UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Sorbonne Université

Diplôme Universitaire **Data Analytics (DU SDA)**

Projet webscrapping-Python Avancé

Donnée Immobilière par Web Scraping et Analyse Statistique-seloger

---

Auteurs : Ibrahima BA
Mahamat Nour MAHAMAT SULTAN
Moustapha MENDY
Formation : DU Sorbonne Data Analytics
Année universitaire : 2025–2026
Enseignant : SLIMI ANISS
Date de rendu : 20/12/2025

Table des matières

Chapitre 1

Introduction

1.1 Contexte et enjeux

Le marché immobilier constitue un pilier fondamental de l'économie française, tant par son poids financier que par son impact social. Il joue un rôle structurant dans l'organisation des territoires, influence directement la mobilité résidentielle et constitue un indicateur clé du pouvoir d'achat des ménages. Les dynamiques immobilières reflètent des phénomènes économiques, démographiques et géographiques complexes, rendant leur analyse essentielle pour les acteurs publics comme privés.

Malgré cette importance, le marché immobilier demeure marqué par une forte opacité, notamment en ce qui concerne l'accès à des données détaillées, fiables et régulièrement mises à jour. Les bases institutionnelles existantes sont souvent partielles ou difficilement exploitables pour des analyses fines à l'échelle locale.

À l'inverse, les plateformes immobilières en ligne telles que *SeLoger*, *Leboncoin* ou *Citya* publient quotidiennement un volume considérable d'annonces immobilières. Ces données, bien que non structurées, contiennent des informations essentielles telles que le prix affiché, la surface, la localisation ou encore le type de bien. Exploitées correctement, elles représentent une source d'information majeure pour analyser les dynamiques du marché immobilier.

1.2 Objectifs du projet

Ce projet vise à concevoir et implémenter une chaîne complète de traitement de données reposant sur un pipeline **ETL (Extract, Transform, Load)** appliqué au domaine de l'immobilier. L'objectif est de démontrer la capacité à transformer des données brutes issues du web en informations exploitables pour l'analyse statistique et la visualisation.

1.3 Problématique

Comment extraire de manière automatisée et fiable des données immobilières à grande échelle à partir de sites web fortement protégés par des systèmes anti-bot (tels que Datadome), puis transformer ces données brutes en informations exploitables pour une analyse statistique et cartographique du marché immobilier ?

Chapitre 2

État de l'art et choix techniques

2.1 État de l'art du web scraping

Le web scraping désigne l'ensemble des techniques permettant d'extraire automatiquement des données à partir de pages web. Il est largement utilisé dans des contextes de veille, d'analyse de marché et de recherche académique. Toutefois, la diversité des architectures web et la généralisation des protections anti-bot rendent cette pratique de plus en plus complexe.

On distingue généralement les sites statiques, dont le contenu est directement accessible dans le code HTML, et les sites dynamiques, reposant sur des frameworks JavaScript modernes tels que React ou Angular. Cette distinction conditionne directement les outils de scraping utilisables.

2.2 Outils de scraping

La combinaison *Requests + BeautifulSoup* constitue une solution rapide et efficace pour les sites statiques. En revanche, elle se révèle inefficace face aux sites dynamiques et protégés.

Les outils comme *Selenium* ou *Playwright* permettent de piloter un navigateur réel et de simuler un comportement humain, au prix d'une complexité et d'une consommation de ressources plus élevées.

2.3 Défi des protections anti-bot

Les plateformes immobilières modernes utilisent des solutions avancées telles que *Datadome*, reposant sur l'analyse du fingerprint navigateur, du comportement utilisateur et des requêtes réseau. Ces mécanismes bloquent efficacement les tentatives de scraping automatisé via des erreurs HTTP ou des CAPTCHA.

2.4 Architecture retenue

Face à ces contraintes, une architecture hybride a été retenue :

— Scraping via Selenium et Remote Debugging

— Nettoyage via Pandas et Regex

— Visualisation via Streamlit et Folium

Chapitre 3

Méthodologie et réalisations

3.1 Stratégie de scraping semi-automatique

Les méthodes classiques reposant sur des navigateurs automatisés ont rapidement été bloquées. La solution retenue repose sur le pilotage d'une instance Chrome déjà ouverte en mode débogage, permettant de bénéficier d'une session utilisateur légitime.

3.2 Algorithme d'extraction adaptatif

Un algorithme en cascade a été développé afin de s'adapter aux variations fréquentes du DOM. Plusieurs niveaux d'extraction sont utilisés, allant de sélecteurs précis à une analyse textuelle basée sur des expressions régulières.

3.3 Nettoyage et structuration

Les données extraites contenaient de nombreux artefacts (encodage, bruit textuel, informations fusionnées). Un pipeline de nettoyage basé sur *Pandas* a permis de structurer les données et de calculer des indicateurs dérivés tels que le prix au mètre carré.

3.4 Enrichissement géographique

L'API *Nominatim* a été utilisée pour transformer les adresses en coordonnées géographiques, permettant la visualisation cartographique.

Chapitre 4

Résultats et analyses

4.1 Analyse descriptive

Les résultats mettent en évidence une forte hétérogénéité des prix au mètre carré selon la localisation. Une corrélation positive, mais non linéaire, est observée entre la surface et le prix total.

4.2 Tableau de bord interactif

Un tableau de bord Streamlit permet une exploration dynamique des données via filtres, indicateurs clés, cartes interactives et graphiques statistiques.

4.3 Module de prédiction

Un modèle de régression linéaire a été implémenté afin d'estimer le prix théorique d'un bien en fonction de sa surface. Bien que simple, ce modèle illustre le potentiel prédictif des données collectées.

Chapitre 5

Discussion et limites

5.1 Difficultés rencontrées

Les principales difficultés concernent l'instabilité du DOM et les limitations imposées par les systèmes anti-bot. Des pauses aléatoires et des sélecteurs tolérants ont permis d'améliorer la robustesse du scraping.

5.2 Limites de l'étude

Les données correspondent aux prix affichés et non aux prix de transaction. L'analyse repose sur un échantillon ponctuel, limitant l'étude des dynamiques temporelles.

5.3 Perspectives

Des améliorations futures incluent une collecte périodique automatisée, l'intégration de données exogènes et l'utilisation de modèles prédictifs plus avancés.

Conclusion

Ce projet a permis de couvrir l'ensemble du cycle de vie de la donnée, depuis l'extraction complexe de données web protégées jusqu'à leur valorisation analytique et visuelle. Il met en évidence l'importance d'une approche méthodologiquement rigoureuse pour exploiter des données non structurées dans un contexte réel.

L'approche semi-automatisée retenue s'est révélée être un compromis efficace entre robustesse et coût. Les perspectives ouvertes par ce travail soulignent le potentiel des données immobilières web pour des analyses avancées et des applications d'aide à la décision.