure 14 : Régression de l'âge sur la hauteur totale

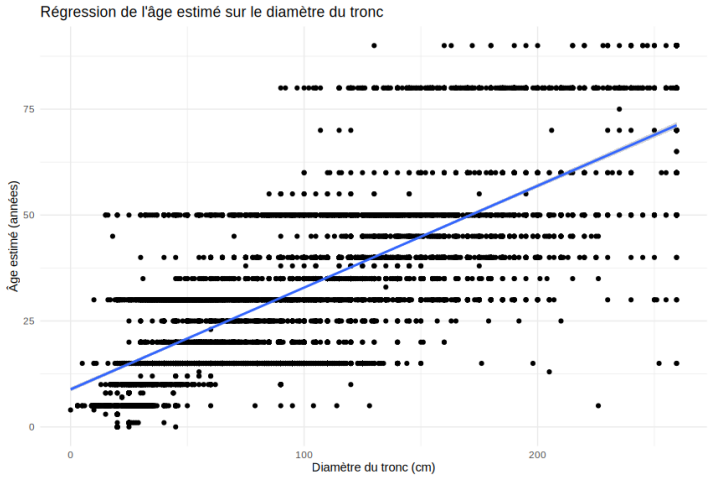

Figure 15 : âge estimé par rapport au diamètre du tronc

Régression logistique :

Afin de savoir quels sont les arbres à abattre nous devons utiliser de la régression logistique. Cela permet d'avoir un résultat binaire, oui ou non pour chaque arbre afin de savoir s'il doit être abattu.

Nous avons utilisé notre tableau de corrélation de la figure 16 afin de déterminer les différentes variables utiles pour notre régression logistique. Après analyse nous avons déterminer que les paramètres utiles étaient la hauteur du tronc, le diamètre du tronc, et le stade développement de l'arbre.

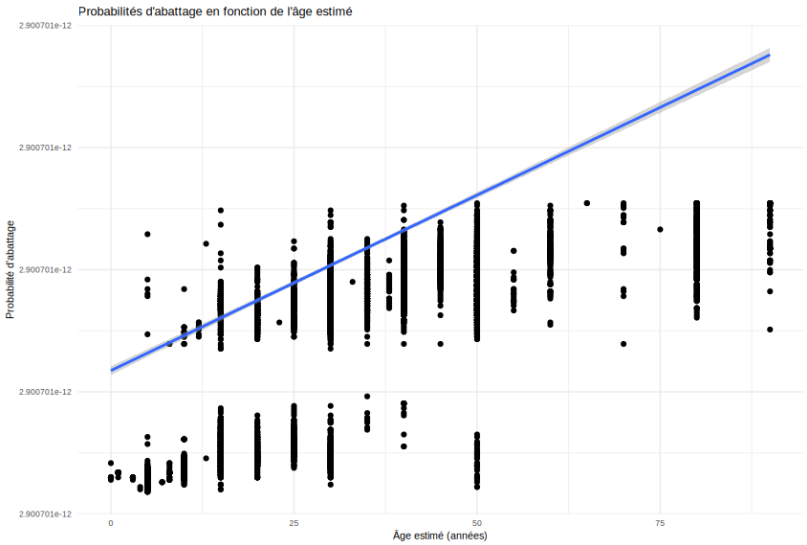

Figure 16 : Probabilités d'abattage en fonction de l'âge estimé

Malheureusement les p-values sont de 1, ce qui montre que les variables utilisées ne sont pas significatives, et le score de fisher est de 25, ce qui est très élevé pour une régression logistique.
De ce fait les prédictions de notre modèle ne possèdent pas des valeurs très élevés comme nous pouvons le voir sur la figure 17

```
glm(formula = abattu ~ haut_tronc + tronc_diam + fk_stadedev,
    family = "binomial", data = patrimoine_w)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.657e+01  1.417e+04  -0.002    0.999
haut_tronc           7.074e-15  3.481e+03   0.000    1.000
tronc_diam           7.913e-16  8.592e+01   0.000    1.000
fk_stadedevjeune    -1.997e-13  9.729e+03   0.000    1.000
fk_stadedevsenescent -6.669e-14  4.979e+04   0.000    1.000
fk_stadedevvieux    -5.508e-14  4.389e+04   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 10144  degrees of freedom
Residual deviance: 5.8857e-08  on 10139  degrees of freedom
AIC: 12

Number of Fisher Scoring iterations: 25
```

Figure 17 : Retour de la fonction de regression logistique

Conclusion :

Le projet visait à concevoir une application pour l'étude du patrimoine arboré de la ville de Saint-Quentin. Les objectifs incluaient l'extraction, la visualisation, et le nettoyage des données, ainsi que l'application de modèles statistiques.

Nous avons commencé par un diagramme de Gantt pour répartir les tâches. Le jeu de données initial, issu du patrimoine arboré de Saint-Quentin, a été exploré et nettoyé. Les données inutiles et les valeurs aberrantes ont été supprimées, et les valeurs manquantes ont été imputées.

Les visualisations montrent une répartition des arbres selon leur situation, quartier, et stade de développement. La majorité des arbres sont alignés le long des rues, avec une forte présence dans les quartiers de Saint-Martin - Oëstres et Remicourt. Les arbres adultes dominent, suivis par les jeunes, ce qui reflète une politique de renouvellement arboré proactive.

Les corrélations ont été étudiées pour comprendre les relations entre variables. Les analyses ont révélé des dépendances significatives, notamment entre la hauteur totale et la hauteur du tronc des arbres. Les arbres remarquables sont majoritairement adultes et sénescents.

La carte interactive permet de visualiser la distribution des arbres, avec une concentration d'arbres remarquables dans les parcs centraux. Les stratégies de plantation de la ville et de l'agglomération montrent une répartition équitable.

La régression logistique, utilisée pour prédire les arbres à abattre, n'a pas produit des résultats significatifs en raison des p-values élevées et d'un score de Fisher élevé.

En conclusion, ce projet a permis de développer une compréhension détaillée du patrimoine arboré de Saint-Quentin. Les résultats fournissent une base solide pour la planification et la gestion des arbres, bien que des modèles prédictifs plus robustes soient nécessaires pour améliorer la précision des recommandations d'abattage.