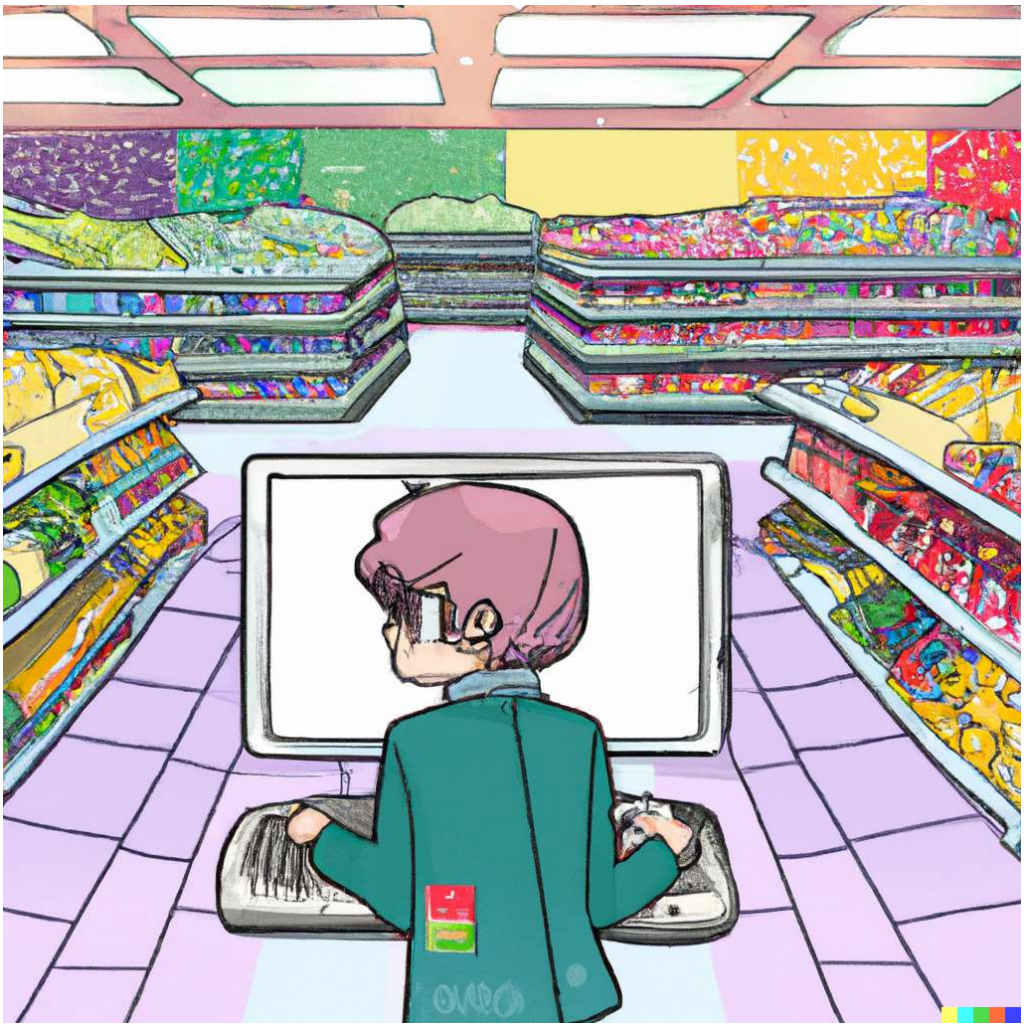
# Data Analysis & AI

# 2022-2023

# HELBFour



*DALL.E : A programmer with a computer in a supermarket, in the center of an aisle, filled with products of all colors. 2D Cartoon style. Anime style.*

## Description :

La société HelbFour possède une petite vingtaine de supermarchés en Belgique. Dans chaque magasin, chaque jour, une centaine de clients viennent faire leurs achats du quotidien. Toutefois la société n'est pas très à jour avec les technologies du numérique. Dans l'optique de rester compétitive, elle souhaite développer son parc informatique et pouvoir générer de la valeur en tirant profit des données qu'elle génère.

Vous avez reçu un jeu de données (disponibles sur eCampus) qui sont les transactions effectuées dans un même magasin pendant plusieurs semaines (depuis 2014). Chaque ligne du fichier décrit soit une indication temporelle, soit le panier d'achat d'un consommateur pour le jour correspondant. Chaque produit est identifié par un numéro. Le magasin compte une centaine de produits. La gamme de produits vendu n'a pas évoluée depuis 2014.

Il vous est demandé d'écrire un programme pour analyser le jeu de données et expliquez vos conclusions sur les habitudes de vente de ce magasin. La société est tout particulièrement intéressée par les différences qu'il existe dans les habitudes de consommation des clients du magasin en fonction de la période de l'achat. Elle se pose notamment les questions suivantes :

- Quels sont les produits les plus populaires/ les moins populaires du magasin ?
- Y a-t-il des jours de la semaine, ou des semaines, plus lucratives que d'autres ?
- Comment ont évolués les habitudes d'achats des clients en fonction des années ?
- Y a-t-il des groupes de produits plus souvent achetés ensembles ?
- Existe-t-il des règles d'association pour ces groupes de produits ?
- Ces règles d'association dépendent t'elles de la période d'achat ?
- Quel est la pertinence de ces règles d'association ?
- Y a-t-il d'autres informations pertinentes quant aux habitudes d'achats dans le magasin ?

**Données :**

Vous avez un jeu de données disponibles au format .txt .
La structure du fichier est la suivante :

```
YEAR:2014**********************************************************
WEEK:1++++++++++++++++++++++++++++++++++++++++++++++++++
DAY:Monday----------
['p_3', 'p_7']
['p_19', 'p_57']
['p_55', 'p_45', 'p_31']
DAY:Tuesday----------
['p_1', 'p_43', 'p_64']
['p_10', 'p_18']
```

Cela signifie que dans ce magasin, le lundi de la première semaine de 2014 il y a eu 3 clients (transactions). Le premier client a acheté les articles 3 et 7. Le deuxième client a acheté les articles 19 et 57. Mardi il y a eu 2 clients (transactions). Le premier client a acheté les articles 1, 43 et 64. Le deuxième a acheté les articles 10 et 18.

La société HELBFour vous met toutefois en garde, le fichier de logs des transactions est utilisé depuis maintenant plusieurs années. Elle soupçonne que certaines erreurs aient pu se glisser dans le fichier. Elle vous demande de faire au mieux votre analyse de façon que vos conclusions restent pertinentes.

Note : Remarquez qu'un client ne peut pas acheter deux fois les mêmes articles. Ne vous en inquiétez pas, il s'agit simplement d'une simplification dans l'énoncé.

Attention, si vous désirez explorer le fichier manuellement, n'ouvrez pas les fichiers avec le bloc note Windows, celui-ci ne prend pas en compte les retours à la ligne. Privilégiez un IDE comme Notepad++

## Groupes :

Ce travail devra être réalisé **individuellement**.

## Choix des technologies :

Vous êtes libres d'utiliser le langage et les technologies qui vous paraissent les plus appropriées. Il vous est toutefois recommandé d'utiliser le langage Python pour analyser le fichier. La méthode « eval() » pourra notamment vous être utile au parsing du fichier.

## Rapport de projet :

Il vous est demandé de rédiger un rapport contenant au minimum les sections suivantes :

**Introduction** : Cette section devra introduire votre projet. Décrire ce qui a été réalisé et présenter brièvement la structure de votre rapport.
**Description des technologies utilisées** : Cette section devra introduire les différentes technologies utilisées et motiver leur utilisation.
**Nettoyage des données :** Cette section devra décrire les processus de nettoyage des données afin de les rendre utilisables pour l'analyse. Expliquez notamment les différents problèmes rencontrés dans le dataset et justifiez vos choix de résolution pour ces différents problèmes.
**Processus et Résultats d'analyse des données** : Cette section devra décrire vos processus d'analyse de données ainsi que les résultats sur le dataset fourni pour le projet. Quels ont été les techniques mises en œuvre pour analyser les données ? Quels sont les conclusions de votre analyse ? **Une attention particulière sera portée sur la clarté de leur présentation. Toute visualisation appropriée de ces résultats est donc la bienvenue.**
**Limitations et développement futur** : Les limites de votre application, par exemple : dans quels cas d'utilisation votre application pourrait ne pas fonctionner comme prévu ? Y a-t-il des aspects techniques du site qui n'ont pas été traité ? Si vous aviez plus de temps pour le projet, qu'auriez-vous amélioré ? Plusieurs points de vue sont possibles, il revient au groupe d'étudiant de choisir les points qu'il considère les plus pertinents pour réaliser son autocritique.
**Conclusion** : Votre conclusion sur le projet. Ce qui a été vu dans le rapport, ce que vous avez réussi à faire ou non durant le projet et les apprentissages que vous en tirez.

**N’hésitez pas à mettre l’accent sur les défis techniques et les solutions apportées.** Le rapport sera notamment évalué sur la qualité écrite et l’effort de présentation ainsi que la pertinence et la complétude des points abordés.

Faites attention à bien référencer vos sources documentaires ainsi que les codes potentiellement utilisés.

## Respect des consignes :

Le non-respect des consignes énoncées dans ce document entrainera automatiquement, sans possibilité de recours, une pénalité au niveau des points pour le projet.

## Développement et Triche

- Tout acte de triche sera sanctionné par **une note de fraude au bulletin et sera notifié à la direction qui pourra possiblement décider de sanctions supplémentaires**. Des parties de code réutilisées d’un projet existant (d’un autre étudiant ou disponible sur le net) sans références dans votre rapport et sans mention de l’utilité du code utilisé est considéré comme une fraude.

- Pour ce projet, **vous ne pouvez pas reprendre des parties du code d’un autre étudiant,** que ce soit de cette année ou d’une année antérieure.

- Pour ce projet, vous pouvez vous inspirer d’un code disponible sur internet mais le projet ainsi que les parties réutilisées ou inspirées doivent être correctement référencées. **Votre apport personnel doit également être suffisant.** Si vous avez un doute, contactez l’enseignant le plus tôt possible afin d’éviter du refactoring inutile, ou pire, **un zéro/fraude**.

## Deadlines et remises :

**23/12/22 23h59** : Remise finale du projet contenant, le rapport final et le code source de l’application.

## Evolution des consignes :

**Les consignes du projet sont susceptibles d’évoluer. Veillez à consulter les annonces sur e-campus afin de vous tenir à jour.**

## FAQ :

- **Puis je ajouter d’autres sections ou sous-sections dans le rapport ?**

Oui. La partie rapport de ce document donne seulement la structure minimum.

- **Le rapport est-il important ?**

**Oui**. Le rapport est une **pièce centrale de votre projet** et c'est le premier outil de communication qui me servira à juger de la bonne réalisation du projet, pas seulement du point de vue du code mais également de la méthodologie utilisée.

- **Je n'ai pas réussi à tout réaliser. Est-ce que ça vaut la peine de vous rendre le projet ?**

Oui veillez toutefois à être claire sur les parties non implémentées. Il est très déconseiller de dissimuler ou d'« oublier » de mentionner qu'une partie n'a pas été réalisée. Veillez toutefois à bien respecter les consignes.