L'objectif de notre analyse était de procéder à un rapprochement entre les données extraites de notre système d'Enterprise Resource Planning (ERP) et celles extraites de notre Content Management System (CMS). Ces données étaient présentes dans deux fichiers, un fichier issu de l'ERP nommé erp.csv et un fichier issu du CMS nommé web.csv. Ces deux systèmes n'étant pas intégrés, l'accès aux données et la gestion des rapprochements entre ces deux systèmes sont plutôt « artisanaux », très manuels, et nous avons également des problèmes d'intégrité sur les données exportées par ces deux systèmes. Un troisième fichier nommé liaison.csv avait été préparé par notre stagiaire mais pour des raisons qui seront explicitées plus loin, on a choisi de ne pas utiliser ce fichier, dans la mesure où il n'apparaît pas ajouter aux informations contenues dans les fichiers ERP et WEB, que ce soit en termes de quantité ou de qualité.

Ces notes présentent la démarche qui a présidé au rapprochement des deux fichiers WEB et ERP, dans un premier temps en détaillant l'import et la vérification du typage des données, puis dans un second temps en expliquant comment la qualité des données a été vérifiée et améliorée, depuis l'analyse des doublons jusqu'à l'identification des clés primaires dans les tables en passant par l'identification et l'analyse des valeurs manquantes. Enfin dans une troisième partie, on passera en revue les indicateurs spécifiques qui ont été requis par le COPIL de la mission, à savoir les calculs du chiffre d'affaires et l'analyse de la structure de prix de nos produits. On conclura enfin avec quelques pistes pour l'amélioration des informations et indicateurs gérés par nos systèmes.

Les références [entre crochets] renvoient aux cellules du notebook Jupyter.

Étape 1 - Import des données et vérification du typage des colonnes

[1]

Dans cette première étape, nous avons importé nos 3 fichiers .csv dans des dataframes et inspecté les colonnes d'informations qu'ils contiennent. Nous avons également vérifié la cohérence de ces informations avec le data type des colonnes et nous avons modifié les data types là où c'était nécessaire.
Nous avons également renommé certaines colonnes en vue de faciliter l'identification des informations dans les jointures de tables dont nous aurons besoin plus tard pour calculer le chiffre d'affaires. [7]

Enfin, nous avons vérifié le nombre d'identifiants produit et d'identifiants produit WEB présents dans chaque fichier. [17]

Nous avons constaté que :

- 825 références produit uniques sont présentes dans le fichier de liaison ;
- 94 de ces références n'ont pas de référence produit WEB correspondante ; seules 731 réf érences produit WEB sont présentes. Ces références manquantes ne concernent-elles qu e des produits non-vendus en ligne, i.e. n'ayant pas de référence sku? Dans le cas contra ire, le fichier de liaison est potentiellement incomplet ;
- 715 références produit WEB uniques sont présentes dans le fichier WEB [18]; et

825 références produit uniques sont présentes dans le fichier ERP [19], dont 717 sont indiquées comme étant vendues en ligne et 108 sont indiquées comme étant vendues offline.

Cette analyse nous donne une première idée du nombre de valeurs manquantes que nous devrons essayer de compléter avant de pouvoir croiser les fichiers avec des jointures. Nous y reviendrons en deuxième partie.

Étape 2 - Étude de la qualité des données, duplicatas, imputations, suppressions d'indicateurs

2.1 - Identification, analyse & suppression des doublons[20]

Dans un premier temps, nous avons donc recherché la présence de doublons dans les 3 tables. Aucun n'a été trouve dans les tables ERP et liaison, en revanche 82 ont été trouves dans la table WEB. [21] En les inspectant plus précisément, on s'aperçoit qu'il s'agit de lignes contenant seulement des cellules vides[22] , des zéros et des NaNs, et en particulier sur toutes ces lignes le sku_web est manquant et il n'y a aucun descriptif produit, ni information sur les quantités vendues, donc elles sont inutilisables pour notre calcul de chiffres d'affaires [23] . Toutes ces lignes dupliquées ont donc été supprimées du fichier web.

2.2 - Identification des valeurs manquantes

La table ERP ne semble pas en contenir. [24] La table liaison [25] en contient un nombre significatif dans la colonne sku (ce sont les 84 identifiants WEB manquants dont nous avons parlé plus tôt). Le fichier WEB quant à lui contient des colonnes entièrement vides [26], d'autres à moitié vide, et également des références sku_web manquantes. Ce dernier point est très étrange et mérite d'être étudié plus en détail, nous y reviendrons donc.

2.3 - Suppression des lignes et des colonnes vides

Pour réduire le fichier WEB aux données utilisables [27], nous avons donc supprime les colonnes entièrement vides, et également les lignes ne contenant que des zéros. Il nous reste à identifier les clés primaires dans nos 3 tables avant d'effectuer des jointures. [28]

2.4 - Vérification des clés primaires

Nous avons donc défini une fonction nous permettant d'identifier si une colonne d'un dataframe est une clé candidate. [29]

[30] Dans les tables ERP et liaison, on voit que les colonnes des product ids sont des clés candidates, en revanche sur la table WEB, ce n'est pas le cas pour la colonne sku_web **[31]**; il y a des doublons dans la colonne sku_web **[32]**. Nous avons donc cherché si on pouvait lui adjoindre une autre colonne pour former la clé primaire, ou trouver un moyen d'éliminer ces doublons. Il n'y a en effet que 715 valeurs uniques dans la colonne sku_web sur un total de 1431 lignes. [33] Inspectons les autres colonnes. [34] La colonne post_mime_type ne contient que 2 valeurs uniques, et la colonne post_type 3 valeurs uniques. [35] La source des duplicatas pourrait dont être la présence d'une photo pour chaque référence produit sur le site, ce qui doublerait dont le nombre de lignes – et c'est l'ordre de grandeur que nous cherchons (x 2 – de 715 à 1,431 lignes comme on l'a mentionné). [36 à 39] Nous avons donc essayé de retirer ces duplicatas en excluant les lignes où le post_mime_type est une image/jpeg et en ne gardant que les lignes où le post_type est un produit. Il reste 716 lignes dans le fichier après cette manipulation, dont 715 ont des identifiants sku_web uniques, donc sku_ web n'est toujours pas une clé candidate après cette manipulation et il nous faudra identifier le sku_ web en double pour en exclure un impact possible sur le calcul du chiffre d'affaires (nous y reviendrons). Il faut également préciser que dans le même temps, nous avons retiré de la table WEB les colonnes inutiles pour le calcul du chiffre d'affaires, telles celles contenant des mots de passe ou des commentaires clients. [40] Nous avons alors inspecté les autres colonnes du fichier WEB pour rechercher une clé candidate. L'inspection de la colonne ['guid'] en

particulier a livré des données intéressantes, [41] puisque les chaînes de 4 caractères à droite des strings ['guid'] ressemblent aux références produit présentes dans le fichier de liaison. Vérifions si elles sont cohérentes dans tout le fichier et pourraient éventuellement constituer une clé candidate dans la table web. [42]

[43] Nous avons donc créé une nouvelle colonne dans la table web_trim contenant une extraction de ces 4 caractères et le product_id_web ainsi crée s'avère être une clé candidate de la table web_trim. Notre hypothèse est donc vérifiée. [44]

[45] Si on essaie de vérifier l'hypothèse sur la table web_trim avant retraitement pour la présence d'images, on s'aperçoit [46] que le product_id_web que nous avons créé cesse d'être une clé candidate. Dans la pratique, il faudrait donc à l'avenir scinder l'export de la table WEB en 2 fichiers (un pour les photos et un pour les produits) pour éviter ces doublons et rendre les données plus directement exploitables avec un minimum de manipulations.

Les clés primaires de nos tables sont donc :

- Pour la table ERP => product_id_erp
- Pour la table web_trim => product_id_web

[47] Poursuivons à present comme on l'a dit précédemment l'analyse des doublons sku_ web dans la table web_trim. [48] Le sku_web dupliqué est celui du produit 5075 ; le duplicata est en fait une cellule vide (le sku_web de ce produit est en fait manquant). [49] Nous avons donc recherché les autres sku_web manquants dans la table web. Nous en trouvons 2. Une vérification du fichier de liaison nous indique que les identifiants sku sont également manquants pour ces 2 produits (5070 et 5075) dans la table de liaison. [50] En tout état de cause, on ne peut pas utiliser le sku_web dans la table WEB comme clé primaire car il contient des doublons ; certes ces doublons sont des valeurs NULL (i.e. des cellules vides) mais il n'est pas possible de remplacer ces références comme elles n'existent dans aucun autre fichier pour les 2 Product id concernés. On aurait pu choisir de supprimer ces lignes, dont les ventes et in-stock sont à zéro, mais nous avons préféré les garder dans la table et utiliser une autre colonne que le sku_ web comme identifiant primaire de la table retraitée, c'est-à-dire le product_id_web. La fiabilité de la colonne sku_ web est en outre discutable, car elle contient ces mêmes valeurs vides (un fichier WEB avec des identifiants WEB manquants, ça remet en cause la fiabilité de l'import de données et/ou des informations présentes dans le CMS qui gère le site web,mais cela sort du périmètre de notre analyse). [51] Si on vérifie la table ERP, ces 2 produits ont leur indicateur onsaleweb à True, donc ils doivent bien *a priori* figurer dans les extractions du CMS. On remarque aussi que ces produits ont un indicateur « out-of-stock » - nous avons donc essaye de vérifier si cela avait pu avoir un impact sur l'absence de leur sku_web – c'est-à-dire, l'export des données a-t-il pu pour une raison donnée « planter » entre guillemets parce que le stock était à zéro ? Pour vérifier cette hypothèse, nous avons donc listé les produits destinés à être vendus en ligne d'après la table ERP tout en ayant un stock à zéro [52] : on en trouve 143 qui sont uniques. [53] En les listant et en inspectant visuellement la liste, on ne note pas d'erreurs apparentes (telles des irrégularités de formatage ou des erreurs lexicales par exemple). [54]

[55] On a donc ensuite fait une jointure entre cette liste de 143 produits vendus en ligne d'après l'ERP mais avec un stock à zéro, avec les données de ventes de ces produits dans la table web, et on a ensuite restreint le résultat aux produits pour lesquels l'identifiant sku_web est manquant. On ne trouve que 3 produits. [56] On retrouve les produits références 5070 et 5075 déjà identifiés plus haut.

On peut donc exclure l'hypothèse qu'un état de stock à zéro ait provoqué une erreur de référencement web, 143 produits vendus en ligne étant out-of-stock mais seulement 3 références sku_web étant manquantes en ligne dans le fichier web_trim. Il s'agit plus probablement d'erreurs de saisie dans l'un des 2 fichiers (le plus probablement dans l'ERP, i.e. produits indiqués comme devant être

vendus en ligne alors qu'ils ne le sont pas dans le cas des produits 5070 et 5075). Ces 3 produits sont out-of-stock avec zéro total_sales => là encore, s'agit-il erreurs de saisie? On peut les ignorer car pas d'impact sur le CA (ces 3 produits ayant des ventes à zéro). (Note : en pratique, il faudrait en plus vérifier s'il n'y a pas eu de problème d'enregistrement des ventes en lignes pour justifier qu'elles soient à zéro - il est en effet peu probable qu'un produit destine à être vendu en ligne ait un stock à zéro en n'ayant pas été vendu du tout - à moins qu'il ne s'agisse d'un problème d'approvisionnement, à vérifier également. Ces vérifications sont hors du champ de cette analyse, les données n'étant pas disponibles, mais il est néanmoins bon de le mentionner). Nous avons donc trouvé 716 product_id_web uniques dans le fichier web, or il y a 717 produits vendus en ligne d'après le ficher ERP. On retrouve le produit 4594 déjà identifié à l'étape précédente, et que nous pouvons donc ignorer comme on l'a dit car ses ventes sont à zéro. [57 à 59] => Les autres produits vendus sur le WEB mais out of stock ayant tous des sku_web renseignés à l'exception de 3, donc on peut définitivement écarter l'hypothèse du problème d'import de données. Vérifions à present les références WEB manquantes dans le fichier de liaison. [61] Comme nous l'avons dit précédemment, 94 produits ont un identifiant WEB manquant dans le fichier de liaison. Si on croise ces 94 produits avec la table ERP, on s'aperçoit que 5 d'entre eux sont destinés à être vendus sur le web. On a déjà identifié 3 d'entre eux (4594, 5070 et 5075) et décidé de les ignorer pour l'analyse. [62]

[63] On note par ailleurs une erreur sur le produit 4954 => stock_quantity = 0 mais stock_status = instock. Il semblerait logique que ces 2 indicateurs soient liés dans le système pour éviter ce type d'erreur mais ça ne semble pas être le cas (**Note:** ici encore, il faudrait s'assurer d'une absence d'erreur de données sur les imports depuis le SKU ou sur les enregistrements des variations de stock, mais ces vérifications sont hors du champ de cette analyse, les données n'étant pas disponibles). En affichant l'information complète de ce produit, on s'aperçoit que ce produit est un bon-cadeau, donc pas un "vrai" produit physique (le "stock" est "crée et détruit" au moment de la vente). Il faudrait néanmoins vérifier si d'autres produits ne sont pas concernés par cette anomalie, qui pourrait indiquer un problème avec le SKU (Note: hors-sujet de notre analyse ici encore). Les ventes de ce bon-cadeau doivent potentiellement être retirées du CA total afin de ne pas dupliquer les valeurs (si on suppose que tous les bons vendus ont déjà été utilisés en totalité). [64] Il nous reste donc le produit 7247. En faisant une jointure avec la table web-trim, on s'aperçoit que ce produit a bien un sku_web renseigne dans notre fichier web_trim. Notre fichier web_trim semble donc relativement complet. Nous nous passerons donc définitivement d'utiliser le fichier de liaison, eu égard aux réserves sur son intégrité émises plus haut. Il nous reste à present à aborder les résultat des analyses requises pour le COPIL

Étape 3 - Analyses requises pour le COPIL : indicateurs statistiques, filtrage, jointures & visualisation

3.1 - Rapprochement des fichiers WEB et ERP

Nous avons commencé par procéder à une jointure externe entre les fichiers web_trim et ERP. On s'aperçoit que la jointure ne se fait que d'un côté sur 109 lignes, toutes situées à droite, donc 109 lignes de la table ERP n'ont pas de correspondance dans la table web_trim – or on devrait n'en trouver que 108 (souvenons nous somme on l'a dit plus haut que sur les 825 références produits uniques du fichier ERP 108 ne sont pas vendues sur le site web, donc logiquement on ne s'attend pas à les trouver dans le fichier web) – mais il nous faut identifier la 109ème. [65 à 69]

[70] On retrouve le produit 4594 déjà identifié dans l'étape précédente qui n'a pas d'identifiant sku dans la table liaison ni dans la table web, et qui peut donc être ignoré. Comme nous cherchons le chiffre d'affaires des ventes en ligne, un left joint sur la table web_trim s'impose. [71]

3.2 - Calcul du chiffre d'affaires par produit et du chiffre d'affaires total

[72] Pour le calcul du chiffre d'affaires par produit, il suffit de créer une colonne dans le résultat de notre jointure gauche entre ERP et web_trim (que nous avons appelée CA_source ici) et d'y mettre le produit des ventes totales en nombre de bouteilles par le prix de vente unitaire d'une bouteille. Ensuite, pour simplifier la lecture, nous n'avons garde que 2 colonnes, la référence du produit et son chiffre d'affaires en euros. [73] En l'absence d'informations pertinentes dans les fichiers ERP et WEB, il est impossible de savoir si les bons-cadeau vendus ont été utilisés, totalement ou partiellement ou pas du tout. Le chiffre d'affaires total a donc été calculé de 2 manières différentes, en les incluant et en les excluant. Il serait recommandé d'inclure un flag sur les transactions WEB payées avec des bons cadeau, afin d'éviter une augmentation artificielle du chiffre d'affaires qui pourrait en résulter, ainsi que d'éventuels problèmes de gestion de stock.

[74 à 76] Le chiffre d'affaires total s'élève à 70,568.6 € bons-cadeau inclus.
Le chiffre d'affaires des bons-cadeau s'élève à 250.0 €.
Le chiffre d'affaires total s'élève à 70,318.6 € bons-cadeau déduits.

3.3 - Identification des outsiders prix dans le fichier ERP

[77] Nous avons tout d'abord cherché à décrire la distribution des prix des vins présents dans le fichier ERP par le biais d'un histogramme, afin de déterminer quelle méthode d'identification des outsiders serait plus pertinente. [78] Nous avons donc calculé la taille de notre échantillon, les 1er et 3ème quartiles, l'écart interquartile et le nombre de barres de l'histogramme en utilisant la règle de Freedman-Diaconis pour la largeur des intervalles. [79 & 80] Nous sommes arrivés à l'histogramme suivant, qui montre clairement que la distribution des prix ne semble pas suivre une Loi Normale. Au lieu du Z-score, nous utiliserons donc ici un box plot pour visualiser les outsiders, ce qui nous donne le graphique suivant [81] – avec une valeur de la moustache haute à 83.1 euros, on a donc pu identifier 37 vins sur 825 dont les prix semblent anormalement élevés et qu'il serait bon de vérifier. [82 & 83] En l'absence d'informations plus détaillées sur les types de vins, il est impossible de déterminer si les prix de ces vins "outsiders" sont des valeurs atypiques (vins haut-de-gamme par exemple) ou aberrantes (telles qu'auraient pu en créer des erreurs de saisie, par exemple). Les vins concernés ont donc tous été conservés dans l'analyse, aucun critère discriminant n'étant disponible pour savoir s'ils doivent être retirés ou corrigés.

Conclusion

Comme nous en avons discuté dans cette présentation, notre jeu de données de départ était donc affecté d'un certain nombre de problèmes d'intégrité qu'il a fallu identifier, analyser et résoudre. Faute d'informations suffisantes disponibles, certains de ces problèmes n'ont pas pu être résolus ou seulement partiellement. La plupart de ces problèmes semblent liés à une mauvaise intégration des systèmes, ainsi qu'à des erreurs de saisie et d'export des données qu'il serait bon de corriger.

***