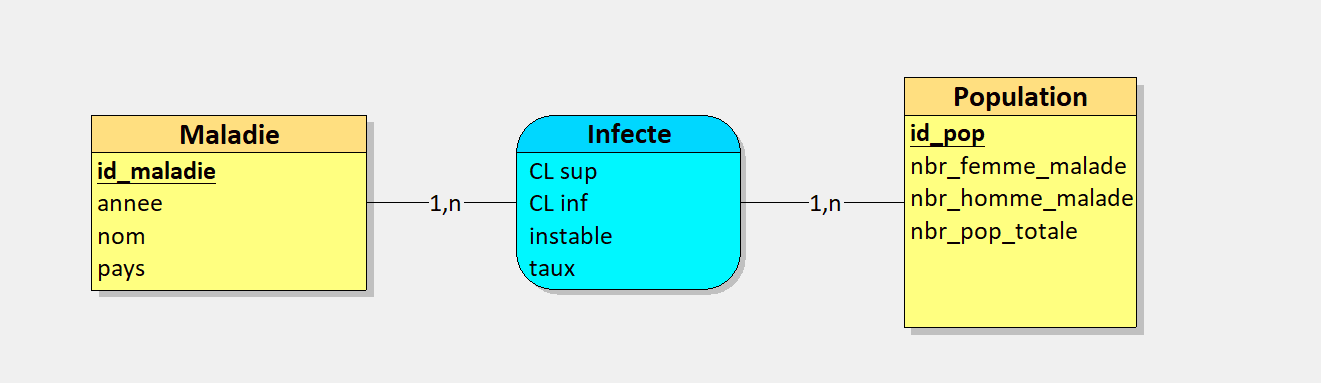
BD – Projet n°2

**ETAPE 1 : Structuration des données, lien avec Python**

Rétro conception

1. **Faites un retro engineering pour déterminer le modèle EA (utilisez Looping)**



(Nous ne sommes tout à fait convaincus concernant nos cardinalités.

1 maladie infecte 1 population

1population est infectée par n maladies

(Mais plusieurs maladies peuvent infecter une population non, ou même plusieurs populations ? Doit-on poser des hypothèses ?)

1. **Ecrire les relations et schémas relationnels**

Nos relations sont :

* Maladie
* Infecte
* Population

Nos schémas relationnels :

* Maladie (id\_maladie, année, nom, pays)
* Population ( id\_population, nbr\_femme\_malade, nbr\_homme\_malade, nbr\_pop\_totale)
* Infecte ( id\_pop, id\_maladie, cl sup, cl inf, instable, taux)

Nous avons obtenu ce dernier schéma relationnel grâce à la règle liée au cardinalités n-m/n-m.

1. **Veillez à ce qu’ils soient normalisés en 3FN**

* Maladie (id\_maladie, année, nom, pays)

Cette relation est bien en troisième forme normale car :

* Tous les attributs sont atomiques
* Toute colonne dépend de toute la clé
* Il n’y a aucune DF entre les attributs non clés.
* Population ( id\_population, nbr\_femme\_malade, nbr\_homme\_malade, nbr\_pop\_totale)

Cette relation est bien en troisième forme normale car :

* Tous les attributs sont atomiques
* Toute colonne dépend de toute la clé
* Il n’y a aucune DF entre les attributs non clés.
* Infecte ( id\_pop, id\_maladie, cl sup, cl inf, instable, taux)

Cette relation est bien en troisième forme normale car :

* Tous les attributs sont atomiques
* Toute colonne dépend de toute la clé
* Il n’y a aucune DF entre les attributs non clés.

1. **Faire un modèle physique**

Voici notre modèle physique :

Infecte

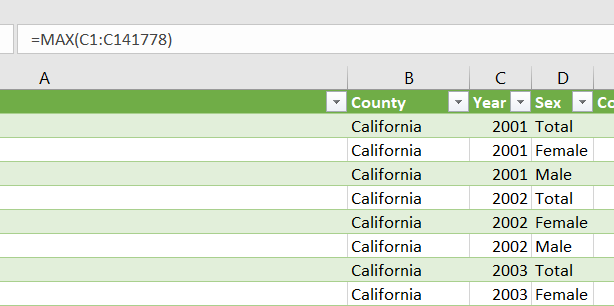
Population

Maladie

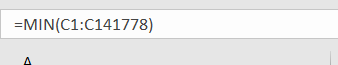
Data analyse, lien avec Excel

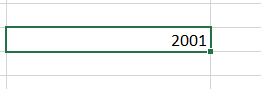
1. Utilisation des fonctions mathématiques simples d’Excel

Parmi les fonctions simples disponibles sur Excel, on retrouve le MAX & MIN :









Utiliser les fonctions MAX et MIN nous permet de voir rapidement jusqu’à quelle année les valeurs de notre dataset s’étendent.

On peut voir que notre dataset s’étend sur une durée allant de 2001 à 2014.

On remarque que dans notre dataset, nous avons une colonne count, qui est étroitement liée avec la colonne « sex ».

Nous avons donc fait la moyenne des femmes malades et des hommes malades comme suit :



Commentaires : On sélectionne les données où les cases allant de D1 à D141778 continent ‘female’, puis ont fait la somme.

On obtient :

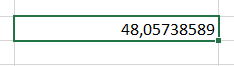


En moyenne, 74 femmes ont été infectés par une maladie.

Afin de comparer cette valeur à celle des hommes, pour voir si le sexe à un impact sur la transmissibilité de la maladie, on entre :



On obtient :



On peut donc voir qu’il y a beaucoup moins d’homme qui ont été touché par la maladie que de femmes.

Lorsqu’on souhaite voir combien de personnes au total ont été touchée par la maladie :



On obtient un résultat égal à la somme de nos deux résultats précédents, ce qui est tout à fait cohérent :



1. Création de figures

Bonus : Tableaux croisés dynamiques

Lien avec Python

Pour faire le lien avec Python, nous aurons besoin de la librairie Pandas.

Pour cela, nous avons utilisés Spyder, où pandas est déjà installé.

Dans un premier temps, on importe la librairie :



Ensuite, on importe notre dataset :



(j’ai renommé notre dataset pour des raisons de rapidité et d’efficacité.)

Puis, on crée notre nouvelle table :

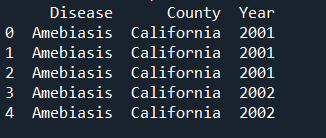


Celle-ci contient les colonnes Disease, Country, et Year.

Ensuite, on la met sous le format csv.

Nous pouvons utiliser la commande head() pour avoir accès aux cinq premiers enregistrements :



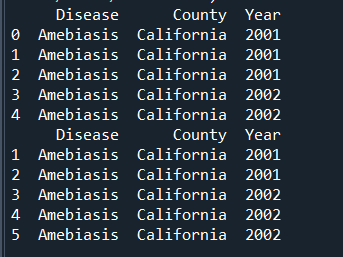


b) Suppression d’un enregistrement

Pour cela, on utilise la fonction drop comme suit :



On affiche de nouveau les cinq premiers enregistrements afin de voir si notre enregistrement 0 s’est bien supprimé.

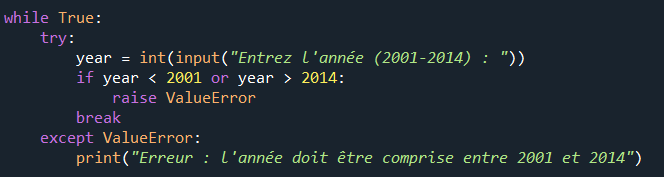


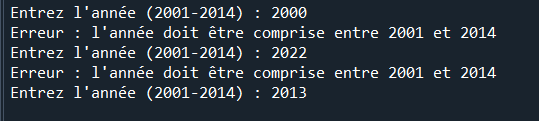
On peut voir que notre enregistrement 0 s’est bien supprimé.

c) Contraintes d’entrées

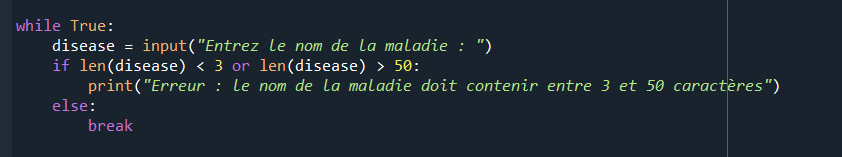
Afin que l’utilisateur ne puisse pas entrer n’importe quelle valeur dans notre dataset, il nous faut définir des contraintes.

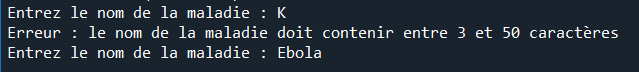
Par exemple, nous avons vu précédemment que nos données s’étalent les années allant de 2001 à 2014, ainsi nous avons crée une restriction qui ne permet d’entrer des données en dehors de cet intervalle :





Nous avons aussi créé une restriction sur le nom des maladies :





Pour finir, nous avons fait un sorte qu’un utilisateur ne puisse pas insérer une valeur vide :

Etape 2 : SQL

1. Ajout des données Kaggle
2. Commentaires nécessaires à la compréhension