Rapport Projet N4 Traitement et visualisation des données

1) Comment vous avez optimiser les performances lors de l'import?

Pour optimiser les performances j'ai vu plusieurs petites optimisations.

a. En ayant read tout le csv, j'ai pu constater qu'il n'y avait pas de doublons de ligne. J'ai donc j'ai commencé à lire le CSV en enlevant la colonne ID. ça a permis de réduire la consommation en RAM de 10.5kb à 9.5kb ; c'est infime comme gain à cette échelle mais sur des très gros jeux de donnés, ça doit permettre d'économiser pas mal.
b. recommended_product C'est une donné texte avec très peu de variante -> ça devient donc un type category.

2) Ce que vous pensez de premier abord sur la qualité de la donnée et de sa pertinence ?

Le CSV est plutôt complet, il y a quelque anomalies, comme des lignes avec des cellule vides / pas de bonne donnés, nous avons décidé de les ignorer, et de virer ces lignes-là.
Sinon, Il y a des ligne qui peuvent contenir un nom de canal_recommande avec une capital en trop, auquel cas on réaligne avec la catégorie global plutôt que d'ignorer la ligne. Par exemple « Insta » passe à « insta », « Mail » à « mail ».
Le jeu de donné et ensuite bien pertinent une fois que ces erreurs sont corrigé.

3) Dans le rapport, indiquez et justifiez clairement :

a. Quelles valeurs ont été supprimées ou transformées ?

Je réitère, mais l'id 516 par exemple, contient un recommended_product "Test", ce qui n'est pas aligné avec la trend Fifa, Fortnite ou Instagram pack, donc on vire la ligne.

b. Pourquoi ces choix ont été faits ?

Pour que la ligne constitue un donné pertinente, nous avons besoin de toute les valeurs, si une valeur est manquante dans une colonne, par exemple recommended pour l'ID 512.

c. L'impact de votre nettoyage sur la mémoire (avant/après).

CF 1.a) (gain de 1kb de mémoire sur notre jeu de donné)

4) Dans le rapport, indiquez et justifiez clairement :

a. La ou les méthodes pour détecter les anomalies? Et Pourquoi ces choix ont été faits ?

Le cours propose 2 méthodes pour détecter des anomalies, l'écart type ou l'Interquartile range. Cependant, ces méthodes ne sont pas pertinentes pour notre jeu de donnés car nous savons que dans notre contexte, un score ne peut qu'être compris entre 0 et 100 inclus, et un âge, compris entre 20 et 100 inclus, or une méthode écart type « souligne » des valeurs anomalie de plus de 1.5x la moyenne, donc un âge à -10 ne serait pas considéré anormale par cette méthode.

b. Quelles sont les anomalies retenues ?

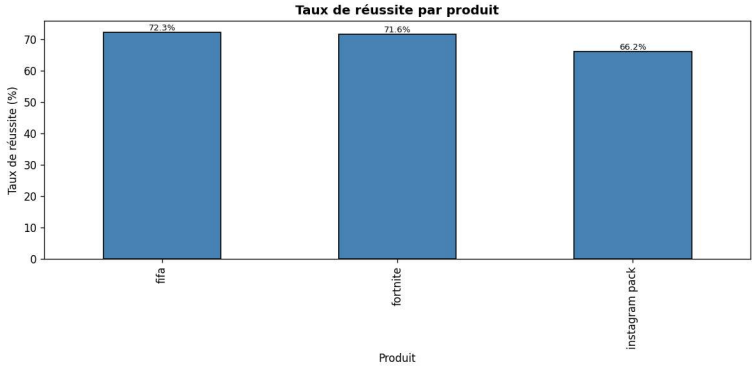
```
================================================================
LISTE DES LIGNES CONTENANT AU MOINS UNE ANOMALIE
================================================================

Nombre total de lignes avec anomalies: 12

Index | Age | Gaming | Insta | Football | Type(s) d'anomalie(s)
----------------------------------------------------------------
   10 |  23 |   -100 |    63 |        2 | Gaming hors plage
   31 |  40 |    -82 |    85 |       90 | Gaming hors plage
   44 |  51 |    -11 |    61 |       76 | Gaming hors plage
  160 |   5 |     60 |    38 |        0 | Age hors plage
  178 |  24 |     81 |    68 |      -52 | Football hors plage
  184 |   4 |      0 |    20 |       54 | Age hors plage
  200 |  53 |     51 |    37 |      -54 | Football hors plage
  243 |   2 |     11 |    46 |        0 | Age hors plage
  250 |  55 |     16 |     8 |     -114 | Football hors plage
  314 |  30 |     58 |   -56 |       53 | Insta hors plage
  422 |  20 |     28 |   -26 |       35 | Insta hors plage
  499 |  20 |    -28 |    36 |       48 | Gaming hors plage
```

NEEDS TO UPDATE

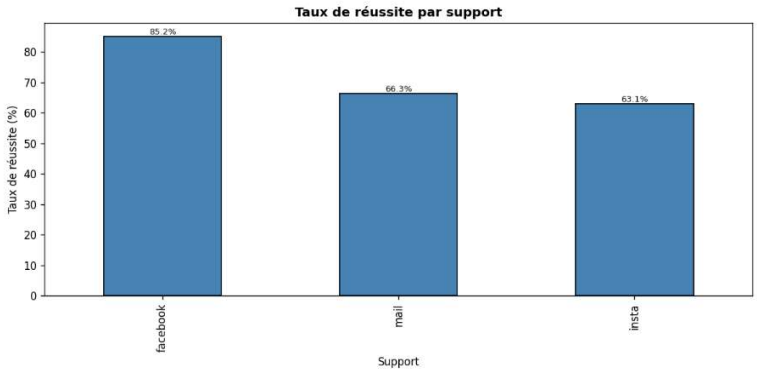
5) Dans le rapport, indiquez et justifiez clairement :

a. Décrivez les tendances ou observations principales que vous avez identifiées.

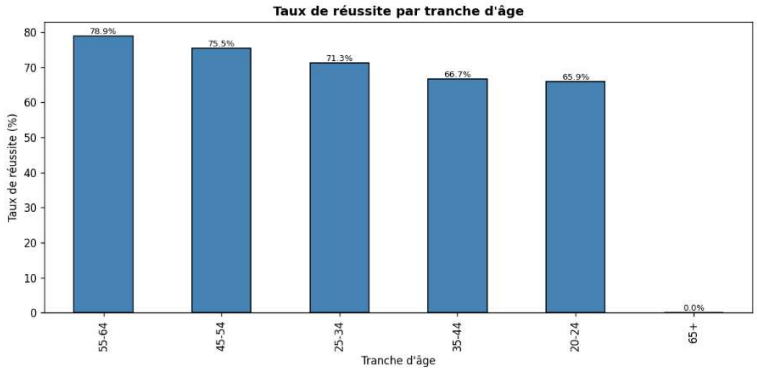
Le produit recommandé ayant rendu le plus d'attaques réussites est Fifa avec 72.3% de réussite, suivi de près par Fortnite à 71.6%, et enfin Instagram avec 66.2%.



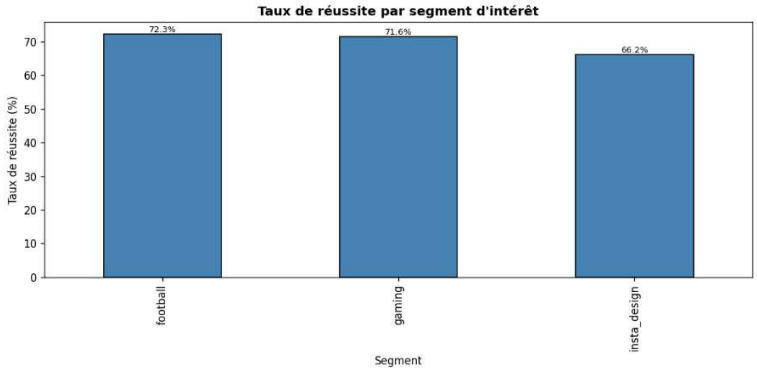
Le meilleur support est Facebook avec un score très élevé de 85.2%, loin des 66.3% par mail et 63.1% sur Instagram.



La meilleure tranche d'âge à cibler est celle des 55 à 64 ans, avec un taux de réussite de 78,9 %. On observe clairement que la naïveté augmente avec l'âge. Cependant, notre jeu de données ne contient aucune information concernant des personnes de plus de 60 ans ; nous ne pouvons donc qu'estimer que le groupe des 65 ans et plus présenterait un taux de réussite supérieur à celui des 55 à 64 ans.



Enfin, le meilleur centre d'intérêt à cibler est le football, avec 72.3%, suivi de près par le gaming avec 71.6%, et enfin Insta_Design avec 66.2%.



b. Soulignez les points forts et les limites de votre première approche.

On obtient des valeurs précise, et concrète. Cependant, le jeu de données est incomplet et pourrait être complété par davantage de données, pouvant affiner l'analyse et possiblement révélé de meilleurs axes sur lesquels basé ses attaques de phishing.

c. Indiquez les hypothèses ou pistes à approfondir pour l'étape suivante.

Pas certain de l'étape concerné par cette question ? Si c'est celle d'avant, oui il faudrait idéalement plus de données.

À partir des premières analyses de vos données, rédigez une section de Data Telling montrant comment les tendances observées permettent d'anticiper les types d'attaques auxquels un groupe de population pourrait être exposé. Rédiger clairement votre Data Telling argumente à l'aide de chiffre clé et de vos courbes.

Votre récit doit :

d. **Présenter les données clés identifiées.**
e. **Mettre en évidence les comportements ou vulnérabilités propres au groupe étudié.**
f. **Expliquer comment ces éléments permettent de prévoir les méthodes d'attaque susceptibles de les viser.**
g. **Justifier clairement la cohérence entre les données et les scénarios d'attaque anticipés.**

Exemple (non détaillé) d'un data telling :

Les données montrent que 59 % des enfants âgés de 0 à 3 ans utilisent régulièrement des applications ou contenus inspirés de jeux vidéo populaires tels que FIFA, avec un score moyen d'intérêt avoisinant les 80 %. Par ailleurs, leur niveau de vulnérabilité apparaît élevé : la campagne de simulation de phishing menée dans ce cadre pédagogique indique que 78 % des enfants de cette tranche d'âge ont été trompés par le message testé. Le moyen de support utilisé pour lancer l'attaque sera Instagram.

Story telling ici