RAPPORT DE TRAVAIL

EXTRACTION DE DONNEES D'UN SITE DE E-COMMERCE ET ANALYSE DE DONNEES A BASE DU SQL

Dans le cadre de notre travail sur le webscrapping et l'analyse de données, les objectifs poursuivis étaient les suivants :

1) Planification et organisation du travail en groupe

2) Extraction des données d'un site de E-commerce à l'aide de la bibliothèque python Beautiful Soup

3) Reconstitution et mise en place de la base de données à l'aide du SQL

4) Faire une analyse descriptive sur les données disponibles dans votre base de données ou sur un fichier csv

5) Visualisation :

- Les 10 produits les plus couteux
- Les 10 produits les moins couteux
- Les 10 produits les plus recommandés
- Les 10 produits les moins recommandés
- Les 5 catégories ayant le plus de produits en stocks
- Les trois catégories ayant les produits les plus couteux

Pour atteindre nos objectifs, nous avons procédé à la réponse des différentes questions, suivant une procédure de travail organisée de manière participative des différents collaborateurs.

PLAN DE TRAVAIL

I-Importation des bibliothèques

-BeautifulSoup : librairie permettant d'extraire les informations sur les pages web

-Request : effectuer les requêtes

-Pandas : permet l'analyse des données

-Word2num : pour convertir les lettres en numériques

-Url lib : pour combiner les composants dans une chaine d'URL

II- Reconstitution et mise en place à l'aide du SQL

Les données ont été extraits du site web propose puis sauvegarde dans un fichier CSV et dans une base de données SQL.

III- Analyse descriptive de données et visualisation

-Les 10 produits les plus couteux sont représentés par un diagramme circulaire qui répartit les produits les plus couteux et leurs pourcentages. Voir graphique 1.

-Les 10 produits les moins couteux sont visualises sur un diagramme circulaire donnant une répartition en pourcentage. Voir graphique 3.

-Les 10 produits les plus recommandes sont ceux qui ont le score le plus élevé 5. Visualises sous forme de barres horizontale.

- Les 10 produits les moins recommandes sont ceux qui ont le score le plus bas 1. Visualises sous forme de barres horizontale.

-les cinq catégories ayant le plus de produits en stock représentés par les Barres verticales : Default : 1345 ; Nonfiction : 975 ; Séquential Art : 686 ; Fiction : 588 ; Add a comment : 516

Les trois catégories ayant les produits les plus couteux. Représentés sous forme de diagramme circulaire. Suspense : 58 ; Novels : 54 ; Politics : 53

Participants :

-Fondja Elton

- Waindja Audrey

-Madjoukou Claudette Noëlla

- Tsakam Ludovic