UNIVERSITÉ D'ÉTAT D'HAITI

(UEH)

FACULTÉ DES SCIENCES(FDS)

Formation pour le Renforcement du Secteur Technologique (FRST)

Projet final : Introduction au language de programmation Python

Auteur : Lovenson JEUDINOR

Titre : ***Analyse de données de l'entreprise EduMart***

Date : 10 Janvier 2025

Plan du Rapport : Analyse de Données EduMart

1. Introduction

- **Contexte :** Présentation d'EduMart et du rôle de Data Analyst.
- **Objectifs :** Finalités de l'analyse (nettoyage, étude des ventes et satisfaction).
- **Périmètre :** Description des trois jeux de données sources (`customers`, `products`, `order_lines`).

2. Étape 1 : Audit et Compréhension des Données

- **Méthodologie d'inspection :** Utilisation des outils Pandas (`info`, `describe`, `head`).
- **Diagnostic des colonnes à problèmes :** Identification des erreurs de typage et des données manquantes.
- **Livrable :** Rapport d'audit sur l'intégrité initiale du patrimoine de données.

3. Étape 2 : Nettoyage et Préparation des Données

- **Normalisation et imputation :** Correction des types (dates, numérique) et gestion des valeurs manquantes (moyenne vs médiane).
- **Traitement des anomalies métiers :** Suppression des quantités négatives et des délais de livraison incohérents.
- **Livrable :** Documentation des 5 règles de nettoyage appliquées pour garantir la fiabilité.

4. Étape 3 : Analyse des Indicateurs Clés de Performance (ICP)

- **Performance financière :** Analyse du Chiffre d'Affaires net, du panier moyen (AOV) et des remises.
- **Analyse des retours :** Étude des taux de retour par catégorie de produit et par canal de vente (Web vs App vs Boutique).
- **Satisfaction client :** Analyse des scores d'avis et corrélation avec les délais de livraison.

5. Étape 4 : Jointures et Enrichissement des Ventes

- **Méthodologie de consolidation :** Processus de fusion des tables (`left join`) et vérification de l'intégrité des clés.

- **Validation et recalcul :** Contrôle de cohérence financière entre les montants sources et calculés.
- **Analyse croisée (Pivot) :** Étude du chiffre d'affaires par segment de clientèle et par catégorie de produit.

6. Conclusion Générale

- **Synthèse technique :** Bilan sur la qualité finale des données.
- **Enseignements stratégiques :** Identification des segments moteurs et des points d'optimisation (retours web, logistique).
- **Recommandations :** Actions suggérées pour améliorer la performance d'EduMart

1. Introduction

Contexte

Dans le cadre de ce projet, j'exerce les fonctions de **Data Analyst pour EduMart**, une enseigne spécialisée dans la distribution multi-canal (boutique physique, site web et application mobile). EduMart propose un catalogue varié de produits technologiques et éducatifs, incluant des cours, des livres, des logiciels, des crédits cloud, des ordinateurs portables et divers accessoires. Ce rôle consiste à transformer des données brutes en informations exploitables pour soutenir les décisions stratégiques de l'entreprise.

Objectif

L'objectif principal de cette mission est de mettre en place un pipeline complet de traitement de données sous Python, en utilisant la bibliothèque Pandas. Il s'agit de :

- Réaliser un audit approfondi pour évaluer la qualité des données sources.
- Nettoyer les jeux de données afin de corriger les anomalies de types, de valeurs manquantes et d'incohérences métiers.
- Analyser les performances de vente et le comportement des clients, notamment à travers l'étude des retours produits et des scores d'avis.
- Calculer des indicateurs clés de performance (ICP) pour mesurer l'efficacité commerciale et la satisfaction client.
- Jointure et vérfication des clés de jointures.

Périmètre

L'analyse s'appuie sur l'exploitation de trois jeux de données fondamentaux fournis au format CSV:

1. **`customers.csv`** : Contient les informations démographiques et les profils des clients (âge, genre, segment, ville, date d'inscription).
2. **`products.csv`** : Répertorie le catalogue des produits, leurs catégories, marques et prix unitaires.
3. **`order_lines.csv`** : Constitue la base transactionnelle détaillant chaque vente, les quantités, les remises appliquées, ainsi que les données logistiques (délais de livraison) et de satisfaction (retours, avis).

2. Étape 1 : Charger et comprendre

Méthodologie d'inspection

Pour débuter cette analyse, la première phase a consisté à importer les trois fichiers sources, à savoir `customers.csv`, `products.csv` et `order_lines.csv`, à l'aide de la bibliothèque Pandas. Afin de garantir une compréhension exhaustive de la structure des données avant tout traitement, j'ai appliqué systématiquement les fonctions d'inspection standards préconisées. La méthode `info()` a permis de dresser un premier bilan technique sur la complétude des colonnes et la nature des types de données détectés automatiquement par l'interpréteur. En complément, la fonction `describe(include="all")` a servi à observer la distribution statistique des variables numériques ainsi que la diversité des variables catégorielles. Enfin, l'affichage des premières lignes via `head()` a permis de valider visuellement le bon formatage du texte et la cohérence du chargement initial.

Diagnostic des colonnes à problèmes

L'audit technique approfondi a révélé plusieurs points de vigilance critiques qui pourraient fausser les résultats si des mesures correctives ne sont pas appliquées. Le fichier des clients présente des types de données inappropriés, notamment pour la variable `age` qui est interprétée comme du texte en raison de la présence de mentions "unknown", ainsi que pour la date d'inscription stockée en format objet. On note également des absences de données pour le genre et la ville de résidence. Concernant le fichier des transactions `order_lines`, la colonne `order_date` nécessite impérativement une conversion au format temporel pour autoriser les calculs de saisonnalité. Ce dataset transactionnel contient aussi des valeurs manquantes pour les scores d'avis et les délais de livraison, ainsi que des anomalies métiers majeures telles que des quantités négatives et des délais d'expédition irréalistes dépassant les trente jours. À l'inverse, le fichier des produits apparaît comme le plus intègre, ne présentant aucune valeur manquante ni incohérence de type majeure.

Livrable : Rapport d'audit de qualité

« Le patrimoine de données d'EduMart présente une structure solide mais nécessite un nettoyage ciblé pour être exploitable. Le fichier `products` est intègre, tandis que `customers` souffre de problèmes de typage (âge et dates) et de données démographiques manquantes. Le fichier transactionnel `order_lines` est le plus critique : il contient des incohérences de types de dates et des anomalies métiers (quantités négatives, délais de livraison aberrants) qui fausseraient les calculs de performance s'ils n'étaient pas corrigés. Une phase de normalisation des types et de filtrage des valeurs extrêmes est donc indispensable avant toute analyse statistique. »

3. Étape 2 : Nettoyage et préparation des données

Normalisation des types et gestion des valeurs manquantes

La première phase du nettoyage a consisté à transformer les données brutes en formats exploitables pour l'analyse statistique. La variable `age` a été convertie en type numérique après avoir traité les mentions textuelles, et les colonnes de dates ont été normalisées au format datetime. Pour traiter les valeurs manquantes, une approche différenciée a été adoptée en fonction de la distribution des données. L'analyse du coefficient d'asymétrie a révélé que la variable `delivery_days` présentait une forte asymétrie positive (2,79), justifiant l'utilisation de la médiane pour l'imputation afin de ne pas biaiser les résultats par des valeurs extrêmes. À l'inverse, le score d'avis (`review_score`) affichant une asymétrie faible (0,33), la moyenne a été privilégiée. Pour les variables catégorielles comme la ville ou le genre, les entrées manquantes ont été systématiquement remplacées par la mention « Unknown » afin de conserver l'intégrité du volume transactionnel tout en signalant l'absence d'information.

Traitement des anomalies métiers et validation de la cohérence

Au-delà de la forme, un filtrage rigoureux a été appliqué pour garantir la fiabilité métier des indicateurs futurs. L'analyse a permis d'identifier et de supprimer six observations présentant des quantités nulles ou négatives, ainsi que six autres lignes affichant des délais de livraison supérieurs à trente jours, jugés irréalistes pour le modèle opérationnel d'EduMart. Concernant les doublons, une vérification sur les clés métiers (identifiants clients et commandes) a été effectuée pour éviter tout double comptage du chiffre d'affaires. Enfin, un test de cohérence financière a été opéré en comparant le montant net enregistré avec un montant net recalculé à partir du prix unitaire, de la quantité et du taux de remise. Ce contrôle a validé la fiabilité de la base transactionnelle, aucune ligne ne présentant d'écart supérieur au seuil de tolérance de 0,01 lié aux arrondis de calcul.

Bilan du processus de nettoyage

Le processus s'est conclu par l'exportation des données vers des fichiers dits « clean », prêts pour l'agrégation. Au total, l'application des règles métiers a conduit à la suppression de 12 lignes jugées aberrantes. Le jeu de données final des ventes comprend désormais 2 188 observations et s'est enrichi de nouvelles colonnes calculées, offrant une base saine et robuste. Ce passage d'une structure de 2 225 lignes initiales à un dataset nettoyé garantit que les analyses de performance et les calculs de taux de retour ne seront pas pollués par des erreurs de saisie ou des incohérences de stockage.

Livrable : Les 5 règles de nettoyage documentées

Règle 1 — Normalisation des types : Conversion systématique des dates au format datetime et des variables numériques stockées en texte (âge, remises) pour permettre les calculs.

Règle 2 — Imputation statistique : Remplacement des valeurs manquantes par la médiane pour les variables asymétriques (délais de livraison) et par la moyenne pour les variables symétriques (avis), tandis que les variables catégorielles sont marquées comme "Unknown".

Règle 3 — Élimination des doublons : Identification et suppression des enregistrements redondants basés sur les clés primaires `customer_id` et `order_id` pour garantir l'unicité des transactions.

Règle 4 — Filtrage des anomalies métiers : Exclusion des lignes présentant des quantités non positives ou des délais d'expédition aberrants excédant les standards logistiques de l'entreprise (30 jours).

Règle 5 — Vérification de l'intégrité financière : Validation de l'exactitude des montants nets par recalcul arithmétique, assurant que les données de facturation sont exemptes d'erreurs de calcul internes.

4. Étape 3 : Analyse des Indicateurs Clés de Performance (ICP)

Performance commerciale et rentabilité globale

L'analyse de la performance commerciale d'EduMart révèle une activité solide avec un chiffre d'affaires net total s'élevant à 2 023 658,44 unités monétaires. Cette performance est portée par un panier moyen global (AOV) particulièrement élevé de 924,89, un chiffre qui s'explique par la présence de produits à forte valeur unitaire comme les ordinateurs portables dans le mix produit. En matière de stratégie promotionnelle, l'entreprise applique un taux de remise moyen de 11,66 % en valeur. Ce niveau de remise semble maîtrisé, permettant de stimuler la conversion sans dégrader excessivement la marge brute, tout en restant dans les standards de la distribution de produits technologiques et éducatifs.

Analyse des retours et performance par canal

La gestion des retours constitue un indicateur critique pour la satisfaction client et l'efficacité logistique. Le taux de retour global s'établit à 4,30 % en volume et 3,77 % en valeur, ce qui témoigne d'une bonne adéquation entre les attentes clients et les produits livrés. Toutefois, une analyse plus fine par catégorie montre que les « Accessoires » et les « Livres » affichent les taux de retour les plus élevés, respectivement 6 % et 5 %. Concernant les canaux de distribution, le canal Web présente un taux de retour de 5 %, soit deux points de plus que les applications mobiles ou les boutiques physiques (3 %). Cette disparité suggère que l'expérience d'achat sur le web

pourrait bénéficier d'une meilleure description des produits ou de visuels plus précis pour réduire les erreurs de choix des clients.

Satisfaction client et impact de la logistique

La satisfaction globale des clients d'EduMart est positive avec un score d'avis moyen de 4,36 sur 5. Par catégorie, les « Accessoires » obtiennent la meilleure note (4,42), tandis que les services « Cloud » ferment la marche avec un score de 4,30, restant néanmoins à un niveau très satisfaisant. Un point particulièrement intéressant concerne la corrélation entre la satisfaction et les délais de livraison. Contrairement aux attentes habituelles, la note moyenne reste remarquablement stable, oscillant entre 4,33 et 4,39, même lorsque les délais dépassent les dix jours. Cette résilience du score de satisfaction, malgré des délais de livraison plus longs, pourrait indiquer que les clients d'EduMart valorisent davantage la qualité du produit reçu ou la fiabilité de la livraison que sa simple rapidité.

4. Étape 3 : Analyse approfondie des Indicateurs Clés de Performance (ICP)

Analyse de la performance financière et des leviers promotionnels

L'activité globale d'EduMart sur la période affiche une performance solide avec un chiffre d’affaires net total de 2 023 658,44. Cette réussite repose sur un panier moyen (AOV) particulièrement élevé de 924,89, un indicateur qui témoigne de la capacité de l'enseigne à vendre des produits à forte valeur ajoutée, comme les ordinateurs portables. Un point crucial de cette analyse réside dans la stratégie de remise : le taux de remise moyen global s'établit à 11,66 % en valeur. Ce chiffre indique que pour chaque centaine d'unités monétaires vendues, EduMart concède environ 11,6 unités en promotions. Ce niveau est stratégiquement équilibré : il est assez incitatif pour stimuler les volumes de ventes sans pour autant éroder de manière critique la rentabilité nette de l'entreprise.

Analyse sectorielle et logistique des retours clients

L'examen détaillé du taux de retour révèle des disparités significatives selon les segments et les canaux. Globalement, les retours sont maîtrisés à 4,30 % en volume, mais l'analyse par catégorie montre que les « Accessoires » et les « Livres » subissent les taux de retour les plus élevés (respectivement 6 % et 5 %). Cela suggère un potentiel problème de conformité ou de déception à la réception pour ces articles. À l'opposé, les produits intangibles comme le « Cloud » ou les « Logiciels » présentent les taux les plus bas (3 %), ce qui est cohérent avec leur nature dématérialisée. Concernant les canaux de distribution, le canal « Web » affiche une fragilité relative avec un taux de retour de 5 %, contre seulement 3 % pour la « Boutique » et l’« App ». Ce décalage souligne la nécessité d'améliorer l'expérience de visualisation sur le site internet pour réduire l'écart entre la perception client et le produit réel.

Satisfaction client et étude de la perception logistique

La satisfaction client est un indicateur de santé majeur pour EduMart, avec un score d’avis moyen global de 4,36 sur 5. Par catégorie, les « Accessoires » dominent avec une note d'excellence de 4,42, suivis de près par les « Cours » (4,38), tandis que le « Cloud » ferme la marche avec 4,30. L'un des enseignements les plus riches de cette étape concerne l'impact des délais de livraison sur la satisfaction perçue. Le tableau croisé entre les délais et les scores d'avis montre une stabilité remarquable : alors que l'on pourrait s'attendre à une chute de la satisfaction pour les livraisons tardives, le score d'avis reste quasiment identique entre une livraison express de 0 à 2 jours (4,34) et une livraison dépassant les 10 jours (4,33). Cette observation suggère que la clientèle d'EduMart privilégie la fiabilité de la livraison et la qualité intrinsèque du produit plutôt que la seule rapidité d'exécution.

5. Étape 4 : Jointures et enrichissement des ventes

Méthodologie et intégrité de la consolidation

L’aboutissement de ce projet a consisté à fusionner les trois sources de données nettoyées pour créer une table d’analyse unique et multidimensionnelle. Pour garantir qu’aucune transaction ne soit égarée lors de cette manipulation, j'ai opéré une jointure à gauche (`left join`) en prenant le fichier des ventes `order_lines` comme base pivot. Les contrôles d'intégrité révèlent une parfaite cohérence du système d'information : les 492 clients ayant effectué des achats ont tous été retrouvés dans le référentiel `customers` (qui compte 500 individus, indiquant que 8 clients n'ont pas encore commandé). De même, les 60 produits présents dans les transactions correspondent exactement au catalogue `products`. Cette absence de lignes "orphelines" confirme la robustesse des clés de jointure et permet d'attribuer avec certitude chaque euro de chiffre d'affaires à un profil client et à une catégorie de produit spécifique.

Recalcul métier et validation de la table enrichie

Une fois la table finale constituée — passant de 20 à 25 colonnes grâce à l'apport des données démographiques et techniques — une étape de vérification arithmétique a été menée. En croisant le prix unitaire issu du catalogue produit avec les quantités et les taux de remise des ventes, j'ai généré les indicateurs calculés `gross_amount_calc` et `net_amount_calc`. La comparaison systématique avec les montants originaux n'a révélé aucune ligne suspecte dépassant le seuil de tolérance des arrondis (0,01). Ce processus d'enrichissement ne se limite pas à une simple fusion ; il sécurise la donnée financière en s'assurant que les prix appliqués lors de la vente sont conformes au référentiel tarifaire de l'entreprise.

Analyse multidimensionnelle et livrable final

Le résultat final de cette consolidation est la production d'un tableau pivot croisant le chiffre d'affaires net par segment de clientèle et par catégorie de produit. Cette vue d'ensemble met en

lumière la concentration stratégique de l'activité d'EduMart : la catégorie « Laptop » domine largement les revenus, portée principalement par les segments des « Étudiants » et des « Professionnels ». Le chiffre d'affaires total consolidé dans cette table finale (2 023 658,44) concorde parfaitement avec les analyses de l'étape précédente. Ce jeu de données final, exporté sous le nom `orders_enriched.csv`, constitue désormais une base de décision fiable, prête à être utilisée pour des rapports de gestion ou des modèles de prédiction du comportement d'achat.

Livrable : 5 lignes de commentaires finaux

La jointure intégrale des données confirme l'absence de perte de ventes et une parfaite correspondance entre les référentiels clients/produits et les transactions. L'analyse enrichie révèle que le segment « Étudiant » est le premier moteur de croissance, particulièrement sur le matériel informatique de haute valeur. Les contrôles de cohérence financière valident l'exactitude des calculs de remises et de montants nets sur l'ensemble du dataset. Aucune anomalie résiduelle n'a été détectée après fusion, garantissant une base de reporting 100 % fiable. Cette table enrichie permet désormais de piloter la stratégie commerciale par segment avec une précision granulaire.

6. Conclusion Générale

Ce projet d'analyse de données pour **EduMart** a permis de transformer un patrimoine de données brutes hétérogènes en un outil d'aide à la décision fiable et structuré. Le passage par les étapes critiques d'audit, de nettoyage et d'enrichissement a été essentiel pour garantir la véracité des indicateurs produits.

Synthèse de la qualité des données

L'audit initial avait révélé des fragilités, notamment sur le typage des variables et des anomalies métiers (quantités négatives, délais de livraison incohérents). Le processus de nettoyage a permis d'écarter 12 lignes aberrantes et d'imputer les valeurs manquantes de manière statistique, assurant une intégrité de 100 % lors de la fusion finale des fichiers. Les contrôles de cohérence financière confirment que le jeu de données enrichi est exempt d'erreurs de calcul, offrant une base saine pour les futurs rapports de gestion.

Enseignements clés et performance commerciale

L'analyse des indicateurs de performance met en lumière plusieurs points stratégiques :

- **Domination du segment matériel** : La catégorie **Laptop** génère la majorité du chiffre d'affaires (1,4M sur 2,02M), portée par une clientèle d'**Étudiants** et de **Professionnels**.
- **Satisfaction et Logistique** : Malgré des délais de livraison parfois supérieurs à 10 jours, la satisfaction client reste stable à un niveau élevé (**4,36/5**), indiquant une forte loyauté envers la marque et une tolérance des clients vis-à-vis des délais si la qualité est au rendez-vous.
- **Optimisation des canaux** : Le canal **Web** présente un taux de retour plus élevé (5 %) que l'application ou la boutique physique. Une amélioration des fiches produits sur le site internet pourrait réduire ce coût logistique.

Recommandations finales

En conclusion, EduMart dispose d'un modèle économique solide avec un panier moyen robuste (924,89). Pour la suite, il est recommandé de capitaliser sur le segment **Étudiant** via des offres groupées (Laptops + Cours/Logiciels) et de surveiller de près la catégorie **Accessoires**, qui, bien que satisfaisante en termes de notation, subit le taux de retour le plus élevé de l'enseigne.

La table finale `orders_enriched.csv` est désormais prête à être intégrée dans un outil de visualisation (type Power BI ou Tableau) ou à servir de base pour un modèle de prédiction de la demande.