Rapport de projet – Data Refinement

Dans le cadre du module *Data Refinement*, j'ai réalisé un projet ayant pour objectif de transformer un dataset brut en un dataset propre et exploitable. Ce travail vise à comprendre le cycle de vie des données, à identifier les problèmes de qualité présents dans un jeu de données réel, puis à appliquer différentes étapes de nettoyage et de transformation afin d'améliorer la fiabilité des informations et de permettre une analyse pertinente.

Dans un premier temps, j'ai mis en place une structure de projet nommée **DATA-REFINEMENT-PROJECT**, en respectant l'organisation recommandée. Le dataset brut a été placé dans le dossier **DATA/RAW/**, tandis que le dataset final nettoyé est stocké dans **DATA/PROCESSED/**. Les notebooks utilisés pour l'exploration, le nettoyage et la transformation des données sont regroupés dans le dossier **NOTEBOOKS**, et le rapport final au format PDF est placé dans **REPORTS**. Cette organisation permet de structurer clairement le projet et de faciliter sa compréhension et sa réutilisation.

Le dataset choisi est **Cafe Sales – Dirty Data**, qui représente les ventes d'un café à partir de données volontairement imparfaites. Ce choix s'explique par le fait que ce dataset contient des problématiques fréquentes dans des contextes réels, telles que des valeurs manquantes, des incohérences ou des erreurs de saisie. Il constitue donc un bon support pour appliquer les concepts de *data quality* et de *data refinement* abordés durant le module.

Exploration des données

Lors de la première phase d'exploration du dataset, plusieurs problèmes de qualité ont été identifiés. Certaines colonnes contiennent des valeurs manquantes, tandis que d'autres utilisent des valeurs génériques telles que *"ERROR"* ou *"UNKNOWN"* à la place d'informations réelles. Des incohérences de format ont également été observées, notamment dans l'écriture de certaines valeurs similaires. Enfin, certaines valeurs semblent incorrectes ou peu plausibles, en particulier dans les colonnes financières.

À ce stade, aucun doublon évident n'a été identifié, mais leur présence restait possible et devait être vérifiée lors des étapes suivantes. Cette première analyse confirme que le dataset est brut et qu'un travail de nettoyage et de transformation est nécessaire avant toute analyse fiable.

Nettoyage des données

Le premier problème de qualité identifié concerne la présence de valeurs non exploitables telles que *"ERROR"*, *"UNKNOWN"* et des valeurs manquantes. Ces valeurs ne correspondent pas à des informations réelles et empêchent une analyse fiable des données. L'ensemble des décisions de nettoyage et de transformation a été guidé par une logique métier et par les principes de data quality. Lorsque l'information était critique et impossible à corriger de manière fiable, les lignes concernées ont été supprimées afin de garantir la fiabilité du dataset. À l'inverse, lorsque l'information était secondaire ou corrigeable, les

données ont été nettoyées sans supprimer les transactions, afin de conserver un maximum d'informations exploitables. Chaque choix a ainsi été réalisé dans l'objectif de privilégier la cohérence et la qualité des données plutôt que la quantité.

Colonnes critiques

La colonne **Item**, qui décrit le produit vendu, est une colonne critique. Les lignes contenant des valeurs *"ERROR"*, *"UNKNOWN"* ou des valeurs manquantes ont donc été supprimées, car il était impossible de déterminer le produit réellement vendu. Conserver ces lignes aurait faussé toute analyse par produit.

De la même manière, la colonne **Quantity** contenait certaines valeurs non numériques ou manquantes. Ces lignes ont été supprimées, la quantité étant indispensable au calcul des montants financiers et à la cohérence des données.

Colonnes financières

La colonne **Total Spent** contenait des valeurs incorrectes ou incohérentes. Afin de garantir la fiabilité des données financières, le montant total a été recalculé à partir de la quantité et du prix unitaire pour chaque transaction.

Concernant la colonne **Price Per Unit**, plusieurs valeurs aberrantes ont été identifiées, notamment des prix unitaires irréalistes ne correspondant pas au contexte d'un café. En analysant la relation entre la quantité achetée et le montant total payé, il a été possible d'identifier des prix unitaires cohérents. Par exemple, pour le produit *Tea*, le prix unitaire réel a été déterminé à partir de transactions indiquant un total de 3 pour 2 articles, soit un prix unitaire de 1.5. Les valeurs aberrantes ont ainsi été corrigées afin d'obtenir des prix réalistes.

Colonnes contextuelles

Les colonnes **Payment Method** et **Location** contenaient également des valeurs *"ERROR"* et *"UNKNOWN"*. Ces colonnes n'étant pas critiques pour le calcul des ventes, les lignes concernées n'ont pas été supprimées. Les valeurs non exploitables ont été remplacées par des valeurs manquantes, ce qui permet de conserver un maximum de transactions tout en nettoyant l'information.

Enfin, la colonne **Transaction Date** a été nettoyée en supprimant les lignes contenant des valeurs non valides ou manquantes, la date étant une information importante pour l'analyse temporelle des ventes.

Vérification de la qualité finale

À l'issue des étapes de nettoyage et de transformation, une vérification globale de la qualité des données a été réalisée. Le dataset ne contient plus de valeurs non exploitables telles que *"ERROR"* ou *"UNKNOWN"* dans les colonnes critiques. Les montants financiers sont désormais cohérents avec les quantités et les prix unitaires, et les valeurs aberrantes ont été corrigées.

Le dataset final est ainsi propre, cohérent et exploitable pour une analyse fiable.

Conclusion

Ce projet de *Data Refinement* a permis de mettre en pratique les notions de qualité des données et de transformation abordées durant le module. Il met en évidence l'importance d'un travail rigoureux sur les données brutes avant toute analyse ou prise de décision. Grâce aux différentes étapes d'exploration, de nettoyage et de transformation, le dataset initialement brut a pu être transformé en un jeu de données fiable et exploitable, prêt à être utilisé dans un contexte d'analyse ou de reporting.