

Big Data :

Exploration et Analyse du Patrimoine Arboré de la ville de Saint-Quentin

SOMMAIRE

<u>I- Introduction</u>	p.3
<u>II- Fonctionnalité 1 : Description et exploration des données</u>	p.4
a) Description du jeu de données	p.4
c) Statistiques descriptives univariées, bivariées	p.5
e) Nettoyage des données	p.7
<u>III- Fonctionnalité 2 : Visualisation des données sur des graphiques</u>	p.8
<u>IV- Fonctionnalité 3 : Visualisation des données sur une carte</u>	p.16
a) Cartes des arbres répertoriés	p.16
b) Carte de la quantité d'arbres par quartier et secteurs	p.17
<u>V- Fonctionnalité 4 : Etude des corrélations entre variables</u>	p.20
a) Quels sont les liens entre les variables ?	p.20
b) Conduire des analyses bivariées	p.23
c) Etude des relations entre variables qualitatives	p.24
i - tableau croisés et tests d'indépendance	p.24
ii- graphiques des tableaux	p.26
<u>VI- Fonctionnalité 5 : Etude des corrélations entre variables</u>	p.27
a) Plantation des arbres à Saint-Quentin	p.27
b) Prédiction de la variable âge	p.28
c) Abattage des arbres à Saint-Quentin	
<u>VII- Conclusion</u>	

I- Introduction

Dans le cadre de notre projet, nous nous lançons dans la conception et le développement d'une application dédiée à l'étude approfondie du patrimoine arboré. Nous avons pour objectif de concevoir une application exhaustive qui permettra d'explorer et d'analyser le patrimoine arboré de la ville de Saint-Quentin (Aisne). Notre projet est de mettre en place une approche pluridisciplinaire qui combine des méthodes de collecte, de traitement et de visualisation de données afin de fournir des données précieuses sur les arbres de la ville.

Dans cette première semaine du projet, nous entamons notre exploration par la partie Big Data, plongeant ainsi dans l'analyse approfondie des données. Afin d'obtenir une vision complète du patrimoine arboré, nous allons collecter des données à partir de différentes sources, comme des fichiers CSV issues du site data.gouv, des bases de données publiques ou des API en ligne. Nous procéderons à une analyse minutieuse des données afin de repérer et rectifier les valeurs manquantes, les doublons ou les éventuelles erreurs, assurant ainsi la fiabilité des informations utilisées dans notre application. Nous développerons des fonctionnalités avancées comme la visualisation interactive des données, l'utilisation de modèles statistiques et la création de rapports détaillés afin de faciliter une exploration approfondie du patrimoine arboré. Notre objectif est de fournir aux autorités locales et aux décideurs des renseignements précieux pour une gestion efficace et durable du patrimoine arboré, ce qui contribue à la préservation de l'environnement urbain.

PLANIFICATION PROJET

SEMAINE : BIG DATA : DU 03/06 au 07/06				
Lundi 03/06/2024	Mardi 04/06/2024	Mercredi 05/06/2024	Jeudi 06/06/2024	Vendredi 07/06/2024
16h30 : présentation de l'avancement du groupe OBJECTIF : bien comprendre le sujet, se l'approprier : VALIDE	10h30 : présentation de l'avancement du groupe Agathe : Fonctionnalité 1 et 2 Clémence : nettoyage des données Mélissa : statistiques descriptives univariées, bivariées Agathe : description du jeu de données	10h30 : check up Agathe : Fonctionnalité 1 et 2 Clémence : nettoyage des données Mélissa : aide sur le nettoyage des données Agathe : aide sur le nettoyage des données Mélissa : réalisation des graphiques	10h30 : check up Mélissa : fin de fonctionnalité 4 Clémence : Liens entre les variables Mélissa : tableaux croisés dynamique Agathe : carte	10h30 : check up Mélissa : fin de fonctionnalité 5 + 6 Clémence : rapport Mélissa : fonction S Agathe : finition
16h30 : visualisation de l'avancement du groupe OBJECTIF : bien comprendre le sujet, se l'approprier : VALIDE	10h30 : check up Agathe : Fonctionnalité 2 Mélissa : liaison des variables (fonctionnalité 2) Clémence : commencement de la carte (fonctionnalité 2) Agathe : réalisation des histogrammes	10h30 : check up Agathe : Fonctionnalité 3-4 Clémence : fin de feuille des relations Mélissa : idem Agathe : finition carte	10h30 : check up Agathe : fin de fonctionnalité 5 Clémence : suite Mélissa : suite Agathe : suite	10h30 : check up final DEFINITION ISH35 PRÉSENTATION ORALE
16h30 : réunion d'avancement OBJECTIF : finir fonctionnalité 1 et 2 : VALIDE	16h30 : réunion d'avancement OBJECTIF : finir fonctionnalité 3 : VALIDE	16h30 : réunion d'avancement OBJECTIF : finir fonctionnalité 3 et 4 : NON VALIDE	16h30 : réunion d'avancement travail chez soi à faire, vendredi matin jusqu'à 15h35	
REALISATION DU RAPPORT				

II- Fonctionnalité 1 : Description et exploration des données

Dans cette première partie du rapport, l'accent est mis sur la description approfondie du jeu de données relatif au patrimoine arboré de Saint-Quentin. Elle inclut une étude statistique descriptive univariée et bivariée, qui met en évidence la répartition et les liens entre les diverses variables. En outre, une attention particulière est accordée au nettoyage des données, qui comprend l'identification et le traitement des valeurs manquantes, des valeurs aberrantes et des doublons, dans le but de garantir la qualité et la fiabilité des données pour les analyses futures.

La procédure a débuté par l'extraction des données d'un fichier CSV, suivi de leur visualisation sur Excel avec une séparation des colonnes. Après une brève analyse préliminaire, les données ont été importées dans R à l'aide de la librairie "readr". Les fonctions R telles que "summary" et "str" ont permis d'obtenir un premier aperçu du type de données présent dans le fichier. Avant le processus de nettoyage, le jeu de données se compose de 11 421 lignes et 37 colonnes.

a) description du jeu de données

Voici un aperçu détaillé des différentes colonnes présentes dans le jeu de données :

- **X et Y** : Correspondent aux coordonnées de l'arbre. X pour l'axe Est-Ouest et Y pour l'axe Nord-Sud. Ils sont initialement considérés comme une chaîne de caractères en raison des guillemets.
- **OBJECTID** : Identifiant unique de l'arbre dans la base de données.
- **created_date** : Date et heure de création de l'enregistrement.
- **created_user** : Utilisateur ayant créé l'enregistrement.
- **src_geo** : Source d'information géographique des données.
- **clc_quartier** : Quartier de la ville où se trouve l'arbre.
- **clc_secteur** : Secteur de la ville où se trouve l'arbre.
- **id_arbre** : Identifiant de l'arbre.
- **haut_tot** : Hauteur totale de l'arbre (tronc et feuillage), initialement considéré comme chaîne de caractère.
- **haut_tronc** : Hauteur du tronc de l'arbre, initialement considéré comme chaîne .
- **tronc_diam** : Diamètre du tronc de l'arbre.
- **fk_arb_etat** : État de l'arbre (abattu, en place).
- **fk_stadeDev** : Stade de développement de l'arbre.
- **fk_port** : Port de l'arbre (forme et structure).
- **fk_pied** : Type de sol au pied de l'arbre.
- **fk_situation** : Situation de l'arbre (en alignement, isolé).
- **fk_revetement** : Revêtement autour de l'arbre.
- **commentaire_environnement** : Commentaires sur l'environnement autour de l'arbre.
- **dte_plantation** : Date de plantation de l'arbre.
- **age_estim** : Âge estimé de l'arbre.
- **fk_prec_estim** : Précision de l'estimation de l'âge.
- **clc_nbr_diag** : Nombre de diagnostics effectués sur l'arbre.
- **dte_abattage** : Date d'abattage de l'arbre si abattu.
- **fk_nomtech** : Nom technique de l'arbre.
- **last_edited_user** : Utilisateur ayant effectué la dernière modification.
- **last_edited_date** : Date et heure de la dernière modification.

- **villeca** : Positionnement de l'arbre (ville ou agglomération)
- **nomfrancais** : Nom français de l'arbre.
- **nomlatin** : Nom latin de l'arbre.
- **GlobalID** : Identifiant global unique de l'enregistrement.
- **CreationDate** : Date de création de l'enregistrement.
- **Creator** : Créateur de l'enregistrement.
- **EditDate** : Date de la dernière modification.
- **Editor** : Éditeur de la dernière modification.
- **feuillage** : Type de feuillage de l'arbre.
- **remarquable** : Indique si l'arbre est remarquable ou non.

Certaines colonnes ont été jugées inutiles pour la suite du projet et ont été supprimées à l'aide de la fonction "select".

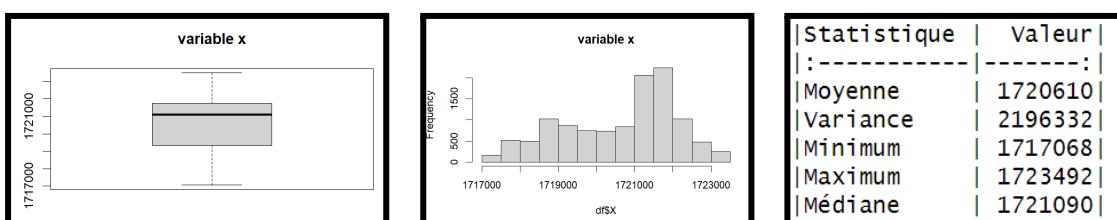
b) Statistiques descriptives univariées, bivariées

Avant de commencer l'analyse, la signification des termes "statistiques descriptives univariées" et "bivariées" a été recherchée pour une meilleure compréhension.

Pour les statistiques descriptives univariées, les mesures telles que la moyenne, la médiane, l'écart-type, le minimum et le maximum ont été calculées pour chaque variable numérique. Les histogrammes ont été utilisés pour visualiser la distribution des valeurs et les fréquences ont été examinées pour les variables catégorielles.

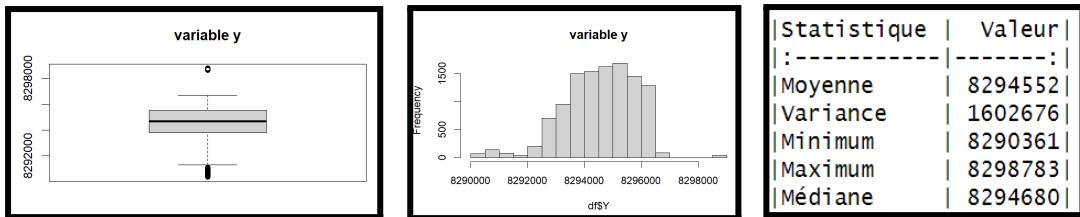
Pour les statistiques descriptives bivariées, la matrice de corrélation a été analysée pour identifier les relations linéaires entre les variables numériques. Des graphiques de dispersion ont été utilisés pour visualiser la relation entre deux variables numériques, et des tests statistiques comme le coefficient de corrélation de Pearson ont été effectués pour évaluer la relation entre deux variables.

Statistiques descriptives univariées:



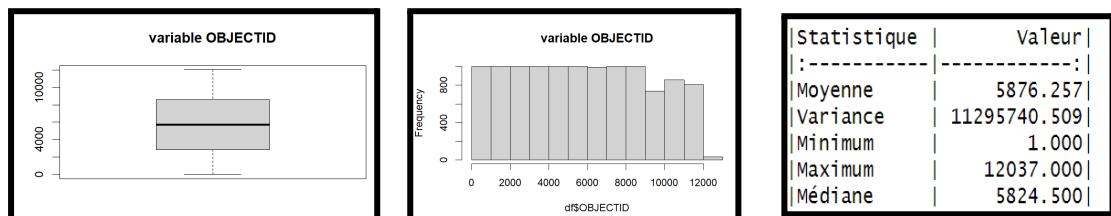
interprétation : boîte à moustaches n'est pas symétrique, suggérant un biais dans les données. Cela indique également une répartition non uniforme des valeurs. Cette asymétrie est notamment visible grâce aux valeurs de la moyenne et médiane qui n'est pas identique.

interprétation : données avec un centre proche de 1722000 et une dispersion comprise approximativement entre 1717000 et 1723497.



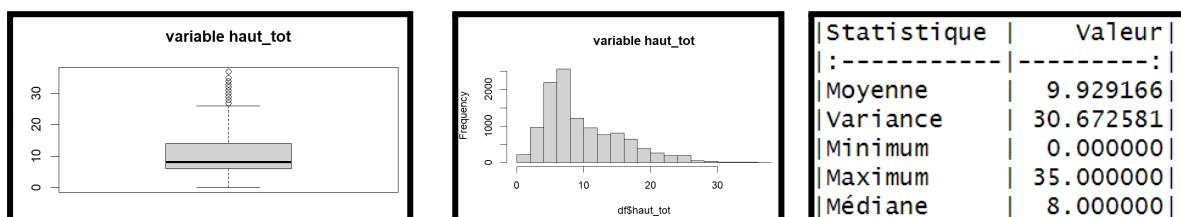
interprétation : boîte à moustaches est symétrique, nos données suivent une distribution normale.

interprétation: presque symétrique visible grâce à la boîte de moustache mais aussi grâce à la moyenne et médiane qui est presque similaire.



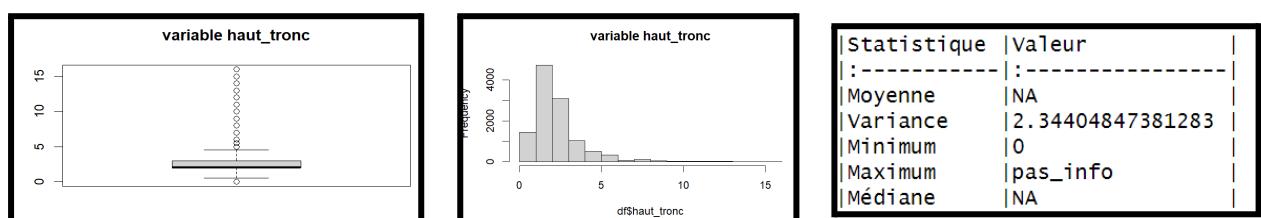
interprétation : symétrique visible grâce à la boîte à moustache et les valeurs de la moyenne et médiane.

interprétation : uniformité partielle



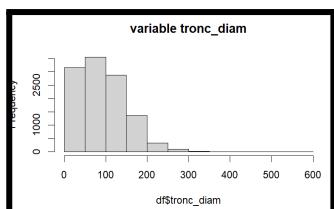
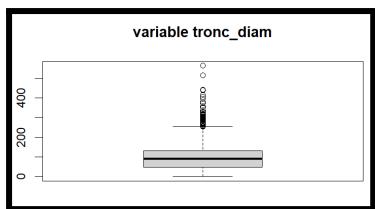
interprétation : asymétrique vers la gauche

interprétation : plage de donnée tourne autour de 8-9



interprétation : asymétrique vers la gauche

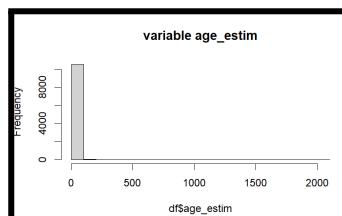
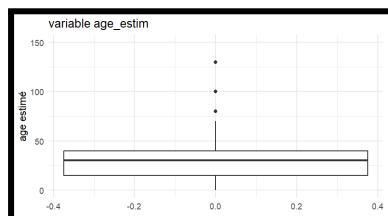
interprétation : plage de donnée tourne autour de 1-2



Statistique	Valeur
Moyenne	89.5028
Variance	2540.7737
Minimum	0.0000
Maximum	212.0000
Médiane	85.0000

interprétation : asymétrique vers la gauche

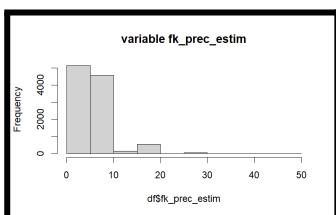
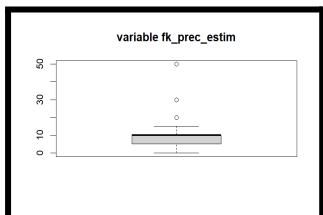
interprétation : plage de donnée tourne autour de 0-150 en majorité



Statistique	Valeur
Moyenne	29.46226
Variance	321.93413
Minimum	0.00000
Maximum	130.00000
Médiane	30.00000

interprétation : symétrique

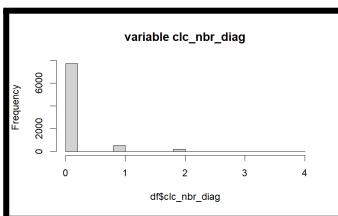
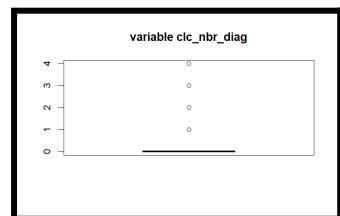
interprétation : valeurs tournent autour de 30



Statistique	Valeur
Moyenne	7.146419
Variance	19.245602
Minimum	0.000000
Maximum	30.000000
Médiane	5.000000

interprétation : asymétrique vers la gauche

interprétation : plage de donnée compris en majorité entre 0-10



Statistique	Valeur
Moyenne	0.0930605
Variance	0.1292825
Minimum	0.0000000
Maximum	4.0000000
Médiane	0.0000000

interprétation : plage de donnée tourne autour de 0

c) Nettoyage des données

Il est crucial de trier rigoureusement un fichier contenant plus de 11 000 données afin d'assurer la pertinence et le réalisme des informations collectées. Ce fichier, rempli manuellement par les personnes étudiant les arbres dans la ville de Saint-Quentin, peut comporter des erreurs telles que des cases vides ou mal renseignées en raison de fautes de frappe. L'objectif a été de trier ces données pour ne conserver que celles utilisables pour les analyses.

La première étape a consisté à homogénéiser le fichier en mettant les titres des colonnes en majuscules et les valeurs en minuscules. De plus, il a été observé que l'importation en

format CSV2 transformait certaines valeurs numériques en caractères. Il a donc été nécessaire de convertir ces valeurs en numérique pour faciliter leur traitement. Certaines données étaient écrites de différentes manières, comme "orthophoto", "ortho", et "plan ortho", bien qu'elles aient la même signification. Pour homogénéiser l'ensemble, toutes ces variations ont été uniformisées sous un même libellé.

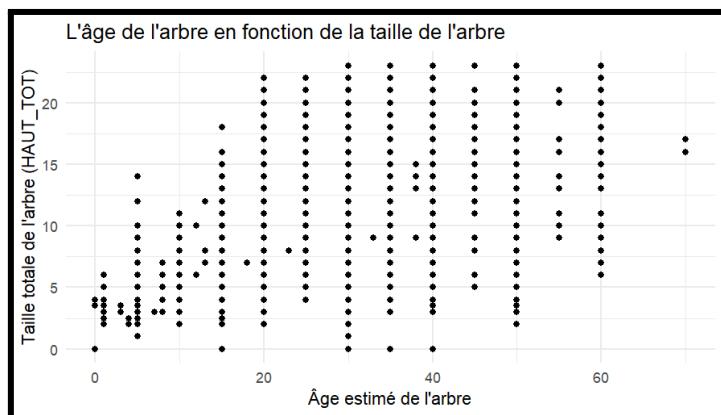
Les lignes comportant plus de six informations manquantes, c'est-à-dire avec plus de six cases vides, ont été supprimées. Cette étape a été nécessaire car, après avoir éliminé les colonnes jugées inutiles, il fallait conserver suffisamment d'informations sur chaque arbre pour optimiser les résultats. De plus, deux lignes sans coordonnées ont été supprimées.

Pour traiter les valeurs aberrantes, deux méthodes différentes mais aboutissant aux mêmes résultats ont été utilisées : La méthode MAD et la méthode interquartiles. La méthode MAD a été retenue pour éliminer les lignes comportant des valeurs aberrantes. Certains préfèrent remplacer ces valeurs par une moyenne, mais il a été jugé inutile de conserver ces lignes en les remplaçant par des valeurs moyennes. La méthode MAD fonctionne en définissant des bornes inférieure et supérieure pour détecter les valeurs aberrantes strictes, basées sur un calcul de la médiane de la colonne, en ignorant les valeurs manquantes.

Malgré le traitement et le nettoyage du fichier, certaines cases restent vides. Il a été observé que certaines cases contiennent la mention "à renseigner". Par conséquent, il a été décidé de remplir toutes les cases vides avec cette phrase, facilitant ainsi la tâche des personnes souhaitant compléter les données du tableau.

Après ces modifications, il est devenu plus facile de détecter la présence de doublons. En utilisant la fonction "duplicated", il a été constaté que le fichier ne contenait aucun doublon.

III- Fonctionnalité 2 : Visualisation des données sur des graphiques



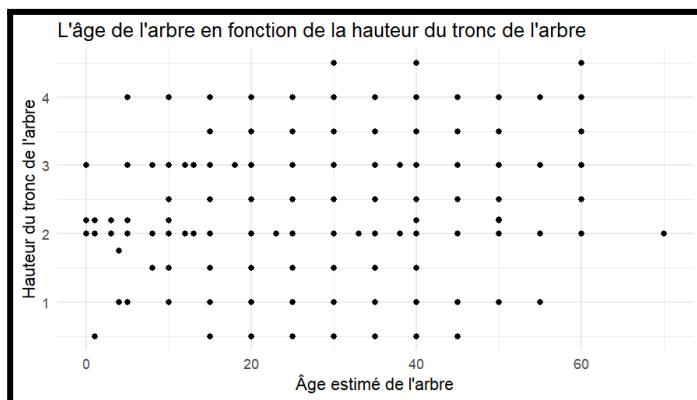
Le graphique présente la relation entre l'âge estimé des arbres et leur taille totale. Chaque point sur le graphique représente un arbre individuel de l'ensemble de données.

A mesure que l'âge estimé des arbres augmente, leur taille totale tend également à augmenter. Cette relation positive suggère que, dans la plupart des cas, les arbres plus âgés sont plus grands.

Cependant, il y a une certaine dispersion des points autour de cette tendance. Cela signifie que l'âge de l'arbre n'explique pas entièrement la variation de la taille des arbres. Plusieurs facteurs pourraient contribuer à cette dispersion, tels que les variations dans les espèces d'arbres, les conditions de croissance, les pratiques de gestion forestière, ou des événements environnementaux passés.

Les valeurs aberrantes ont été filtrées, ce qui permet d'obtenir une représentation plus claire et précise des données. Cela signifie que les points extrêmement éloignés de la tendance générale, qui pourraient fausser l'analyse, ont été supprimés pour améliorer la lisibilité du graphique.

Cette relation entre l'âge et la taille est importante pour plusieurs raisons . Comprendre cette relation peut aider les gestionnaires forestiers à prévoir la croissance future des arbres et à prendre des décisions éclairées concernant l'abattage et la plantation. Les arbres plus grands et plus âgés peuvent avoir une plus grande valeur écologique, fournissant des habitats importants pour la faune et jouant un rôle clé dans l'écosystème forestier.



Le graphique montre la relation entre l'âge estimé des arbres (AGE_ESTIM) et la hauteur du tronc des arbres (HAUT_TRONC).

Avant de tracer le graphique, les indices des valeurs aberrantes pour HAUT_TRONC et AGE_ESTIM ont été combinés pour obtenir une liste unique d'indices aberrants. Les valeurs aberrantes combinées ont été supprimées du jeu de données original.

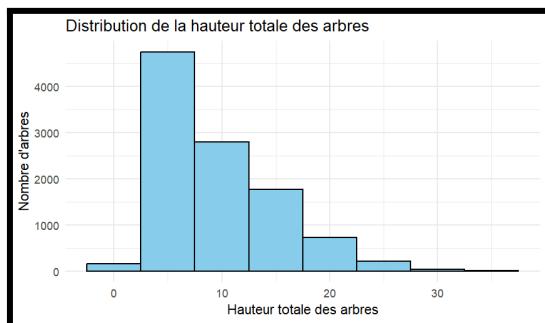
Le graphique est créé en utilisant ggplot2 avec les données filtrées (df_clean_no_outliers).

Le graphique permet de visualiser la relation entre l'âge estimé et la hauteur du tronc des arbres. Une tendance générale peut être observée où la hauteur du tronc augmente avec l'âge de l'arbre, suggérant une corrélation positive entre ces deux variables.

Bien qu'il y ait une tendance générale à l'augmentation de la hauteur du tronc avec l'âge, les points sont dispersés, ce qui indique que l'âge de l'arbre n'explique pas entièrement la variation de la hauteur du tronc.

Cette dispersion peut être due à plusieurs facteurs comme cité au-dessus.

En filtrant les valeurs aberrantes, le graphique devient plus clair et moins influencé par des valeurs extrêmes qui pourraient fausser l'analyse. Cela permet une interprétation plus précise des relations entre les variables.



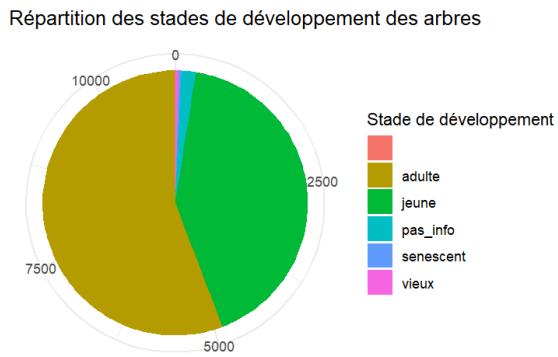
L'histogramme de la hauteur totale des arbres peut fournir plusieurs informations utiles. Comme la distribution de la hauteur des arbres permettant d'observer la répartition de la hauteur des arbres. Cela donne une idée de la diversité des hauteurs des arbres présents dans la région étudiée. De plus, la hauteur moyenne ainsi que la médiane des arbres peut être observée. L'écart entre les valeurs extrêmes et la concentration de valeurs autour de la moyenne peut indiquer la variabilité des hauteurs des arbres. Une distribution large et aplatie peut indiquer une grande variabilité, tandis qu'une distribution étroite et haute peut indiquer une faible variabilité.

interprétation : La distribution de la hauteur des arbres dans la région est inclinée vers la gauche, cela signifie que la majorité des arbres ont une hauteur relativement faible, tandis que quelques-uns ont des hauteurs très élevées.

La plupart des arbres dans la région sont probablement de petite taille. Cela peut être dû à plusieurs facteurs, tels que la jeunesse des arbres, des conditions de croissance défavorables, ou des pratiques de gestion spécifiques.

Bien que la majorité des arbres soient de petite taille, il existe néanmoins quelques arbres exceptionnellement grands dans la région. Cela peut indiquer la présence d'arbres matures, d'espèces d'arbres spécifiques qui poussent à de grandes hauteurs, ou des conditions de croissance favorables pour certains individus.

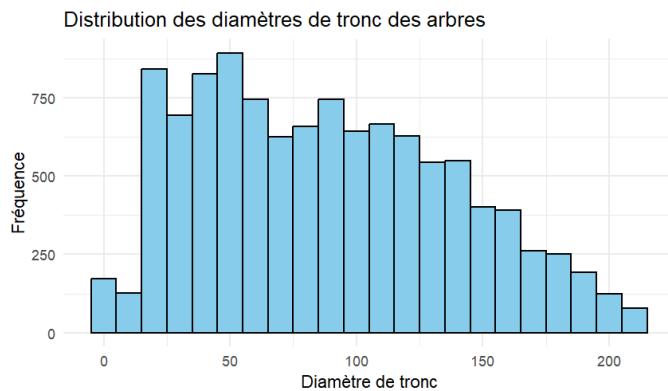
De plus, l'inclinaison vers la gauche de la distribution suggère une asymétrie positive, ce qui signifie que la queue de la distribution est plus étirée du côté gauche. Cela peut indiquer une concentration de valeurs plus basses et une dispersion plus large des valeurs plus élevées.



Ce diagramme à secteurs illustre la répartition des stades de développement des arbres dans notre ensemble de données. Chaque portion du cercle représente un stade de développement spécifique, et sa taille correspond au nombre d'arbres dans ce stade.

En l'observant, nous pouvons constater que le stade de développement le plus représenté est celui où les arbres sont à un stade intermédiaire de croissance.

Cette visualisation permet une comparaison rapide des proportions relatives des différents stades de développement des arbres. Cependant, il est important de noter que les informations sur les valeurs exactes ne sont pas directement fournies par ce type de graphique, mais il est utile pour obtenir une vue d'ensemble de la distribution des stades de développement.

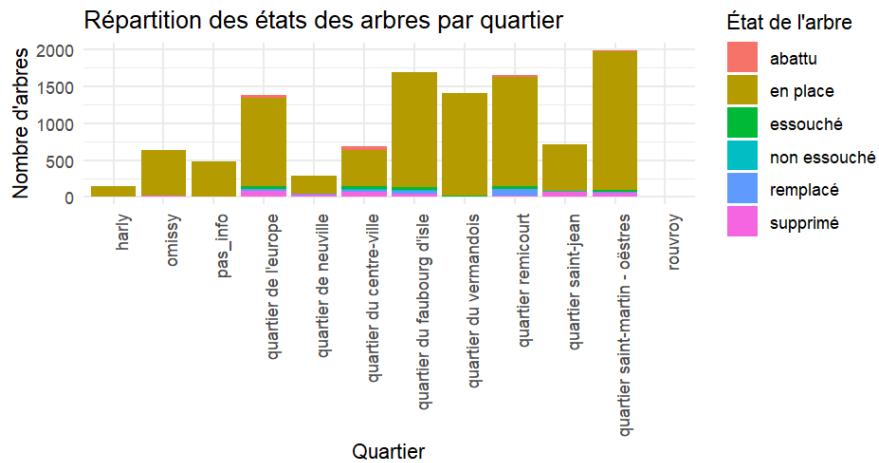


L'histogramme de la distribution des diamètres des troncs d'arbre représente la répartition des diamètres dans votre échantillon. La distribution des diamètres des troncs d'arbre semble être asymétrique et légèrement inclinée vers la droite. Cela suggère que la majorité des arbres ont des diamètres plus petits, tandis qu'il y a quelques arbres avec des diamètres plus grands.

La classe modale, c'est-à-dire l'intervalle avec la fréquence la plus élevée, semble se situer autour de 20 à 50 unités sur l'axe des diamètres de tronc. Cela indique que la plupart des arbres dans votre échantillon ont des diamètres de tronc dans cette plage.

Quant à la distribution, elle semble relativement étroite, ce qui suggère que les diamètres des troncs des arbres ne varient pas énormément dans votre échantillon. Cependant, il y a

une queue de distribution du côté des diamètres plus grands, indiquant la présence de quelques arbres avec des diamètres exceptionnellement grands.



Ce graphique en barres empilées présente la répartition des différents états des arbres selon les quartiers. Chaque barre représente un quartier spécifique, et les segments colorés à l'intérieur de chaque barre reflètent la répartition des états des arbres dans ce quartier.

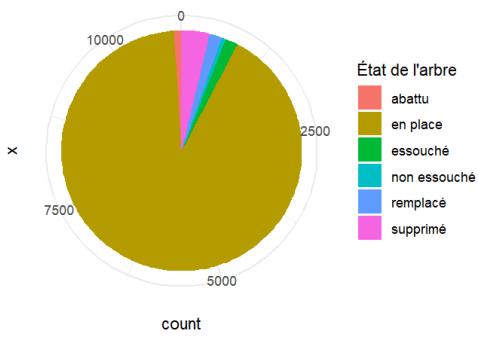
En examinant le graphique, plusieurs observations peuvent être faites :

La répartition par quartier est clairement visible. Chaque barre représente le nombre total d'arbres dans un quartier donné, permettant une comparaison facile entre les quartiers.

Chaque barre est composée de segments colorés représentant les différents états des arbres, tels que enraciné, en bonne santé, endommagé, etc. La hauteur totale de chaque barre indique le nombre total d'arbres dans le quartier, tandis que les segments empilés à l'intérieur montrent comment ces arbres sont répartis entre les différents états.

En comparant les hauteurs des segments colorés à l'intérieur de chaque barre, nous pouvons observer les variations dans la répartition des états des arbres entre les quartiers. Certains quartiers peuvent présenter une prédominance d'arbres en bonne santé, tandis que d'autres peuvent avoir une répartition plus équilibrée entre les différents états ou une concentration d'arbres dans un état spécifique.

Répartition des états des arbres

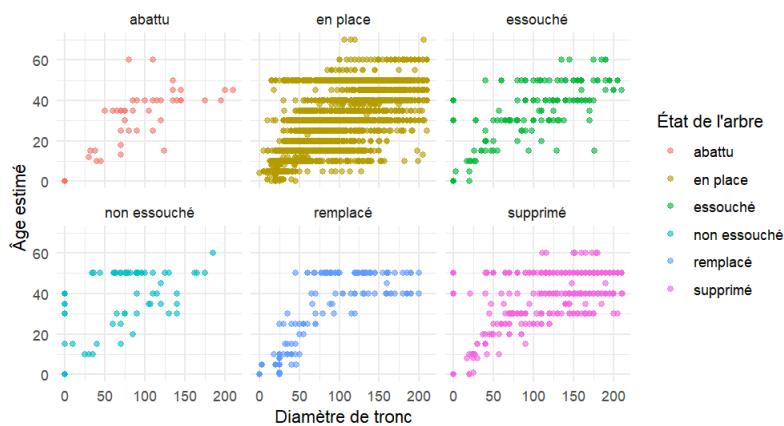


Ce diagramme à secteurs illustre la répartition des divers états des arbres dans notre ensemble de données. Chaque tranche du cercle représente un état spécifique des arbres, sa taille correspondant à la proportion d'arbres dans cet état par rapport au total.

En l'observant, nous pouvons discerner les proportions relatives des différents états des arbres. Les couleurs distinctes désignent les divers états de l'arbre, et la taille de chaque portion reflète la proportion de chaque état parmi tous les arbres de l'ensemble de données.

Cette visualisation permet de saisir rapidement la distribution des états des arbres et de comparer visuellement la prévalence de chaque état. Toutefois, il convient de noter que ce type de graphique ne fournit pas de détails précis sur les valeurs numériques spécifiques de chaque état. Il offre plutôt une vue d'ensemble utile de la répartition des états des arbres.

Relation entre l'âge estimé et le diamètre du tronc, par état de l'arbre

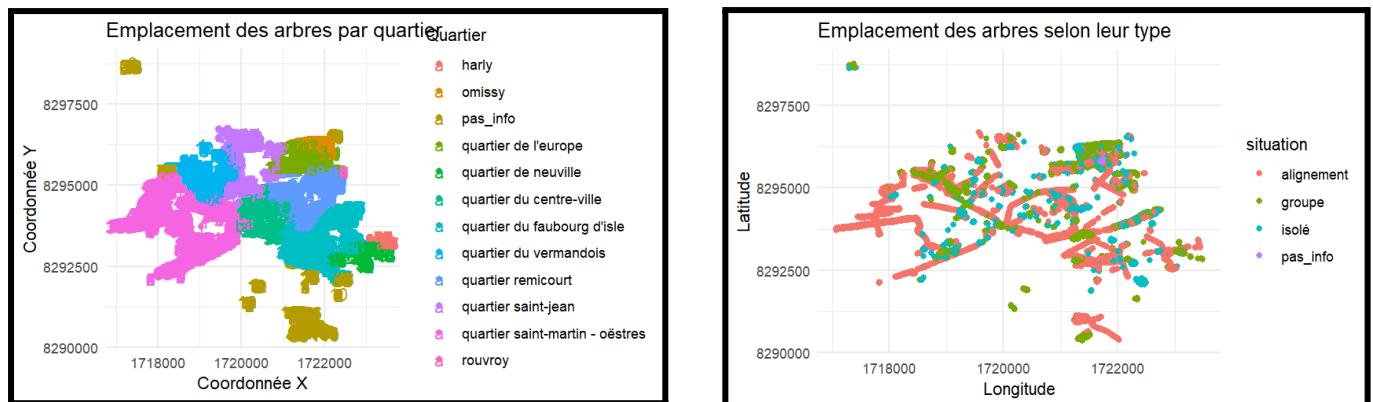


Ce graphique présente la relation entre l'âge estimé des arbres et le diamètre de leur tronc, segmentée par l'état de chaque arbre. Chaque point représente un arbre individuel, avec sa position sur le graphique déterminée par son âge estimé (sur l'axe des ordonnées) et le diamètre de son tronc (sur l'axe des abscisses).

En observant les différentes facettes du graphique, correspondant à différents états d'arbre, on peut remarquer certaines correspondances. Une corrélation est visible, les arbres plus âgés ont tendance à avoir un diamètre de tronc plus grand. Cela est visible par l'inclinaison globalement positive des points dans chaque facette, suggérant une corrélation positive entre l'âge estimé et le diamètre du tronc. De plus, les différentes couleurs représentent les

différents états des arbres. Les points sont répartis de manière variée dans chaque facette, ce qui indique des différences dans la relation âge-diamètre selon l'état de l'arbre. Par exemple, certains états peuvent montrer une corrélation plus forte entre l'âge et le diamètre que d'autres.

Enfin, les axes sont adaptés pour mieux visualiser les données, en limitant les valeurs min et max des diamètres de tronc et des âges estimés aux valeurs observées dans les données filtrées. Cela permet une représentation plus précise de la relation entre l'âge et le diamètre du tronc, en excluant les valeurs aberrantes qui pourraient fausser l'interprétation.



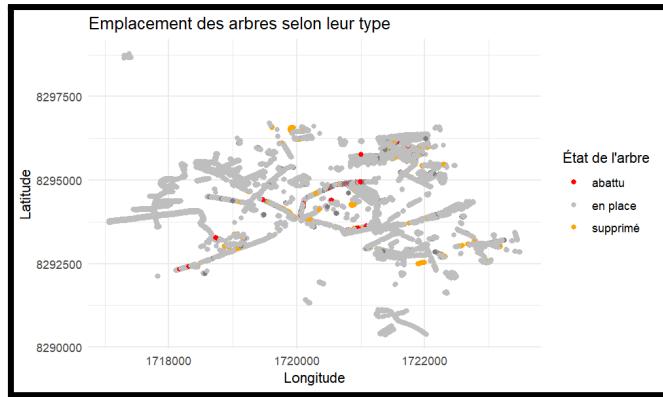
Ce graphique présente l'emplacement des arbres par quartier, avec la couleur des points indiquant le quartier dans lequel chaque arbre est situé. Chaque point sur le graphique représente un arbre individuel, positionné en fonction de ses coordonnées de latitude et de longitude.

En observant le graphique, plusieurs observations peuvent être faites :

La distribution spatiale des points révèle la répartition géographique des arbres dans la région étudiée. Certains quartiers peuvent présenter une densité plus élevée d'arbres que d'autres, ce qui se traduit par des concentrations plus élevées de points dans certaines régions.

La couleur des points permet de distinguer visuellement les quartiers. Chaque quartier est représenté par une couleur différente, ce qui facilite la visualisation de la répartition géographique des différents types d'arbres à travers les quartiers.

En examinant la dispersion des points, il est possible d'observer la diversité des types d'arbres dans la région. Certains quartiers peuvent avoir une variété plus grande d'espèces d'arbres, tandis que d'autres peuvent être dominés par un type particulier.



Ce graphique représente l'emplacement des arbres en fonction de leur état, avec chaque point sur le graphique correspondant à un arbre individuel. La couleur des points indique l'état de chaque arbre, avec une légende pour les différents états.

En observant le graphique, plusieurs observations peuvent être faites :

Les points sont répartis sur toute la zone, montrant la répartition géographique des arbres selon leur état. Certains secteurs peuvent présenter une concentration plus élevée d'arbres dans un état particulier, tandis que d'autres secteurs peuvent être plus diversifiés en termes d'états d'arbres.

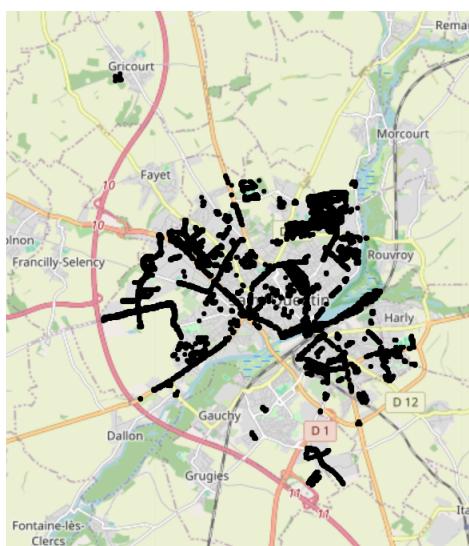
Les couleurs des points permettent d'identifier facilement les différents états des arbres. Par exemple, les arbres en place sont représentés en gris, les arbres abattus en rouge, les arbres supprimés en orange et les arbres dans un autre état en bleu. Cette distinction visuelle facilite l'identification des états prédominants et des tendances générales dans la répartition des arbres.

Cette visualisation peut aider à identifier les tendances ou les modèles dans la répartition des différents états des arbres. Par exemple, elle peut révéler des zones où les arbres sont plus susceptibles d'être abattus ou supprimés, ce qui pourrait être utile pour la planification urbaine ou la gestion des espaces verts.

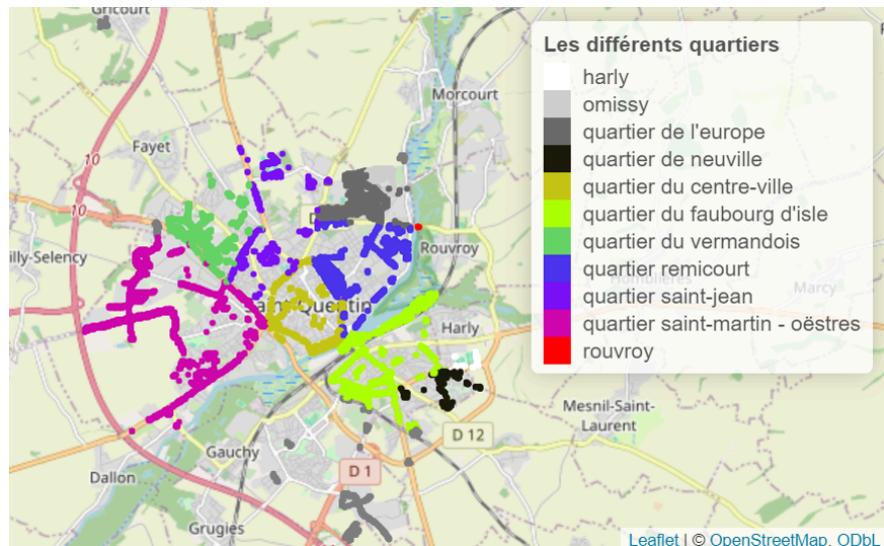
IV- Fonctionnalité 3 : Visualisation des données sur une carte

a) Cartes des arbres répertoriés

Un affichage spatial des données permet une visualisation différente des informations. Tout d'abord il a fallu afficher tous les arbres sur une carte avant d'approfondir la représentation des données.

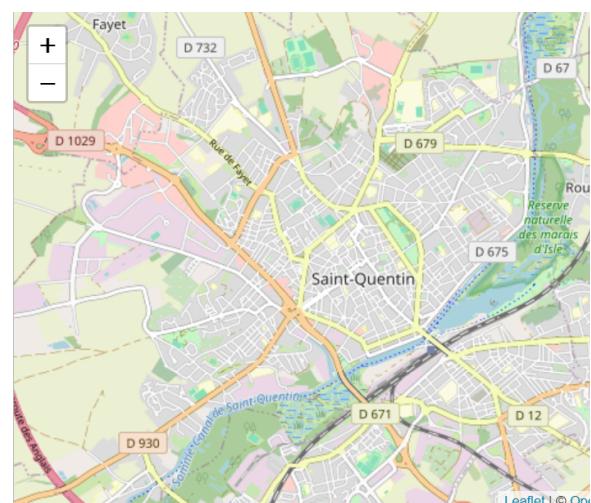
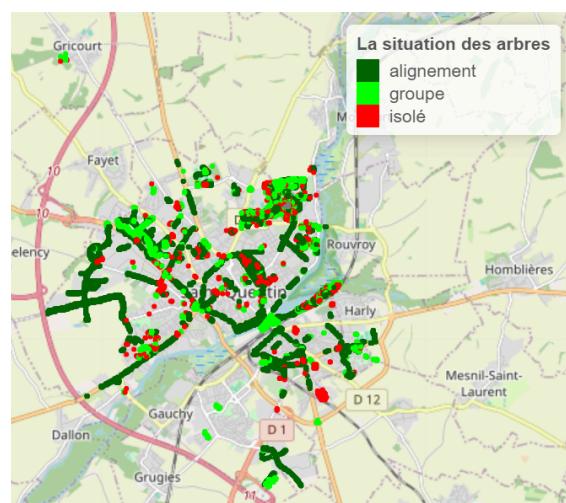


Cartes des arbres de Saint-Quentin



Cartes des arbres selon les quartiers

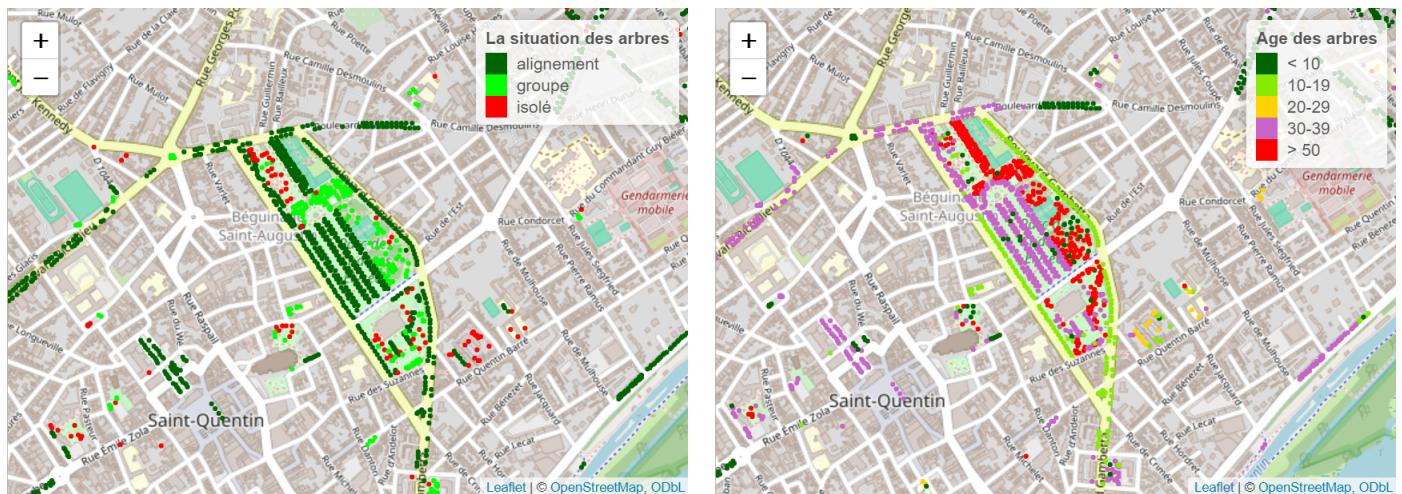
Cette carte nous montre donc la répartition des arbres sur Saint-Quentin, elle permet d'avoir un aperçu de la répartition des arbres. Nous pouvons également observer cela par secteur ou par quartier comme sur la seconde carte. Certains semblent alignés le long des routes, et d'autres groupés en grande quantité dans le parc ou dans des zones plus vertes. Nous pouvons alors vérifier cette hypothèse grâce à une carte sur la situation des arbres à Saint-Quentin.



Carte de situation des arbres

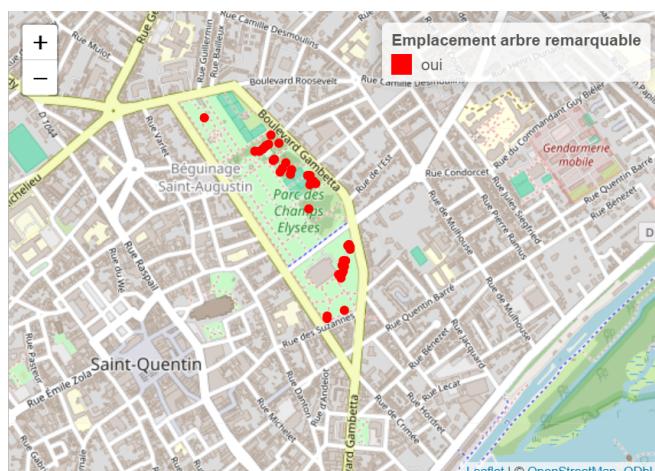
En effet, les arbres qui semblaient être alignés le long des routes sont bien classés comme positionnés en alignment. Avec une vérification sur la carte à vide, ils sont bien situés le long des routes sauf pour le parc des Champs Elysées. L'agencement des arbres dans ce parc peut donner plusieurs informations sur l'impact de l'activité humaine. Il semble intéressant de la coupler à la carte des âges des arbres ainsi qu'à celle des arbres remarquables.

Carte de Saint-Quentin



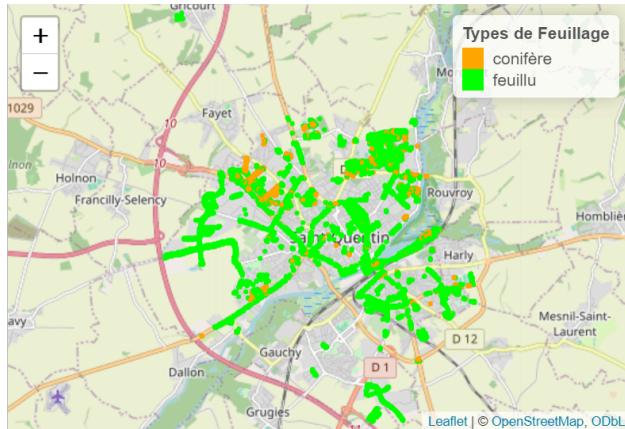
Situation des arbres au parc des Champs Elysées

Age des arbres du parc des champs Elysées

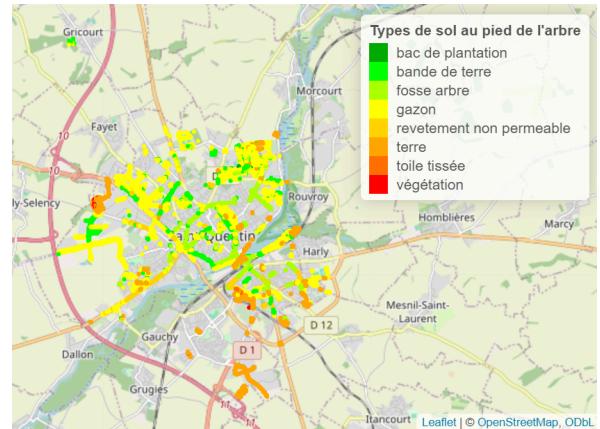


Arbres remarquables du parc des Champs Elysées

Cela nous informe que les arbres placés en alignment ont dû être plantés il y a environ 35, tandis que ceux qui sont placés plus naturellement, soit isolés soit en groupe, sont majoritairement plus vieux, plus de 50 ans. De plus, les arbres remarquables font tous partie de ces arbres plus anciens.

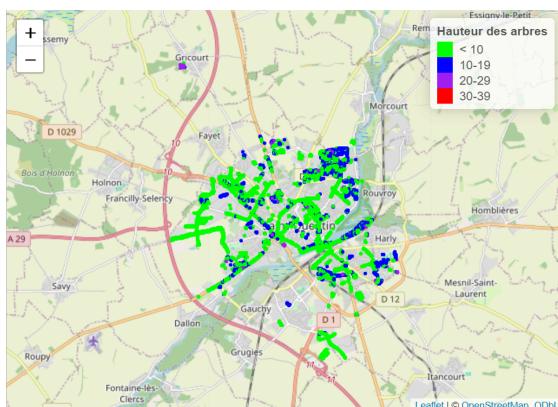


Types de feuillages

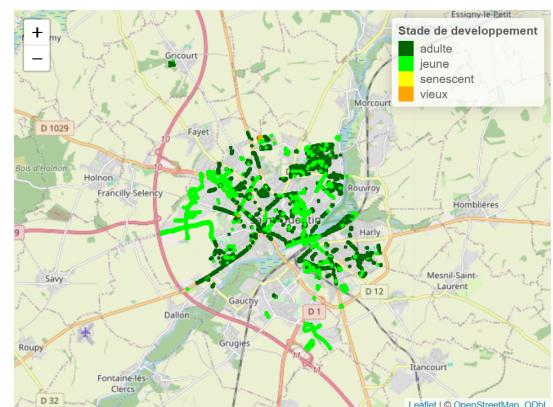


Type de sol

La première carte montre bien la part plus importante de feuillus présents à Saint-Quentin. Il semblerait également que les conifères soient plus présents dans le nord de la ville. Sur la deuxième nous pouvons observer que les arbres sont majoritairement présents dans le gazon ou de la terre et peu d'entre eux sont entourés de végétation.

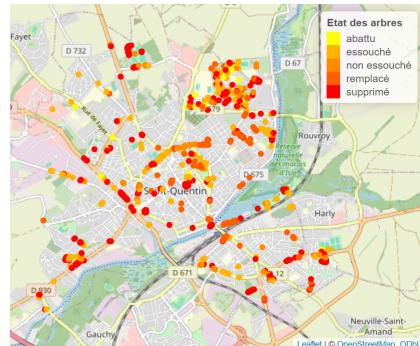


Hauteur des arbres

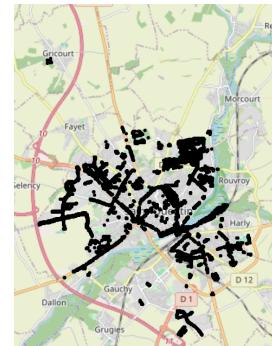


Stade de développement

Bien que les cartes manquent légèrement de précision pour pouvoir poser une corrélation, il semblerait que la hauteur des arbres soit liée à son stade de développement. Majoritairement, nous pouvons observer que les arbres les plus hauts semblent positionnés au même endroit que les arbres les plus vieux. Nous pouvons notamment observer ce phénomène accentué au niveau de Grincourt et du nord de Saint-Quentin.



Carte des arbres supprimés

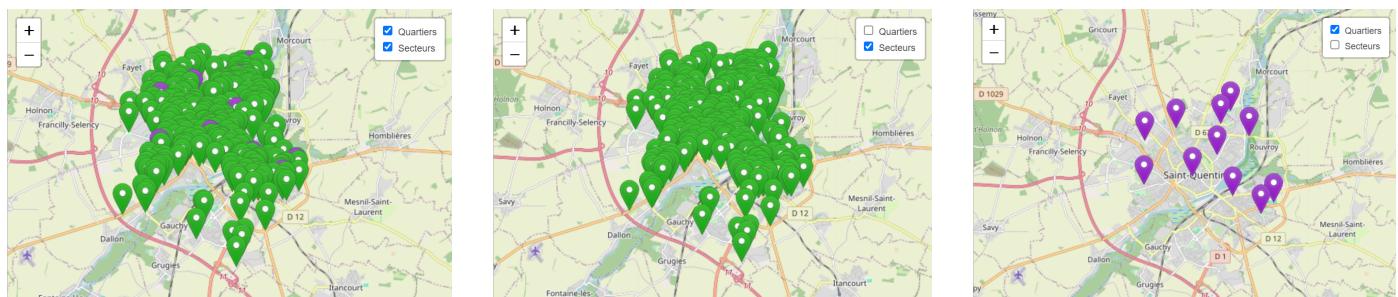


Carte des arbres de Saint-Quentin

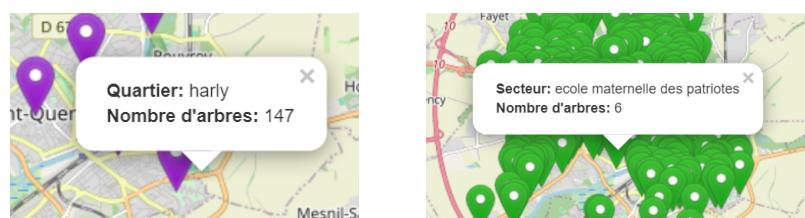
Les zones présentant un taux d'arbre abattu plus important correspondent avec les zones ayant une densité d'arbre en place élevée.

b) Carte de la quantité d'arbres par quartier et secteurs

Cette carte permet de voir le nombre d'arbres par secteur et par quartier. Il suffit de cliquer sur un marqueur pour avoir le nom du quartier ou du secteur ainsi que le nombre d'arbres présents dans ce dernier. Il y a également la possibilité d'afficher soit les quartier, soit les secteur, soit les deux.



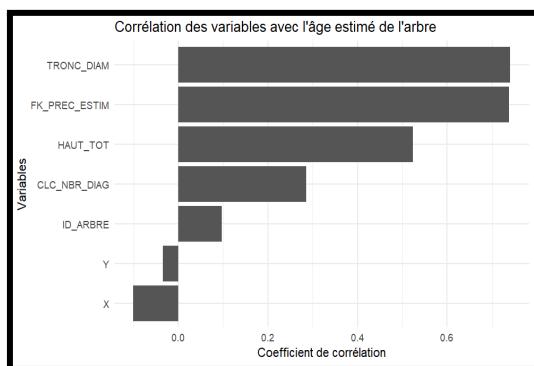
Cartes de la quantité d'arbre par quartier/secteur



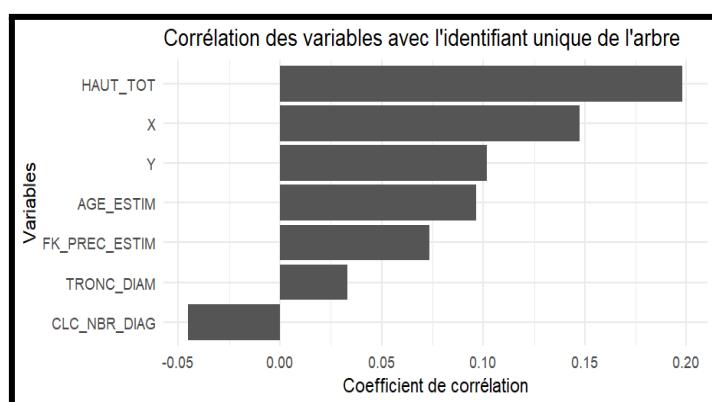
Affichage des données

V- Fonctionnalité 4 : Etude des corrélations entre variables

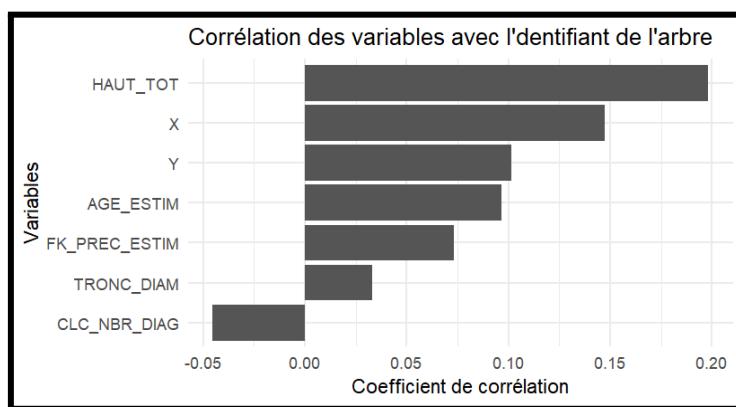
a) Quels sont les liens entre les variables ?



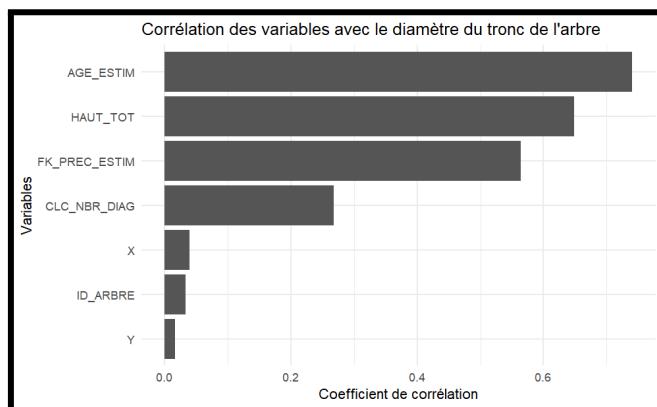
Le graphique de corrélation visualise les coefficients de corrélation entre l'âge estimé des arbres (AGE_ESTIM) et d'autres variables du dataset. Chaque barre représente la force et la direction de la corrélation entre AGE_ESTIM et une autre variable. Les variables avec des barres plus longues et positives sont celles qui ont une forte corrélation positive avec l'âge estimé de l'arbre. Ici dans ce premier graphique, c'est le cas de la variable TRONC_DIAM FK_PREC_ESTIM et HAUT_TOT. Cela signifie que, à mesure que ces variables augmentent, l'âge estimé de l'arbre a tendance à augmenter également. En effet, plus l'âge estimé sera grand est plus le diamètre ainsi que la taille de l'arbre sera grand. De plus, plus l'âge estimé sera grand est plus la précision de l'estimation de l'arbre sera grande. Les variables avec des barres plus longues et négatives ont une forte corrélation négative avec l'âge estimé de l'arbre. C'est le cas pour la variable OBJECTID par exemple. Cela signifie que, à mesure que ces variables augmentent, l'âge estimé de l'arbre a tendance à diminuer. Les variables avec des barres courtes proches de zéro ont peu ou pas de corrélation avec l'âge estimé de l'arbre, indiquant qu'elles n'ont pas beaucoup d'influence sur l'estimation de l'âge de l'arbre. C'est le cas des variables CLC_NBR_DIAG, ID_ARBRE, Y et X.



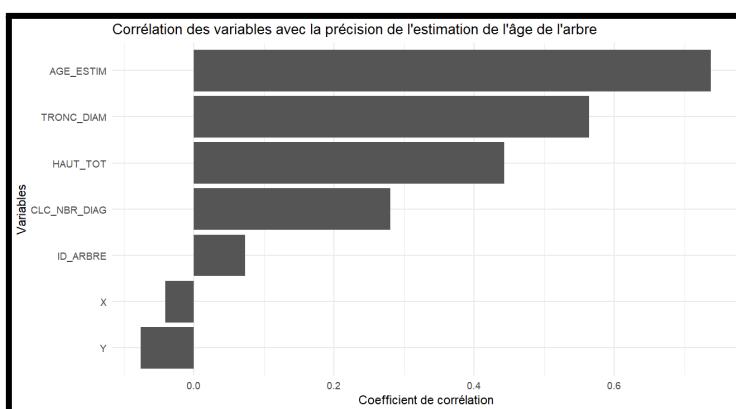
Dans ce deuxième graphique les variables ont une très faible corrélation positive en majorité et une variable à corrélation négative avec l'identifiant unique de l'arbre. Ainsi, ces variables ont très peu voire pas du tout d'impact.



Sur ce troisième graphique on peut voir que la variable de l'identifiant est très peu corrélée avec les autres variables. Ce qui paraît évident puisque ces variables n'ont aucune influence sur ce dernier.



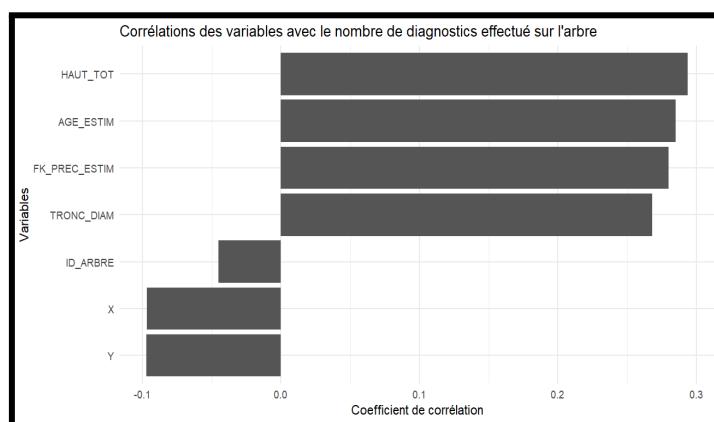
Sur ce graphique on peut voir que le diamètre du tronc d'arbre est corrélé avec la variable AGE_ESTIM, HAUT_TOT et PREC_ESTIM car les barres sont plus longues. De plus il s'agit d'une corrélation positive donc à mesure que ces variables augmentent, le diamètre du tronc de l'arbre a tendance à augmenter également. Quant aux autres variables, elles ne présentent aucune corrélation, donc aucun lien avec le diamètre du tronc. En effet, les coordonnées X et Y ou encore l'identifiant de l'arbre n'ont aucun impact sur le diamètre du tronc de l'arbre et ne permettent pas d'avoir un diamètre plus grand ou plus petit.



Sur ce graphique on peut voir que la précision de l'estimation de l'arbre est corrélée avec la variable AGE_ESTIM, TRONC_DIAM et HAUT_TOT. En effet, les barres sont plus longues. De plus il s'agit d'une corrélation positive donc à mesure que ces variables augmentent, la précision de l'estimation de l'âge de l'arbre a tendance à augmenter également. Quant aux autres variables, elles ne présentent aucune corrélation, donc aucune liaison avec la précision de l'estimation de l'arbre. En effet, les coordonnées X et Y ou encore l'ID_ARBRE n'ont aucun impact sur le diamètre du tronc de l'arbre et ne permettent pas d'avoir un diamètre plus grand ou plus petit. Toutefois, la variable OBJECTID est corrélé négativement avec la variable étudiée. En effet elle tend plus vers -1 que 0. Cela vient à dire que cet identifiant unique de l'arbre a un lien avec la variable. Il s'agit donc d'une relation linéaire inverse. Lorsque la valeur de la variable OBJECTID augmente, la valeur de la précision de l'estimation de l'âge étudiée tend à diminuer.

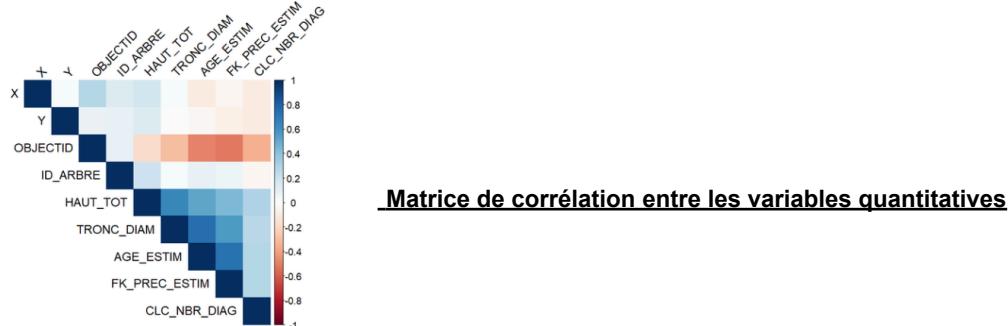
Il est plus facile d'estimer l'âge des arbres jeunes avec précision en utilisant des méthodes directes comme le comptage des anneaux de croissance ou les enregistrements récents de plantation. En revanche, pour les arbres plus âgés, ces méthodes deviennent souvent moins précises en raison de la décomposition des parties internes ou de la difficulté à accéder aux coeurs des arbres. Ainsi, il est logique de constater une corrélation entre la précision de l'estimation et l'âge estimé.

Par exemple, un arbre jeune pourrait avoir une estimation très précise parce que les données de plantation sont récentes et les caractéristiques de croissance sont plus uniformes et moins sujettes aux variations environnementales sur une courte période. Tandis que l'estimation de l'âge d'arbre pourrait être moins précise en raison de l'usure du tronc, de la perte de données historiques, ou de méthodes de datation indirectes. Ainsi la variable OBJECTID est liée avec la variable étudiée puisque lorsque la précision de l'estimation de l'âge est précise, cela voudrait dire que l'arbre est jeune et donc son OBJECTID est petite car ce numéro est attribué dès l'arbre planté.



Ce graphique étudie la corrélation des variables avec le nombre de diagnostic effectué sur l'arbre. La seule variable corrélée, négativement, plus que les autres est l'OBJECTID ayant une valeur proche de -1. Ainsi cela voudrait dire que les arbres ayant un identifiant plus petit aura tendance à plus être diagnostiqués contrairement aux arbres plus anciens qui ont un risque d'être abattu ou supprimé.

b) Conduire des analyses bivariées



Une matrice de corrélation est un tableau carré affichant les coefficients de corrélation entre plusieurs variables, allant de -1 à 1 pour indiquer la force et la direction de la corrélation. Cette représentation permet de visualiser et d'analyser rapidement les relations entre les différentes variables d'un ensemble de données. Nous avons créé une matrice de corrélation avec les variables suivantes : X, Y, OBJECTID, ID_ARBRE, HAUT_TOT, HAUT_TRONC, TRONC_DIAM, AGE_ESTIM, FK_PREC_ESTIM et CLC_NBR_DIAG. Les variables sont listées à la fois horizontalement et verticalement, formant ainsi une grille où chaque intersection de lignes et de colonnes affiche la corrélation entre les deux variables correspondantes. Nous aurions pu enlever les variables X, Y et OBJECTID car elles n'ont aucune corrélation avec les autres variables.

Les cellules de cette matrice sont colorées pour représenter les valeurs de corrélation entre les différentes variables, avec une échelle de couleurs allant du bleu foncé (indiquant une corrélation négative) au rouge foncé (indiquant une corrélation positive). Plusieurs observations peuvent être faites à partir de cette matrice. OBJECTID présente une corrélation positive modérée avec certaines variables, tandis qu'une forte corrélation positive est observée entre les variables liées à la hauteur et au diamètre des troncs. Des corrélations négatives significatives sont également visibles, comme entre AGE_ESTIM et CLC_NBR_DIAG.

Les corrélations positives indiquent des relations directes entre les mesures, tandis que les corrélations négatives signalent des relations inverses. Par exemple, une forte corrélation positive entre la hauteur totale et le diamètre du tronc suggère que les arbres plus grands ont tendance à avoir des troncs plus larges, tandis qu'une corrélation négative entre l'âge estimé des arbres et le nombre de diagnostics pourrait indiquer que les arbres plus jeunes sont plus souvent diagnostiqués.

Plusieurs variables sont liées entre elles.

- AGE_ESTIM et TRONC_DIAM

Coefficient de corrélation de $r = 0.6$

Interprétation : Il existe une corrélation positive modérée entre l'âge estimé de l'arbre et le diamètre du tronc. Cela signifie que les arbres plus âgés tendent à avoir un tronc de plus grand diamètre.

AGE_ESTIM et HAUT_TOT :

Coefficient de corrélation $r = 0.4$

Interprétation : Il y a une corrélation positive modérée entre l'âge estimé de l'arbre et la hauteur totale. Les arbres plus âgés tendent à être plus grands.

- TRONC_DIAM et HAUT_TOT :

Coefficient de corrélation $r = 0.5$

Interprétation : Il existe une corrélation positive modérée entre le diamètre du tronc et la hauteur totale de l'arbre. Les arbres avec des troncs plus larges tendent également à être plus hauts.

- FK_PREC_ESTIM et AGE_ESTIM :

Coefficient de corrélation $r = -0.3$

Interprétation : Il y a une corrélation négative faible entre la précision de l'estimation et l'âge estimé de l'arbre. Cela pourrait indiquer que la précision de l'estimation diminue légèrement à mesure que l'âge de l'arbre augmente.

- FK_PREC_ESTIM et TRONC_DIAM :

Coefficient de corrélation : $r = -0.2$

Interprétation : Il y a une corrélation négative faible entre la précision de l'estimation et le diamètre du tronc. Les estimations pourraient être moins précises pour les arbres avec des troncs plus larges.

c) Etude des relations entre variables qualitatives

i- tableau croisés et tests d'indépendance

Un tableau de contingence initial a été calculé pour différentes variables qualitatives. Ce tableau permet de visualiser la distribution des données et d'identifier les fréquences des différentes combinaisons de variables. Pour maintenir la fiabilité des analyses, en particulier pour les tests statistiques sensibles aux faibles effectifs dans certaines catégories, des adaptations ont dû être réalisées.

Les tableaux de contingence avec des valeurs inférieures à 5 peuvent introduire des biais dans les résultats, notamment dans le calcul du test du Chi-deux. Pour éviter ce biais, des actions spécifiques ont été prises :

- Exclusion des valeurs "à renseigner" : Les lignes contenant la valeur "à renseigner" dans les colonnes des variables explicatives ont été exclues. Cette étape a permis d'éliminer les données incomplètes qui pourraient fausser les analyses.
- Regroupement des catégories à faible fréquence : Les combinaisons de catégories avec des fréquences inférieures à 5 ont été regroupées sous une catégorie "autres". Ce regroupement est crucial pour assurer des effectifs suffisants dans chaque catégorie et éviter les biais dans les tests statistiques.
- Automatisation du processus : Une fonction a été développée pour automatiser l'identification des combinaisons à faible fréquence et leur regroupement sous la catégorie "autres". Cela a permis de systématiser et de simplifier le processus de nettoyage des données, garantissant ainsi une distribution plus équilibrée des données dans le tableau de contingence.

Ces ajustements ont permis de créer un tableau de contingence fiable, prêt pour des analyses statistiques.

	<th>groupe</th> <th>isolé</th>	groupe	isolé
harly	130	26	0
omissy	145	475	22
quartier de l'europe	476	708	237
quartier de neuville	163	112	27
quartier du centre-ville	619	71	50
quartier du faubourg d'isle	910	570	202
quartier du vermandois	674	685	56
quartier remicourt	1099	507	161
quartier saint-jean	482	201	63
quartier saint-martin - oëstres	1576	280	140
rouvroy	6	0	0

	<th>groupe</th> <th>isolé</th>	groupe	isolé
autres	281	501	22
quartier de l'europe	476	708	237
quartier de neuville	163	112	27
quartier du centre-ville	619	71	50
quartier du faubourg d'isle	910	570	202
quartier du vermandois	674	685	56
quartier remicourt	1099	507	161
quartier saint-jean	482	201	63
quartier saint-martin - oëstres	1576	280	140

Tableaux de contingence des quartiers en fonction de la situation des arbres, avant et après regroupement des classes

Calcul du khi2 :

```
Pearson's Chi-squared test

data: table_port_situation
X-squared = 1485.2, df = 16, p-value < 2.2e-16
```

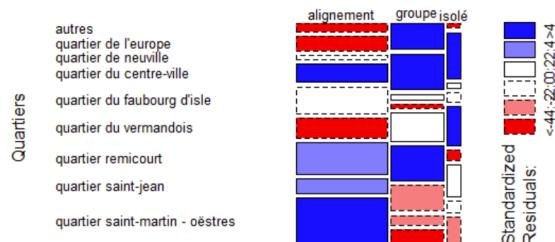
Le test du chi-carré de Pearson affiche les résultats suivants : $X^2=1485.2$, $X^2=1485.2$, $df=16$, $p\text{-valeur} < 2.2e-16$. Ces résultats indiquent une forte dépendance entre les quartiers et les situations (alignement, groupe, isolé).

La valeur $X^2=1485.2$, $X^2=1485.2$, $df=16$, $p\text{-valeur} < 2.2e-16$ (très proche de zéro) signifie que la probabilité que les différences observées soient dues au hasard est extrêmement faible.

Ces résultats permettent de conclure que la répartition des situations varie significativement selon les quartiers. Cette forte association indique que certains quartiers sont plus susceptibles de présenter certaines situations (alignement, groupe, isolé) que d'autres. Cela corrobore les observations du tableau mosaïque, où des surreprésentations et sous-représentations notables ont été identifiées pour différentes situations dans divers quartiers.

ii- graphiques des tableaux

Tableau mosaïque: Quartier et Situation



LÉGENDES :

Bleu foncé : Résidus positifs élevés, ce qui signifie que la fréquence observée est beaucoup plus élevée que la fréquence attendue.

Rouge foncé : Résidus négatifs élevés, ce qui signifie que la fréquence observée est beaucoup plus faible que la fréquence attendue.

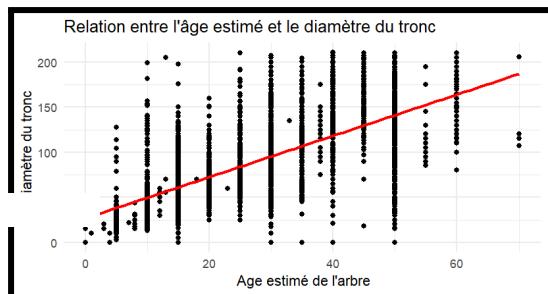
Blanc : Fréquence observée proche de la fréquence attendue.

Largeur des blocs : Proportionnelle à la fréquence de chaque catégorie de la variable sur l'axe horizontal (par exemple, "Situation").

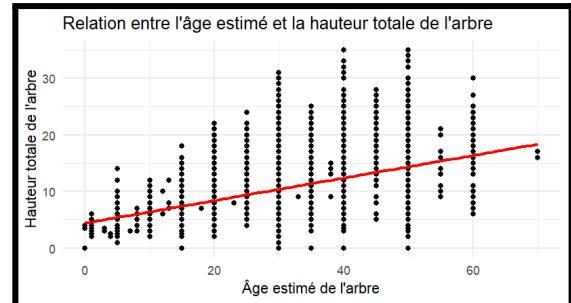
Hauteur des blocs : Proportionnelle à la fréquence de chaque catégorie de la variable sur l'axe vertical (par exemple, "Quartiers").

Le tableau mosaïque illustre la relation entre différents quartiers et les situations ("alignement", "groupe", "isolé"). Les barres colorées indiquent les proportions, avec le rouge signifiant une surreprésentation et le bleu une sous-représentation par rapport à l'attendu. Par exemple, le "quartier de l'Europe" et le "quartier de Neuville" montrent une forte surreprésentation de l'alignement (rouge), tandis que ces mêmes quartiers sont sous-représentés pour la situation de groupe (bleu). À l'inverse, le "quartier du Centre-ville" a une surreprésentation pour la situation de groupe et une sous-représentation pour l'alignement. Certains quartiers comme le "quartier du Faubourg d'Isle" et le "quartier Saint-Jean" montrent une distribution équilibrée sans sur ou sous-représentation notable. Ce tableau est utile pour visualiser les disparités entre quartiers et orienter les décisions en matière d'urbanisme et de politiques sociales.

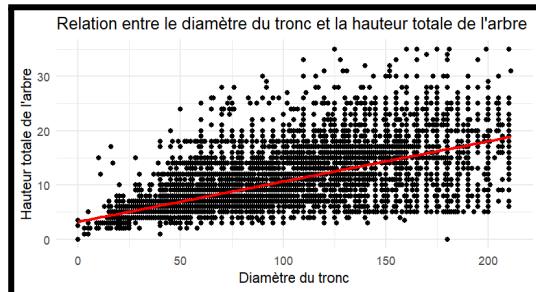
Liaison entre variable :



$$R^2 = 0.53$$



$$R^2 = 0.44$$

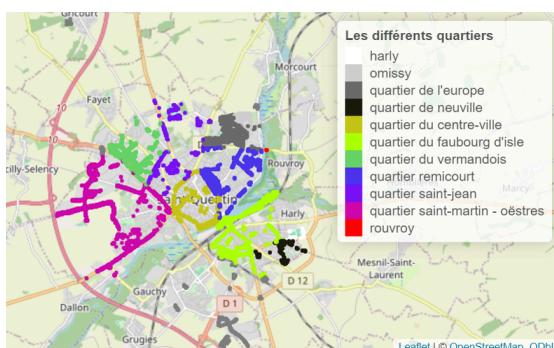


$$R^2 = 0.53$$

Ainsi, les valeurs des coefficients de régression initialement obtenues étaient plutôt faibles. Pour améliorer le modèle, une régression linéaire a été effectuée en incluant toutes les variables. Le coefficient de détermination R^2 obtenu est de 0.83, ce qui est très proche de 1. Cela indique que le modèle est performant et que l'âge de l'arbre peut être estimé de manière fiable en utilisant les variables TRONC_DIAM, HAUT_TOT, FEUILLAGE, FK_PIED et CLC_SECTEUR.

VI- Fonctionnalité 5 : Etude des corrélations entre variables

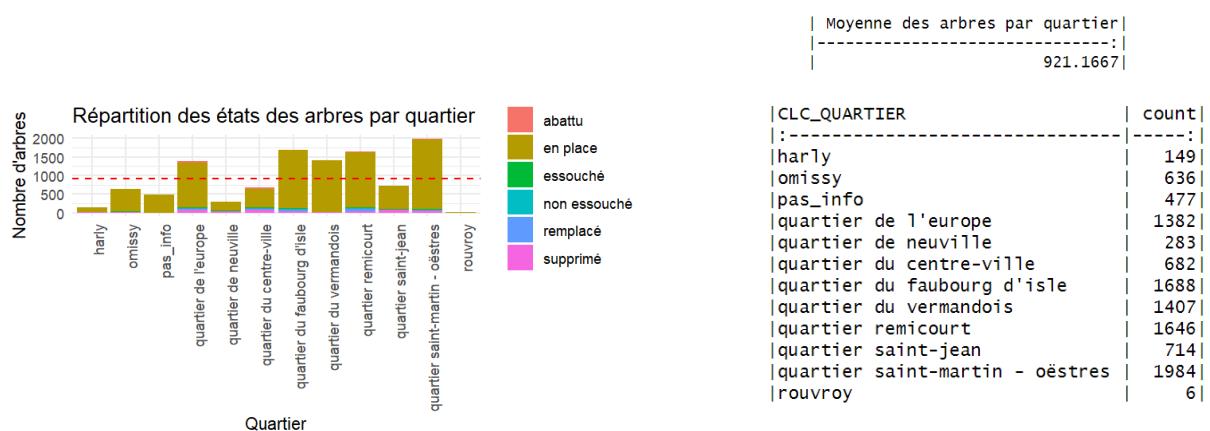
a) Plantation des arbres à Saint-Quentin



Répartition des arbres selon les quartiers

Carte de Saint-Quentin (google maps)

Il est possible d'avoir un premier aperçu des possibilités de plantation des arbres à Saint-Quentin grâce aux cartes. Il semblerait qu'il y ait une zone sans arbres dans le quartier de Saint-Martin-Oëstres, cependant sur l'autre carte de Saint-Quentin il semblerait que ce soit une zone de végétation. L'une des possibilités était que les arbres de cette zone n'aient pas été enregistrés. Néanmoins, après vérification, il s'agirait d'une zone agricole avec un bocage très faible voir inexistant. Ce serait donc un premier point d'amélioration possible pour la plantation d'arbres à Saint-Quentin. Il semblerait également que dans les quartiers de Saint-Jean et de Remicourt, des zones aient une plus faible densité de population.



Cependant l'étude de la répartition des arbres par quartier, nous apporte des précisions qui remettent en question nos idées de départ. Ainsi, pour harmoniser le développement global de la ville en termes de couverture arboricole, plusieurs quartiers nécessitent une attention particulière. Les quartiers harly, omissy, quartier de neuville, quartier du centre-ville, quartier saint-jean et enfin le quartier rouvroy ont actuellement une densité d'arbres bien inférieure à la moyenne de la ville. En plantant davantage d'arbres dans ces zones, nous pourrions non seulement équilibrer la répartition des arbres dans la ville mais aussi améliorer la qualité de l'air, réduire les îlots de chaleur urbains et fournir des espaces verts supplémentaires pour les résidents.

b) Régression logistique

Dans cette étude, nous avons cherché à comprendre les facteurs influençant la décision d'abattre des arbres. Pour ce faire, nous avons utilisé une régression logistique, une méthode statistique adaptée à la modélisation de relations entre une variable dépendante binaire et un ensemble de variables explicatives.

La variable cible de notre étude est binaire et indique la situation de l'arbre : un arbre est en place (noté 1) ou abattu (noté 0). Pour prédire cette variable, nous avons identifié deux variables explicatives principales : la distinction entre les arbres remarquables et non remarquables, ainsi que l'âge estimé des arbres, en particulier ceux de plus de 80 ans.

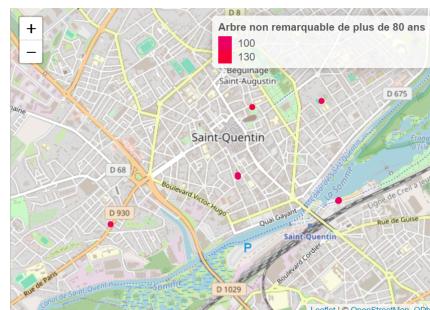
La régression logistique nous permet de modéliser la probabilité qu'un arbre soit en place en fonction de ces variables explicatives. La formule du modèle est la suivante :

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 \times \text{Remarquable} + \beta_2 \times (\text{Âge estimé de l'arbre} > 80 \text{ ans})$$

Les résultats de la régression logistique fournissent des coefficients pour chaque variable explicative ainsi que des tests de significativité pour évaluer leur impact. Un coefficient positif pour la variable "Remarquable" suggérerait que les arbres remarquables ont une plus grande probabilité d'être en place, tandis qu'un coefficient négatif pour la variable "Âge estimé de l'arbre > 80 ans" indiquerait que les arbres de plus de 80 ans ont une plus grande probabilité d'être abattus.

En interprétant ces résultats, nous pouvons mieux comprendre les facteurs influençant la décision d'abattre des arbres. Malheureusement nous obtenons seulement un taux d'erreur de 10% de notre test mais pas de courbe. Il est donc difficile d'interpréter les résultats. Avec un peu plus de temps, nous aurions pu développer ce point.

Voici les potentiels arbres à abattre dans la ville de St quentin, selon les critères définis.



CARTE DES ARBRES À ABATTRE

VII- Conclusion

En conclusion, les analyses effectuées ont permis de mettre en évidence des relations significatives entre les différentes variables étudiées et l'âge des arbres. Malgré des coefficients de régression initialement faibles, l'utilisation d'un modèle de régression linéaire multiple a abouti à un coefficient de détermination élevé de 0.83. Cela indique que le modèle est robuste et que les variables TRONC_DIAM, HAUT_TOT, FEUILLAGE, FK_PIED et CLC_SECTEUR sont des prédicteurs importants pour estimer l'âge des arbres. Ces résultats soulignent l'importance de ces variables dans la compréhension de la croissance et du développement des arbres, ce qui pourrait avoir des implications significatives pour la gestion et la conservation des écosystèmes forestiers. En plus des résultats obtenus, il serait intéressant d'approfondir l'analyse en examinant d'autres variables potentiellement pertinentes pour estimer l'âge des arbres. Par exemple, l'inclusion de données sur le type de sol, le climat local, ou encore l'histoire de la parcelle forestière pourrait enrichir le modèle et améliorer sa précision. De plus, une validation externe du modèle serait nécessaire pour évaluer sa performance sur des ensembles de données indépendants. Enfin, il serait

judicieux d'explorer d'autres techniques de modélisation, telles que les méthodes d'apprentissage automatique, pour comparer et optimiser les performances du modèle dans la prédiction de l'âge des arbres.