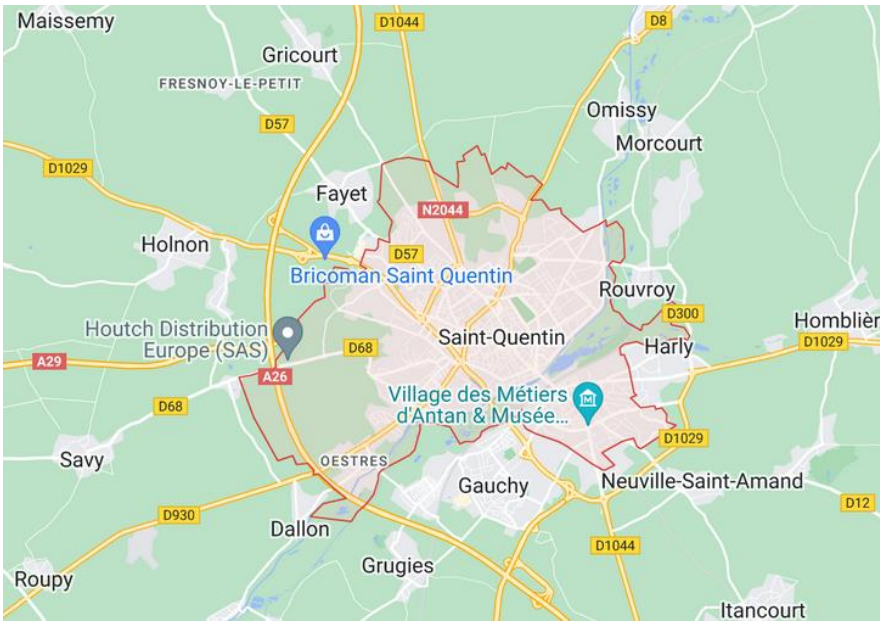
Rapport
Projet de
Big Data
3ème année



Groupe 4

Arthur Grossmann--Le Mauguen

Enzo Guillard

Lucas Bercegeay

CSI 3 Nantes

Table des matières

Fonctionnalité 1 : Description et exploration des données

Nous avons créé un tableau sur Excel pour la description du jeu de données. Pour chaque variable, nous avons expliqué à quoi elle correspond, si elle est quantitative ou qualitative et si nous avons décidé de la conserver pour la suite de notre étude. Voici le résultat :

Qualitative :

Created_User, src_geo, clc_quartier, clc_secteur, fk_arb_etat, fk_staddev, fk_port, fk_pied, fk_situation, fk_revetement, commentaire_environnement, fk_nomtech, last_edited_user, villeca, nomfrancais, nomlatin, Creator, editor, feuillage, remarquable

Quantitative :

X, Y, ObjectID, Created_date, id_arbre, haut_tot, haut_tronc, tronc_diam, dte_plantation, age_estim, fk_prec_estim, clc_nbr_diag, dte_abattage, last_edited_date, GlobalID, CreationDate, EditDate,

Les variables qui concernent la date de création de l'instance, savoir qui la crée, le nom en latin, la date d'édition et autres sont des variables que nous avons décidé de ne pas garder pour la suite puisqu'elles ne nous semblaient ni pertinentes pour l'étude des corrélations et la création de nos histogrammes et graphiques ni pour l'entrainement de notre modèle d'intelligence artificielle.

Pour effacer les doublons, nous avons simplement regardé les arbres qui possédaient exactement les mêmes coordonnées X et Y. En revanche, nous avons également regardé l'état de l'arbre car il y a des cas ou deux arbres possèdent exactement la même coordonnée mais ont un état différent. Il y a des exemples où un arbre a été ajouté sur des coordonnées où un arbre a été détruit. Donc lorsqu'un arbre possédait les mêmes coordonnées X et Y et le même état qu'un autre arbre nous le supprimions.

Pour les colonnes 'clc_quartier' et 'clc_secteur', nous avons regardé les coordonnées de chaque quartier, secteur. Si un arbre n'avait pas de secteur ou de quartier attribué, nous regardions ses coordonnées et nous lui attribuions le quartier ou le secteur correspondant aux coordonnées.

Pour le feuillage, nous avons remplacé les NA par la variable 'inconnu' car nous estimons qu'il est impossible de déduire le feuillage de l'arbre avec le jeu de données, de même pour le revêtement.

Enfin, pour la colonne 'villeca', lorsqu'il y avait des valeurs manquantes, nous regardions le quartier associé. Nous prenions la valeur de 'villeca' la plus fréquente dans le quartier et nous l'associons à l'arbre.

Pour le calcul de quelques statiques, nous avons également calculé quelques moyennes univariés comme la moyenne d'âge estimé. Nous trouvons une moyenne d'environ 31.9 ans, on peut en conclure que cette moyenne peut sembler assez jeune.

En calculant l'écart type qui est de 28.5, nous pouvons constater une grande variabilité entre les âges des arbres de Saint-Quentin.

Nous avons également calculé la médiane qui est de 30, nous avons donc une valeur très proche de la moyenne ce qui implique qu'il n'y a que peu de valeurs extrêmes qui influencent la moyenne.

Par la suite, nous avons aussi calculé des moyennes bivariés, comme certaines moyennes en fonction du quartier où nous nous situons. Par exemple, on observe que dans le quartier d'Harly la moyenne de la hauteur totale d'un arbre est d'environ 14,2 m alors que dans le quartier Saint-Jean elle est de 9,07 m.

On utilise la fonction "group_by" pour regrouper les données par la variable "clc_quartier", cela nous permet de faire des calculs séparément pour chaque quartier. Nous utilisons la fonction "summerise" afin de créer un autre dataframe avec les lignes qui correspondent aux quartiers et les colonnes qui correspondent aux moyennes. On utilise "as.numeric" pour toutes les moyennes car pour certaines nous avions des problèmes, et "na.rm=True" permet d'ignorer les valeurs manquantes. Voici un exemple du résultat que nous obtenons en calculant des moyennes bivariées en fonction du quartier où l'on se situe :

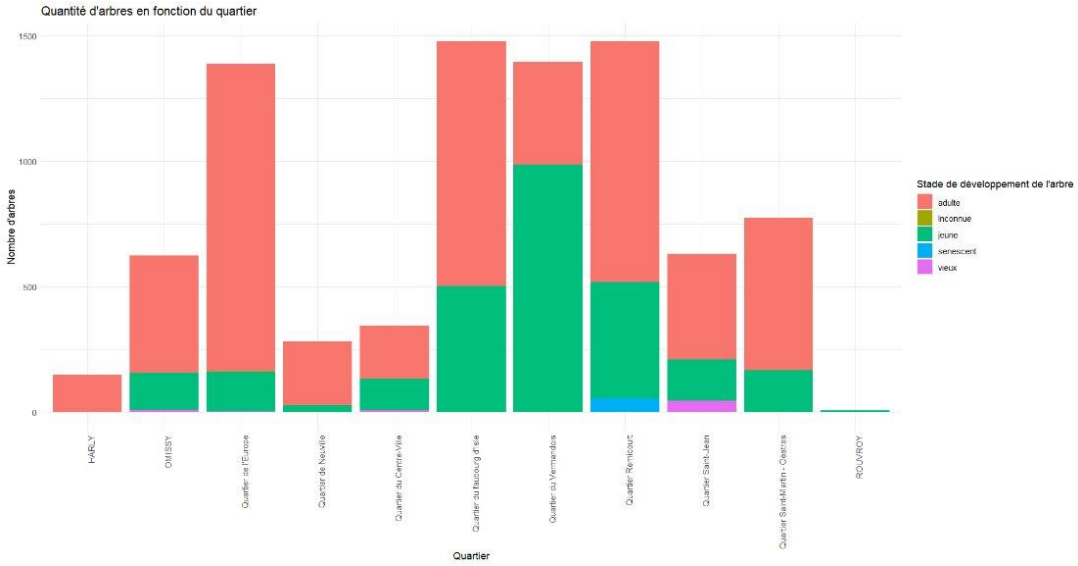
	clc_quartier	moyenne_age_estim	median_age_estim	q1_age_estim	q3_age_estim	iqr_age_estim	moyenne_haut_tot	moyenne_haut_tronc	moyenne_tronc_diam
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	HARLY	34.7	30	30	40	10	14.9	2.46	140.
2	OMISSY	31.7	30	30	35	5	14.7	4.04	96.9
3	Quartier Remicourt	41.2	40	15	50	35	12.4	3.53	112.
4	Quartier Saint-Jean	29.7	30	15	50	35	9.68	2.09	96.8
5	Quartier Saint-Martin - Oëstres	37.4	40	20	50	30	10.3	2.60	100.
6	Quartier de Neuville	37.0	40	30	45	15	13.5	2.63	133.
7	Quartier de l'Europe	35.0	40	30	40	10	13.2	2.61	120.
8	Quartier du Centre-Ville	38.0	50	15	50	35	9.41	2.32	91.9
9	Quartier du Vermandois	18.4	15	15	15	0	8.08	1.71	87.6
10	Quartier du faubourg d'Isle	26.8	30	10	40	30	10.1	2.63	89.8
11	ROUVROY	5	5	5	5	0	3	1	10

>

Nous remarquons ci-dessus, lorsque l'on prend en exemple le quartier d'Harly, que la moyenne est de 34.7 pour une médiane de 30. La proximité de ces deux valeurs nous montre une certaine symétrie dans notre jeu de données et qu'il n'y a pas de valeurs aberrantes qui faussent la moyenne. L'écart interquartile ici est assez faible ce qui signifie que nos valeurs sont regroupées et assez proches de la médiane.

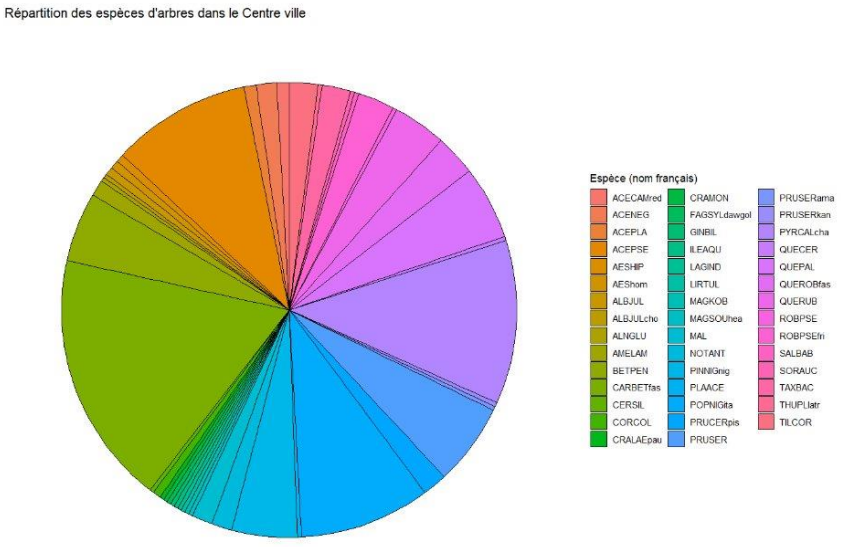
Fonctionnalité 2 : Visualisation des données sur des graphiques

Pour cette partie, nous avons donc construit des histogrammes entre différentes variables à notre disposition. Pour commencer, nous avons réalisé des histogrammes selon la quantité d'arbres, suivant leur statut de développement mais cela ne nous semblait pas le plus pertinent. Nous avons donc tracé la quantité d'arbres en fonction du quartier avec différentes couleurs suivant leur statut de développement comme représenté ci-dessous :



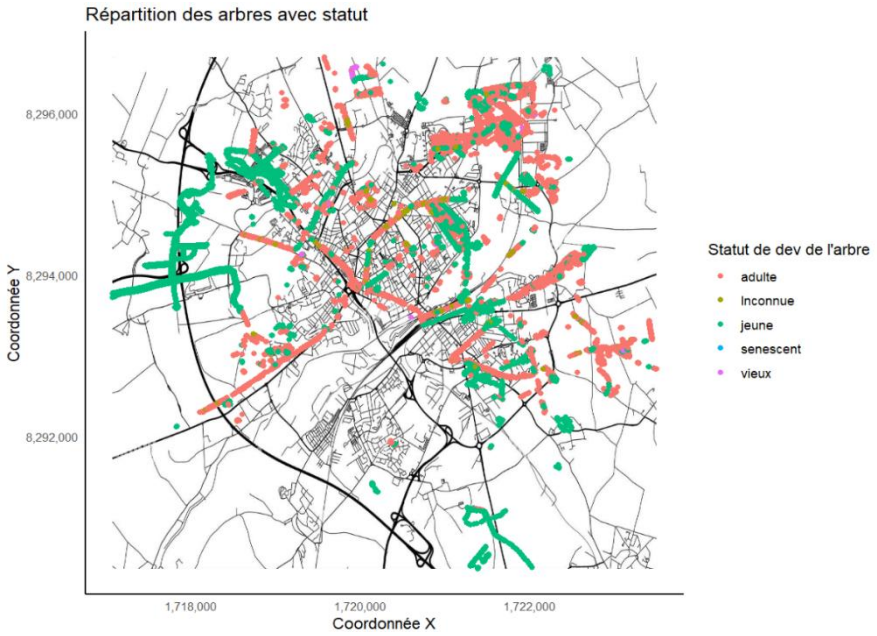
Nous pouvons donc voir que des quartiers ont plus d'arbres jeunes que d'arbres vieux, ce qui peut nous induire que le quartier est plus récent que les autres par exemple. De même, plus le quartier possède d'arbres, plus le quartier est grand ou dense.

Parallèlement, nous avons réalisé un histogramme du nombre d'instance des espèces d'arbres par quartier pour chaque quartier qu'on avait à disposition. Malheureusement, nous trouvions l'histogramme illisible, c'est pour cela que nous l'avons affiché sous la forme d'un diagramme circulaire (diagramme en camembert). Un exemple ci-dessous pour le centre-ville :



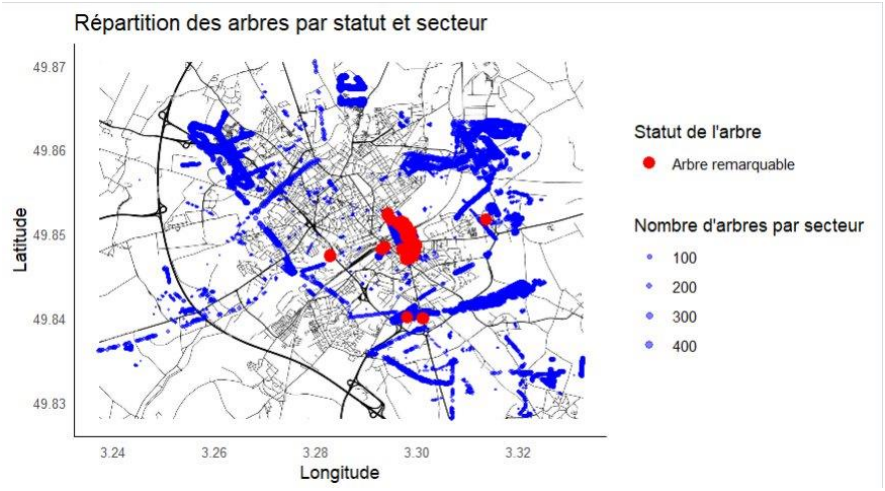
Fonctionnalité 3 : Visualisation des données sur une carte

Dans un premier temps, nous avons fait une représentation graphique de la répartition des arbres dans la ville de Saint-Quentin en fonction de leur statut. Nous avons utilisé une carte que nous avons pris sur internet que nous avons mis en fond, puis nous avons superposé un point par arbre sur la carte et fait une légende en fonction de leur statut. Nous obtenons le résultat suivant :

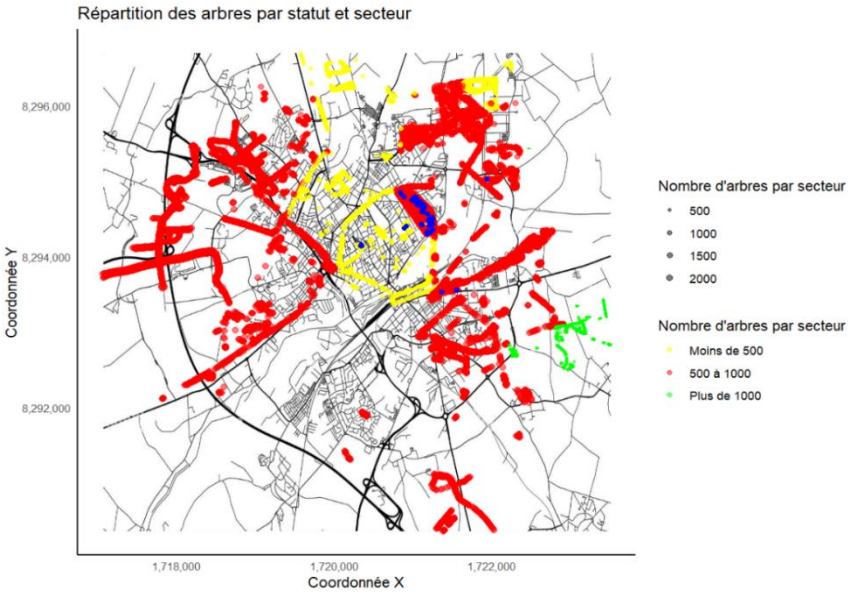


Ce graphique nous permet de voir où sont situés les arbres jeunes, et donc hypothétiquement les quartiers récents par rapport aux quartiers plus anciens avec les arbres adultes, voir vieux.

Nous avons fait un autre graphique avec un code assez similaire que celui-là, mais ici on peut voir quels arbres sont remarquables ou non, sur ce graphique nous avons aussi fait varier la taille des points par rapport au nombre d'arbre dans chaque secteur:



Nous avons fait une deuxième représentation graphique de la répartition des arbres dans la ville en fonction du statut et du secteur. Dans le même principe que la première, nous superposons les points qui représentent les arbres sur la carte. En fonction du nombre d'arbres dans le secteur le point est plus ou moins grand. Nous avons également une couleur différente en fonction du nombre. Nous avons le graphique ci-dessous :

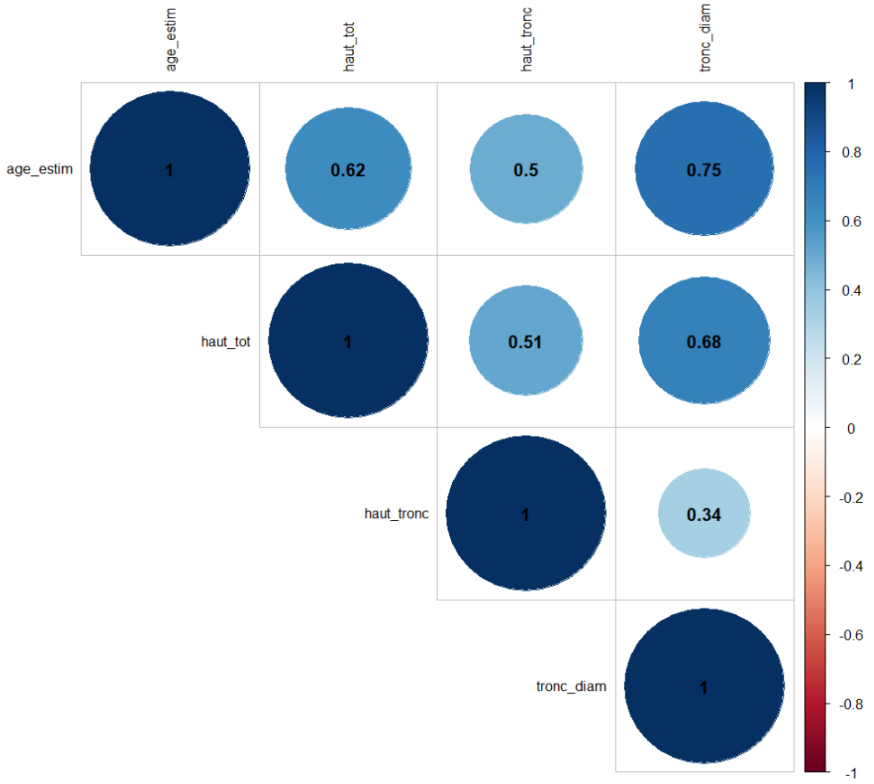


Fonctionnalité 4 : Etude des corrélations entre variables

Nous avons calculé la matrice de covariance afin d'identifier les relations (linéaires ou non) entre les variables et si ces relations sont positives ou négatives :

	age_estim	haut_tot	haut_tronc	tronc_diam
age_estim	381.00167	72.230902	17.119149	829.44877
haut_tot	72.23090	35.470263	5.377363	228.48562
haut_tronc	17.11915	5.377363	3.116002	33.69513
tronc_diam	829.44877	228.485620	33.695125	3189.51878

Nous avons également calculé la matrice de corrélation de Pearson pour déterminer si nos corrélations sont positives ou négatives, les valeurs variant entre -1 et 1. Dans notre cas, toutes les valeurs sont positives, ce qui indique que toutes nos relations linéaires sont positives. La corrélation de Pearson la plus forte, de 0.75, se trouve entre le diamètre du tronc et l'âge estimé. Cette forte corrélation positive suggère qu'une régression linéaire entre ces deux variables serait appropriée.



```
$fk_arb_etat

        Pearson's Chi-squared test

data:  table(var, t6$age_estim)
X-squared = 1178.2, df = 160, p-value < 2.2e-16


$fk_stadedev

        Pearson's Chi-squared test

data:  table(var, t6$age_estim)
X-squared = 11487, df = 128, p-value < 2.2e-16


$fk_situation

        Pearson's Chi-squared test

data:  table(var, t6$age_estim)
X-squared = 870.86, df = 64, p-value < 2.2e-16


$clc_quartier

        Pearson's Chi-squared test

data:  table(var, t6$age_estim)
X-squared = 10939, df = 320, p-value < 2.2e-16


$clc_secteur

        Pearson's Chi-squared test

data:  table(var, t6$age_estim)
X-squared = 70932, df = 8160, p-value < 2.2e-16
```

Le test du Chi2 est utilisé pour examiner les relations entre deux variables catégorielles. Il permet de déterminer si une association observée entre les variables dans un tableau croisé(contingence) est statistiquement significative ou si elle est due au hasard.

Interprétation d'un test du Chi-carré :

- Valeur p (p-value) : Si la valeur p est inférieure à un seuil alpha (souvent 0.05), on rejette l'hypothèse nulle d'indépendance. Cela signifie qu'il existe une association statistiquement significative entre les variables.
- Statistique du Chi-carré : La valeur de la statistique du Chi-carré indique l'écart entre les fréquences observées et attendues. Plus cette valeur est élevée, plus il y a de chances que l'hypothèse nulle soit rejetée.

Mosaic Plot : une visualisation graphique qui permet de représenter les relations entre deux (ou plusieurs) variables catégorielles. Il est particulièrement utile pour illustrer les résultats d'un test du Chi-carré. Dans un mosaic plot, les tailles des cases sont proportionnelles aux fréquences des combinaisons de catégories des variables.

Interprétation d'un mosaic plot :

- Taille des rectangles : La taille des rectangles représente les fréquences des combinaisons de catégories des deux variables. Des rectangles plus grands indiquent des fréquences plus élevées. Dans notre cas, nous pouvons en conclure que la majorité des arbres n'ont pas de revêtement, mais que ceux qui en ont un sont majoritairement en ville.

Fonctionnalité 5 : Etude des corrélations entre variables

Harmoniser le développement global de la ville

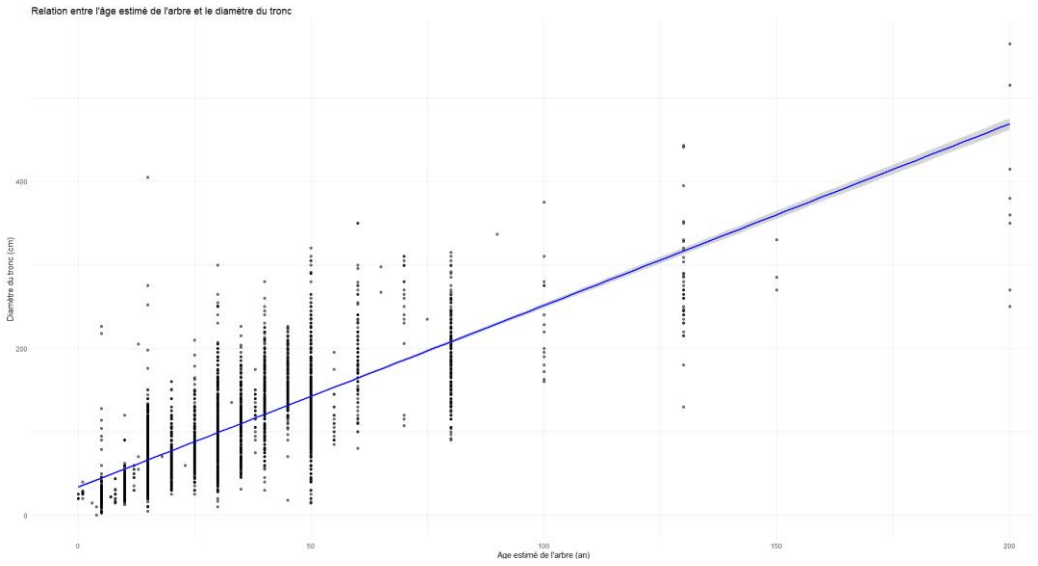
Pour harmoniser le développement global de la ville, nous envisageons de calculer la densité d'arbres par secteur et de planter des arbres dans les secteurs présentant la densité la plus faible. De plus, nous pourrions harmoniser les espèces en plantant celles qui sont les moins présentes dans les quartiers où il faut planter. Nous n'avons malheureusement pas eu le temps d'obtenir un code fonctionnel pour cette partie.

Etude d'une régression pour prédire l'âge d'un arbre

Une valeur proche de 1 dans la matrice de corrélation de Pearson (cf. fonctionnalité 4) entre deux variables indique une forte corrélation positive entre ces variables. Cela signifie que lorsque la valeur d'une variable augmente, la valeur de l'autre variable augmente également de manière proportionnelle, et vice versa lorsque la valeur d'une variable diminue.

Cette forte corrélation positive est un indicateur clé qui justifie l'utilisation de la régression linéaire pour modéliser la relation entre ces deux variables.

Le résultat de la régression linéaire pour prédire l'âge de l'arbre est le suivant :



Obtenue de cette manière :

```
# Créer un modèle de regression pour prédire l'age d'un arbre
# Construire le modèle de prédiction en utilisant tronc_diam comme prédicteur
modele <- lm(age_estim ~ tronc_diam, data = t6)
```

Nous cherchons maintenant à estimer la qualité de notre modèle :

```
> summary(modele)

Call:
lm(formula = age_estim ~ tronc_diam, data = t6)

Residuals:
    Min      1Q  Median      3Q     Max
-95.338  -7.043  -1.517   5.976 129.971

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.015571   0.288248    17.4   <2e-16 ***
tronc_diam  0.260055   0.002462   105.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.86 on 8550 degrees of freedom
  (5 observations effacées parce que manquantes)
Multiple R-squared:  0.5661,    Adjusted R-squared:  0.5661
F-statistic: 1.116e+04 on 1 and 8550 DF,  p-value: < 2.2e-16
```

```
> # Erreur standard des résidus (RSE) :
> summary(modele)$sigma
[1] 12.85765
> # Coefficient de détermination (R**2)
> summary(modele)$r.squared
[1] 0.5661442
```

Estimate: C'est l'estimation du coefficient pour chaque variable.

Std. Error: L'erreur standard de l'estimation du coefficient.

t value: Le rapport de la valeur estimée du coefficient sur son erreur standard.

Pr(>|t|) : La valeur de p associée au test de nullité du coefficient (généralement utilisé pour tester si le coefficient est significativement différent de zéro)

Résidus : Cette partie du résumé fournit des statistiques sur les résidus du modèle (erreurs résiduelles)

RSE (Root Mean Square Error) : Un RSE plus faible indique que les prédictions du modèle sont en moyenne plus proches des valeurs réelles de la variable dépendante.

R-squared (Coefficient de détermination) : mesure la proportion de variance expliquée par le modèle de régression par rapport à la variance totale des données. Il varie de 0 à 1 et est souvent exprimé en pourcentage.

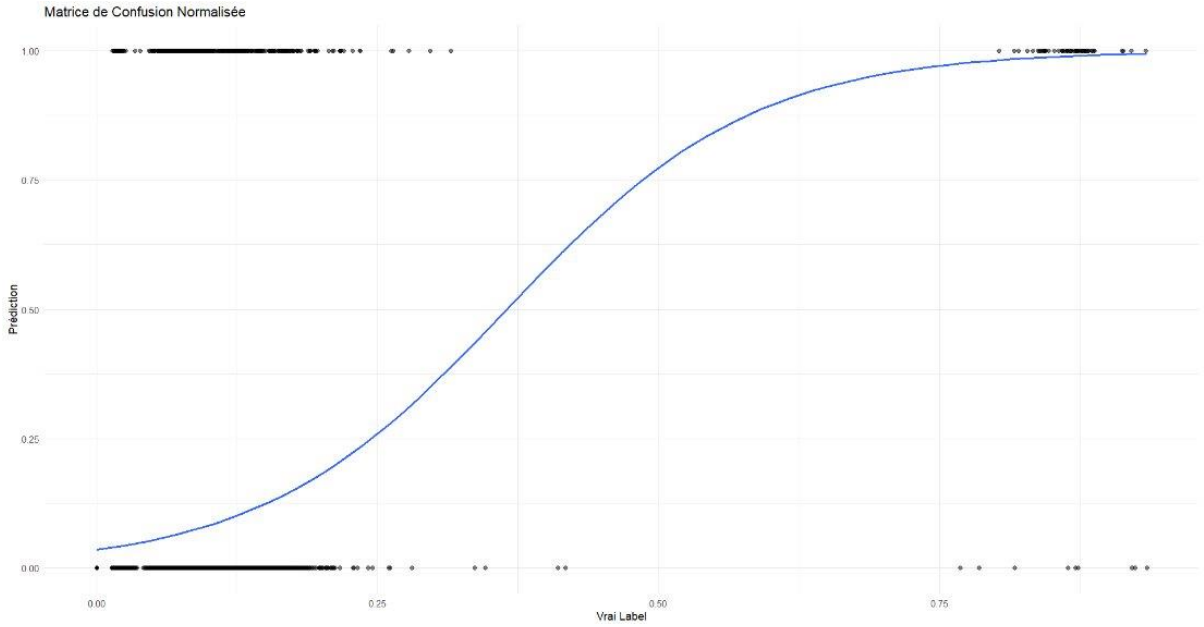
Interprétation : Proportion de variance expliquée : Un R-squared proche de 1 indique que le modèle explique une grande partie de la variance des données observées. Cela signifie que les variables indépendantes dans le modèle expliquent efficacement la variation de la variable dépendante. Dans notre cas, une valeur R-squared de 0.56 signifie que 56% de la variance totale de la variable dépendante est expliquée par les variables indépendantes incluses dans le modèle.

Etude d'une régression logistique pour prédire les arbres à abattre

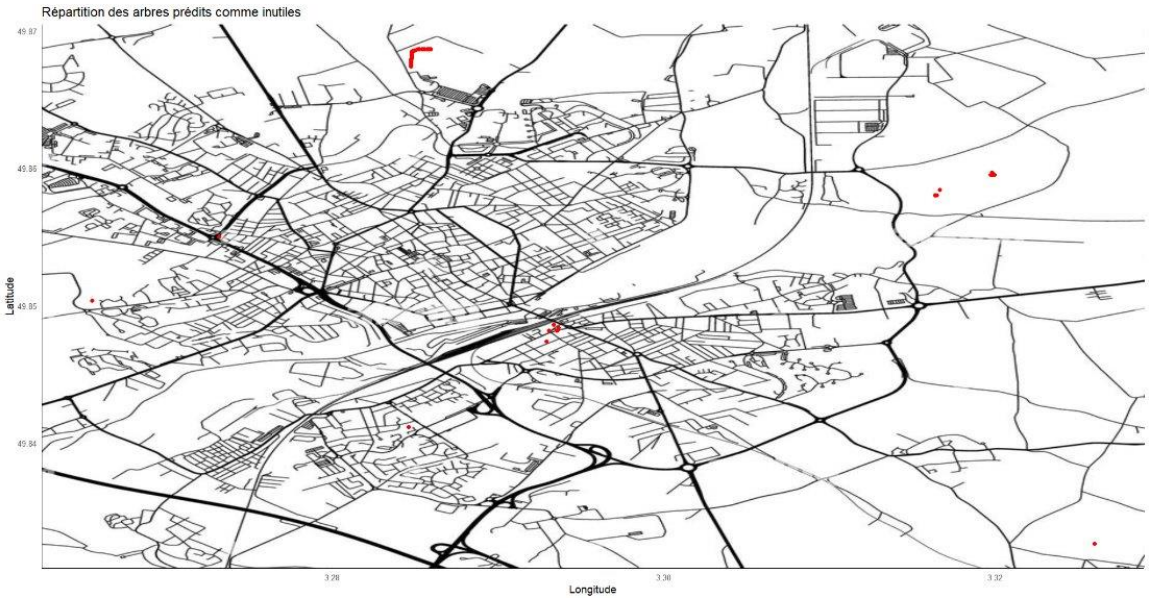
Pour prédire les arbres à abattre dans la ville, nous avons fait une régression logistique. On a tout d'abord traité les cas "Essouché", "Non essouché" et "ABATTU" comme "ABATTU". Ensuite, après cela, on a lancer notre régression logistique et on a affiché la matrice de confusion qui est la suivante :

```
> print(conf_matrix)
   pred
      0    1
0 7915    9
1  570   58
```

L'interprétation est simple : 58 des arbres devants être abattus et 7915 arbres ne devant pas être abattus ont bien été prédit par le modèle tandis que 570 devant être abattus et 9 arbres ne devant pas être abattus ont été mal prédit par le modèle. Nous avons donc pu calculer la précision qui est de 0.86 et l'exactitude (Accuracy) qui est de 0.92. On en déduit donc que notre modèle à une précision plutôt correct, il prédit bien les arbres à abattre et l'exactitude est correcte, même si le fait qu'il y est beaucoup d'arbres non abattus, cela peut influencer l'exactitude. Nous avons donc représenté la régression logistique sur un graphique, sachant que les points correspondent à la matrice de confusion :



Nous avons bien une forme sigmoïde et une bonne séparation des classes. Nous avons donc représenté les arbres à abattre sur une carte :



Fonctionnalité 6 : Export pour l'IA

Pour faciliter l'intégration des données dans la partie du projet dédiée à l'intelligence artificielle, nous avons créé et rempli un fichier .csv contenant les données nettoyées. Dans cette étape préliminaire, nous avons traité les valeurs manquantes (NA ou vides) en les remplaçant lorsque cela était possible à partir d'autres données disponibles, ou en supprimant les objets concernés lorsque cela n'était pas faisable.

À ce stade, nous disposons d'un tableau de 8552 objets et de 20 variables pertinentes pour notre analyse (clc_quartier, clc_secteur, haut_tot, haut_tronc, tronc_diam, fk_arb_etat, fk_stadedev, fk_port, fk_pied, fk_situation, fk_revetement, age_estim, fk_prec_estim, clc_nbr_diag, fk_nomtech, feuillage, remarquable, longitude, latitude, villeca). Nous avons sélectionné spécifiquement ces variables en vue de leur utilisation pour entraîner notre modèle d'intelligence artificielle la semaine prochaine.

```r
###PARTIE 6: Export pour l'IA
#Supprimer la colonne pred et etat_binaire ajouté au tableau t6
t6 <- t6 %>% select(-pred, -etat_binaire)
t6 <- subset(t6, select = -c(X, Y, count))

write.csv(t6, "Patrimoine_Arboré_Nettoyé.csv", row.names = FALSE)
f1 <- read.csv(file = "Patrimoine_Arboré_Nettoyé.csv")
tab <- data.frame(f1)
View(tab)
```

clc_quartier	clc_secteur	haut_tot	haut_tronc	tronc_diam	fk_arb_etat	fk_stadedev	fk_port	fk_pied	fk_situation	fk_revetement	age_estim	fk_prec_estim	clc_nbr_diag	fk_nomtech	feuillage	remarquable	longitude	latitude	villeca
Quartier du Centre-Ville	"Quai Gayant"	6	2	37	"EN PLACE"	"jeune"	"semi libre"	"gazon"	"Alignement"	"Non"	15	"5"	"0"	"QUERUB"	"Feuillu"	"Non"	3.29326122090351	49.8405025184876	"VILLE"
Quartier du Vermandois	"Stade Cepy"	13	1.160	"EN PLACE"	"adulte"	"semi libre"	"gazon"	"Groupe"	"Non"	50	"10"	"0"	"PINNIGnig"	"ConifÂ¨re"	"Non"	3.27788323626716	49.8629785814885	"CASQ"	
Quartier du Centre-Ville	"Rue Villebois Mareuil"	12	3.116	"REMPLACÃ‰"	"adulte"	"semi libre"	"gazon"	"Alignement"	"Non"	30	"10"	"0"	"ACEPSE"	"Feuillu"	"Non"	3.28984041950138	49.84486993394	"VILLE"	
Quartier de l'Europe	"Square Des Marronniers"	16	3.150	"EN PLACE"	"adulte"	"semi libre"	"gazon"	"Groupe"	"Non"	50	"2"	"0"	"ACEPLA"	"Feuillu"	"Non"	3.31388980094243	49.859262541222	"VILLE"	
Quartier de l'Europe	"Avenue Buffon"	5	2.170	"EssouchÃ©"	"adulte"	"rÃ©duit"	"gazon"	"Isolé"	"Non"	40	"2"	"0"	"SALBAB"	"Feuillu"	"Non"	3.31447582707741	49.8558925883406	"VILLE"	
Quartier de l'Europe	"Rue Laplace"	8	3.103	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"fosse arbre"	"Alignement"	"Oui"	30	"10"	"0"	"ACEPLA"	"Feuillu"	"Non"	3.3174089786037	49.8580210719523	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	6	3.100	"EN PLACE"	"adulte"	"rÃ©duit"	"gazon"	"Alignement"	"Non"	50	"10"	"1"	"TILCOR"	"Feuillu"	"Non"	3.2649339111111	49.8411326134694	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	4.135	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.26390622217511	49.8409445224943	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	4.121	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.26365274400262	49.8408976939192	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	2.107	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.26313528398209	49.8408071283402	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	2.110	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.26124309087253	49.8404346992741	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	5.115	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.26097437904898	49.8403850141292	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	5.103	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.26074061633878	49.8403373348477	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	4.160	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.25517180168126	49.8392184770522	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	4.140	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.25465103970517	49.8391228461668	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	4.160	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.25439387080466	49.8390681679564	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	5.163	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.25412692957558	49.8390080084382	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	4.139	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.25364040680104	49.8389098475865	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	11	3.142	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.25081950379401	49.8383591746301	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	11	4.160	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.24852181325394	49.8378862908989	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	5.110	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.24826408503611	49.8378354456097	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	2.125	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.24798840112795	49.8377751471588	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	9	2.138	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.24778247710816	49.8377323151496	"VILLE"	
Quartier Saint-Martin - OÃ»stres	"Rue De Paris"	10	3.125	"EN PLACE"	"adulte"	"rÃ©duit relÃ¢chÃ©"	"gazon"	"Alignement"	"Non"	50	"10"	"2"	"TILCOR"	"Feuillu"	"Non"	3.24722802113465	49.8376243677027	"VILLE"	

Kanban

A FAIRE	EN COURS	FAIT
	Fonctionnalité 5: Etude des corrélations entre variables : • La ville a une politique urbaine qui consiste à planter des nouveaux arbres. Dans quelle zone faut-il les planter pour harmoniser le développement global de la ville ?	Fonctionnalité 1 : Description et exploration des données: • Description du jeu de données • Statistiques descriptives univariées, bivariées • Nettoyage des données • Valeurs manquantes, valeurs aberrantes • Doublons
		Fonctionnalité 2: Visualisation des données sur des graphiques : • Créer des représentations graphiques Exemple : répartition des arbres suivant leur stade de développement • Créer des histogrammes Exemple : quantité d'arbres en fonction du quartier / secteur, de sa situation
		Fonctionnalité 3: Visualisation des données sur une carte : • Construire des cartes des arbres répertoriés (grâce au latitude et longitude) Exemple : en général, dans quelle partie de Saint-Quentin se situent les arbres remarquables ? • Proposer une représentation graphique sous formes de cartes de la quantité d'arbres par quartier / secteur
		Fonctionnalité 4: Etude des corrélations entre variables : • Quels sont les liens entre les variables ? Exemple : si on veut estimer la variable âge de l'arbre, quelles sont les variables importantes dans son estimation ? • Conduire des analyses bivariées • Étude des relations entre variables qualitatives Faire des tableaux croisés et des tests d'indépendance du chi2 sur les tableaux entre les différentes variables Représenter graphiquement ces tableaux (mosaicplot) et les analyser

		Fonctionnalité 5: Etude des corrélations entre variables : • On souhaite prédire la variable âge de l'arbre. Faire une étude de régression. • On souhaite savoir quels sont les arbres à abattre. Faire une étude à l'aide de régression logistique.
		Fonctionnalité 6: Export pour l'IA • Exporter le fichier nettoyé en format csv pour une utilisation dans la partie Intelligence Artificielle.
		Présentation
		Rapport

Sources

Pour nous aider dans notre étude nous avons utilisés des informations présentes sur les sites suivants :

https://hub.arcgis.com/maps/aaf5c6a2a3cc49da84c8cc60b97c3507/about

https://oseox.fr/langage-r/importer-fichier-csv.html

https://www.developpez.net/forums/d1042486/general-developpement/algorithme-mathematiques/statistiques-data-mining-data-science/r/erreur-read-table-ligne-x-n-avait-p-elements/

https://sites.google.com/site/rgraphiques/home/les-objets-r/les-tableaux-data-frames

https://juba.github.io/tidyverse/10-dplyr.html

https://www.delftstack.com/fr/howto/r/remove-duplicates-in-r/#utilisez-la-fonction-distinct-du-package-dplyr-pour-supprimer-les-lignes-en-double-par-colonne-dans-r

https://delladata.fr/nettoyer-et-valider-les-donnees-avec-r/

https://thinkr.fr/abcdr/comment-supprimer-les-na-valeurs-manquantes-dans-r-avec-dplyr/

https://thinkr.fr/abcdr/comment-remplacer-chaine-caracteres/

https://www.developpez.net/forums/d1858125/general-developpement/algorithme-mathematiques/statistiques-data-mining-data-science/r/remplacer-cellules-vides-na/

https://r-graph-gallery.com/120-plot-with-an-image-as-background.html