PROJET MACHINE LEARNING

Marlene CHEVALIER et Olga SILVA

3/12/2020

1- Cadrage

L'objectif du projet est de mettre en oeuvre des méthodes d'apprentissage statistique dans un cadre essentiellement prédictif

Le projet porte sur l'analyse de deux fichiers de données concernant une campagne marketing conduite auprès d'un ensemble de clients. Chaque ligne d'un fichier décrit un client. Le fichier projet-app-13-learn.csv contient en outre les résultats de la campagne pour les clients concernées. Les variables sont les suivantes :

- **age** : âge ;
- **sex** : genre ;
- **f_name** : prénom ;
- **last name** : nom;
- **commune** : nom de la commune de résidence ;
- **insee code** : code insee de la commune de résidence ;
- **city type** : type de la commune de résidence ;
- **department** : numéro du département de résidence ;
- **reg** : code de la région de résidence ;
- **catégorie** : code de catégorie socio-professionnelle, selon la table 1 ;
- **revenue** : salaire mensuel en équivalent temps plein (attention, cette information n'est pas disponible pour tous les clients) ;
- **cible** : résultat de la campagne codé par success pour un résultat considéré comme positif et failure dans le cas contraire.

Il est important de noter que Lyon, Paris et Marseille sont découpées en arrondissements et que le nom de commune est alors celui de la ville suivi de l'indication d'arrondissement.

Certaines variables seront naturellement lues comme des variables numériques (par exemple reg et catégorie) alors qu'elles ne contiennent pas des nombres mais des codes. Il est vivement conseillé de convertir les variables concernées au format factor en R pour faciliter la suite des traitements.

Il est aussi vivement conseillé d'enrichir les données fournies par des données annexes, notamment liées à la géographie de la France.

Contenu du Rapport (Index à construire)

1. Analyse exploratoire minimale des données (statistiques univariées, dépendances, etc.) ;
2. Justification du modèles prédictif choisi ;
3. Description précise de la chaîne de traitement : prétraitements éventuels, ajustement des modèles, choix du modèle, évaluation de ses performances attendues (le rapport doit impérativement contenir un tableau indiquant la qualité numérique attendue pour les prévisions sur le fichier test) ;

4. Analyse de l'importance des variables : cela peut être fait avant l'ajustement des modèles, pendant celui-ci ou après le choix du modèle final. Dans tous les cas, le rapport doit discuter de l'opportunité de construire des modèles sur une partie seulement des variables. Si c'est le cas, les prévisions finales et les performances attendues doivent concerner les modèles n'utilisant que les variables pertinentes ;
5. interprétation du modèle retenu : si cela est possible, une interprétation de la façon dont les décisions du modèle retenu sont prises fournira un complément très important au reste de l'analyse.

1.Preparation des données

1.1 - Charger les données. Voici la structure du dataset d'entraînement. Nous avons 10 000 observations et 12 variables

```
## Observations: 10,000
## Variables: 12
## $ age        <int> 42, 28, 28, 55, 39, 76, 50, 32, 29, 59, 85, 58, 67, 54, ...
## $ sex        <fct> Female, Female, Male, Male, Male, Female, Male, Male, Fe...
## $ f_name     <fct> WANDA, MAURICETTE, MARTIAL, PHILIPPE, CLÉMENT, EVELYNE, ...
## $ last.name  <fct> FAHD, LE BIHAN, DE ALMEIDA, QUILLERE, STANGL, CAVALERI, ...
## $ commune    <fct> Marseille 5e  Arrondissement, Champeaux, Lancié, Beaune,...
## $ insee.code <fct> 13205, 50117, 69108, 21054, 34151, 26106, 67482, 67034, ...
## $ city.type  <fct> Préfecture de région, Commune simple, Commune simple, So...
## $ department <fct> 13, 50, 69, 21, 34, 26, 67, 67, 75, 56, 91, 59, 94, 51, ...
## $ reg        <int> 93, 28, 84, 27, 76, 84, 44, 44, 11, 53, 11, 32, 11, 44, ...
## $ catégorie  <int> 5, 5, 10, 11, 7, 13, 3, 5, 12, 5, 13, 11, 13, 6, 10, 3, ...
## $ revenue    <dbl> 1260, 1320, NA, NA, 2460, NA, 4500, 1230, NA, 1440, NA, ...
## $ cible      <fct> failure, success, failure, success, failure, failure, su...
```

1.2- Changement du format: nous allons renommer la variable catégorie, pour enlever l'accent et convertir celle-ci et la région au format nominal. Il vont nous rester uniquement deux variables en format numérique, l'âge et le revenue

1.3- Completer avec les données géographiques

Ajoutons les coordonnées géographiques et la population. Nous n'avons pas besoin d'ajouter le nom du département ni de la région pour l'instant.

1.4- Traitement des données manquantes

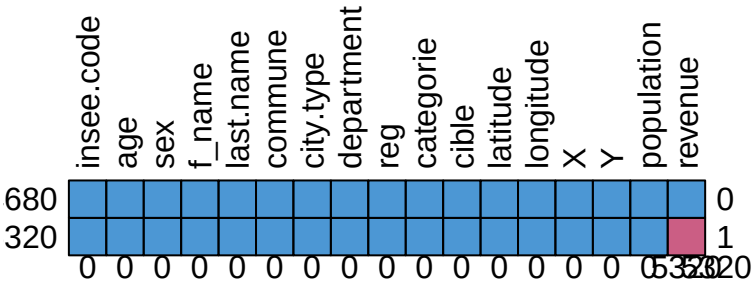
La variable revenue a 5320 valeurs manquantes pour le dataset de learn (53%) et 5175 pour le dataset de test (51%)

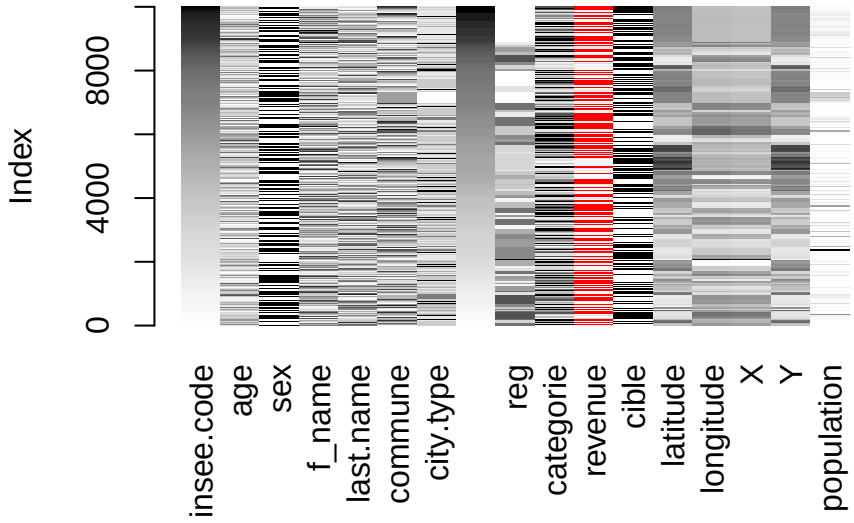
```
## insee.code        age        sex     f_name  last.name    commune  city.type
##          0          0          0          0          0          0          0
## department        reg  categorie    revenue      cible   latitude  longitude
##          0          0          0       5320          0          0          0
##          X          Y population
##          0          0          0

## insee.code        age        sex     f_name  last.name    commune  city.type
```

```
##          0          0          0          0          0          0          0
## department        reg  categorie    revenue   latitude  longitude          X
##          0          0          0       5175          0          0          0
##          Y population
##          0          0
```

A garder ces graphiques? En annexe? Voici deux visualisations des données manquantes, qui nous confirmemt qu'uniquement la variable revenus a des valeurs manquantes et ils sont repartis tout au long du dataset





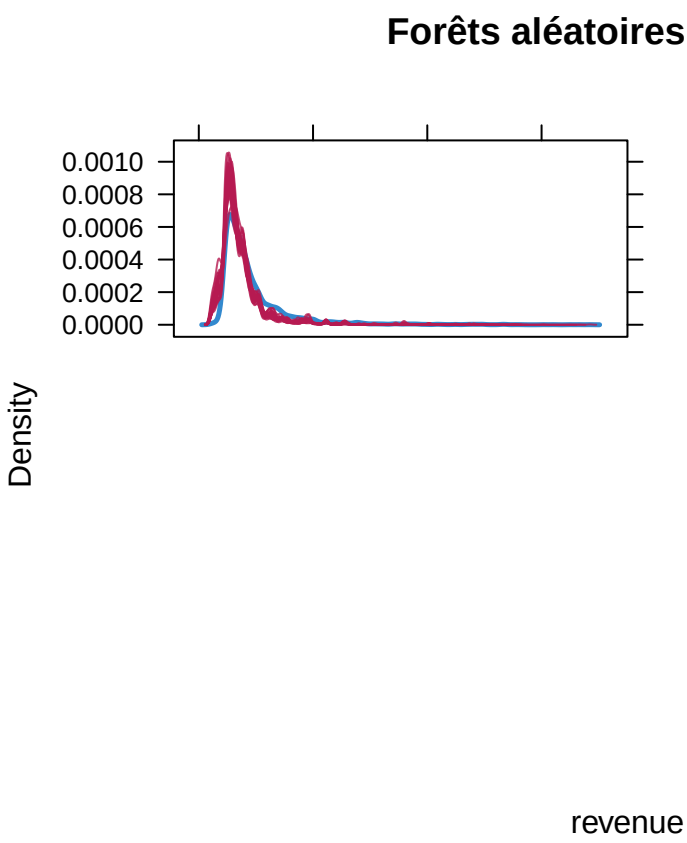
Pour que l'imputation fonctionne, il faut enlever des colonnes avec trop de categories, et celles qui ne devraient pas apporter plus d'information comme le prénom et le nom.

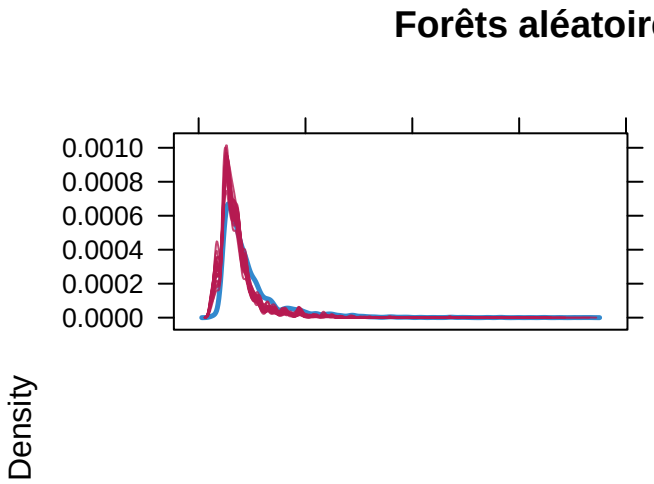
L'imputation avec mice, sera faite avec rf : "Random forest imputation", en utilisant les données d'age, sex, region, categorie et population pour trouver le revenue correspondant.

```
## Warning: Number of logged events: 150

## Warning: Number of logged events: 150
```

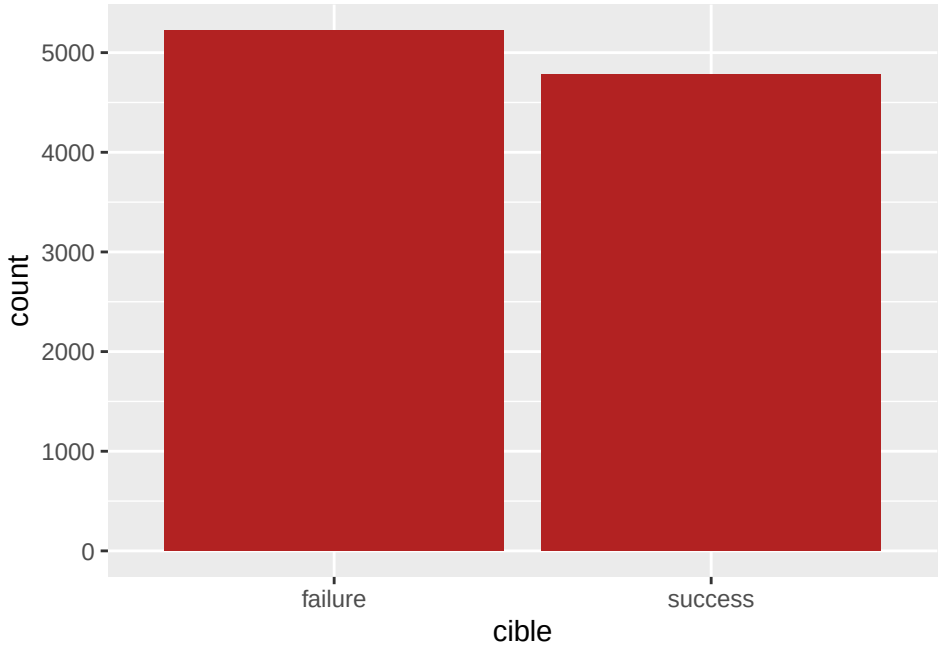
Nous verifions que l'imputation respecte bien la structure des données orginal, et c'est bien le cas :





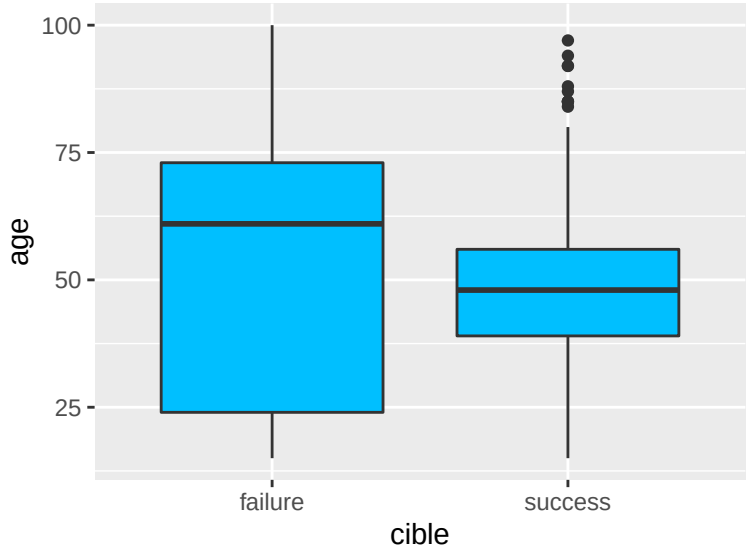
2 - Analyse exploratoire des données

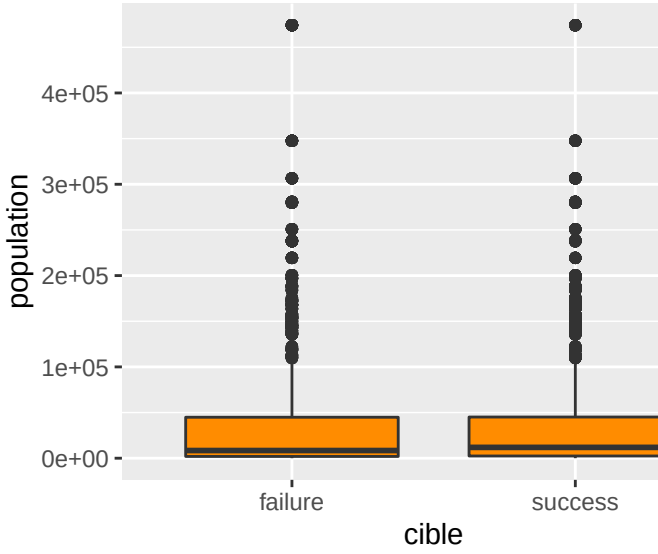
Les données semblent bien reparties entre les deux catégories à predire : Failure et Success

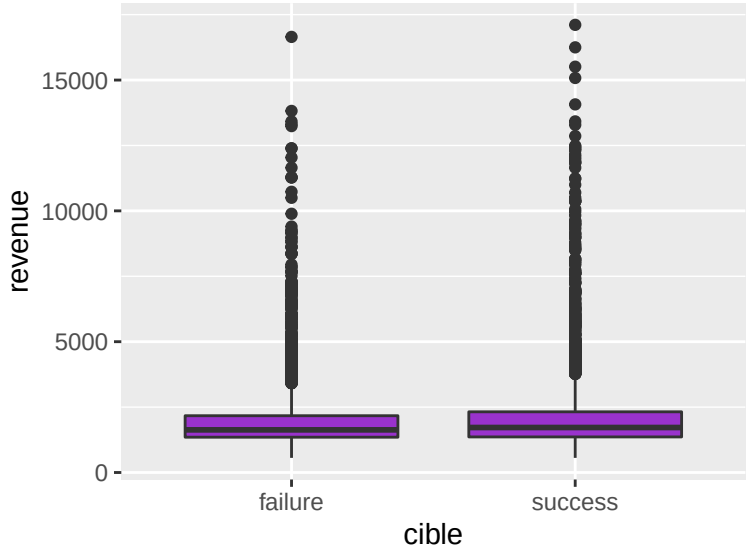


Nous observons une distribution très différente selon l'âge : l'échec est bien répartie entre 25 et 75 ans, par contre le succès est concentré entre 30 et 55 ans, avec quelques valeurs extrêmes superieurs à 75 ans. Une variable qui sera pertinente pour la prediction.

Pour la population et les revenues, il ne semble pas avoir des différences.

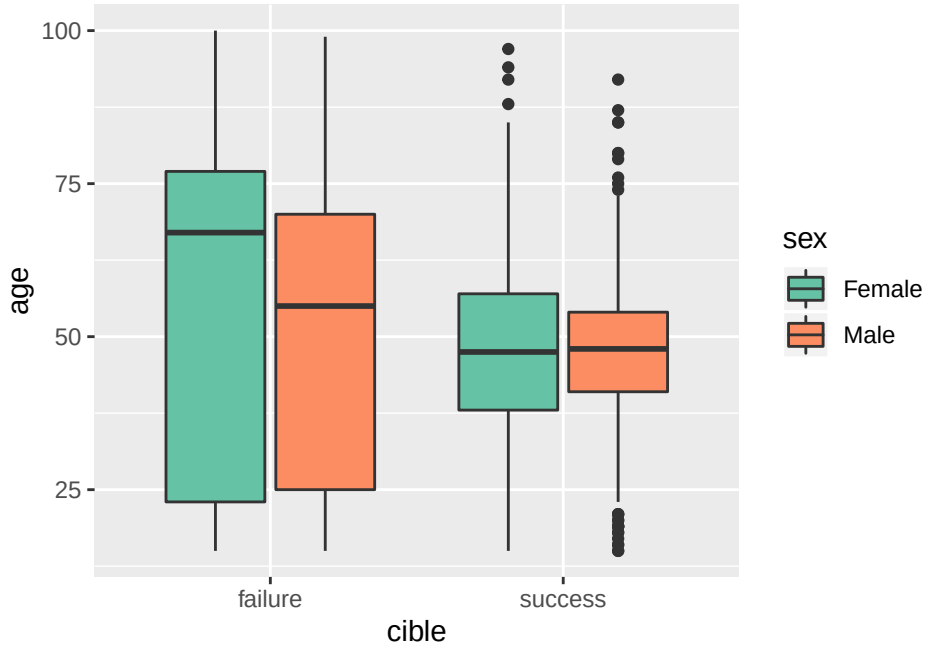






Les hommes ont un taux d'échec (60%) plus élevé que les femmes (45%). Pour les âges, la medianne de l'échec des femmes est autour de 70 ans et pour les hommes 50 ans.

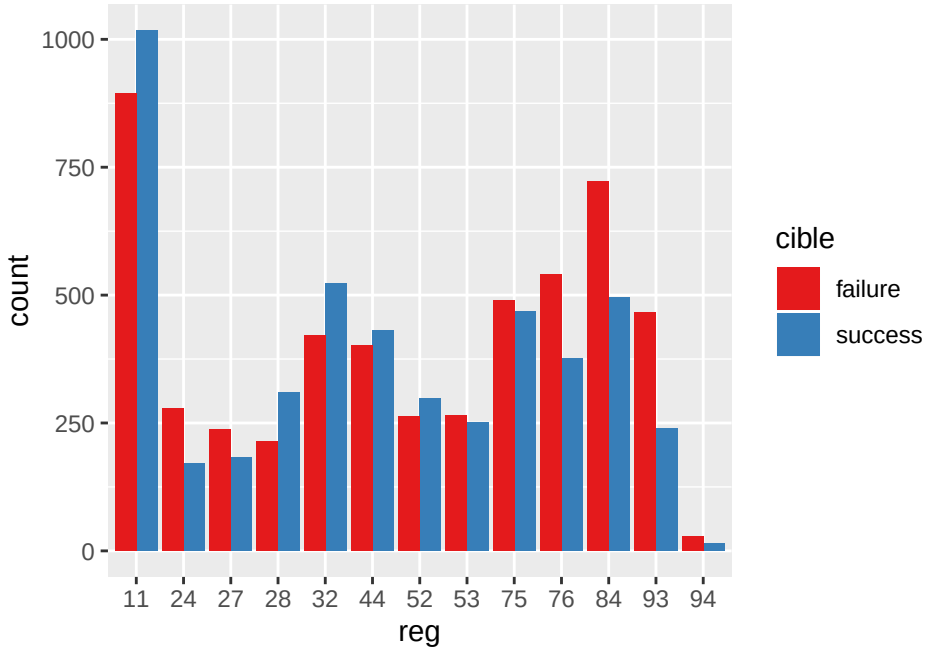
L'échantillon est composé d'un plus grand nombre de femmes que d'hommes (5259 vs 4741)




```
##
##          failure success
##   Female    2351    2908
##   Male      2870    1871
```

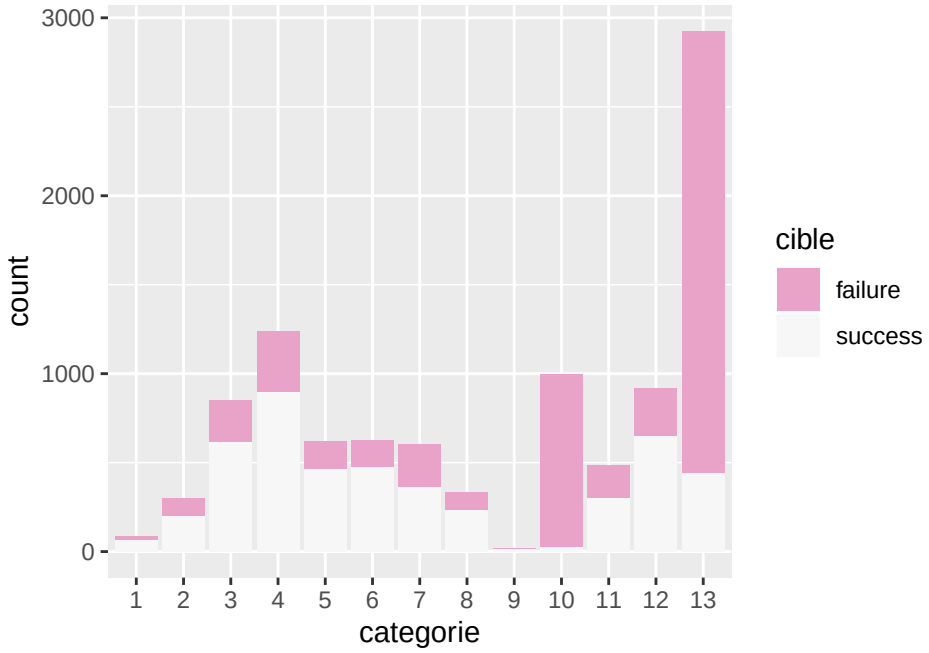
La region 11 (Ile de France) cumule une bonne partie des resultats, mais en global ils semblent bien equilibrés, sauf pour la region 76 (OCCITANIE), 84 (AUVERGNE RHONE ALPES) et 93 (PROVENCE ALPES COTE D'AZUR). Le succes se concentrent au nord : ILE DE FRANCE(11), HAUTS DE FRANCE (32) ET NORMANDIE (28).

La région 94 (Corse) a une participation très basse



```
##
##      failure success
##   11     894    1017
##   24     278     171
##   27     237     183
##   28     214     309
##   32     421     522
##   44     402     432
##   52     262     299
##   53     264     251
##   75     489     469
##   76     541     377
##   84     723     495
##   93     467     240
##   94      29      14
```

Les retraités et les étudiants (catégorie 13 et 10), enregistrent la plupart des échecs, tandis que les catégories 2 a 8, 11 et 12 (chomeurs et inactifs) ont la plupart des succès, surement il s'agit d'une des variables à garder pour la prédiction.

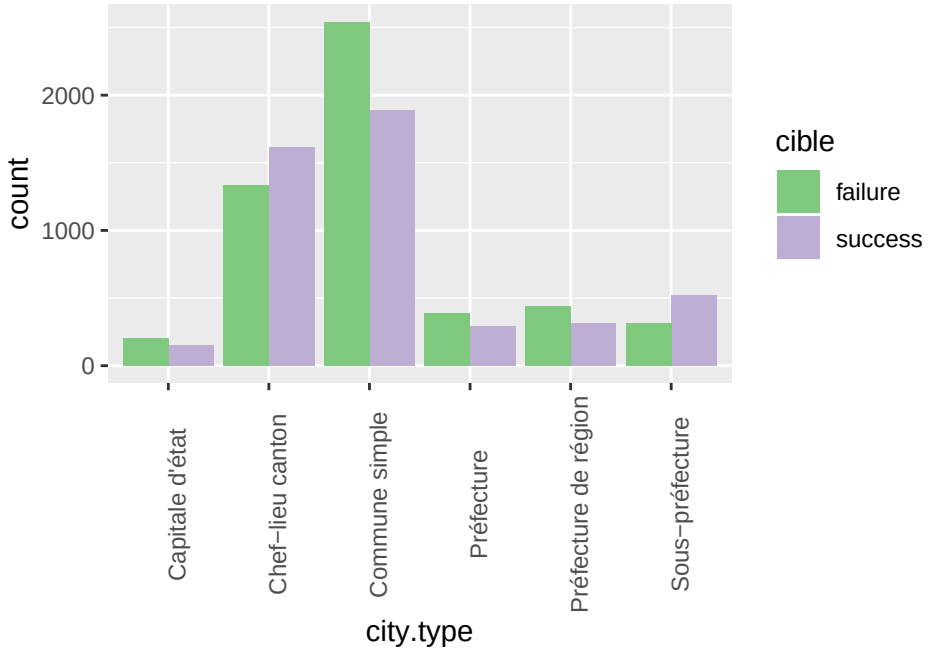



```
##
##      failure success
##   1       19      68
##   2       97     205
##   3      233     617
##   4      338     901
##   5      154     466
##   6      149     478
##   7      237     366
##   8       96     239
##   9        4      13
##   10     968      26
```

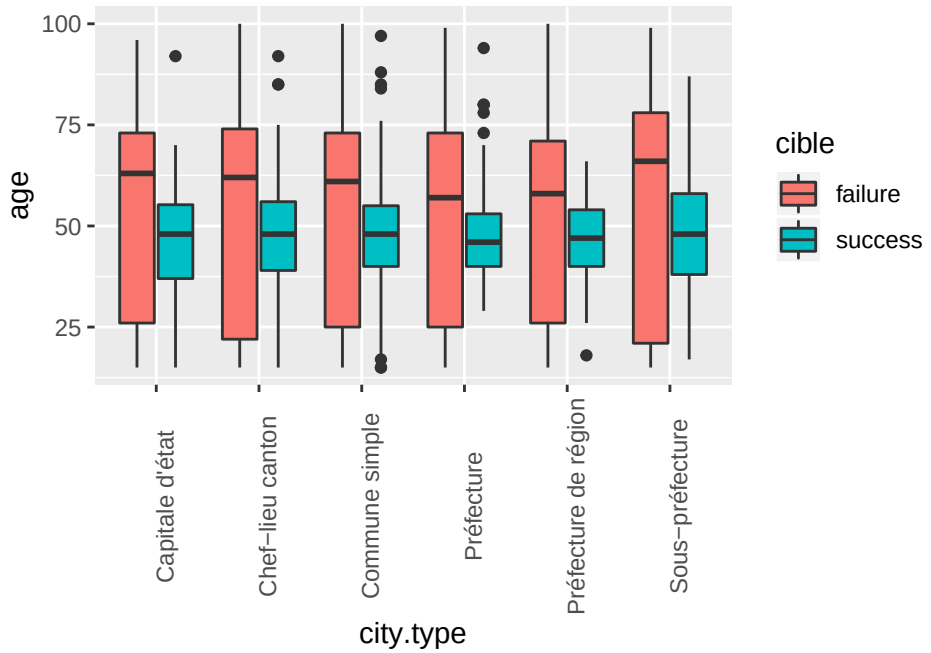
```
##    11      180      305
##    12      266      651
##    13     2480      444
```

Paris, en tant que capitale, enregistre très peu de résultats, en sachant que la région Ile de France c'est la plus représentée dans l'échantillon.

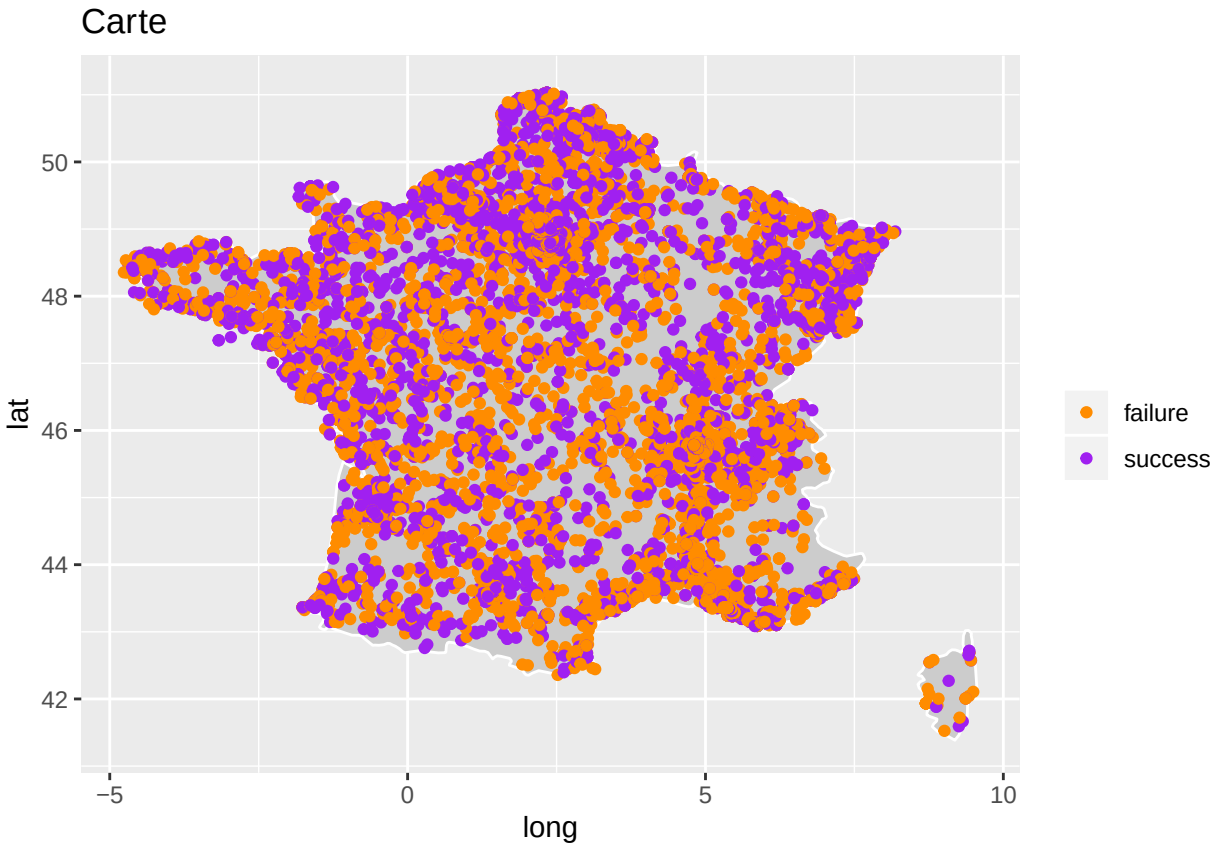
Les ages sont bien reparties entre toutes les types de communes.




```
##
##                        failure success
##   Capitale d'état          206     148
##   Chef-lieu canton        1331    1615
##   Commune simple          2541    1890
##   Préfecture               389     293
##   Préfecture de région     437     312
##   Sous-préfecture          317     521
```



Voici la repartition des reponses sur toute la france, on observe plus de succèss au nord et plus des echecs au sud (region Auvergne-Rhône-Alpes):



2 Modèles prédictives (à suivre)