Analyse préalable au nettoyage

Voici une liste des questions mentionnées dans le sujet :

- `Ces données nécessitent-elles un nettoyage ?`
- `Faut-il écarter certaines instances non liées au problème ?`
- `Y a-t-il des valeurs manquantes ou aberrantes ?`
- `Y a-t-il des attributs redondants ou superflus ?`
- `Les valeurs numériques correspondent-elles réellement à des attributs de nature numérique, ordinale ou catégorielle ?`
- `Comment traiter ces différents problèmes ?`

Pour répondre à ces questions en analysant les données de démissionnaires, voici les étapes à suivre et quelques premières pistes de réflexion :

1. **Ces données nécessitent-elles un nettoyage ?**
    - Il est fort probable que les données contiennent des valeurs aberrantes ou manquantes, et qu'un nettoyage soit nécessaire. On peut vérifier cela en explorant les données, notamment en examinant les valeurs minimales, maximales, moyennes et les éventuelles valeurs nulles ou incohérentes.
2. **Faut-il écarter certaines instances non liées au problème ?**

- Des instances qui ne sont pas liées aux démissions ou qui présentent des données incohérentes ou incomplètes peuvent biaiser l'analyse. Il est donc nécessaire d'identifier ces instances et, selon le cas, les exclure.

3. **Y a-t-il des valeurs manquantes ou aberrantes ?**
    - Pour identifier les valeurs manquantes, nous allons compter les valeurs nulles ou aberrantes dans chaque colonne. Les valeurs aberrantes peuvent inclure des dates incohérentes, des codes incorrects, ou des montants de revenus extrêmes.
4. **Y a-t-il des attributs redondants ou superflus ?**
    - Il est nécessaire de vérifier si certaines colonnes contiennent des informations redondantes ou non pertinentes pour la prédiction des démissions. Par exemple, si deux attributs fournissent la même information mais sous des formes différentes.
5. **Les valeurs numériques correspondent-elles réellement à des attributs de nature numérique, ordinale ou catégorielle ?**
    - Certains attributs peuvent sembler numériques (comme les revenus) mais être en réalité catégoriels ou ordonnés (par exemple, des tranches de revenus). Il est important de classifier correctement chaque attribut pour appliquer des méthodes statistiques adaptées.
6. **Comment traiter ces différents problèmes ?**
    - **Nettoyage des données manquantes**
    - **Élimination des valeurs aberrantes**

Voici une analyse détaillée des attributs présents dans notre jeu de données :

```
file_path = 'table1.csv'
data = pd.read_csv(file_path)

summary_stats = data.describe(include='all')

# valeurs manquantes
#missing_values_detail = data[data.isnull().any(axis=1)]
```

```
# valeurs manquantes pour un attribut (ex : CDSEXE)
missing_values_detail = data['CDSEXE'].isnull().sum()

print(summary_stats, missing_values_detail)
```

Code Sexe

Cet attribut numérique de nature catégorielle ne semble pas nécessiter d'être écarté. Il n'y a pas de valeurs manquantes ou aberrantes, les valeurs sont comprises dans un intervalle précis : entre 2 et 4 (que des sous-classes).

C'est un attribut catégoriel qui pourrait être utile pour l'analyse.

mean	2.681656
std	0.670356
min	2.000000
25%	2.000000
50%	3.000000
75%	3.000000
max	4.000000

```
Données manquantes :  0
```

Montant revenus

Cet attribut numérique présente quelques particularités à prendre en compte :

- Il contient beaucoup de zéros :
  - 27050 / 30332 = 89.18% de zéros.

```
Nombre de zéros :  27050
```

  - On interprète ça comme des valeurs manquantes ?
    - Oui
- Il présente moyenne de 289 avec une grande variance et des revenus allant de 0 à 1 4 490.
  - Traitement comme valeurs extrêmes ?
    - Oui

```
mean    2.890012e+02
std     8.959197e+03
min     0.000000e+00
25%     0.000000e+00
50%     0.000000e+00
75%     0.000000e+00
max     1.524490e+06
```

Nombre d'enfants

Il s'agit d'un attribut numérique :

- Il contient aussi beaucoup de zéros :
  - 27017 / 30332 = 89% de zéros.
  - En ce qui concerne le salaire, on peut se dire qu'un salaire de 0$ n'a pas de sens, mais dans ce cas-là, est-ce qu'on considère que ces personnes n'ont pas d'enfant ou que ce sont des valeurs manquantes ?
    - Pas d'enfant

```
Nombre de zéros :  27017
```

  - Le nombre varie de 0 à 6 → Pas de valeurs extrêmes.

```
mean    0.201998
std     0.636844
min     0.000000
25%     0.000000
50%     0.000000
75%     0.000000
max     6.000000
```

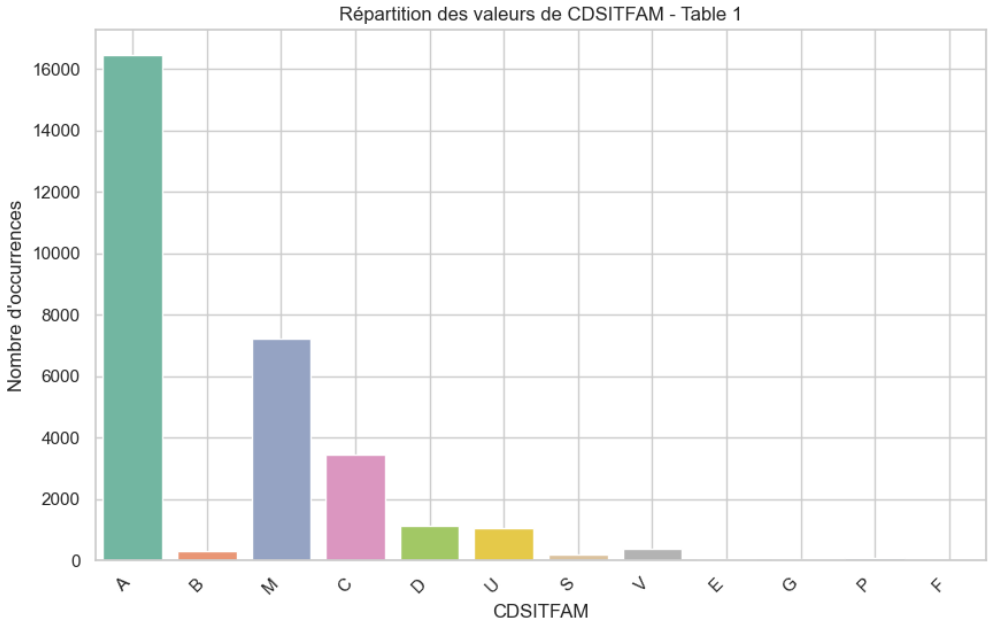
Situation familiale

Il s'agit d'un attribut catégoriel :

- La catégorie A est la plus représentée avec 16463 démissionnaires dans cette situation.

```
Nombre de A :  16463
```

- Les catégories B,S sont peu représentées et V,E,G,P,F sont très peu représentées et peuvent-être considérées comme des valeurs extrêmes.
    - Type de valeurs à écarter pour l'analyse ?
        - Probablement
- Pas de valeurs manquantes/abberantes.



Date d'adhésion

- Il faudrait vérifier le format des dates et s'assurer qu'il n'y a pas de valeurs aberrantes (dates dans le futur ou trop anciennes).

Voici le code qui permet de le vérifier.

```
import pandas as pd

file_path = 'table1_modified.csv'
data = pd.read_csv(file_path)

summary_stats = data.describe(include='all')

data['DTADH'] = pd.to_datetime(data['DTADH'], format='%d/%m/%
```

```
start_date = pd.Timestamp('1950-01-01')
end_date = pd.Timestamp('2007-01-01')

aberrant_dates = data[(data['DTADH'] < start_date) | (data['D

print(aberrant_dates[['ID', 'DTADH']])
```

- Après avoir exécuté ce code, nous pouvons voir qu'il n'y a pas de valeurs manquantes ou aberrantes concernant cet attribut.

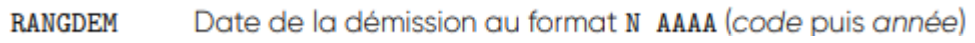
Code statut et Code de démission

Ces attributs sont numériques de nature catégorielle et contiennent une très grande majorité de 0 ou de 2, ce qui fait que la variance est extrêmement faible.

On peut envisager des méthodes pour le traitement des valeurs dominantes.

Date de démission, Année de démission, et Date démission (N AAAA)

- Ces attributs sont redondants. Il faudrait choisir le format le plus approprié pour l'analyse et potentiellement écarter les autres pour éviter la redondance, surtout en ce qui concerne les deux dates de démission.

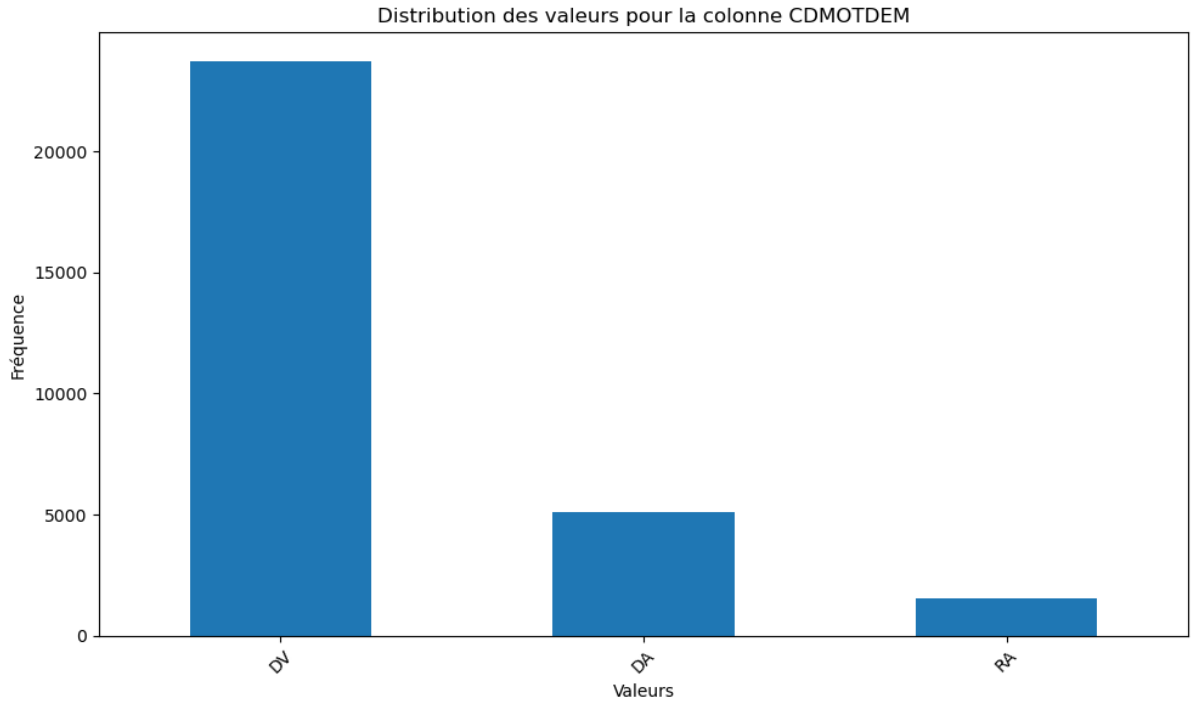


  - A quoi fait référence ce code ? Allant de 1 à 8 dans l'ordre de 1999 à 2006 ?
    - Au numéro de l'intervalle

Motif de démission

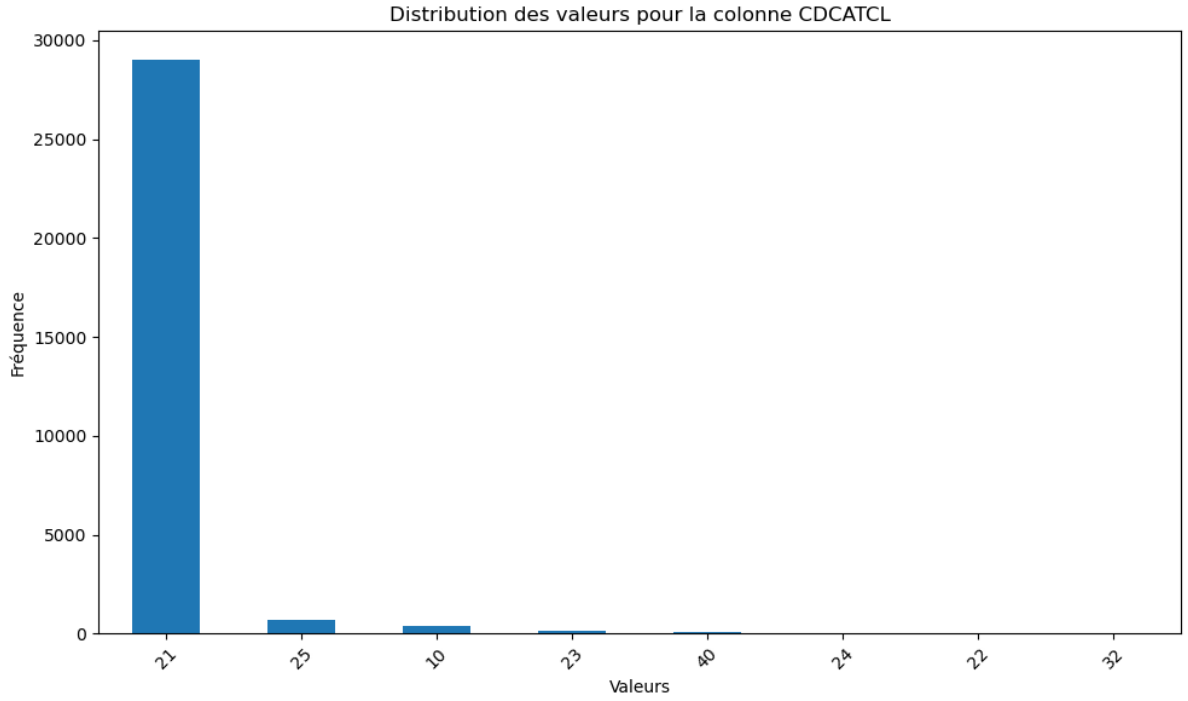
Cet attribut catégoriel est probablement crucial pour l'analyse.

- DV est la valeur la plus représentée.
  - Probablement "démission volontaire", "démission administrative/anticipée" et "retraite anticipée



Cet attribut numérique de nature catégorielle contient beaucoup de valeurs extrêmes.

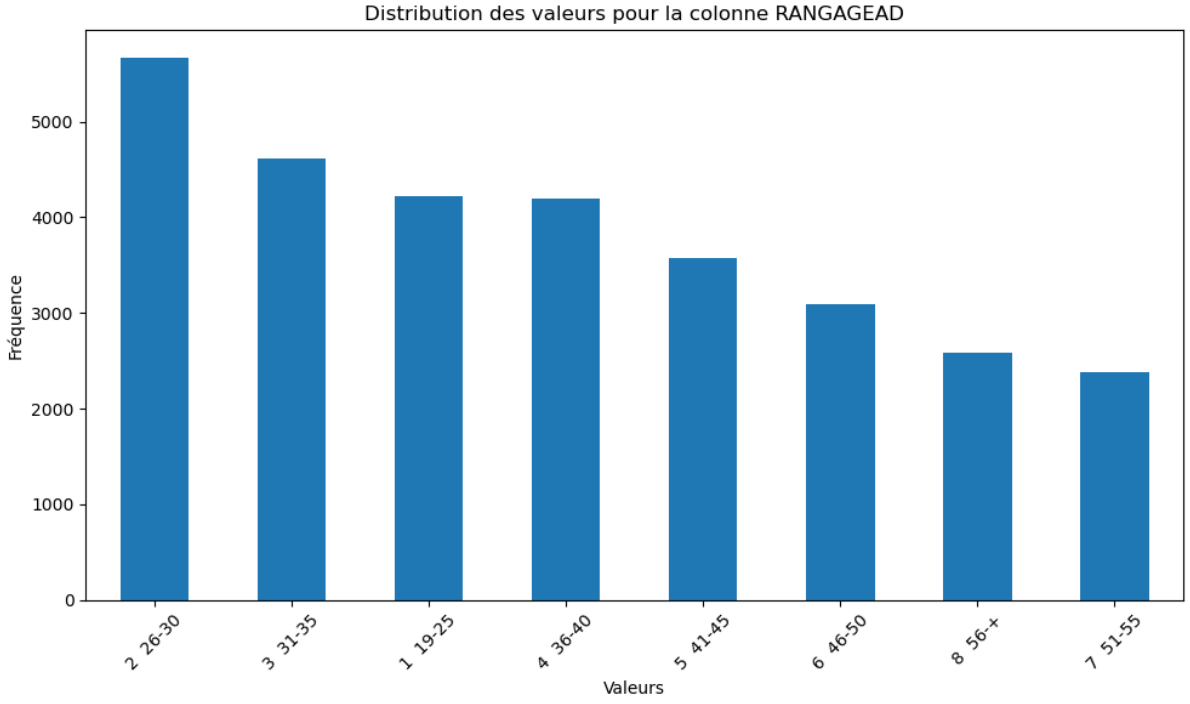
- Plusieurs solutions :
  - On écarte l'attribut.
  - On le garde en rééquilibrant les poids de manière à faire ressortir les valeurs extrêmes sous-représentées (23, 40, 24 etc...).
  - On supprime les valeurs sous-représentées.



Age à l'adhésion, Tranche d'âge (adhésion), Age démission, Tranche d'âge (démission)

Ces attributs semblent partiellement redondants.

- Il faudrait extraire la plage de valeurs sans le code qui est encore listé dans l’ordre pour le nettoyage.



Durée adhésion

- Cet attribut numérique a déjà été calculé, il faut s'assurer que les calculs correspondent bien à la durée d'adhésion.
  - Voici le code qui permet de le vérifier.

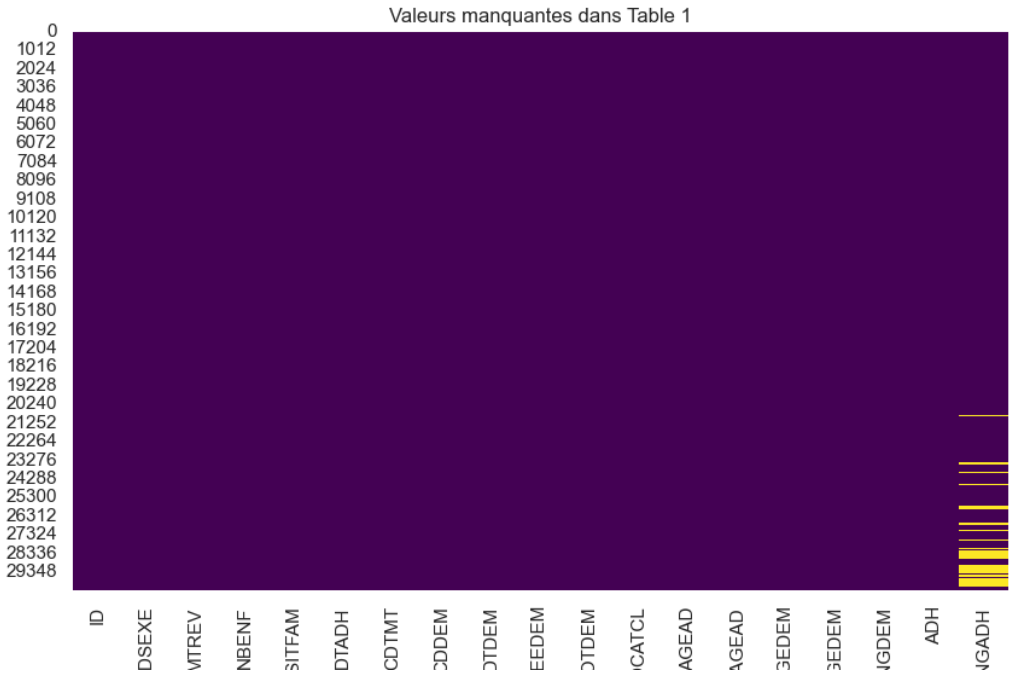
```
import pandas as pd

file_path = 'table1_modified.csv'
data = pd.read_csv(file_path)

data['DTADH'] = pd.to_datetime(data['DTADH'], format='%d/%
data['DTDEM'] = pd.to_datetime(data['DTDEM'], format='%d/%

data['Duree_adhesion_jours'] = (data['DTDEM'] - data['DTAD

data['Duree_adhesion_annees'] = data['Duree_adhesion_jours

# Au moins un an d'écart
tolerance = 1
erreurs = data[abs(data['ADH'] - data['Duree_adhesion_anne

# Afficher les lignes avec des erreurs
print(erreurs[['ID', 'DTADH', 'DTDEM', 'ADH', 'Duree_adhes
```

- Après vérifications, il n'y a pas d'erreurs sur les calculs ni de valeurs manquantes ou aberrantes.

Tranche durée d'adhésion

Cet attribut catégoriel contient un bon nombre de valeurs manquantes (2584 valeurs manquantes) dans le fichier table1.

- Même problème qu'avec la tranche d'âge d'adhésion, il nous faut la tranche sans le code.



Annexe

1. Valeurs problématiques et solutions de remplacement

   a. Valeur problématiques

   b. Solutions de remplacement

- Traitements simples mais brutaux
  - Suppressions des lignes (individus)
  - Suppressions des colonnes (variables)
- Remplacement des valeurs manquantes
  - Moyenne ou médiane pour les variables numériques
  - Valeur la plus fréquente
  - Valeur par défaut (0 par exemple pour les variables numériques)
  - Modalité spécifique (NA pour les variables catégorielles)

- Imputation
  - Estimation des valeurs manquantes par des statistiques
  - Exemples
    - Modèles de régressions linéaires pour prédire une valeur de la variable manquante à partir des autres variables
    - Modèles des k plus proches voisins pour estimer la valeur manquante en fonction des points les plus proches
- Valeurs manquantes acceptées par certains classificateurs : arbre de décision, classification bayésienne naïve