Soutenance projet Big Data

Lelièvre Tom, Le Boulch Antoine, Simon Nathan

Nesma Settouti
Abir El Haj

Jeu de données

```
> colnames(data)
 [1] "X"                          "Y"                  "OBJECTID"           "created_date"       "created_user"       "src_geo"
 [7] "clc_quartier"               "clc_secteur"        "id_arbre"           "haut_tot"           "haut_tronc"         "tronc_diam"
[13] "fk_arb_etat"                "fk_stadedev"        "fk_port"            "fk_pied"            "fk_situation"       "fk_revetement"
[19] "commentaire_environnement" "dte_plantation"     "age_estim"          "fk_prec_estim"      "clc_nbr_diag"       "dte_abattage"
[25] "fk_nomtech"                "last_edited_user"   "last_edited_date"   "villeca"            "nomfrancais"        "nomlatin"
[31] "GlobalID"                  "CreationDate"       "Creator"            "EditDate"           "Editor"             "feuillage"
[37] "remarquable"
> length(colnames(data))
[1] 37
```

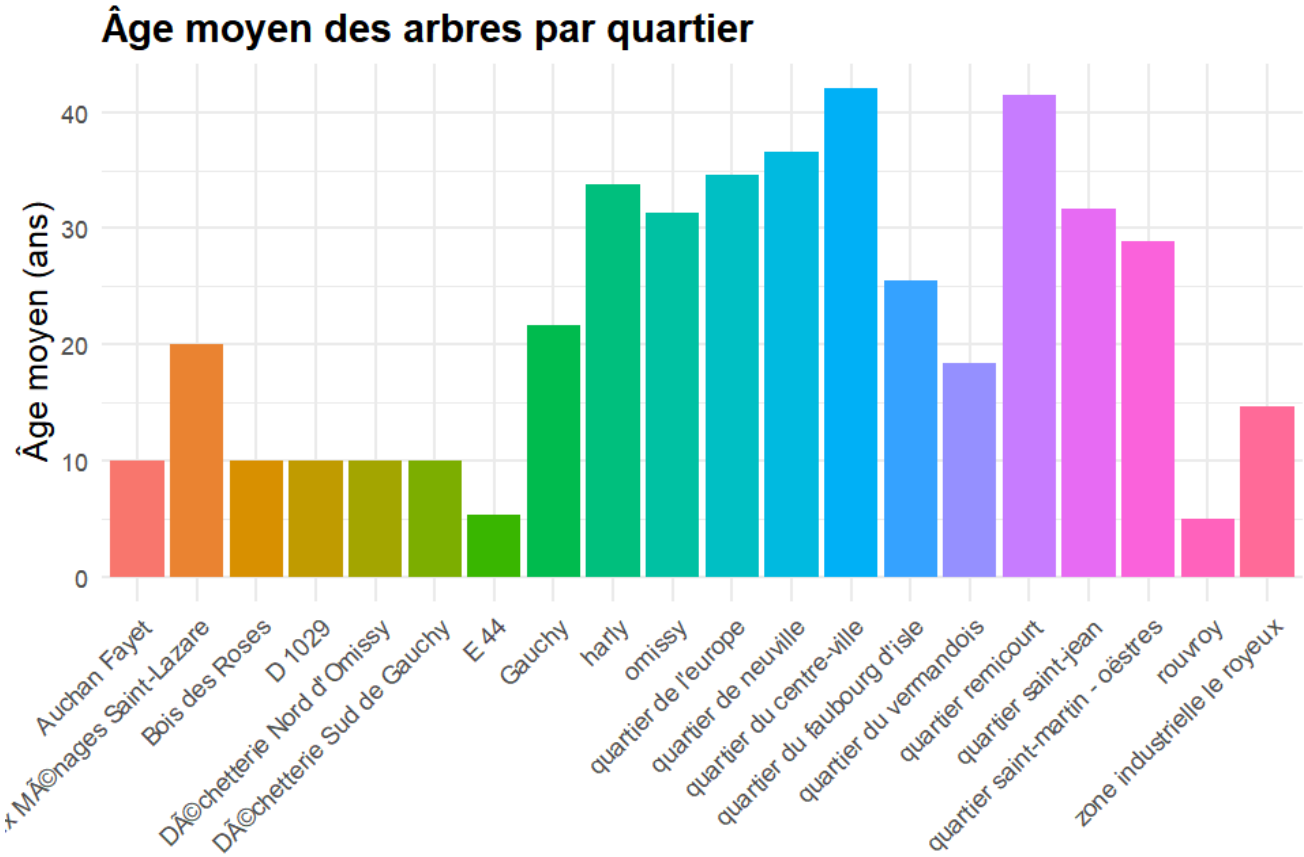
Nettoyage des données

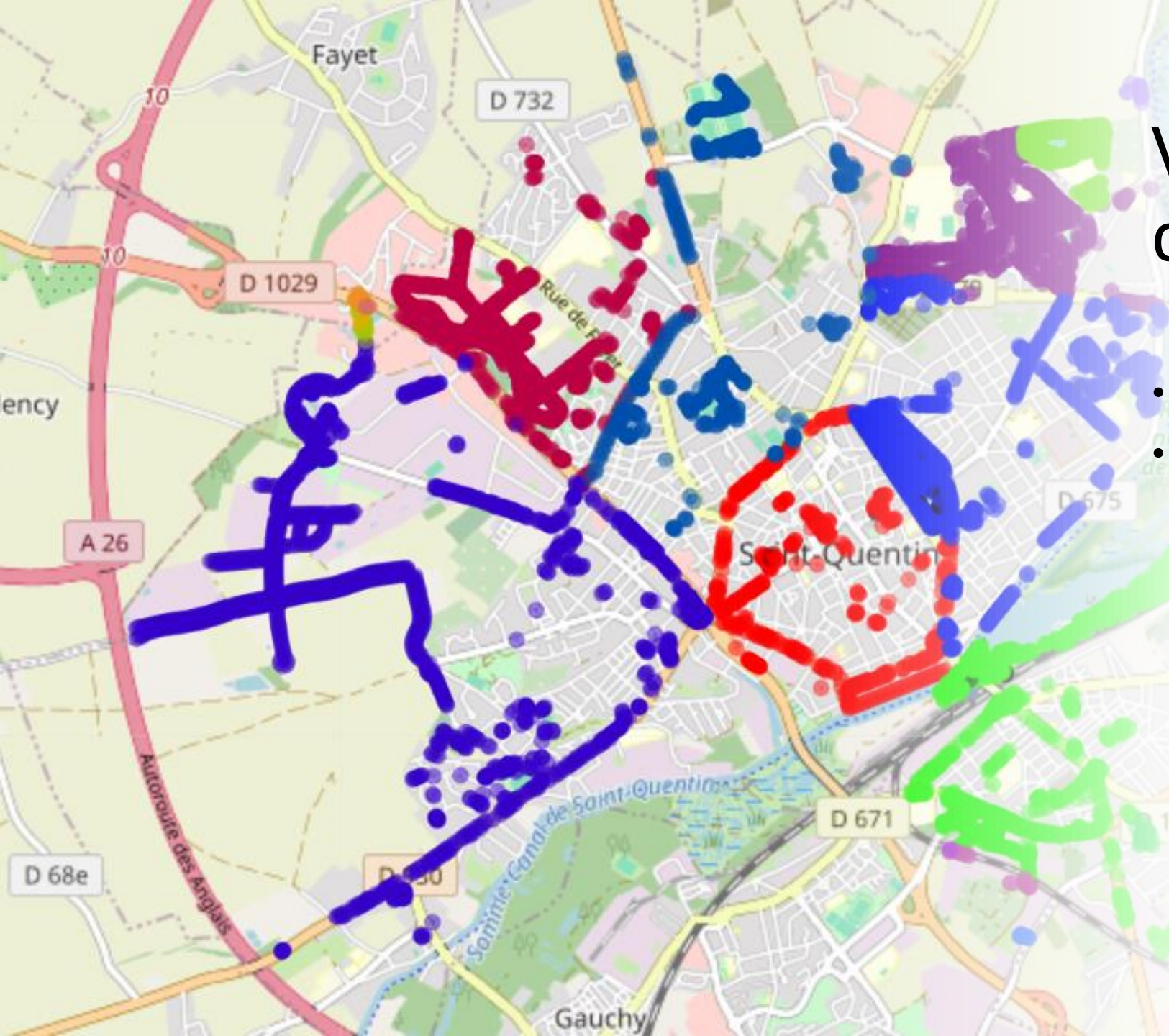
- Encodage UTF-8
- to_lower
- Conversion des types données
- Remplissage case vide ""
- Remove les NA sur colonne des coordonnées
- Remove les doublons GlobalID
- Adapter zone industrielle royeux
- Remove empty line (colonne NA * 14)
- Get quartier by coords
- Convert data (pour remettre le bon type)
- to_factor
- Conversion coordonnées EPSG3949 -> EPSG4326

Visualisation des données : plot

- Création de plots permettant de visualiser nos données

- Permettent de mieux comprendre notre jeu de donnée et de faire des hypothèses sur des corrélations possibles entre nos variables


Âge moyen des arbres par quartier
Âge moyen (ans)
0
10
20
30
40
Auchan Fayet
x MÃ©nages Saint-Lazare
Bois des Roses
D 1029
DÃ©chetterie Nord d'Omissy
DÃ©chetterie Sud de Gauchy
E 44
Gauchy
harly
omissy
quartier de l'europe
quartier de neuville
quartier du centre-ville
quartier du faubourg d'isle
quartier du vermandois
quartier remicourt
quartier saint-jean
quartier saint-martin - oëstres
rouvroy
zone industrielle le royeux
Quartier



Visualisation des données : carte

- Utilisation du package leaflet
- Cartes par
  - Quartiers
  - Feuillages
  - Stade de développement
  - Etat

Visualisation des données : carte

- Démo
  - https://maps-bigdata-groupe-12.netlify.app/

Lien entre variables

Corrélation entre variables quantitatives

Variable 1	Variable 2	Coefficients de pearson
Hauteur totale	Âge estim	0.589
Diamètre du tronc	Âge estim	0.767
Hauteur tronc	Diamètre du tronc	0.685
Hauteur totale	Diamètre du tronc	0.368

Lien entre variable

```
        Pearson's Chi-squared test

data:  tab
X-squared = 2752.1, df = 4, p-value < 2.2e-16
```

Analyse bivariée mixte

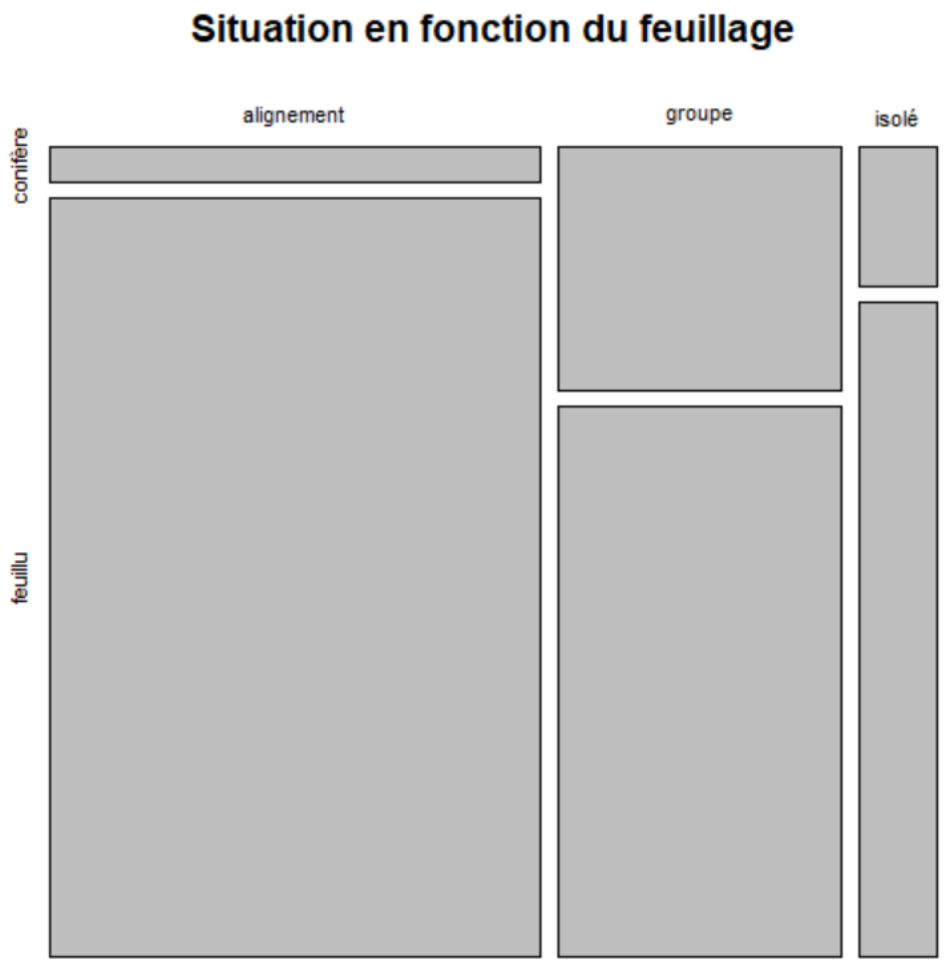
Analyse qualitative entre `remarquable` et `fk_stadedev`

Pourcentage en fonction de l'effectif des arbre par chaque catégories

Remarquable/stade dev	Adulte	Jeune	senescent	vieux
non	0.98	0.99	0.33	0.97
oui	0.01	0.0006	0.71	0.02
Nb d'arbre	6396	4421	53	68

En général, les arbres sénescent sont le plus souvent des arbres remarquables.

Lien entre les variables




```
            conifère feuillu
  alignement      285    6022
  groupe         1118    2526
  isolé           173     816

              conifère    feuillu
  alignement 0.02605119 0.55045704
  groupe     0.10219378 0.23089580
  isolé      0.01581353 0.07458867

        Pearson's Chi-squared test

data:  tab2
X-squared = 1290.4, df = 2, p-value < 2.2e-16
```

Prédiction tronc_diamètre

```
lm(formula = tronc_diam ~ haut_tronc + haut_tot + fk_stadedev +
    feuillage + age_estim, data = data_temp)

Residuals:
    Min      1Q  Median      3Q     Max
-155.56  -18.92   -4.72   18.32  330.06

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.47731   23.13453   0.194    0.847
haut_tronc           -3.66405    0.23820 -15.382  < 2e-16 ***
haut_tot              3.53034    0.07593  46.492  < 2e-16 ***
fk_stadedevadulte    26.38571   23.14810   1.140    0.254
fk_stadedevjeune     15.61901   23.12899   0.675    0.500
fk_stadedevsenescent  2.62513   23.68555   0.111    0.912
fk_stadedevvieux     29.38702   23.53165   1.249    0.212
feuillagefeuillu     -4.47731    0.91191  -4.910 9.26e-07 ***
age_estim             1.55597    0.02527  61.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.69 on 10105 degrees of freedom
  (13 observations effacées parce que manquantes)
Multiple R-squared:  0.6722,    Adjusted R-squared:  0.6719
F-statistic:  2590 on 8 and 10105 DF,  p-value: < 2.2e-16
```

Prédiction âge estim

```
Call:
lm(formula = age_estim ~ tronc_diam + haut_tot + haut_tronc,
    data = data_temp)

Residuals:
    Min      1Q  Median      3Q     Max
-91.152  -7.280  -1.889   5.282 135.529

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.271316   0.280736  -0.966   0.3338
tronc_diam   0.231658   0.002887  80.228   <2e-16 ***
haut_tot     0.056554   0.030660   1.845   0.0651 .
haut_tronc   3.046354   0.083061  36.676   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.29 on 10151 degrees of freedom
Multiple R-squared:  0.6362,    Adjusted R-squared:  0.6361
F-statistic:  5916 on 3 and 10151 DF,  p-value: < 2.2e-16
```

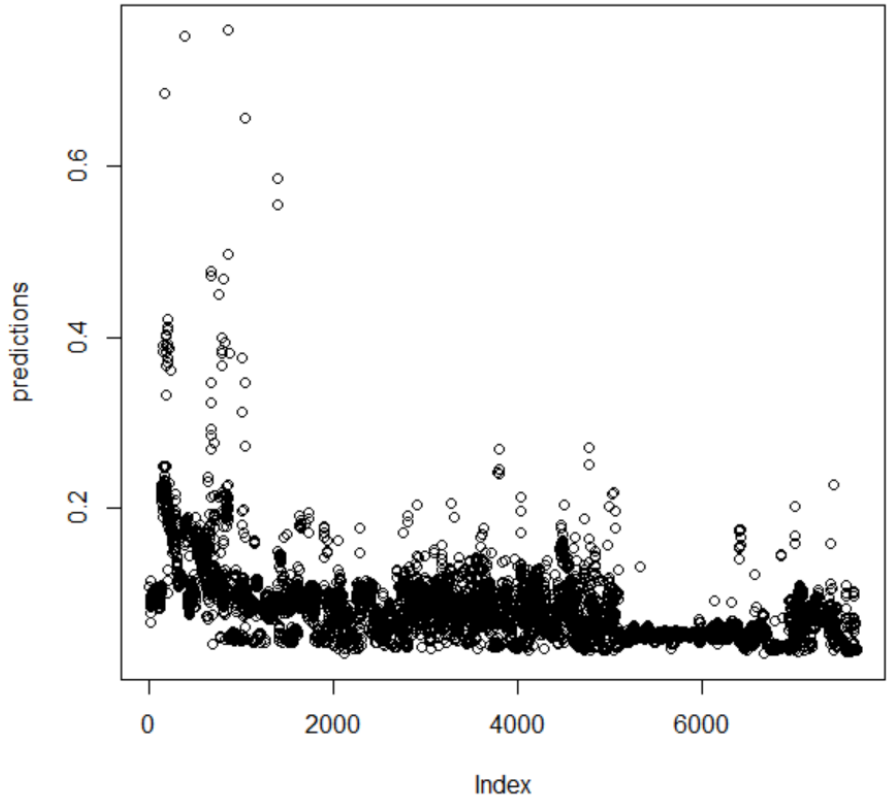
Prédiction remarquable

Régression logistique

tronc_diam + haut_tot + age + fk_stadedev + nomfrancais

2 levels oui non

Prédiction des arbres à supprimé



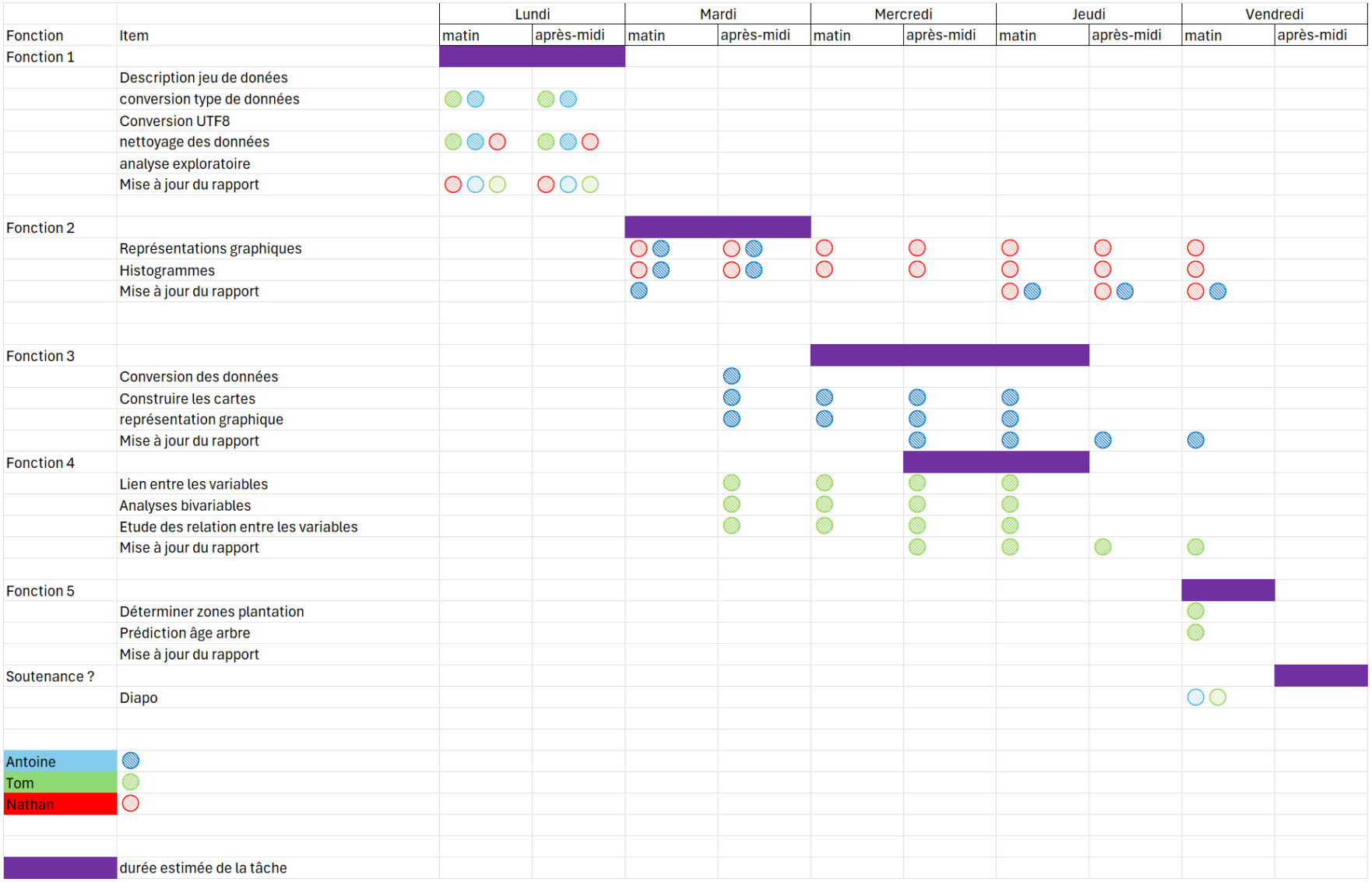
Régression logistique

tronc_diam + haut_tot + age_estim

"supprimé"
"en place"

38 arbres

Organisation



Fin

https://fr.wikipedia.org/wiki/S%C3%A9quoia_g%C3%A9ant


Questions