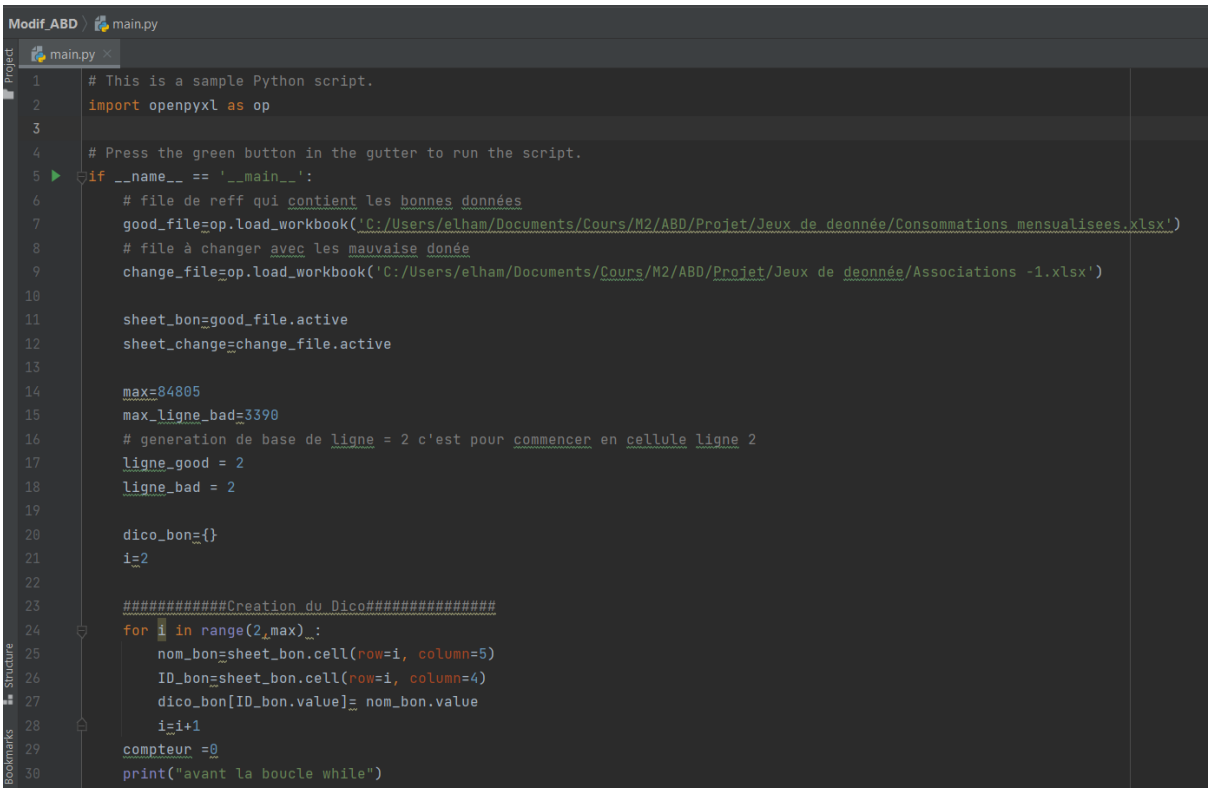
PS : Tous les scripts sont dans le zip insert_sql avec une note.txt à lire. Il y'a aussi le fichier pdf des Classes et attribut demander en exercice 1.

1. Homogénéisation des données
    a. Les macros créer

Des scripts d'homogénéisation ont été créés sous python. Celui-ci marche néanmoins il n'est pas possible de l'utiliser car les données sont trop importantes et il faudrait l'améliorer. Notamment en modifiant le script pour stocker toutes les données des excès en variable pour augmenter la rapidité. De plus, nous avons aussi créé un script (macro VBA) sous excel. Il reprend un algorithme semblable à celui précédent, le problème est que excel n'est pas optimiser pour enchainer des boucles aussi grande et 90% du temps la macro plante.
Script python :



```
while ligne_good < max_:
    compteur=compteur+1
    if compteur == 10000:
        print(compteur)
        compteur=0
    ##Ici on met les colonnes que l'on veut homogénéiser
    nom_change=sheet_change.cell(row=ligne_bad, column=2)
    ID_change=sheet_change.cell(row=ligne_bad, column=1)

    if ligne_bad != max_ligne_bad:
        if dico_bon.get(ID_change.value) is not None:
            # En gros quand nos ID dans chacun des fichiers concorde on rentre dans cet boucle
            nom_change.value=dico_bon.get(ID_change.value)
            change_file.save(
                'C:/Users/elham/Documents/Cours/M2/ABD/Projet/Jeux de deonnée/Associations -1_corrige.xlsx')

            ligne_bad=ligne_bad+1
        elif ID_bon.value != ID_change.value_:
            ligne_bad=ligne_bad+1

    ##On risque d'avoir un pb si l'ID du second fichier est déja passer?
    elif ligne_bad == max_ligne_bad:
        ligne_good = ligne_good + 1
        ligne_bad = 2
```

Résumé :

On utilise la librairie openpyxl qui permet de manipuler des fichiers excel. On charge un fichier de référence avec les bonnes valeurs et le second fichier contenant les colonnes à changer. On inscrit le nombre maximal de ligne pour chaque fichier et on initialise chaque ligne à 2 (pour ne pas prendre la 1 contenant les noms de colonnes). On va créer un dico qui va contenir en clé les bons ID et en valeur le nom qui va être à changer. S'ensuit une boucle while qui va regarder chaque ligne du mauvais fichier. Il va regarder s' il y a une concordance entre l'ID du mauvais fichier et la présence du bon ID dans le dicos. Si oui alors le dico va remplacer le nom du mauvais fichier (nom change). Ainsi de suite jusqu'à ce que le nombre maximal de bonnes lignes soit atteint.

Script VBA :

```
Sub Macro1()
'
' Macro1 Macro
'

    Dim i As Long
    Dim a As Long

    For i = CLng(2) To 84802
        a = CLng(2)

        For a = CLng(2) To 3391
            If Cells(i, 4) = Cells(a, 14) Then
                If Cells(i, 12) <> Cells(a, 16) Then


                    Cells.Replace What:=Cells(i, 12), Replacement:=Cells(a, 16), LookAt:= _
                    xlPart, SearchOrder:=xlByColumns, MatchCase:=False, SearchFormat:=False, _
                    ReplaceFormat:=False, FormulaVersion:=xlReplaceFormula2

                End If
            End If
        Next

    Next


End Sub
```

Résumé :

Cette macro ressemble au script précédent. Sauf que cette fois-ci lorsque on rentre dans la boucle lorsque les deux ID correspondent on va utiliser la fonction de remplacer d'excel. Les deux fichiers doivent être au préalable sur la même feuille excel pour faciliter la manipulation.

b. Homogénéisation manuelle

Nous avons donc décidé de faire l'homogénéisation à la main et les colonnes tel que adresse, code postale et ville seront dispatchés sur l'ensemble des tableaux. Cela permet d'homogénéiser nos données.

Pour les Consommation dans les fichiers consommation annuelle je ne sais pas si les consos négatives sont normales. Dans le cas présent il a été pris comme oui correspondant à une conso inférieur à d'estimations décrites dans un contrat. Comme dans la réalité on peut mensuellement nos consommation électrique et à la fin de l'année si la conso est inférieur au payement elle apparaît négativement et nous sommes rembourser.

2. Les erreurs et comptage des insertions

Nous travaillons sur les serveurs de la faculté. Lors de ce projet qui contient un grand nombre de données il s'est avéré que nous avons un espace maximum pour créer nos tables de données. Erreur ci-dessous.

```
ERREUR a la ligne 1 :
ORA-01536: depassement du quota d'espace affecte au tablespace 'USERS'
```

Il a alors été tenté d'effacer tous nos anciens tableaux SQL de notre sessions via : select table_name from all_tables ; puis un drop table pour chaque tableau trouvé.

Mais rien n'y fait, il n'est pas possible de créer la totalité du fichier de consommation annuelle. Nous obtenons en ligne maximal pour la consommation annuelle

```
SQL> SELECT COUNT(*) FROM table_conso_anuelle;

  COUNT(*)
----------
      3682
```

Ni de créer le fichier de facture dans sa totalité. Le nombre maximal de ligne est de :

```
SQL> SELECT COUNT(*) FROM table_factures;

  COUNT(*)
----------
     10744
```

Pour le tableau des informations sur les compteurs nous obtenons :

```
SQL> SELECT COUNT(*) FROM table_info_compteurs;

  COUNT(*)
----------
      3389
```

Pour le tableau de consommation mensuelle nous ne pouvons pas non plus insérer tout le tableau qui est trop volumineux. Nous avons réussi à en insérer :

```
SQL> SELECT COUNT(*) FROM table_conso_mens;

  COUNT(*)
----------
     23800
```

Pour la table association nous avons :

```
SQL> SELECT COUNT(*) FROM table_Associations;

  COUNT(*)
----------
      3389
```

Et pour la table d'informations des sites :

```
SQL> SELECT COUNT(*) FROM table_info_sites;

  COUNT(*)
----------
       350
```

Pour créer les tables d'énergies à savoir Gaz et électricité nous avons utilisé la table facture. Mais sachant que celle-ci ne peut se créer au maximum nous avons limité sont nombre maximal à 1959 ligne. Ce script d'insertions se nomme insert table facture limite

```
SQL> SELECT COUNT(*) FROM table_factures;

  COUNT(*)
----------
      1959
```

On ne travaillera qu'avec ce dernier par la suite.

3. Création de tables d'énergies.

Pour créer le gaz et l'électricité, on va se servir de la table de facture limite. On définit les colonnes d'intérêt. Pour le gaz on prendra les données qui contiennent simplement le fluide Gaz et la même chose pour l'électricité avec fluide électricité.

Ainsi nous obtenons pour le Gaz :

```
SQL> SELECT COUNT(*) FROM Gaz_table;

  COUNT(*)
----------
       772
```

et pour l'électricité :

```
SQL> SELECT COUNT(*) FROM Electricite_table;

  COUNT(*)
----------
      1187
```

4. Les requêtes

Je n'ai pas bien compris la question ou s' il fallait la faire en 2 requêtes ou tout ensemble. C'est pour ça qu'en commentaire il ya 2 requêtes ou en non commenter la totalité. J'ai préféré laisser les 2 requêtes au cas ou.

5. Commentaires sur l'équipe

Parti Elise : Fatiha m'a beaucoup soutenu en s'occupant notamment de la partie 2, 3, 4.a et 4.b. Sans elle, je n'aurais jamais pu finir ce projet. Nous avions décidé de se répartir comme suit le projet : Fatiha la partie schéma et moi informatique. Car j'ai du mal à faire les schémas et elle l'informatique. C'est dans la diversité que l'on se complète comme on dit. Néanmoins des problèmes d'organisation sont observables.

Parti Fatiha : Merci Elise pour ton aide et ton soutien. Elle est vraiment un atout pour notre équipe et pour toute l'organisation. Son éthique de travail et son implication sont admirables,

j’ai de la chance d’avoir travaillé avec elle, sans elle je n’aurais jamais réussie à réaliser ce projet. Comme elle a dit, des problèmes d’organisation sont palpables.

6. Commentaire sur le sujet

Parti Elise : Ce sujet est vraiment très complet il nous refait travailler des notions vu depuis le début de ce master. Je suis très déçu de ne pas avoir pu tester mes requêtes sur la globalités des données (dû à la limite imposée par les serveurs). De plus, j'aurais aimé retravailler mes scripts d’homogénéisation car ça peut être une vraie plus valu au sein d’une entreprise. Enfin je n’ai pas réussi à créer des fonctionnements de clé primaire et secondaire qui complexifié le tout, c’est pour cela que j’ai utilisé beaucoup de tableaux croisés dans mes requêtes. Moralités: il faudrait que je m'achète un serveur. Merci pour ce projet grandissant et en totale immersion avec la réalité.

Parti Fatiha : En principe le sujet de projet est passionnant, et surtout la partie des diagrammes. Grâce à ce projet j’ai réussi à utiliser le logiciel Powerdesigner. Après ce projet, je devrais plus travailler la partie programmation avec les requêtes. Merci professeurs pour ce projet et ces cours qui étaient intéressants .