Projet de Nettoyage et d'Analyse de Données

Flux de Travail Complet de Nettoyage et d'Analyse de Données

Utilisez `morocco_ecommerce_anomalies.xlsx` pour vous entraîner à un pipeline complet d'analyste de données : de l'inspection et du nettoyage jusqu'à l'analyse groupée et au reporting.

Jeu de données e-commerce réaliste · Flux de travail Pandas + Python · Nettoyage, EDA, valeurs aberrantes · Analyse orientée métier

Pipeline de Workflow d'Analyste de Données

APERÇU

Une feuille de route pratique pour nettoyer et analyser des données réelles, en particulier des jeux de données e-commerce comme celui-ci.

1. Comprendre le Contexte des Données

Avant d'écrire la moindre ligne de code, clarifiez ce que représentent les données et quelles décisions elles soutiendront.

- De quoi parle ce jeu de données ?
- Que représente chaque colonne ?
- Quelles questions métier voulons-nous résoudre ?
- Qui utilisera les résultats et comment ?

CONSEILS & ASTUCES

Sans contexte, vous ne pouvez pas décider comment nettoyer ou interpréter les anomalies. La même valeur peut être acceptable dans une entreprise et erronée dans une autre. Reliez toujours votre travail sur les données à sa signification réelle.

2. Charger les Données & Inspecter Leur Structure

Comprendre rapidement comment le jeu de données est structuré avant d'y apporter des modifications.

- Chargez le fichier dans un DataFrame.
- Affichez les premières lignes.
- Vérifiez la forme (lignes, colonnes).
- Inspectez les types de données et les statistiques de base.

CONSEILS & ASTUCES

La plupart des problèmes structurels sont visibles dans les 10–20 premières lignes. Utilisez `df.head()`, `df.info()` et `df.describe()` pour repérer rapidement les types étranges ou des extrêmes suspects.

3. Évaluer la Qualité des Données (Phase de Profiling)

Identifier ce qui est incorrect ou risqué dans le jeu de données avant de corriger quoi que ce soit.

Recherchez :

- Des valeurs manquantes
- Des types de données incorrects
- Des doublons
- Des valeurs aberrantes
- Des catégories incohérentes
- Des valeurs impossibles (quantités négatives, dates invalides, etc.)

CONSEILS & ASTUCES

C'est la phase de diagnostic. Résistez à l'envie de nettoyer immédiatement. Construisez d'abord une vision claire des problèmes des données ; ensuite seulement décidez comment les corriger.

4. Définir une Stratégie de Nettoyage

Planifier votre approche avant d'appliquer des transformations.

Décidez quoi faire pour :

- Les valeurs manquantes (remplir, supprimer, marquer comme inconnues)
- Les entrées texte incorrectes (comme `"free"`, `"two"`, espaces finaux)
- Les dates invalides

- Les types de données mélangés
- Les doublons
- Les valeurs aberrantes

CONSEILS & ASTUCES

Différents analystes peuvent choisir des stratégies de nettoyage légèrement différentes et avoir malgré tout raison s'ils peuvent justifier leurs décisions. Les données réelles ont rarement une seule solution parfaite.

5. Nettoyer les Données Étape par Étape

Appliquer votre plan de nettoyage dans un ordre logique et reproductible.

Un ordre recommandé :

- **5.1 Corriger les types de données :** convertir les chaînes en valeurs numériques et en dates lorsque c'est approprié.
- **5.2 Nettoyer les données catégorielles/texte :** supprimer les espaces, unifier la casse, remplacer les variantes.
- **5.3 Gérer les valeurs manquantes :** remplir, supprimer ou étiqueter selon l'impact.
- **5.4 Supprimer ou corriger les doublons :** s'assurer qu'aucune commande n'est comptée deux fois.
- **5.5 Détecter et traiter les valeurs aberrantes :** à l'aide de l'IQR, du z-score ou de règles métier.

CONSEILS & ASTUCES

Le nettoyage représente souvent 70 % du travail d'un analyste de données. Rendez vos étapes traçables pour que d'autres puissent reproduire et comprendre ce que vous avez fait.

6. Feature Engineering

Ajouter des colonnes dérivées utiles qui simplifient les analyses ultérieures.

Exemples :

- Variables de date (année, mois, jour de la semaine)
- Champs liés au revenu (prix remisé, taxe, montant net)
- Ratios (prix unitaire, taux de remise)
- Catégories regroupées (par exemple fusionner des catégories rares)

CONSEILS & ASTUCES

De bons features rendent les motifs plus faciles à voir et à expliquer. Réfléchissez aux métriques qui comptent le plus pour l'entreprise.

7. Analyse Exploratoire des Données (EDA)

Explorer les motifs et répondre aux questions métier de haut niveau.

Posez des questions telles que :

- Quelle région vend le plus ?
- Quelle catégorie de produits performe le mieux ?
- Dans quelles plages se situent les quantités et les montants ?
- Existe-t-il des corrélations entre les variables ?

CONSEILS & ASTUCES

Utilisez `groupby`, les agrégations et les visualisations. L'EDA est le moment où vous transformez des données propres en insights et hypothèses.

8. Analyse de Séries Temporelles (Si Pertinent)

Utiliser les informations de date pour analyser les tendances dans le temps.

- Reéchantillonnez le revenu par mois ou par semaine.
- Recherchez de la saisonnalité ou des tendances.
- Comparez les périodes avant et après des événements ou campagnes clés.

CONSEILS & ASTUCES

Le temps peut révéler des comportements que les lignes brutes ne montrent pas. Les sauts soudains peuvent indiquer des promotions, des erreurs ou des événements externes.

9. Tout Documenter

Rendre votre travail compréhensible et reproductible.

- Documentez les problèmes de données que vous avez trouvés.
- Enregistrez les règles de nettoyage et les hypothèses.
- Résumez la qualité finale des données.

CONSEILS & ASTUCES

Les futurs analystes (y compris vous) devraient pouvoir lire votre notebook et comprendre non seulement ce que vous avez fait, mais aussi pourquoi vous l'avez fait.

10. Créer un Rapport Final ou un Tableau de Bord

Communiquer vos résultats clairement à des parties prenantes non techniques.

- Résumez les métriques et tendances clés.
- Mettez en avant les risques et limitations.
- Formulez des recommandations claires.

CONSEILS & ASTUCES

Le travail d'un analyste de données se termine par une décision ou une histoire, pas seulement par du code. Utilisez des visuels et un langage simple pour expliquer ce que disent les données.

CE PIPELINE S'APPLIQUE À DE NOMBREUX CAS

Les choix précis de nettoyage peuvent varier d'un jeu de données à l'autre, mais le flux de travail global reste le même. Vous suivrez une variante de ce schéma dans presque tous les projets de données du monde réel.

Note Importante Avant de Commencer

À LIRE D'ABORD

Conseils sur la manière de réfléchir aux différentes solutions valides et aux choix de l'analyste.

LES SOLUTIONS PEUVENT LÉGÈREMENT DIFFÉRER

Les solutions et étapes de nettoyage que vous appliquerez peuvent légèrement différer en fonction :

- des anomalies spécifiques que vous détectez,
- de vos hypothèses métier,
- de la stratégie de nettoyage que vous choisissez,
- ou de votre interprétation de ce qui doit être corrigé, rempli ou supprimé.

C'est normal et attendu. Les données réelles ont rarement une seule solution "parfaite".

Ce qui compte, c'est d'appliquer les bons principes, d'être cohérent et de justifier vos décisions. Les exercices ci-dessous enseignent les étapes fondamentales que chaque analyste de données suit. Même si deux analystes font des choix de nettoyage différents, le flux de travail global reste le même.

AVANT DE GÉNÉRER DU CODE — PENSEZ COMME UN ANALYSTE DE DONNÉES

Le travail d'un analyste de données ne commence pas par le code. Le vrai travail commence avant de toucher à Python, SQL ou Pandas.

- **Comprendre le problème :** Quelle question métier essayons-nous de résoudre ?
- **Regarder les données en premier :** Explorez visuellement avant de nettoyer.
- **S'attendre à des compromis :** Aucun jeu de données n'est parfait.
- **Nettoyer n'est pas supprimer :** L'analyste doit décider quoi corriger, conserver, retirer ou signaler.
- **Documenter chaque décision :** Un autre analyste doit pouvoir suivre votre logique.
- **Vous racontez une histoire avec les données :** Insights > code.
- **Réfléchir avant de coder :** La planification évite les erreurs et fait gagner du temps.

Un excellent analyste se distingue par son raisonnement — pas par la vitesse à laquelle il tape du code.

🎯 Jeu de Données E-Commerce Maroc — Exercices Complets de Nettoyage & Analyse

EXERCICES

Utilisez `morocco_ecommerce_anomalies.xlsx` pour réaliser un projet de bout en bout avec des problèmes de données réalistes.

SECTION ①

Inspection Initiale des Données

Objectif : Comprendre ce avec quoi vous travaillez avant de modifier quoi que ce soit.

1. Chargez le jeu de données et affichez les 15 premières lignes.
2. Indiquez le nombre de lignes et de colonnes.
3. Listez toutes les colonnes avec leurs types de données.
4. Quelles colonnes semblent avoir un type de données incorrect (par ex. valeur numérique stockée en chaîne) ?
5. Générez un résumé avec `df.describe(include="all")` et commentez tout élément suspect.

CONSEILS & ASTUCES

- `df.info()` révèle instantanément la plupart des problèmes structurels.
- Les chaînes stockées comme nombres sont un drapeau rouge pour l'analyse.
- Des min/max inhabituels ou un écart type extrêmement élevé indiquent souvent des données corrompues ou des valeurs aberrantes.

SECTION ②

Analyse des Valeurs Manquantes

Objectif : Détecter les "trous" dans le jeu de données et décider comment ils affectent l'analyse.

6. Comptez les valeurs manquantes dans chaque colonne.
7. Quelles sont les 3 colonnes qui contiennent le plus de valeurs manquantes ?
8. Calculez le pourcentage de valeurs manquantes par colonne.
9. En vous basant sur la logique métier, classez les colonnes en :
   - Champs critiques (ne doivent pas être manquants)
   - Champs semi-critiques
   - Champs optionnels
10. Proposez un plan pour gérer les valeurs manquantes pour chaque catégorie.

CONSEILS & ASTUCES

- Critiques : `order_id`, `order_date`, `total_amount`, `quantity`, `unit_price`.
- Optionnels : codes de réduction, commentaires, champs d'informations supplémentaires.
- Décidez quand remplir, supprimer ou marquer les valeurs manquantes selon l'impact et la logique métier.

SECTION ③

Types de Données Incorrects & Nettoyage des Valeurs

Objectif : Corriger les champs numériques corrompus par des chaînes et du texte.

11. Explorez toutes les valeurs uniques de `quantity` et identifiez toutes les entrées non numériques.
12. Explorez toutes les valeurs uniques de `unit_price` et identifiez les entrées non numériques.
13. Convertissez ces deux colonnes en numérique à l'aide de règles personnalisées (par ex. `"free"` → 0, `"two"` → 2).
14. Indiquez combien de lignes n'ont pas pu être converties, s'il y en a.
15. Recalculez les statistiques descriptives pour ces colonnes numériques nettoyées.

CONSEILS & ASTUCES

- Utilisez `.str.strip()` pour corriger les problèmes d'espaces cachés.
- `pd.to_numeric(..., errors="coerce")` convertit les valeurs invalides en NaN, ce qui facilite leur traitement.
- Gérez les NaN numériques après la conversion, et non avant, pour savoir quelles entrées ont réellement échoué.

SECTION ④

Incohérences Catégorielles & Nettoyage du Texte

Objectif : Nettoyer les champs texte pour des groupements et filtrages fiables.

16. Listez toutes les valeurs uniques de `city`.
17. Identifiez les incohérences (espaces, casse, orthographes différentes).
18. Standardisez les noms de ville avec `str.strip()` et `str.title()`.
19. Remplacez les variantes de Casablanca (par ex. `"Casa"`, `"casa "`) par `"Casablanca"`.
20. Identifiez les incohérences de `region` et standardisez-les également.

CONSEILS & ASTUCES

- `"casa "`, `"Casa"`, `"CASA"` désignent logiquement la même ville mais sont des catégories différentes.
- Nettoyez, puis inspectez, puis consolidez les catégories pour des analyses groupby fiables.
- Le fuzzy matching peut aider pour des variantes complexes dans le monde réel, mais de simples remplacements couvrent souvent la majorité des cas.

SECTION ⑤

Nettoyage & Validation des Dates

Objectif : Valider et convertir les champs de date en objets datetime corrects.

21. Identifiez les entrées invalides de `order_date`.
22. Convertissez `order_date` et `ship_date` en datetime avec `errors="coerce"`.
23. Combien de dates invalides avez-vous trouvées dans chaque colonne ?
24. Décidez de supprimer ou non les lignes avec des dates invalides et justifiez votre choix.
25. Créez de nouvelles colonnes : `year`, `month` et `weekday` à partir de `order_date`.

CONSEILS & ASTUCES

- `NaT` est l'équivalent datetime de NaN. Traitez-le de manière similaire pour l'analyse de valeurs manquantes.
- Ne remplacez pas les dates manquantes/invalides par des moyennes ; les dates ne sont pas des quantités numériques.
- Le tri par date révèle souvent des lacunes, des pics étranges ou des intervalles invalides.

SECTION ⑥

Détection des Doublons

Objectif : Éviter les transactions comptées deux fois et les historiques de commande incohérents.

26. Comptez les lignes totalement dupliquées (doublons exacts sur toutes les colonnes).
27. Supprimez tous les doublons complets et indiquez le nouveau nombre de lignes.
28. Vérifiez les valeurs dupliquées de `order_id`.
29. Supprimez les commandes en double en conservant uniquement la première occurrence de chaque `order_id`.
30. Indiquez le nombre final de lignes et le total de doublons supprimés.

CONSEILS & ASTUCES

- Les doublons complets proviennent souvent de problèmes d'export ou de fusion.
- Au niveau de la commande, vous devez décider ce que "doublon" signifie en termes métier.
- Documentez si vous gardez la première, la dernière ou une version combinée des commandes dupliquées.

SECTION ⑦

Détection & Traitement des Valeurs Aberrantes

Objectif : Identifier les valeurs extrêmes qui peuvent fausser vos métriques.

31. Utilisez la méthode IQR pour détecter les valeurs aberrantes dans `total_amount`.
32. Comptez combien de valeurs aberrantes (IQR) vous trouvez.
33. Utilisez la méthode du Z-score pour les valeurs aberrantes de `total_amount` et comparez les résultats avec l'IQR.
34. Inspectez les 5 commandes avec les valeurs de `total_amount` les plus élevées.
35. Recommandez une stratégie de traitement des outliers (supprimer, plafonner, signaler ou conserver) et expliquez pourquoi.

CONSEILS & ASTUCES

- Les valeurs aberrantes peuvent représenter de vraies grosses commandes, des erreurs humaines, des bugs systèmes ou de la fraude.
- Supprimer des outliers sans réfléchir peut masquer un comportement métier important.
- Parfois, conserver les outliers mais les signaler est le meilleur compromis.

SECTION ⑧

Filtrage & Sous-Ensembles (Questions Métier)

Objectif : Extraire des insights et répondre à des questions métier concrètes à partir des données nettoyées.

36. Sélectionnez toutes les commandes avec `quantity > 3`.
37. Sélectionnez toutes les commandes avec `total_amount > 1000`.
38. Filtrez toutes les commandes de la région `"Casablanca-Settat"`.
39. Sélectionnez toutes les commandes où `payment_method` n'est pas `"Cash on Delivery"`.
40. Sélectionnez toutes les commandes passées après la date `"2023-06-01"`.

CONSEILS & ASTUCES

- Nettoyez les catégories avant de filtrer pour éviter de manquer des valeurs à cause de l'orthographe ou des espaces.
- Utilisez des parenthèses lors de la combinaison de plusieurs conditions avec `&` et `|`.
- Attention à la sensibilité à la casse lors du filtrage des valeurs texte.

SECTION ⑨

Analyses Groupées & KPIs

Objectif : Identifier les motifs et performances selon les régions, catégories et produits.

41. Calculez la moyenne, la médiane, le minimum et le maximum de `total_amount`.
42. Regroupez par `region` et calculez :
    - le revenu total,
    - la valeur moyenne de commande,
    - le nombre de commandes.
43. Regroupez par `product_category` et calculez :
    - le revenu total,
    - la valeur moyenne de commande,
    - le nombre de commandes.
44. Identifiez les 5 `product_id` avec le plus grand revenu total.

45. Déterminez quelle région a la valeur moyenne de commande la plus élevée.

CONSEILS & ASTUCES

- Utilisez `.agg()` avec plusieurs fonctions pour calculer plusieurs KPIs en une seule étape.
- Une région avec peu de commandes très élevées peut avoir une valeur moyenne élevée mais un revenu total faible.
- Vérifiez si les outliers ou valeurs manquantes affectent ces métriques.

SECTION ⑩

Analyse de Séries Temporelles

Objectif : Comprendre le comportement mensuel et les tendances dans le temps.

46. Définissez `order_date` comme index du DataFrame.
47. Calculez le revenu mensuel total avec `resample('M').sum()` sur `total_amount`.
48. Calculez la valeur moyenne mensuelle des commandes avec `resample('M').mean()` sur `total_amount`.
49. Identifiez le mois ayant les meilleures performances en termes de revenu total.
50. Décrivez les tendances ou pics inhabituels que vous observez dans la performance mensuelle.

CONSEILS & ASTUCES

- Utilisez `.resample('M')` sur un DateTimeIndex pour des séries temporelles mensuelles propres.
- Les dates manquantes ou invalides peuvent créer des lacunes artificielles dans la série temporelle.
- Les promotions et jours fériés apparaissent souvent comme des pics visibles dans le revenu mensuel.

SECTION ⑪

Rapport Final d'Analyste

Objectif : Combiner toutes les conclusions dans un récit clair orienté métier.

50. Rédigez un rapport complet qui inclut :
    - le nombre de lignes après nettoyage,
    - les principales anomalies et problèmes de données découverts,
    - comment chaque problème a été résolu ou traité,
    - les principaux insights numériques et motifs,
    - les risques métier ou limitations du jeu de données nettoyé,
    - des recommandations concrètes pour l'entreprise ou les prochaines étapes.

CONSEILS & ASTUCES

- Un excellent rapport répond à : « Que s'est-il passé ? », « Pourquoi est-ce important ? » et « Que devons-nous faire ensuite ? ».
- Utilisez un langage clair et structurez votre rapport avec des titres et des listes à puces si nécessaire.
- Concentrez-vous sur les décisions et actions que les parties prenantes peuvent prendre sur la base de votre analyse.