**PROJET 5**

**Introduction**

Bonjour, je m’appelle Naziha, je suis data analyste freelance au sein de l’entreprise **Bottleneck** spécialisée dans la vente de vins en ligne.

Ma mission est de rendre la gestion des données plus optimale et de mettre en place un système de gestion moderne et automatisé. Nous avions :

* Un ERP non relié au site des ventes en ligne
* Une gestion des stocks complexe
* Visibilité réduite des ventes

**Objectifs**

* Rapprocher les 2 exports : un export de l’ERP contenant les références produit, le prix de vente, l’état du stock avec un export d’une table de l’outil de CMS contenant les informations des produits commercialisés en ligne (nom, description, nombre de ventes, …)
* Obtenir le CA par produit, et le total du CA réalisé en ligne
* Effectuer une analyse univariée sur le prix des produits dans le but de détecter d’éventuelles valeurs aberrantes

**CMS**: système de gestion de contenu, (en anglais : content management system), il permet de créer un site internet, un blog, ou encore un site de ventes en ligne. (ex : WordPress, Wix)

**ERP**: (Enterprise Resource Planning) est un logiciel optimal et performant de gestion que les entreprises utilisent pour gérer leurs activités quotidiennes telles que la comptabilité, les achats, les opérations de supply chain.

*Il permet de gérer l’ensemble des processus opérationnels d’une entreprise en intégrant plusieurs fonctions de gestion : solution de gestion des commandes, solution de gestion des stocks, solution de gestion de la paie et de la comptabilité, solution de gestion e-commerce, solution de gestion de commerce BtoB ou BtoC, etc.*[*dans un système*](https://www.choisirmonerp.com/erp/definition-d-un-erp)*.*

Process de traitement

* Obtenir la dimension de chaque dataset *.shape*
* Renommer la colonne SKU => id\_web
* Comptabiliser le nombre de valeurs manquantes par variable : *.isna().sum()*
* Obtenir le format des variables *.info()*
* Vérifier que les clés « chaînes de caractères » sont dans les 2 exports *.unique()*
* Suppression des variables entièrement vides (100% des valeurs) *.drop()*
* Suppression des 2 clés nan dans le dataset id\_web *.dropna()*
* Filtrer le dataset web\_df sur « Product »
* Jointure interne sur les clés product\_id et id\_web *(pd.merge(how=‘inner’))*

Pour mener à bien cette opération de nettoyage, j’ai filtré sur Product pour écarter les doublons et redondances, j’ai supprimé les 5 colonnes vides ainsi que les valeurs manquantes de la clé « id\_web »

**Analyse Univariée**

**Boxplot sur la variable prix :**

* La boxplot permet de présenter la dispersion des prix
* L’avantage est qu’il n’a pas d’agrégation et que chaque point de donnée est considéré
* La médiane est représentée par la ligne dans la boîte. La médiane est une mesure courante du centrage des données. La moitié des observations lui sont inférieures ou égales et la moitié des observations lui sont supérieures ou égales.
* La boîte de l'étendue interquartile représente la moitié centrale (50 %) des données. Elle indique la distance entre les premiers et troisième quartiles (Q1-Q3).
* Les moustaches s'étendent des deux côtés de la boîte. Elles représentent les plages des 25 % inférieurs et des 25 % supérieurs des valeurs de données, à l'exclusion des valeurs aberrantes.
* La médiane est située à 23,55 EUR
* Concentration des prix entre 14,10 et 42 EUR correspondant à la boîte du boxplot
* Moustache inférieure : de 5,20 à 14,10 EUR
* Moustache supérieure : de 42,18 à 84,10 EUR
* 84 EUR représente le seuil de l’écart interquartile
* 32 Outliers

**Méthode Interquartile**

* L’écart interquartile est une mesure de dispersion qui permet de quantifier la variabilité des données d’un ensemble. Il est souvent utilisé pour détecter les valeurs aberrantes dans un ensemble de données.
* Toutes les billes rouges représentent les prix au-dessus de l’IQR et elles sont de l’ordre de 32.
* La concentration des prix se situe en bleu à moins de 50 EUR.

**Z-SCORE**

Les Z-scores permettent d'évaluer la déviation d'une valeur par rapport à la moyenne d'un ensemble de données, en termes d'écart-type. *Il est souvent utilisé pour identifier les valeurs aberrantes (Outliers).*

* A partir de la variable « price » :
* On calcule la moyenne des prix
* On calcule l’écart-type des prix
* On calcule le Z-score pour chaque valeur de la colonne « price » que nous stockons dans une nouvelle colonne « Z-score »

*Process*

Cela donne une idée de la distance de chaque valeur par rapport à la moyenne en termes d'écart-type. Un Z-score positif entre 2 et 7 indique que la valeur est loin de la moyenne, ce qui peut indiquer une valeur atypique

**Analyse complémentaire**

**Histogramme des meilleures ventes de vins de prestiges**

* 32 spiritueux de plus de 84 EUR
* Seulement 10 vins représentent des ventes
* Meilleure vente : Cognac Frapin VIP XO avec 13 ventes en 2020

**Conclusion**

Respect des bonnes pratiques de l’analyse des données

* + Pas de problème d’encodage
  + Tester l’unicité des clés
  + Vérifier les jointures (type de jointure, nombres de produits jointés)
  + Nettoyage des données