# PROJET TRANSVERSE - MBA BIG DATA - SQL

**Objectif** :

- Manipuler et analyser de la Data sous SQL
- Programmer en SQL
- A rendre avant le 10 mai 23h (un jour de retard = 2 points en moins)

**Rendu :** Un fichier .sql commenté avec les numéros d'exercices + un rapport comprenant les graphiques effectués sur l'outil de data viz de votre choix. Bonus : un lien vers un répertoire GIT avec le doc .sql et le rapport.

**Contexte** :
Une société X a envoyé ces données client ainsi que les achats sur l'année N-2 (2016) et N-1 (2017).

1. **Etude global**
   a. *Répartition Adhérant / VIP ....*
   Constituer un camembert suivant la répartition suivante :
   - VIP : client étant VIP (VIP = 1)
   - NEW_N2 : client ayant adhéré au cours de l'année N-2 (date début adhésion)
   - NEW_N1 : client ayant adhéré au cours de l'année N-1 (date début adhésion)
   - ADHÉRENT : client toujours en cours d'adhésion (date de fin d'adhésion > 2018/01/01)
   - CHURNER : client ayant churner (date de fin d'adhésion < 2018/01/01)

   Note : le critère le plus au-dessus est prioritaire, exemple : un client étant VIP, et ayant adhéré sur l'année N-1 sera compté comme étant VIP

   b. *Comportement du CA GLOBAL par client N-2 vs N-1*
   Constituer une boîte à moustache pour chaque année (N-2 et N-1) comparant le CA TOTAL (TTC) des clients (sommer les achats par client par années)

   c. *Répartition par âge x sexe*
   Constituer un graphique montrant la répartition par âge x sexe sur l'ensemble des clients.

2. **Etude par magasin**

   a. *Résultat par magasin (+1 ligne Total)*
   Constituer un tableau reprenant les données suivantes :
   - MAGASIN
   - NOMBRE DE CLIENT RATTACHE AU MAGASIN (avec une color_bar en fonction de la quantité)
   - Nombre de client actif sur N-2
   - Nombre de client actif sur N-1
   - % CLIENT N-2 vs N-1 (couleur police : vert si positif, rouge si négatif)
   - TOTAL_TTC N-2
   - TOTAL_TTC N-1
   - Différence entre N-2 et N-1 (couleur police : vert si positif, rouge si négatif)
   - indice évolution (icône de satisfaction : positif si %client actif évolue et total TTC aussi, négatif si diminution des 2 indicateurs, moyen seulement l'un des deux diminue)

   Note : on effectuera un trie sur l'indice d'évolution (les positifs en haut, les négatifs en bas.

b. *Distance CLIENT / MAGASIN*

Le but étant de calculer la distance qui existe entre le magasin et le client. Les infos disponible pour le moment sont :

- la ville du magasin
- le code insee du client

Il faut donc télécharger les données GPS des villes et code-insee pour pouvoir calculer la distance :
https://public.opendatasoft.com/explore/dataset/correspondance-code-insee-code-postal/

Une fois les données acquises, il faut lier les données GPS composé de la latitude et de la longitude au client et au magasin. (constituer pour chaque client et chaque magasin 2 colonnes : latitude et longitude).

Créer une fonction qui détermine la distance entre 2 points. La fonction doit prendre 4 variable en compte : latitude1, longitude1, latitude2, longitude2

pour savoir si la fonction est correct : http://www.lexilogos.com/calcul_distances.htm

Constituer une représentation (tableau ou graphique --> au choix) représentant le nombre de client par distance : 0 à 5km, 5km à 10km, 10km à 20km, 20km à 50km, plus de 50km

## 3. Etude par univers

a. *ETUDE PAR UNIVERS*

Constituer un histogramme N-2 / N-1 évolution du CA par univers

b. *TOP PAR UNIVERS*

Afficher le top 5 des familles les plus rentable par univers (en fonction de la marge obtenu) (tableau ou graphique -> au choix)