**Questions projet 4**

**22-05-2020**

1. Dans chapitre ‘Création du tableau ‘data’ pour l’analyse’ :
   * Il y a des colonnes ‘Unnamed :0’ lors de l’import de chaque csv qui semblent correspondre à la colonne index de chaque tableau. J’ai tenté de résoudre le problème en rajoutant le paramètre ‘index=False’ lors de l’export. Ça fonctionne, sauf avec ‘customers’ et j’ai finalement droppé la colonne restante.
   * Comment éviter l’apparition de ces colonnes ?
2. Dans sous-chapitre ‘Chiffre d’affaire mensuel’ :
   * Dates : Les mois sont numérotés de 01 à 12 dans la colonne ‘month’. La période étudiée va de mars 2021 à février 2022. Si je crée un graphique avec les mois en abscisse, les mois de janvier et février 2022 vont apparaître avant les mois de 2021. J’ai donc remplacé les valeurs de mois 1 et 2 par 13 et 14 pour respecter l’ordre.
   * Y a-t-il une autre méthode ?
3. Dans sous-chapitre ‘Chiffre d’affaire par catégorie et par mois’ :
   * Histogramme : comment afficher les colonnes de chaque catégorie empilées ? **Stacked bar graph**
4. Dans sous-chapitres ‘Top 10 mensuel des meilleures ventes et meilleurs clients’ :
   * Est-ce possible de créer une fonction ?
   * Je fais une tentative dans le sous-chapitre ‘Top 10 des meilleurs clients par mois’ (passée en Markdown).
5. Chaque fois que je crée un graphique (premier graphique dans sous-chapitre ‘Chiffre d’affaire mensuel’) :
   * Un message d’erreur apparaît même si l’affichage du graphique fonctionne.
   * Est-ce normal ?
6. Dans sous-chapitre ‘Courbe de Lorenz et indice de Gini’ :
   * L’export de la courbe de Lorenz donne un graphique sans courbe.
   * Quel est le paramètre à renseigner pour que l’export se fasse correctement ? **Essayer d’autre format (png…)**

Notes projet 4

**Contradiction au niveau de la distribution empirique des ventes :**

* 1. Avec la comparaison Mode – Médiane – Moyenne : je déduis que la distribution empirique des ventes est étalée à gauche
  2. Avec Skewness : je déduis que la distribution empirique des ventes est étalée à droite

REVOIR CONDITION MODE MEDIANE MOYENNE LIGNE 28 – FAIT

**Rajouter dans analyse de dispersion :**

[coefficient de variation](https://openclassrooms.com/fr/courses/4525266-decrivez-et-nettoyez-votre-jeu-de-donnees/4730346-comprenez-les-mesures-de-dispersion) (CV = écart-type empirique divisé par moyenne)

Écart moyen absolu

Médiane des écarts absolus (mesure plus robuste que EMA)

**Ne plus afficher warning :**

import warnings

warnings.filterwarnings('ignore')

**Afficher mois dans le bon ordre :**

dataCAParMois = data  
#dataCAParMois['month'].replace({1:13,2:14}, inplace = True)  
dataCAParMois = dataCAParMois.groupby(['month', 'year'])['price'].sum().reset\_index()  
dataCAParMois.sort\_values(['year', 'month'], ascending = [False, False])

**Feedback général**

NETTOYAGE :

Faire un code pour identifier les données à nettoyer :

* Fonction faisant des print ‘pas de valeurs nulles’, ‘pas de doublons’
* Compléter isnull par isnull().any()

Faire un autre code pour nettoyer les données :

* Si doublons, supprimer doublons…
* Faire une fonction pour chaque fichier csv
* Faire une fonction qui crée nouveau fichier avec name +’\_clean’

ANALYSE : bien mais **trop exploratoire**

Faire Referential integrity : cad vérifier si les données de ‘transactions’ sont bien dans ‘products’ Fait

Pas de fil conducteur

Exemple : je mets en lumière la baisse de CA mais je ne l’interprète pas et je ne poursuis pas par une étude du mois d’octobre en particulier (par jour, par catégorie de produits…)

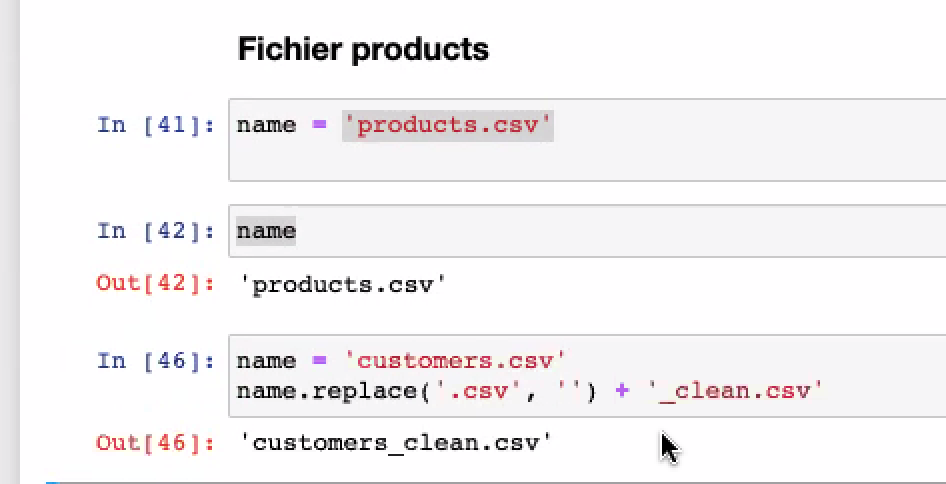
Utiliser des graphes appropriés pour mettre en lumière l’info souhaitée

Pas de métrique identifiée :

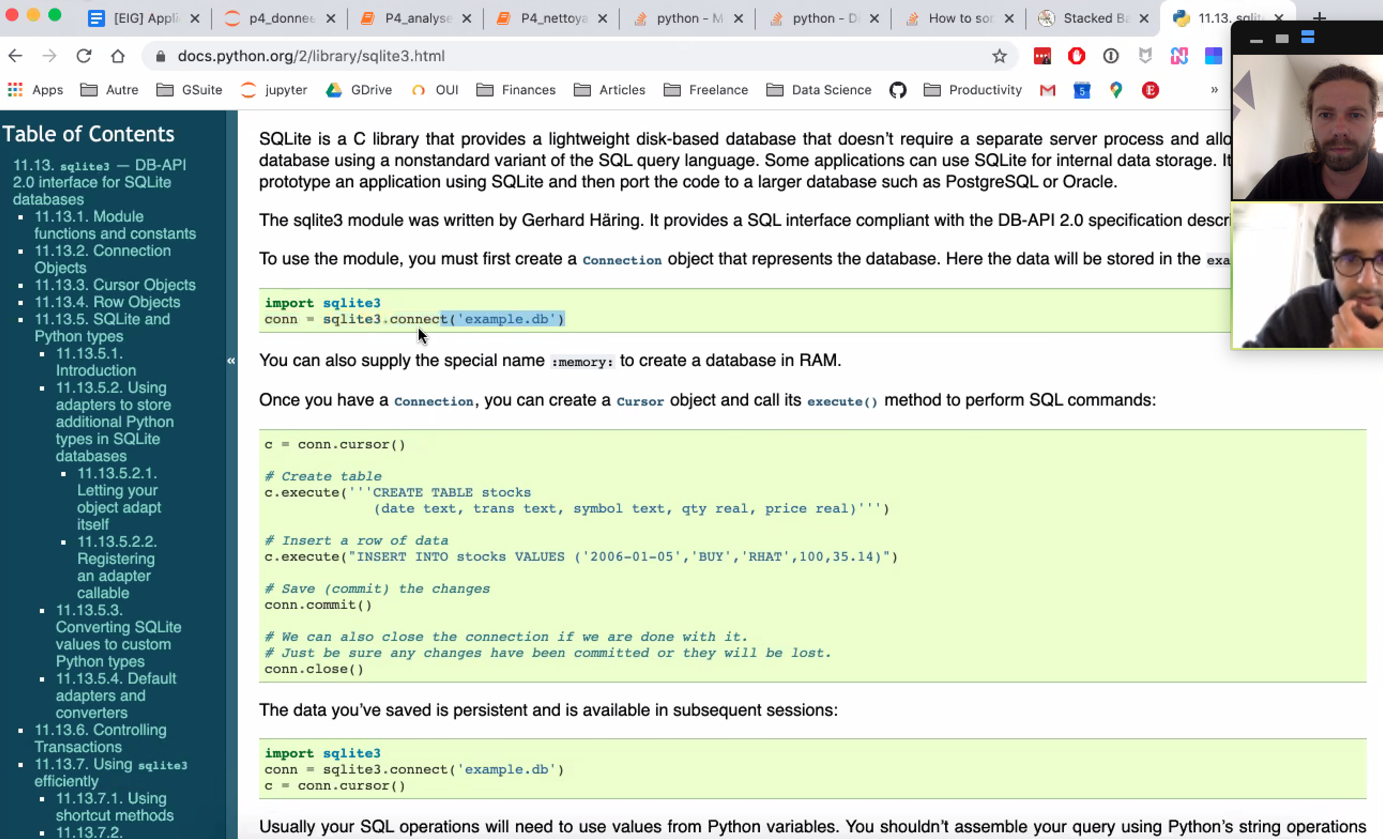
Il faut absolument déterminer les métriques essentielles à l’activité étudiée. Ici :

* Panier moyen par client Fait
* Nombre d’achat par client Fait
* Pourcentage de clients ayant acheté tant de produit Fait
* Pourcentage de client ayant dépensé tant d’argent Fait
* Identifier les clients trop occasionnels Fait
* Identifier les top clients Fait
* Identifier les top produits Fait
* Faire une distribution du panier moyen par client Fait
* Créer des tables pour chacune de ces métriques
* Les analyser puis les analyser entre elles (cross table)





S’entrainer à faire du SQL exécuté dans Python



Session mentorat du 12 juin :

NETTOYAGE :

* Préciser si les doublons identifiés sont de réels doublons ou bien des mêmes achats de la même personne à la même date. Il manque effectivement une colonne ‘quantity’ pour les achats de plus d’une quantité et répertorier des achats sans cette colonne peut aboutir à l’apparition de données identiques mais qui ne sont pas des doublons.
* Reprendre le fichier envoyé par Nico

**L’affinité entre produits**

Cette **KPI e commerce** est souvent oubliée et pourtant, elle peut vous aider à faire décoller vos ventes et à augmenter le panier moyen. Elle consiste à analyser les produits de votre boutique qui sont souvent achetés par le même client.

Déterminer les articles complémentaires, qui plaisent au même type de consommateurs, vous permet alors de peaufiner vos fiches produits et d’augmenter le taux de vente liée.

Si vous avez une boutique de vaisselle et que vous remarquez que les mugs sont souvent achetés en même temps que des bols, peaufinez vos descriptifs en suggérant automatiquement aux personnes qui parcourent les mugs d’opter aussi pour des bols.

Pour améliorer vos performances commerciales, vous devez absolument analyser le trafic vers votre boutique et le comportement des internautes. C’est la clé pour constamment peaufiner vos stratégies et vous assurer de meilleurs résultats.

Ces 7 **KPI e commerce** constituent une bonne base pour évaluer l’environnement de votre site marchand et entrer dans une démarche d’amélioration des performances.

Chercher un boxplot pertinent : **panier moyen par categ**

**Corrélations :**

Âge client / Montant total achats : forte

Âge clients / Nombre d’achats : forte

Âge clients / Montant panier moyen : faible

Âge clients / catégorie : faible

Modifs :

Nettoyage : maintien des 126 doublons

Analyse :

* réorganisation avec fil rouge
* nouveau boxplot
* suppression des distributions de ventes de produits de chaque catégorie (il y a déjà celle avec les trois catégories ensemble)
* suppression catégories ‘client à oublier’ (fusionnée avec ‘Clients à relancer’) et ‘très bons clients’ (fusionnée avec ‘Clients stars’) dans RFM