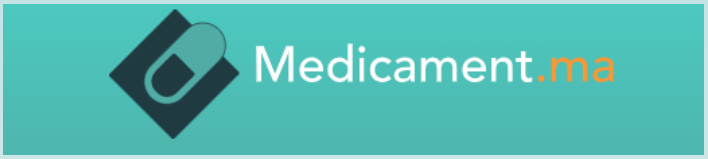
الكلية متعددة التخصصات تازة
ⵜⴰⵖⵉⵍⴰⵢⵜ ⵜⴰⵖⵜⵉⵣⵓⵜⵜ ⵏ ⵜⴰⵣⴰ
FACULTÉ POLYDISCIPLINAIRE DE TAZA

جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⴰⵎⴻⴷ ⴱⵏ ⵄⴰⴱⴷⵍⵍⴰⵀ ⴱⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

Rapport du mini projet du web scraping

Scraper un site web de pharmacie « Medicament.ma »



Réalisé par :

OUGNI Imane

RIFI Maroua

Encadré par :

Mr. Ben Yakhlef

Table des matières

1. Introduction

Ce projet porte sur l'utilisation du web scraping pour extraire des données relatives aux médicaments disponibles sur le marché marocain. L'objectif principal est de collecter, nettoyer, analyser et visualiser les informations issues du site https://medicament.ma, une plateforme en ligne dédiée à l'information sur les médicaments au Maroc. À travers ce projet, nous visons à fournir une analyse approfondie des produits pharmaceutiques disponibles en ligne, en nous concentrant sur des aspects tels que les prix, les dosages, et les caractéristiques des médicaments.

Le projet se divise en plusieurs étapes, dont le **web scraping**, qui consiste à extraire les données depuis le site web choisi, le **nettoyage** des données pour garantir leur qualité et leur fiabilité, l'**analyse** statistique des données pour en tirer des conclusions pertinentes, et enfin la **visualisation** des résultats, permettant d'illustrer les tendances observées. Cette approche fournit non seulement une compréhension des données, mais elle ouvre également la voie à des études de marché plus poussées, notamment pour les professionnels de santé, les patients et les consommateurs à la recherche d'informations sur les médicaments.

2. Web scraping

Le web scraping est une technique utilisée pour extraire des données à partir de sites web. Cela permet de collecter automatiquement des informations présentes sur des pages web de manière structurée. Cette méthode est souvent utilisée dans des domaines tels que l'analyse de données, la collecte d'informations en temps réel, ou l'extraction de contenu pour des études de marché.

Dans ce projet, nous avons utilisé le web scraping pour collecter des informations provenant d'un site web spécifique, que nous avons ensuite traitées, nettoyées et analysées.

3. Le site web scrapé

Dans ce projet, nous avons choisi de scraper un site web spécialisé dans le domaine de la santé, plus précisément dans la pharmacie. Le site en question est https://medicament.ma, une plateforme dédiée à l'information sur les médicaments au Maroc. Ce site propose une large gamme de médicaments, avec des informations détaillées sur leurs prix, leurs dosages, et leurs effets secondaires.

Le choix de ce site a été motivé par la richesse des informations qu'il propose et la structure claire de ses pages, ce qui permet une extraction efficace des données. Le site contient des informations essentielles pour l'analyse, telles que les noms des médicaments, les prix, les dosages, ainsi que d'autres détails utiles pour une analyse approfondie des produits pharmaceutiques disponibles.

Caractéristiques du site :

- Domaine : Pharmacie en ligne et informations sur les médicaments.
- Contenu : Produits pharmaceutiques, leur prix, leurs dosages, et leurs caractéristiques.
- Public cible : Professionnels de santé, patients, et utilisateurs à la recherche d'informations sur les médicaments.

En choisissant medicament.ma, nous avons pu accéder à un large éventail de données, ce qui nous a permis de réaliser une analyse significative du marché des médicaments disponibles en ligne au Maroc. Le site a été scrappé en utilisant des techniques de scraping adaptées pour extraire les données de manière automatisée et structurée.

4. Les étapes suivies

Le projet a consisté en quatre étapes principales : le scraping des données à partir du site https://medicament.ma, suivi du nettoyage des données pour les rendre exploitables. Ensuite, une analyse statistique a été réalisée, complétée par des visualisations pour mieux comprendre les tendances des données collectées.

4.1. Scraping

Le scraping a été effectué sur le site https://medicament.ma, spécialisé dans les informations sur les médicaments. Pour ce faire, nous avons utilisé les bibliothèques **requests** et **BeautifulSoup** de Python.

La première étape consistait à récupérer les pages de médicaments classées par lettres de A à Z. Pour chaque lettre, nous avons envoyé des requêtes HTTP afin de récupérer le contenu HTML des pages du site. Ensuite, nous avons extrait le nombre total de pages disponibles pour chaque lettre. En parcourant toutes les pages de chaque lettre, nous avons utilisé BeautifulSoup pour analyser le contenu HTML et extraire les informations pertinentes, à savoir le **nom**, la **dose**, le **format** et le **prix** des médicaments.

Le processus a été automatisé pour passer par chaque page de chaque lettre, garantissant ainsi l'extraction de toutes les données disponibles. Les informations collectées ont été stockées dans une liste et ensuite converties en un **DataFrame pandas**, avant d'être exportées dans un fichier Excel pour une analyse ultérieure. Le scraping a permis d'extraire un total de [nombre de médicaments extraits].

4.2. Nettoyage

Après avoir extrait les données du site web, un processus de nettoyage a été effectué pour garantir la qualité et la cohérence des informations. Tout d'abord, les colonnes du DataFrame ont été renommées afin de les rendre plus explicites, facilitant ainsi la compréhension des données. Ensuite, les lignes contenant des valeurs "None" dans les colonnes "Dosage", "Forme", ou "Prix (MAD)" ont été supprimées pour éviter toute confusion lors des analyses. En ce qui concerne la colonne "Prix (MAD)", les valeurs ont été extraites sous forme numérique, avec une conversion en nombres à virgule flottante et un arrondi à deux chiffres après la virgule. Après cette conversion, les lignes contenant des valeurs manquantes dans la colonne des prix ont également été éliminées. Enfin, des espaces inutiles dans les colonnes "Nom medicament", "Dosage", et "Forme" ont été supprimés pour uniformiser les données. Ces étapes ont permis d'obtenir un ensemble de données propre et prêt pour l'analyse.

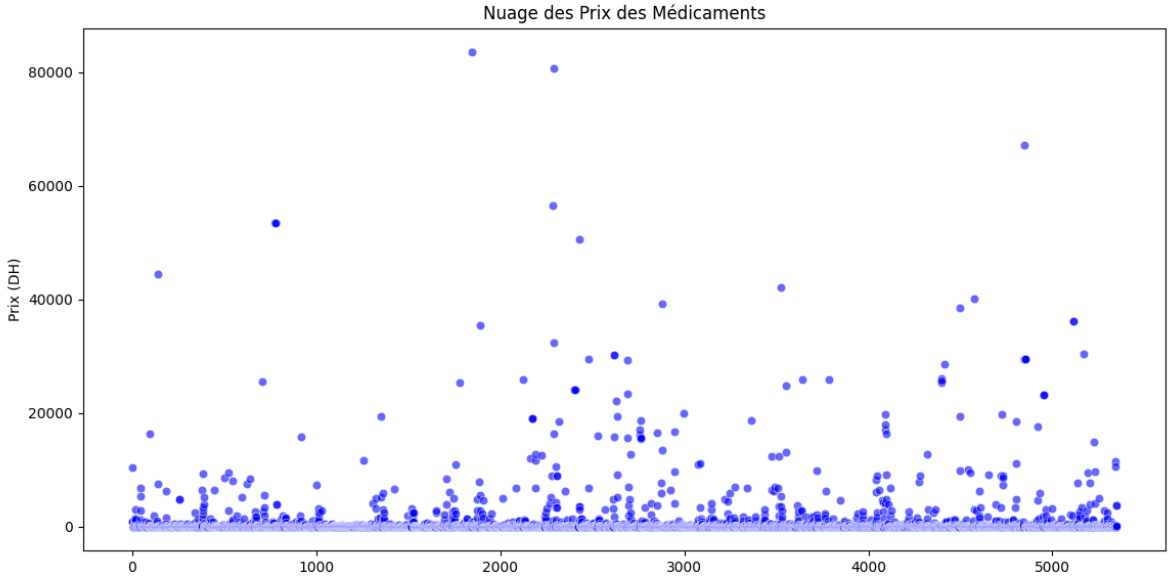
4.3. Analyse

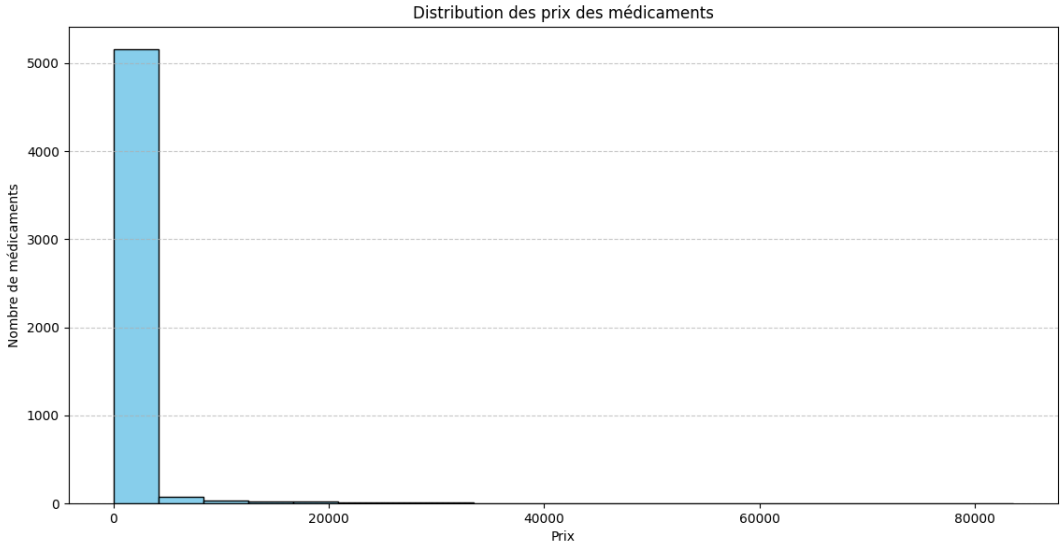
L'analyse des données nettoyées a été réalisée pour extraire des informations pertinentes et mieux comprendre la distribution des prix, des médicaments et des dosages. Tout d'abord, une analyse statistique de la colonne "Prix (MAD)" a été effectuée, fournissant des mesures de tendance centrale et de dispersion, telles que la moyenne, l'écart-type, et les valeurs minimales et maximales. Ensuite, les cinq médicaments les plus chers et les cinq moins chers ont été identifiés, permettant de repérer les extrêmes dans la gamme de prix. En parallèle, une analyse des dosages a permis de déterminer les cinq dosages les plus fréquents parmi les médicaments collectés. Ces informations fournissent une première vue d'ensemble sur la répartition des prix et des dosages, utiles pour des études ultérieures sur les tendances du marché pharmaceutique.

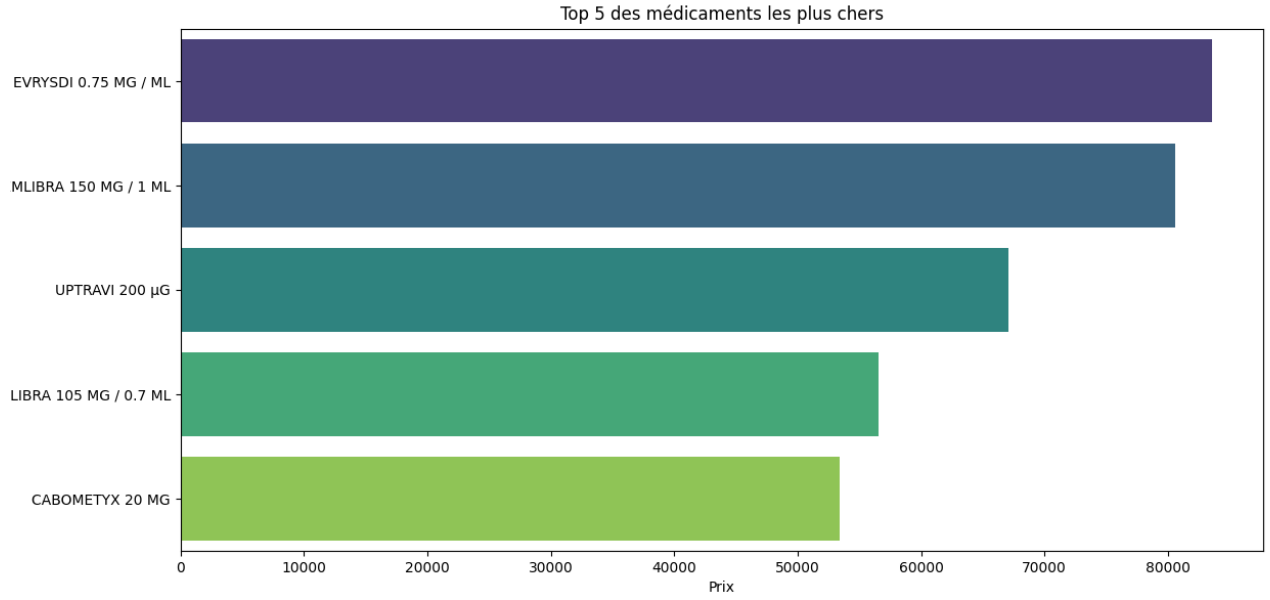
4.4. Visualisation

Pour mieux comprendre la répartition et les tendances des prix des médicaments, plusieurs visualisations ont été réalisées. Un **nuage de points** a été utilisé pour représenter la distribution des prix en fonction de l'index des médicaments, permettant d'observer la dispersion des prix. Ensuite, un **histogramme** a été tracé pour illustrer la fréquence des prix des médicaments, avec une répartition des prix sur 20 intervalles, offrant une vue d'ensemble sur la concentration des prix. Un **boxplot** a été ajouté pour visualiser les statistiques des prix, incluant les quartiles et les valeurs aberrantes.

Les **cinq médicaments les plus chers** et les **cinq médicaments les moins chers** ont été présentés sous forme de **graphique à barres** afin de mieux comparer les prix extrêmes dans l'ensemble des données. Enfin, une analyse des **mots les plus fréquents** dans les noms des médicaments a été réalisée. Après nettoyage des données, les mots ont été comptabilisés, et les **dix mots les plus fréquents** ont été affichés à l'aide d'un graphique à barres. Cette visualisation permet d'identifier les termes récurrents dans les noms des médicaments, ce qui peut être utile pour une analyse plus approfondie des produits présents dans la base de données.







5. Documentation

5.1. Instructions pour Exécuter le Projet

Pour exécuter ce projet, il suffit de suivre quelques étapes simples. Ce projet repose sur l'utilisation de Python pour effectuer le web scraping, le nettoyage des données, l'analyse et la visualisation des données.

5.2. Pré-requis

- On doit installer Python.
- Bibliothèques Python nécessaires

requests : pour envoyer des requêtes HTTP et récupérer les pages web.

BeautifulSoup : pour analyser et extraire des informations de pages HTML.

pandas : pour le traitement des données et la manipulation des DataFrames.

matplotlib : pour la création de graphiques et de visualisations.

seaborn : pour des visualisations avancées et esthétiques.

Pour installer ces bibliothèques, on doit exécuter la commande suivante dans le terminal :

pip install requests beautifulsoup4 pandas matplotlib seaborn

5.3. Exécution du Script

Une fois les bibliothèques installées, on peut exécuter les différents scripts Python pour effectuer les différentes étapes du projet.

Les scripts doivent être exécuté dans l'ordre suivant :

python scraping.py

python nettoyage.py

python analyse.py

python visualisation.py

6. Conclusion

Ce projet a permis de démontrer la puissance du web scraping en tant qu'outil de collecte et d'analyse de données, en utilisant des technologies simples mais efficaces telles que Python, Requests, et BeautifulSoup. Grâce à l'extraction de données depuis le site https://medicament.ma, nous avons pu obtenir des informations précieuses sur les médicaments disponibles au Maroc, qui ont ensuite été nettoyées, analysées et visualisées pour en tirer des conclusions pertinentes.

L'analyse des données a mis en évidence des tendances intéressantes, telles que la répartition des prix des médicaments et la fréquence des différents dosages. Les visualisations, notamment les nuages de points, histogrammes et boxplots, ont permis de mieux comprendre la distribution des prix et des caractéristiques des médicaments, offrant ainsi un aperçu détaillé du marché pharmaceutique en ligne au Maroc.

Ce projet ouvre la voie à de futures améliorations, comme l'intégration d'autres sources de données ou l'analyse de données supplémentaires pour une meilleure compréhension des dynamiques du marché pharmaceutique. En conclusion, cette étude souligne l'importance de l'automatisation de la collecte de données dans le domaine de la santé et montre comment le web scraping peut être utilisé pour extraire des informations essentielles pour des analyses approfondies.