Rapport de projet Big Data

Arthur Ollivier - Le Hénaff Youenn – Rochefort Aubin
Trinôme 3

Dans le cadre de notre projet de troisième année en Big Data/IA/Web, nous avons réalisé une étude approfondie du patrimoine arboré de la ville de Saint-Quentin. L'objectif de ce projet est de concevoir et développer une application pour analyser et visualiser les données des arbres, afin de mieux comprendre et gérer (l'implantation des arbres dans le milieu urbain) les arbres urbains. Ce projet nous permet de mettre en pratique les compétences acquises dans les cours de Big Data, Intelligence Artificielle, Développement Web et Base de Données.

Le patrimoine arboré de la ville est très important pour la biodiversité, la qualité de l'air et le bien-être des habitants. Cependant, la gestion de ce patrimoine demande une analyse précise des données disponibles. Ces données, souvent collectées manuellement, peuvent contenir des erreurs et des incohérences. Pour garantir la fiabilité et la précision des analyses, il est essentiel de nettoyer soigneusement les données.

Ce projet se décompose en plusieurs étapes clés : l'extraction et la visualisation des données, le nettoyage des données, l'utilisation de modèles statistiques pour analyser les données, et le développement d'une interface utilisateur pour faciliter l'accès et la compréhension des informations par les gestionnaires et les citoyens. À travers ce projet, nous visons à créer un outil complet et facile à utiliser pour la gestion durable des arbres urbains.

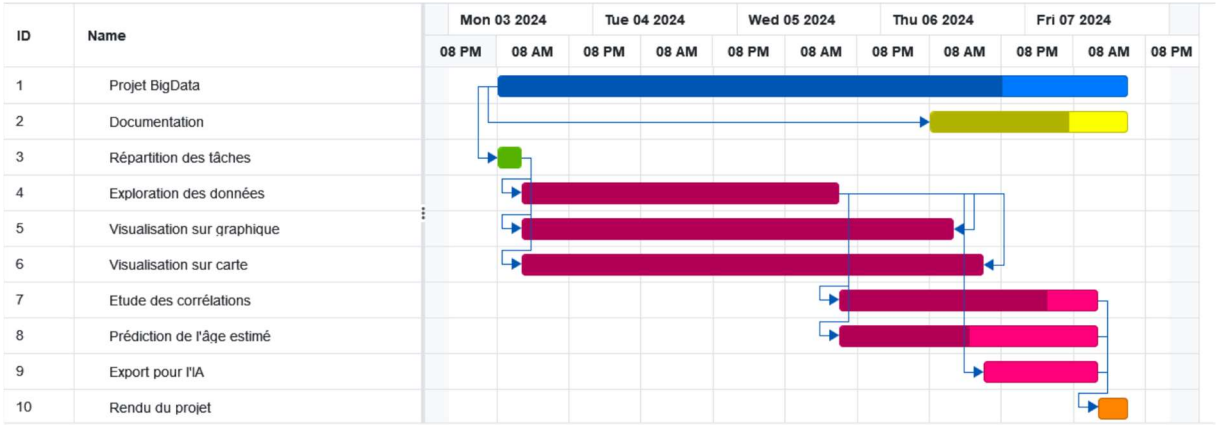
Objectifs du Projet Big Data

Le projet Big Data vise à concevoir et développer une application complète d'analyse et de visualisation des données concernant le patrimoine arboré de la ville de Saint-Quentin. Les principaux objectifs sont :

- Extraction des données : Recueillir des données à partir de fichiers ou de sites web pour constituer une base de données complète et utilisable.
- Nettoyage des données : Éliminer les données incomplètes ou erronées pour améliorer la qualité et la fiabilité des analyses.
- Visualisation des données : Créer des représentations graphiques et cartographiques permettant d'analyser visuellement les grands volumes de données.
- Analyse des données : Appliquer des modèles statistiques, tels que la régression linéaire, ou logistique, et étudier les corrélations entre les différentes caractéristiques des arbres.

Ces objectifs permettent d'approfondir les compétences en Big Data, Intelligence Artificielle et Développement Web, tout en fournissant un outil utile pour la gestion et la conservation du patrimoine arboré urbain.

Organisation



Pour la réalisation ce projet, nous avons structuré notre travail à l'aide d'un diagramme de Gantt détaillant les différentes tâches et leur répartition parmi les membres de l'équipe. Ce diagramme nous a permis de planifier et suivre l'avancement des tâches de manière efficace.

Le projet a commencé par la répartition des tâches, où chacun a pris en charge des parties spécifiques du projet. Nous avons tous contribué à la documentation initiale et à la répartition des tâches pour assurer une compréhension commune des objectifs et des méthodes de travail.

Arthur s'est concentré sur la visualisation sur graphique puis à l'étude des corrélations. Aubin a pris en charge la visualisation sur carte et la prédiction de l'âge estimé des arbres. Youenn s'est de l'exploration des données, a rejoint Arthur sur la visualisation graphique puis sur la rédaction du compte-rendu.

La dernière étape, le rendu du projet, a été une tâche collective où chacun a contribué à la finalisation et à la présentation des résultats. Cette organisation rigoureuse a permis une progression fluide et coordonnée du projet, assurant que toutes les tâches étaient complétées dans les temps impartis.

Détails Base de Données du Patrimoine Arboré

Cette base de données contient des informations sur le patrimoine arboré de la ville. Elle est composée de 11421 lignes et 37 colonnes. Chaque ligne représente un arbre spécifique et contient des informations détaillées sur cet arbre. Voici une explication plus détaillée des colonnes présentes dans cette base de données :

1. **X** : Coordonnée X de l'emplacement de l'arbre sur le système de coordonnées Lambert-93 utilisé par la ville.
2. **Y** : Coordonnée Y de l'emplacement de l'arbre sur le système de coordonnées Lambert-93 utilisé par la ville.
3. **OBJECTID** : Identifiant unique de l'objet dans la base de données, utilisé pour différencier chaque arbre.
4. **created_date** : Date et heure de création de l'enregistrement au format ISO 8601.
5. **created_user** : Nom de l'utilisateur ayant créé l'enregistrement.
6. **src_geo** : Source géographique utilisée pour la localisation de l'arbre, par exemple Orthophoto.

7. **clc_quartier** : Nom du quartier où se situe l'arbre.
8. **clc_secteur** : Nom du secteur spécifique au sein du quartier où se situe l'arbre.
9. **id_arbre** : Identifiant unique de l'arbre.
10. **haut_tot** : Hauteur totale de l'arbre en mètres.
11. **haut_tronc** : Hauteur du tronc de l'arbre en mètres, mesurée depuis la base jusqu'à la première branche.
12. **tronc_diam** : Diamètre du tronc de l'arbre en centimètres.
13. **fk_arb_etat** : Code représentant l'état de l'arbre (par exemple abattu ou en place).
14. **fk_stadedev** : Code représentant le stade de développement de l'arbre (par exemple jeune, adulte, sénescent).
15. **fk_port** : Code représentant le port ou la forme de l'arbre (par exemple libre ou semi-libre).
16. **fk_pied** : Ce qu'il y a au pied de l'arbre (par exemple gazon, revêtement).
17. **fk_situation** : Code décrivant la situation de l'arbre (par exemple isolé, en alignement).
18. **fk_revetement** : S'il y a un revêtement ou non
19. **commentaire_environnement** : Commentaires additionnels sur l'environnement immédiat de l'arbre.
20. **dte_plantation** : Date de plantation de l'arbre, si connue.
21. **age_estim** : Âge estimé de l'arbre en années.
22. **fk_prec_estim** : Précision de l'estimation de l'âge, indiquant le niveau de confiance dans l'estimation.
23. **clc_nbr_diag** : Nombre de diagnostics effectués sur l'arbre.
24. **dte_abattage** : Date d'abattage de l'arbre, si applicable.
25. **fk_nomtec** : Nom technique ou scientifique de l'arbre.
26. **last_edited_user** : Nom du dernier utilisateur ayant modifié l'enregistrement.
27. **last_edited_date** : Date et heure de la dernière modification de l'enregistrement.
28. **villeca** : Nom de la zone urbaine qui a planté l'arbre.
29. **nomfrancais** : Nom commun français de l'arbre.
30. **nomlatin** : Nom scientifique latin de l'arbre.
31. **GlobalID** : Identifiant global unique pour l'arbre, souvent utilisé pour l'interopérabilité entre systèmes.
32. **CreationDate** : Date et heure de création de l'enregistrement dans le système, souvent redondante avec created_date.
33. **Creator**t: Créateur de l'enregistrement, souvent redondant avec created_user.
34. **EditDate** : Date et heure de la dernière modification de l'enregistrement, souvent redondante avec last_edited_date.
35. **Editor** : Dernier modificateur de l'enregistrement, souvent redondant avec last_edited_user.
36. **feuillage** : Type de feuillage de l'arbre (par exemple feuillu, conifère).
37. **remarquable** : Indicateur binaire (Oui/Non) si l'arbre est considéré comme remarquable en raison de son âge, sa taille, son histoire ou sa rareté.

Cette base de données est essentielle pour le suivi et la gestion du patrimoine arboré de la ville. Elle permet de surveiller la santé et le développement des arbres, ainsi que de planifier des actions de maintenance et de conservation. Grâce à ces informations détaillées, les gestionnaires peuvent prendre des décisions éclairées pour assurer la durabilité et la diversité du couvert arboré urbain.

Nettoyage des Données du Patrimoine Arboré

Dans le cadre de notre projet de Big Data, nous avons travaillé sur des données relatives au patrimoine arboré de la ville. Ces données, souvent collectées manuellement, peuvent

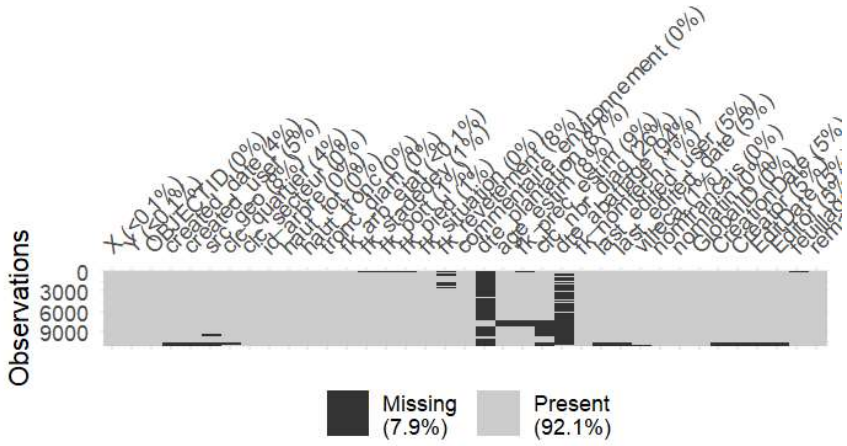
contenir des erreurs et des incohérences. Pour garantir la fiabilité et la précision des analyses, il est crucial de passer par un processus de nettoyage rigoureux. Ce compte rendu présente les étapes essentielles de ce processus.

Suppression des Doublons

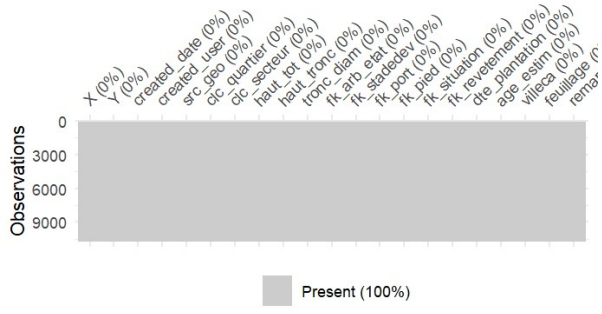
La première étape de notre nettoyage a consisté à supprimer les doublons. Les doublons peuvent biaiser les résultats des analyses en introduisant des répétitions inutiles. Nous avons d'abord éliminé les enregistrements parfaitement identiques. Ensuite, nous avons traité les doublons basés sur les coordonnées géographiques (X, Y). Si deux arbres étaient signalés aux mêmes coordonnées et indiqués comme "en place" ou “remplacé”, nous avons conservé un seul enregistrement. De plus, nous avons supprimé les doublons identifiés par le même GlobalID pour garantir que chaque arbre soit représenté de manière unique.

Gestion des Valeurs Manquantes

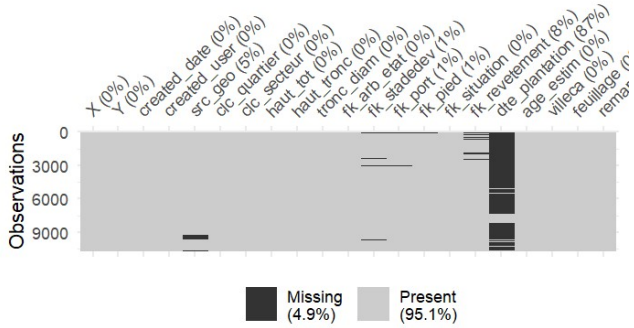
Ensuite, nous avons abordé la gestion des valeurs manquantes, un défi majeur dans tout ensemble de données. Nous avons d'abord supprimé les colonnes jugées inutiles pour l'analyse afin de simplifier la base de données. Pour identifier les colonnes avec des valeurs manquantes, nous avons utilisé la fonction `vis_miss`, qui permet de visualiser clairement les lacunes dans les données.



Pour les colonnes moyennement utiles avec beaucoup de données manquantes, nous avons remplacé les cases vides par la valeur "Temporaire". Pour les colonnes importantes, les cases vides ont été remplies par "NA". Nous avons également appliqué une règle stricte : les lignes contenant plus de trois "NA" ont été supprimées pour maintenir la qualité des données. D’où l’importance de mettre une valeur temporaire dans certaines colonnes, cette approche nous a permis d'éviter la suppression massive de lignes, ce qui aurait pu entraîner une perte significative d'informations.

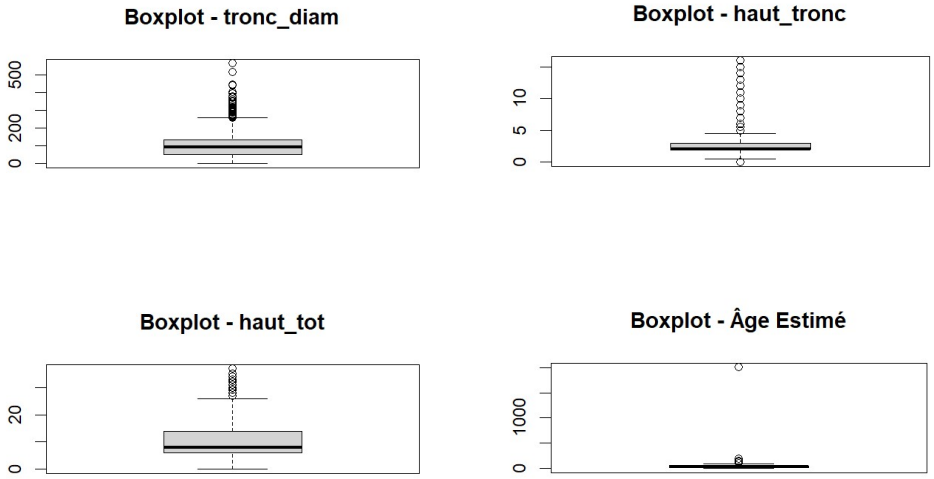


Afin de compléter les valeurs manquantes dans les colonnes importantes, nous avons utilisé des valeurs calculées (moyenne ou médiane) pour éviter de perdre des informations essentielles. Une fois cette imputation effectuée, nous avons remplacé les "Temporaire" par "NA" pour signaler clairement les données manquantes.

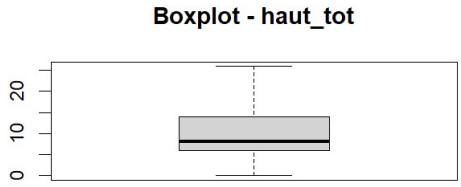


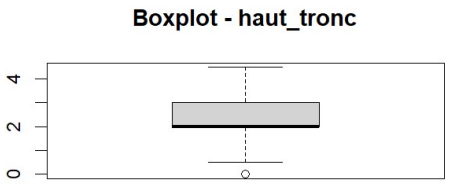
Gestion des Valeurs Aberrantes

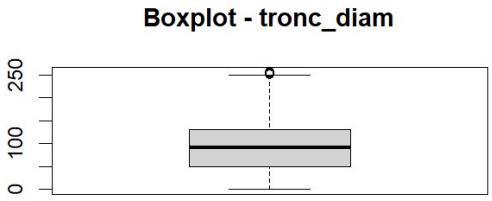
Enfin, nous avons traité les valeurs aberrantes, ces valeurs extrêmes qui peuvent fausser les résultats de l'analyse. À l'aide de boxplots, nous avons identifié les valeurs aberrantes. En appliquant la méthode des quartiles à 25%, nous avons déterminé les limites acceptables et identifié les valeurs qui dépassaient ces limites.

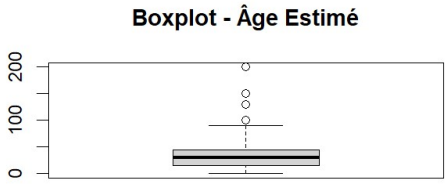


Les valeurs aberrantes ont été remplacées par "NA" pour ne pas supprimer les enregistrements d'arbres réels, mais pour signaler que les données associées étaient suspectes. Cette approche garantit que nous ne perdons pas d'informations précieuses sur les arbres, tout en maintenant la précision de notre base de données.





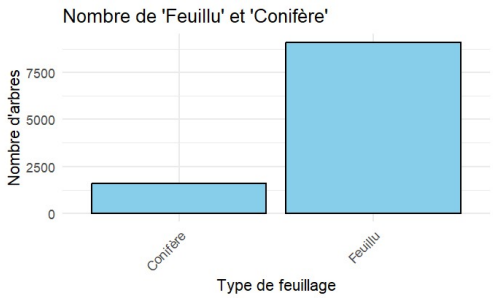




Le nettoyage des données du patrimoine arboré était une étape cruciale pour assurer la fiabilité de nos analyses. En supprimant les doublons, en gérant les valeurs manquantes et en traitant les valeurs aberrantes, nous avons considérablement amélioré la qualité de notre base de données en passant de 11421 lignes à 10676 lignes, soit 745 valeurs de moins (6,5%). Maintenant que les données sont propres, nous pouvons les utiliser pour des analyses et des visualisations graphiques plus fiables qu'auparavant, essentielles pour la gestion et la conservation des arbres urbains.

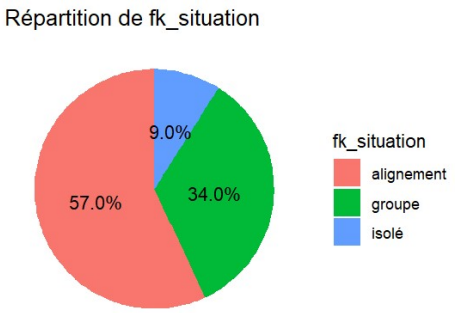
Visualisation Graphique du jeu de données

Pour explorer les données, nous les approprier et les comprendre, nous avons fait une fonction dans laquelle il suffit de taper le nom de la colonne que l'on souhaite voir dans la console et un histogramme de cette colonne apparait, cela nous a permis de visualiser les rapports de force entre les différentes catégories au sein d'une même colonne pour savoir quels graphiques ou quelles cartes pourraient être intéressants et quelles données ne méritent pas d'apparaitre. Nous avons ensuite réalisé les graphiques ci-dessous en prenant en compte une à trois colonnes, afin de comprendre quelles stratégies la ville pourrait adopter pour boiser ses rues.

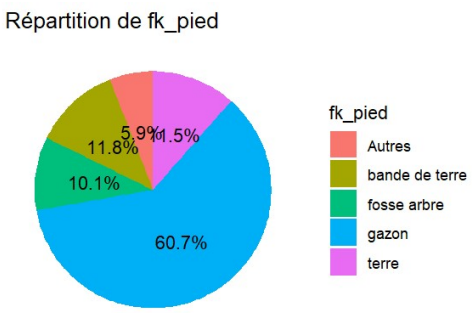


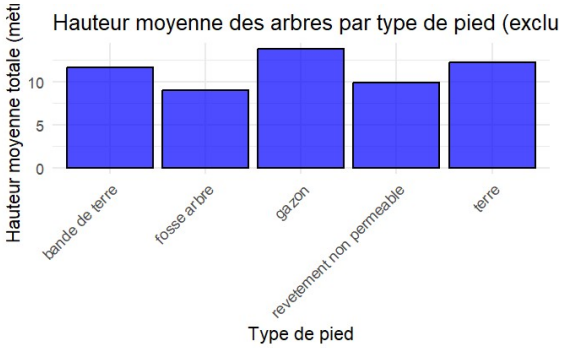
La majorité des arbres sont des feuillus (environ 9103) comparés aux conifères (environ 1573). Les feuillus sont souvent préférés en milieu urbain pour leur capacité à offrir de l'ombre et leur esthétisme saisonnier.

Ce graphique montre la répartition des arbres selon leur situation : en alignement (6084), en groupe (3632), et isolé (960). La majorité des arbres sont en alignement, probablement le long des rues ou des avenues, ce qui est typique dans les zones

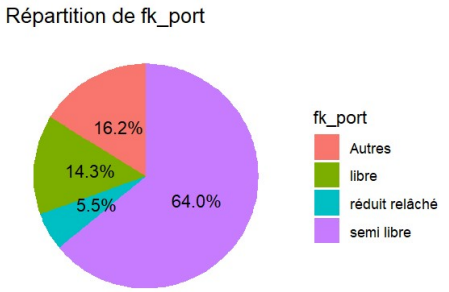


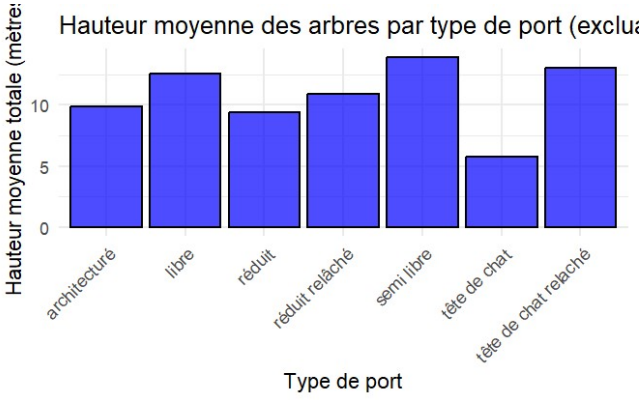
urbaines pour des raisons esthétiques et fonctionnelles. Les arbres en groupe peuvent être dans des parcs ou des jardins, tandis que les arbres isolés pourraient être des spécimens remarquables ou plantés individuellement pour des raisons spécifiques. Saint-Quentin est donc une ville bien arborée dans ses rues, car malgré la présence d'un parc la majorité des arbres se situent en ville.





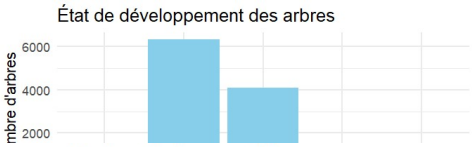
La majorité des arbres sont plantés sur du gazon (6484), suivis par des fosses d'arbres (1074) et des bandes de terre (1246). Cela reflète une préférence pour des plantations sur des surfaces perméables qui favorisent la croissance des arbres. Les autres types de pieds représentent une minorité, ce qui indique une diversité dans les conditions de plantation mais avec une préférence marquée pour les méthodes traditionnelles et pratiques. Comme indiqué sur le second graphique (arbres jeunes exclus), le gazon est un bon revêtement pour l'arbre.



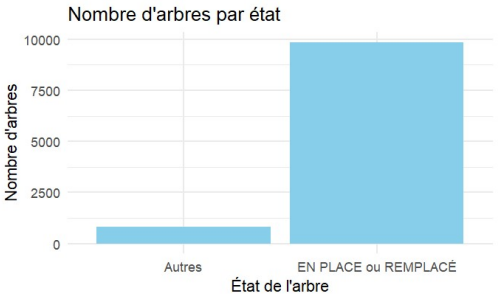


Ce graphique indique que la majorité des arbres ont un port semi-libre (6796) (partiellement taillé), suivi par les arbres en port libre (1528) (non taillé) et réduit relâché (588) (fortement taillé). Le port semi-libre est souvent préféré pour les arbres urbains car il permet un bon équilibre entre esthétisme et gestion de l'espace.
Comme nous pouvons le voir sur le 2ème graphique représentant la hauteur moyenne des arbres selon leur sol (en excluant les jeunes arbres), les méthodes 'libre' et 'semi libre' permettent une plus forte croissance.

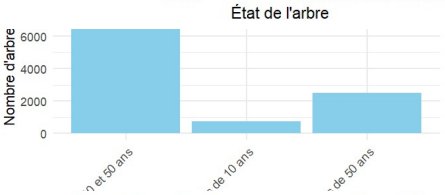
La répartition des stades de développement des arbres montre que 59.1% des arbres sont adultes, 38.4% sont jeunes, et une petite proportion sont vieux ou sénescents. Cela suggère que



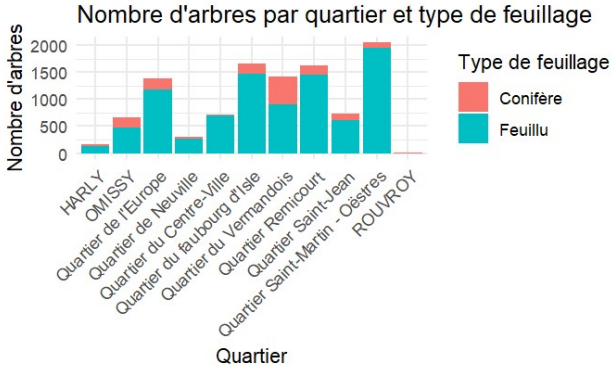
la majorité des arbres sont dans un stade de développement mature, ce qui est positif pour la stabilité écologique urbaine, et pour la surveillance des arbres à risques qui sont peu nombreux.



La majorité des arbres sont "EN PLACE" ou "REMPLACÉ", indiquant une bonne gestion et un suivi rigoureux des arbres urbains. La faible proportion d'arbres dans l'état "Autres" indique qu'ils ont été abattus, montrant ainsi une maintenance efficace et la gestion des arbres qui ne pouvaient plus être conservés.



Les arbres âgés de 10 à 50 ans sont les plus nombreux, ce qui reflète une période de plantation intensive il y a quelques décennies. Les arbres de moins de 10 ans et de plus de 50 ans sont moins nombreux dû à leur remplacement au cours du temps.



Ce graphique montre une répartition par quartier et type de feuillage (conifère et feuillu). Certains quartiers ont une proportion plus élevée de feuillus, ce qui pourrait être dû à des politiques de plantation spécifiques aux préférences locales ou à une prolifération des conifères. Les variations en nombre d'arbres entre quartiers peuvent indiquer des différences dans les stratégies de gestion de la végétation.

Les analyses des différents graphiques montrent une gestion diversifiée et bien planifiée des arbres urbains. La majorité des arbres sont en alignement et plantés sur du gazon, indiquant une méthode de plantation uniforme et efficace. Les stades de développement montrent une population d'arbres majoritairement adultes, ce qui est bon pour la stabilité écologique et l'entretien. Les répartitions par type de port, de pied, et de feuillage indiquent des choix stratégiques pour renforcer l'esthétisme et la santé des arbres. Enfin, les différences entre quartiers montrent des approches de gestion adaptées aux spécificités locales.

Il est important de savoir que de nombreuses valeurs ne peuvent être étudiées, comme la taille et le diamètre des troncs, ou encore la taille totale et l'âge de l'arbre.
Cela est dû aux différences entre les espèces et à leurs spécificités qui empêchent toute comparaison.

Visualisation Géographique des Données du Patrimoine Arboré

Dans le cadre de notre projet, nous avons développé une série de cartes interactives pour visualiser le patrimoine arboré de la ville. Ces cartes permettent de mieux comprendre la répartition et l'état des arbres urbains en utilisant différents critères et catégories. Voici un résumé des étapes et des principales fonctions utilisées pour créer ces visualisations.

Recherche des Packages

Pour commencer, il a fallu trouver les bon packages, ceux nécessaires pour la manipulation des données géospatiales et la création de cartes interactives. Les packages principaux utilisés sont `leaflet` pour la création de cartes interactives, `sf` pour la manipulation de données géospatiales, et `proj4` et ' PROJ' pour les transformations de systèmes de coordonnées.

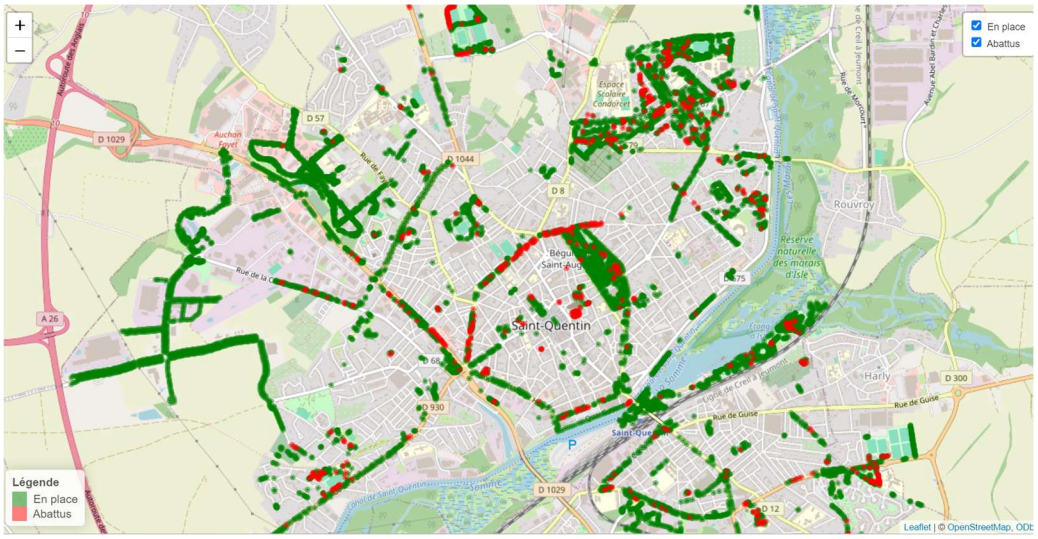
Conversion des Coordonnées

Les coordonnées des arbres dans la base de données étaient initialement au format Lambert-93. Nous avons utilisé les bibliothèques `proj4` et 'PROJ' pour convertir ces coordonnées en latitude et longitude (WGS84), ce qui est nécessaire pour les visualiser sur des cartes interactives. La conversion des coordonnées a été réalisée en définissant la projection Lambert-93 et en utilisant la fonction `project` pour transformer les coordonnées en format WGS84.

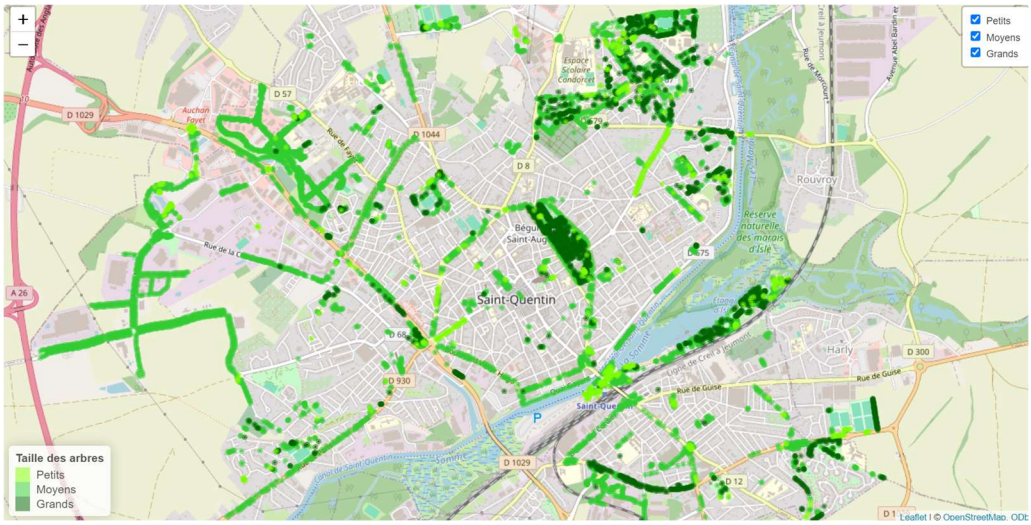
Création des Cartes

Nous avons créé plusieurs cartes interactives à l'aide du package `leaflet`, en ajoutant des marqueurs pour représenter les différentes catégories d'arbres, une légende et des contrôles de couches pour permettre aux utilisateurs de sélectionner les types d'arbres à afficher :

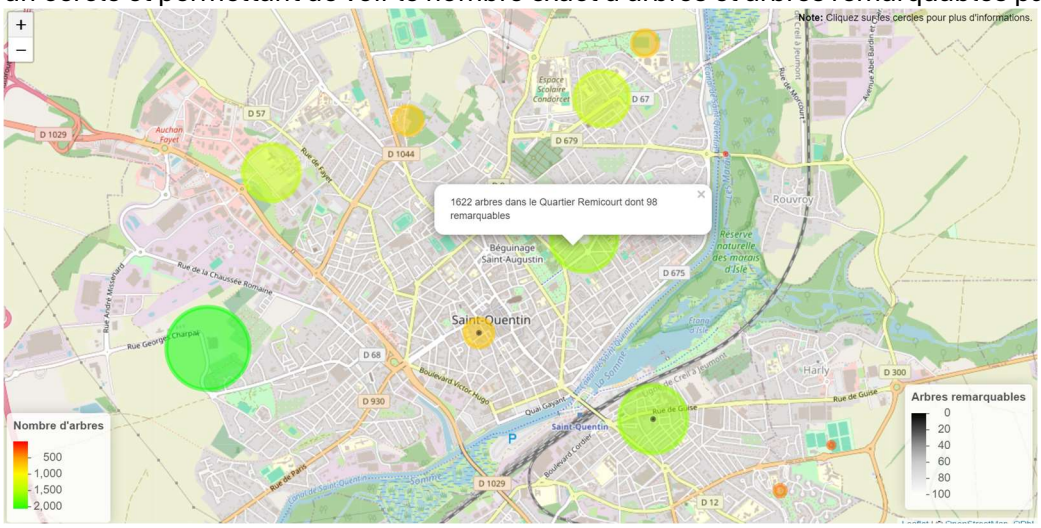
- Carte des Types de Feuillage : Affichage des arbres feuillus et conifères avec des marqueurs de couleurs différentes.
- Carte de l'État des Arbres : Affichage des arbres en place et des arbres abattus, supprimés, essouchés ou non, avec des couleurs différentes.



- Carte des Arbres par Taille : Affichage des arbres petits, moyens et grands avec des marqueurs de couleurs différentes.



- Carte des Arbres Remarquables par Quartier : Calcul du nombre d'arbres et d'arbres remarquables par quartier, et affichage de cercles proportionnels au nombre d'arbres avec des couleurs représentant le nombre d'arbres remarquables. Ajout de pop-ups s'affichant au clic sur un cercle et permettant de voir le nombre exact d'arbres et arbres remarquables par quartier.



- Carte des Stades de Développement : Affichage des arbres jeunes, adultes, vieux et sénescents avec des marqueurs de couleurs différentes.
- Carte des Nécessités d'Entretien : Affichage des différents types de port des arbres (libre, semi-libre, réduit relâché, tête de chat, autres) avec des marqueurs de couleurs différentes.

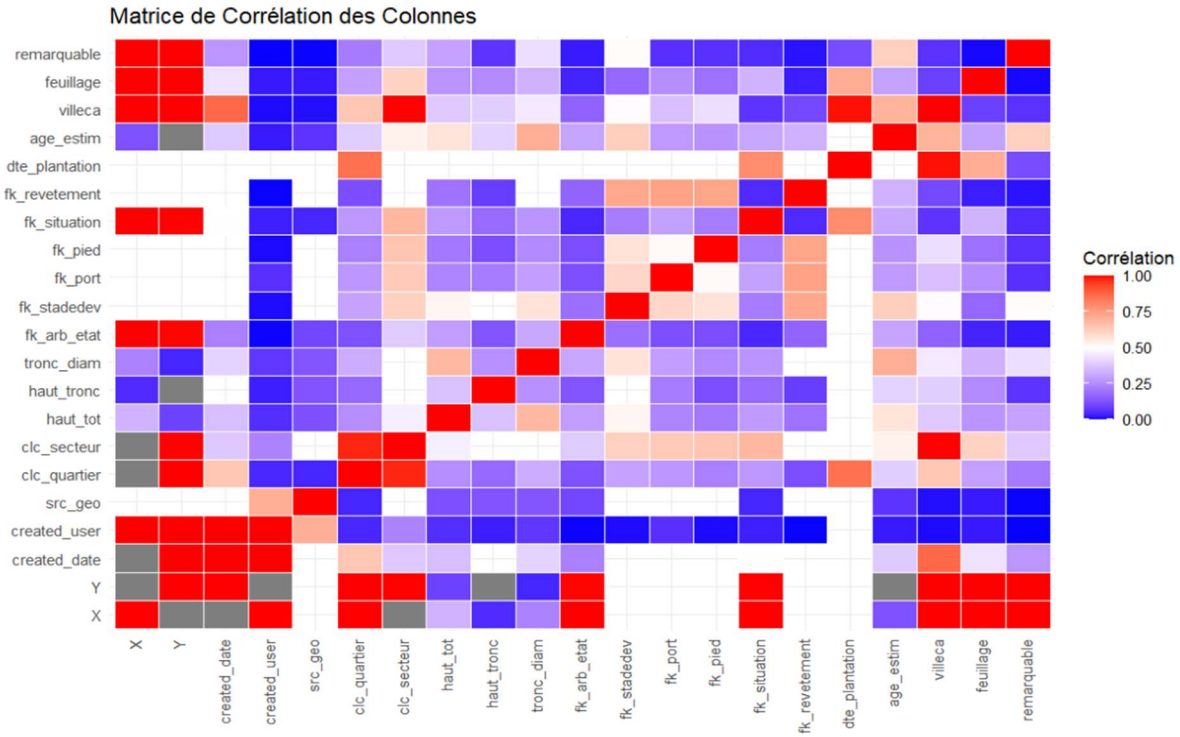
La création de ces cartes interactives a permis de visualiser concrètement les données du patrimoine arboré de la ville. Grâce à l'utilisation de `leaflet` et `sf`, nous avons transformé des données brutes en informations géospatiales précises et exploitables. Ces cartes offrent une vue d'ensemble des différents aspects des arbres urbains, tels que leur type de feuillage, leur état, leur taille, leur stade de développement et leurs nécessités d'entretien. De telles visualisations peuvent être utile aux agents d'entretien pour planifier au mieux leurs interventions.
Cet aspect nous a également aidé à mieux comprendre certaines colonnes de la base de données, en visualisant directement les informations géographiques associées. Cela démontre l'importance des outils de visualisation pour la compréhension des données.

Corrélation

Pour comprendre les liens entre les différentes variables du patrimoine arboré, nous avons réalisé une série d'analyses de corrélation. L'objectif principal de cette fonctionnalité est d'identifier les variables les plus influentes entre elles et d'explorer les relations entre différentes variables qualitatives.

Prenons pour exemple l'estimation de l'âge des arbres.

En réalisant le tableau de corrélation suivant :



Nous avons pu déduire que l'âge des arbres était particulièrement corrélé avec le diamètre de son tronc. Nous pouvons également noter une bonne corrélation avec sa hauteur totale, son stade de développement, son statut remarquable et s'il a été planté par la ville ou l'agglomération.

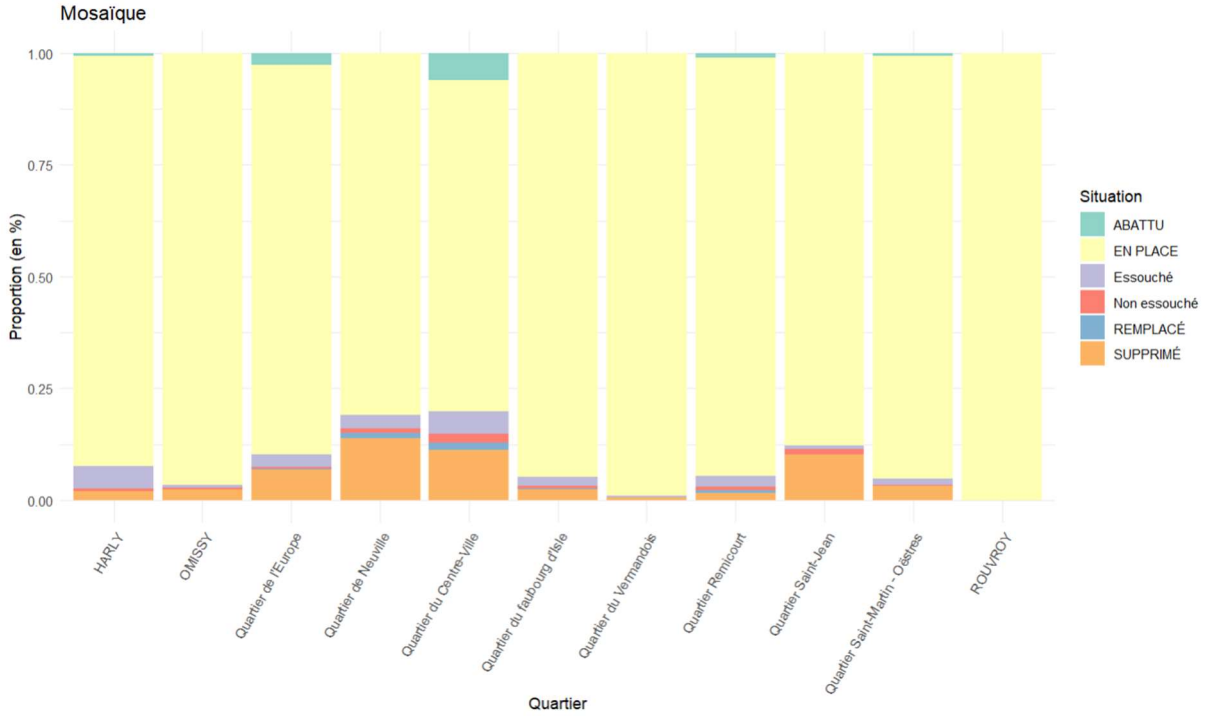
Les analyses bivariées nous ont permis de mettre facilement en relation deux données sélectionnées et de les visualiser efficacement. Il y a une fonction permettant d'analyser des données numériques, l'autre met en relation une donnée numérique et une catégorielle.

Nous avons ensuite fait un tableau croisé qui sera utile par la suite pour le test de chi2 et la visualisation via mosaïque.

Le test du chi2 nous permet d'évaluer si deux variables qualitatives sont indépendantes l'une de l'autre ou s'il existe une association significative entre elles.

Dans l'exemple du test entre clc_quartier et fk_arbre_etat, la valeur du chi2 est 835,19, ce qui indique une corrélation plutôt élevée, le degré de liberté est égal à 50 et le p-value est très proche de 0, le test est donc extrêmement fiable.

Le tableau croisé a donc permis de réaliser la mosaïque suivante :



Dans cet exemple, nous voyons la proportion des situations des arbres en fonction de chaque quartier, il apparait que la majorité sont en place.

Nous pouvons également visualiser toutes sortes de mosaïques pour nous aider à mieux comprendre comment nous pouvons harmoniser le développement écologique de la ville.

Prédiction

Le dernier objectif de ce projet était d'estimer l'âge d'un arbre à partir des autres données à notre disposition, ainsi que de dire s'il doit âtre abattu ou non. Pour ce faire, nous avons utilisé des modèles de régression linéaire et logistique.

Estimation de l'âge :

Afin d'avoir la meilleure estimation de l'âge de l'arbre possible, nous avons utilisé les différents graphiques et tableaux montrant la corrélation des variables créés précédemment pour créer un modèle de régression linéaire. Nous nous sommes également basés sur le coefficient R-carré et le VIF (Variance Inflection Factor). Le R-carré mesure la précision des prédictions du modèle, alors que le VIF permet de quantifier la corrélation entre les variables utilisées lors de la création du modèle.
L'étude de ces coefficients a permis de confirmer les variables à utiliser pour le modèle. Après de nombreux essais, nous avons conclu que le meilleur modèle est celui utilisant le diamètre du tronc, la hauteur totale, le quartier, le port et le fait que l'arbre soit remarquable. Le R-carré du modèle est d'environ 0.691, et le VIF est inférieur à 3.5, ce qui est tout à fait acceptable.

Evaluation de la nécessité d'abattre un arbre :

De la même façon que pour l'estimation de l'âge, nous nous sommes appuyés sur l'étude des corrélations pour créer un modèle de régression logistique répondant au mieux à la problématique. Pour commencer, il a été nécessaire de transformer les valeurs qualitatives en valeurs quantitatives. En effet, la régression logistique ne fonctionne que pour prédire des valeurs entre 0 et 1 (pas de texte).
Nous avons ensuite créé le modèle en lui-même. Pour confirmer les meilleurs paramètres à utiliser, nous avons utilisé les coefficients du pseudo R-carré, qui permet à nouveau de juger la précision du modèle, ainsi que la log-vraisemblance qui permet de comparer les modèles entre eux. Le meilleur modèle que nous avons trouvé est celui utilisant le secteur, la hauteur et le diamètre du tronc, le stade de développement et le type de revêtement.

Conclusion

Notre projet Big Data sur le patrimoine arboré de la ville de Saint-Quentin s'est déroulé en cinq parties : l'appropriation des données, le nettoyage, la visualisation, l'analyse et enfin la prédiction. A travers ces étapes, nous avons développé des outils permettant une représentation visuelle des données grâce à des graphiques et cartes. Cela permet une analyse rapide et efficace des données à disposition, ainsi que de mieux comprendre les besoins du patrimoine arboré de Saint-Quentin. De plus, la mise en place de modèles de régression linéaire et logistique facilite la prise de décision quant à l'entretien requis par les arbres. Pour finir, le nettoyage des données, permettra de les utiliser pour la création de modèles d'intelligence artificielle.