ESCEN
International & Digital Business School
NEXT-U

Projet : Traitement et Visualisation des Données

Analyse de cibles — import, nettoyage, exploration et analyse

Travail de Recherche en Équipe

Équipe : Bentifraouine Imène et Lyam Matic

Objectif global : Transformer les données brutes en informations claires et exploitables pour anticiper les comportements des utilisateurs et préparer une analyse pédagogique.

Date de remise : 06 décembre 2025

Année académique : 2025/2026

Escen - Bachelor Web & Technologies

Exercice pédagogique de sensibilisation au phishing. Données synthétiques uniquement.

Table des matières

1 Introduction

Dans ce projet, nous ne nous contentons pas de manipuler des chiffres. Nous voulons *comprendre les personnes derrière les données*. L'objectif ici est de montrer comment importer, nettoyer et analyser un dataset pour produire des informations fiables et exploitables sur les comportements des utilisateurs lors de campagnes de phishing simulées.

2 Optimisation et import des données

2.1 Optimisation des types de données

Pour que les analyses soient rapides et précises, nous avons choisi les types les plus adaptés pour chaque colonne (`int32`, `float32`, `category`). Cela réduit la mémoire utilisée et permet à Pandas de travailler efficacement, surtout pour les colonnes volumineuses ou catégorielles.

2.2 Première impression sur la qualité des données

Les données sont comme une première rencontre : elles donnent des indices sur ce qu'il faudra explorer. Nous avons observé :

— Quelques doublons et valeurs manquantes, qui peuvent fausser les conclusions.
— Des colonnes à transformer en catégories pour mieux gérer la mémoire.
— Une pertinence correcte : chaque colonne apporte une information utile pour comprendre les comportements.

3 Nettoyage et mise en forme

3.1 Étapes réalisées

Pour que nos données soient fiables, nous avons suivi un processus rigoureux :

— **Suppression des doublons** : éviter que les répétitions ne biaisent nos conclusions.
— **Suppression des valeurs manquantes** : garantir que chaque ligne contient toutes les informations nécessaires.
— **Transformation des colonnes en catégories** : réduire la mémoire utilisée tout en conservant l'information.

3.2 Justification des choix

Ces décisions sont guidées par un principe simple : *un nettoyage efficace produit des résultats fiables et exploitables*. Les doublons et valeurs manquantes peuvent fausser les analyses, tandis que la transformation en `category` optimise les performances.

3.3 Impact du nettoyage sur la mémoire

Moment	**Mémoire utilisée**
Avant nettoyage	0.04 Mo
Après suppression et transformation	0.02 Mo

4 Détection d'anomalies

Méthode choisie : Z-score

Pour repérer les comportements atypiques, nous avons utilisé le Z-score, qui mesure à quel point une valeur s'écarte de la moyenne. La formule du Z-score est :

$$Z = \frac{X - \mu}{\sigma}$$

où :

— X : valeur observée

— μ : moyenne de la colonne

— σ : écart-type de la colonne

Une valeur dont le Z-score absolu est supérieur à 3 ($|Z| > 3$) est considérée comme une **anomalie statistique**.

Pourquoi cette méthode ?

— Simple, rapide et transparente.

— Détecte efficacement les valeurs extrêmes.

— Parfaite pour nos colonnes numériques et l'âge des utilisateurs.

4.1 Colonnes analysées

Colonne	Description
gaming_interest_score	Score d'intérêt pour le gaming
insta_design_interest_score	Score d'intérêt pour le design sur Instagram
football_interest_score	Score d'intérêt pour le football
age	Âge des utilisateurs

4.2 Résultats

Après analyse, les anomalies détectées sont :

Colonne	Nombre d'anomalies détectées
gaming_interest_score	5
insta_design_interest_score	2
football_interest_score	2
age	0

4.3 Dataset final

Après suppression des anomalies, nous disposons de **506 lignes** prêtes pour l'analyse statistique.

5 Phase d'analyse statistique

Nous entrons maintenant dans le vif du sujet : comprendre ce qui motive les individus et identifier les groupes les plus sensibles.

5.1 Indicateurs clés (KPI)

Les KPI essentiels sont :

— Taux de réussite global

— Taux de réussite par produit recommandé

— Taux de réussite par segment d'intérêt (faible/moyen/fort)

— Taux de réussite par canal de communication

— Taux de réussite par tranche d'âge

5.2 Corrélations entre variables

Une matrice de corrélation révèle les relations entre intérêts, âge et succès des campagnes. Ces insights permettent de prévoir quelles caractéristiques influencent le plus la réaction des individus.

5.3 Observations et tendances

- Les produits populaires avec un fort intérêt tendent à générer un taux de réussite plus élevé.
- Les jeunes adultes (18–25 ans) sont particulièrement réactifs.
- Certains canaux de communication sont plus efficaces que d'autres.

5.4 Points forts et limites

- **Points forts :** identification rapide des groupes sensibles, visualisation claire des KPI.
- **Limites :** segmentation simple et corrélations linéaires qui ne capturent pas toutes les interactions complexes.

6 Analyse des comportements et datatelling

6.1 Présentation des données clés

- **Taux de réussite global :** 69% des campagnes suscitent une réaction positive.
- **Performance par produit :** Fifa (71%), Fortnite (70%), Instagram Pack (66%), Test (100%, mais peu de données).
- **Performance par canal :** Facebook (85%) et Mail (66%) sont les plus efficaces, Instagram (62%) moins performant, canaux non définis (40%) inefficaces.
- **Répartition par âge :** Les 18–60 ans sont les plus sensibles, avec un pic à 76% pour les 45–60 ans.
- **Influence des centres d'intérêt :** un intérêt élevé pour certains produits augmente la probabilité de succès.

6.2 Comportements et vulnérabilités du groupe étudié

- Les adultes (18–60 ans) constituent la cible prioritaire.
- Facebook et Mail sont les canaux les plus efficaces pour toucher ce public.
- La connaissance des centres d'intérêt permet de personnaliser les messages et maximiser l'impact.

6.3 Datatelling et justification des méthodes d'attaque

Pour illustrer un scénario concret, prenons un utilisateur fictif issu du dataset :

— **Profil :** 25 ans, fort intérêt pour Fifa et Fortnite, moyen intérêt pour Instagram Pack, utilise Facebook et Mail.

— **Observation :** Le taux de réussite sur des utilisateurs similaires pour Fifa et Fortnite est supérieur à 70%, ce qui indique une forte sensibilité.

— **Scénario d'attaque :** Un message ciblé mettant en avant une promotion ou contenu lié à Fifa/Fortnite envoyé via Facebook ou Mail maximiserait la probabilité de succès.

Justification : Les KPI et corrélations analysés confirment que ce type de profil est particulièrement réceptif aux produits Fifa/Fortnite et aux canaux Facebook/Mail. Ainsi, l'approche est cohérente avec les données : intérêts + âge + canal = forte probabilité de réaction positive.

7 Conclusion

Ce projet a été enrichissant : il nous a permis de passer de données brutes à une compréhension claire des comportements des utilisateurs. Nous avons expérimenté tout le processus d'analyse, du nettoyage à l'interprétation des résultats.

— L'optimisation des types et le nettoyage nous ont appris à traiter les anomalies et doublons, et à préparer un dataset fiable et exploitable.

— La détection d'anomalies nous a permis de mieux comprendre les comportements atypiques et leur impact sur l'analyse.

— L'analyse statistique nous a montré l'importance des KPI : produits, canaux, âge et centres d'intérêt influencent directement la réceptivité aux campagnes.

— Le datatelling nous a permis de relier chiffres et comportements, de créer des scénarios réalistes pour des profils types et de justifier des choix stratégiques.

Grâce à ce projet, nous avons acquis une vision globale et pratique du traitement des données et de leur interprétation. Nous comprenons désormais comment extraire des informations pertinentes à partir de données brutes, anticiper des comportements et proposer des stratégies basées sur l'analyse, tout en gardant un regard critique et éthique.