**Rapport de Nettoyage des Données - Projet "Target Brilliance"**

**Introduction**

**Dans le cadre de mes responsabilités sur le projet "Target Brilliance", j'ai été chargé de réaliser un nettoyage complet des données relatives aux commandes clients. Ce nettoyage avait pour but de préparer ces données pour une analyse plus approfondie. J'ai utilisé deux outils principaux pour accomplir cette tâche : Excel pour une première phase manuelle, puis Python pour automatiser et garantir la précision du processus.**

**1. Vérification et Nettoyage des Données avec Excel**

**J'ai commencé par ouvrir le fichier orders.csv dans Excel afin de mieux comprendre les données et d'identifier rapidement les anomalies. Voici les étapes principales effectuées :**

* **Vérification des dates d'achat :**
  + **J'ai activé un filtre sur la colonne order\_purchase\_timestamp pour vérifier qu'aucune commande ne manquait de date d'achat. J'ai ensuite supprimé les lignes où cette information était absente.**
* **Suppression des commandes avec un statut "inconnu" :**
  + **En filtrant la colonne order\_status, j'ai pu isoler les commandes marquées avec un statut "inconnu" et les supprimer. Cela a permis de ne conserver que les commandes valides dans l'ensemble de données.**

**Cette étape m'a permis de réaliser une première phase de nettoyage tout en me familiarisant avec la structure des données.**

**2. Automatisation du Nettoyage avec Python**

**Après avoir effectué un premier tri dans Excel, j'ai poursuivi avec Python pour automatiser l'ensemble du processus. Cela m'a permis de gagner en efficacité et de garantir une meilleure fiabilité des données. Voici ce que j'ai fait dans cette étape :**

* **Suppression des lignes avec des dates d'achat manquantes :**
  + **J'ai utilisé Pandas, une librairie Python dédiée à la manipulation des données, pour supprimer toutes les lignes où la colonne order\_purchase\_timestamp était vide.**
* **Élimination des commandes avec un statut inconnu :**
  + **J'ai également automatisé la suppression des commandes ayant un statut "unknown" dans la colonne order\_status, afin de s'assurer que seules les commandes valides étaient conservées.**
* **Vérification de la cohérence des dates de livraison :**
  + **Enfin, j'ai vérifié que la date de livraison (order\_delivered\_customer\_date) n'était jamais antérieure à la date d'achat (order\_purchase\_timestamp). Toutes les commandes avec une incohérence entre ces dates ont été supprimées.**

**Conclusion**

**En combinant les capacités d'Excel et de Python, j'ai pu effectuer un nettoyage complet et efficace des données. L'utilisation d'Excel m'a permis d'obtenir une vue d'ensemble rapide et intuitive, tandis que Python a permis d'automatiser le processus, éliminant ainsi les erreurs humaines potentielles et garantissant une cohérence dans les données finales. Le fichier cleaned\_orders.csv est maintenant prêt pour une analyse plus approfondie.**

**------------------------------------------------------------------------------------------------------------------**

**1. Vue d'ensemble du Dataset**

* Total des commandes : 99 441

**2. Données manquantes**

* order\_id : 0 valeurs manquantes
* customer\_id : 0 valeurs manquantes
* order\_status : 0 valeurs manquantes
* order\_purchase\_timestamp : 0 valeurs manquantes
* order\_approved\_at : 160 valeurs manquantes
* order\_delivered\_carrier\_date : 1 783 valeurs manquantes
* order\_delivered\_customer\_date : 2 965 valeurs manquantes
* order\_estimated\_delivery\_date : 0 valeurs manquantes