16e Conférence internationale EAI sur les
infrastructures et services Internet en Afrique

EAI AFRICOMM

Web Scraping for Competitive Intelligence: State of the Art and Case Study

[1]**Edouard Ngor SARR**, [2]Oumar DIAGNE, [3]Abel DIATTA and [4]Lamine FATY
[1,2,3,4] University Assane SECK de Ziguinchor - UASZ
SENEGAL

27-29 novembre 2024, Abidjan, Côte d'Ivoire

Summary

- Introduction
- Definitions
- Research Problem
- State of the art
- Challenge
- Difficulties
- Case Study
- Results and Discussions
- Conclusion

Introduction

- Data collection is a critical technique that involves extracting, organizing, and storing raw data from various sources, whether manual or automated.
- This information was then leveraged to formulate business recommendations aimed at enhancing the competitiveness of the companies studied.
- In This article, we presents
  - A state-of-the-art review
  - A case study on the use of web scraping for competitive intelligence.
    - We collected and merged data from different sources (E-commerce website
    - Performed cross-analyses
    - Visualized the results in informative dashboards to support decision-making
    - We discuss the challenges associated with the use of web scraping for economic intelligence, while highlighting the potential contribution of artificial intelligence to enhance these
- The study demonstrated that integrating web scraping into business strategies can improve market understanding and provide valuable insights to outcompete rivals. systems.

Definitions

- **Competitive Intelligence**
  - Is defined when an organization X, which has similar activities to those of another organization Y, monitors the relevant activities of its competitors.
    - It is a powerful tool used by companies to analyze their environment.
    - It is very essential for e-commerce businesses to remain competitive in a constantly changing market.
    - The success will necessarily depend on a good knowledge of its ecosystem and the implementation of data collection strategies in the websites of competitors.
- **Web scraping**
  - Systematically and automatically collecting relevant information about competitors' activities, products, services, strategies, and performance from their web platforms
  - In a nutshell, the objective of data scraping is to help industry actors to establish autonomous data acquisition systems to facilitate informed decision-making
  - This information extraction process can be manual, automated, or semi-automatic and can include databases, digital files, or web pages.

Research Problem

- In a constantly changing world where the majority of Commercial activities are carried out online through web platforms (online stores/shops) more and more.

- The economical competitive intelligence is very essential for e-commerce businesses to remain competitive in a constantly changing market [3].

- The success of such monitoring will necessarily depend on a good knowledge of its ecosystem and the implementation of data collection strategies in the websites of competitors.

- But, In the context of Web 2.0 or web 3.0, has become
    - Very complex
    - Time-consuming.

State of the art

- **Approaches to web scraping**

Table 1. Advantages and Disadvantages of Different Web Scraping Methods

Web Scraping Methods	Benefits	limit
Regular Expression-Based Web Scraping	- Simple to implement	- Limited to text formats
DOM-Based Web Scraping	- Allows data extraction from HTML structure	- Requires knowledge of HTML/CSS
API-Based Web Scraping	- Provides structured and easy-to-use data	- Dependent on the existence of an API
Machine Learning-Based Web Scraping	- Ability to extract complex data	- Requires expertise in machine learning
AI-Based Visual Scraping	- Can extract data from visual content	- High technical complexity

State of the art

- **Web scraping libraries**

Table 2. Web scraping libraries and frameworks across different languages

Language	Library or Framework	Advantages	Disadvantages
Java	Jsoup	-Easy HTML parsing -Handles malformed HTML well	- Does not support complex JavaScript pages
	Selenium WebDriver	- Supports dynamic content - Works with multiple browsers	- Slower due to reliance on browsers
Python	BeautifulSoup	- Simple syntax - Ideal for small projects	- May be slow on large datasets
	Scrapy	- Fast and scalable - Excellent for large-scale scraping	- Slower learning curve
JavaScript	Puppeteer	- Ideal for dynamic content - Support for headless browsers	- Consumes a lot of resources
	Cheerio	- Lightweight - Syntax similar to jQuery	- Limited to static content
PHP	Goutte	- Easy to use - Lightweight for simple tasks	- Limited features for complex projects

State of the art

- **Forms of competitive intelligence**
- **Active and Passive**
  - **Passive Competitive Intelligence:** This involves collecting information about competitors without a specific predefined objective. It involves observing and monitoring competitors' activities and performance without directly interacting with them.
  - **Active Competitive Intelligence**: In contrast to passive intelligence, active competitive intelligence aims to obtain specific information to produce knowledge and guide concrete actions. It involves direct interaction with competitors or their online resources.
- **Direct and Indirect**
  - **Direct Competitive Intelligence:** This targets competitors offering similar products and services. The focus is on activities and performance of these competitors since they share the same target market.
  - Indirect Competitive Intelligence: This focuses on competitors offering products or services that meet different customer needs, even if they are not direct competitors in the same market.
- **Mixed Competitive Intelligence:**
  - Combines passive and active or direct and indirect approaches to gather comprehensive and in-depth information about competitors.

Challenges to use web scraping in competitive intelligence

EAI AFRICOMM

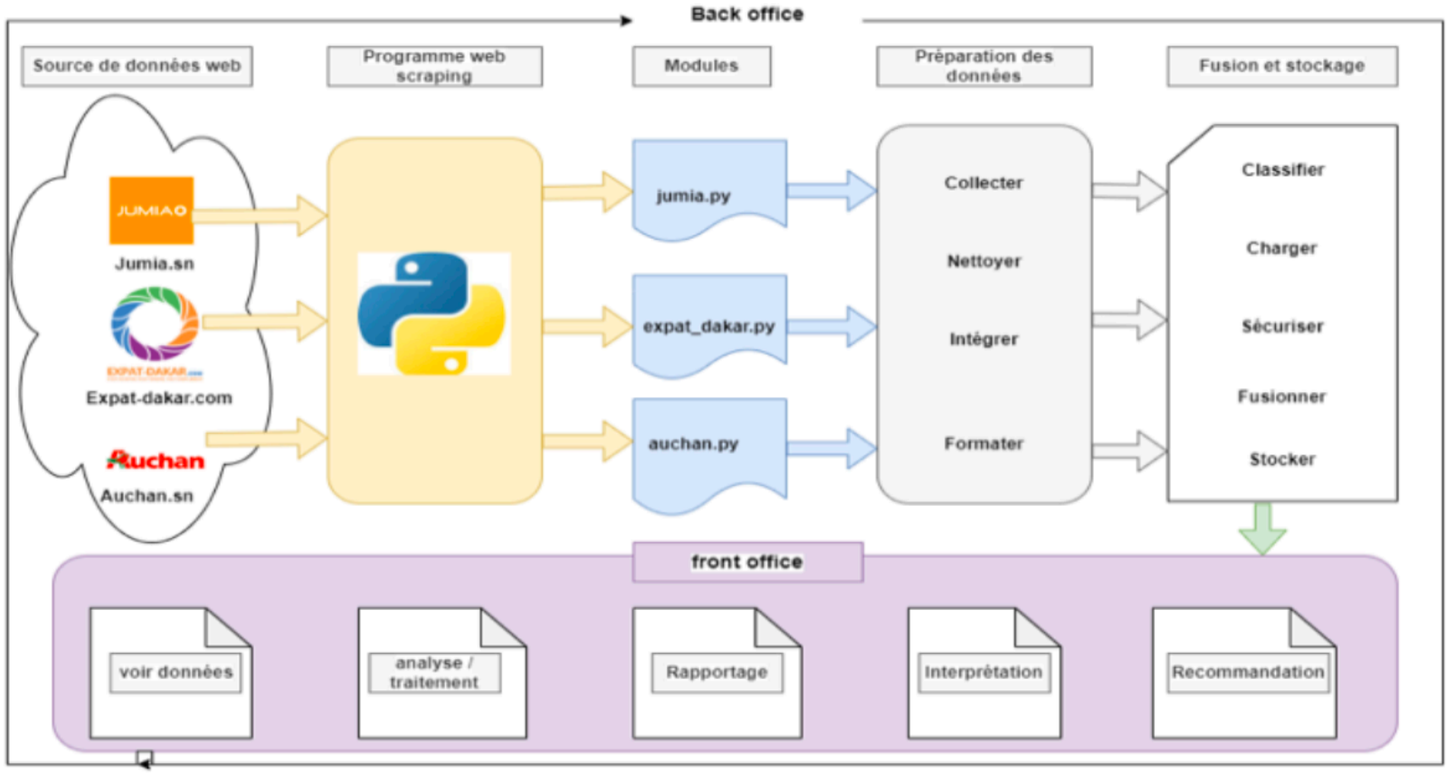
- **Challenges**
  - **Web scraping for good market trend monitoring**
    - By analyzing historical product data and predicting future movements, companies can proactively adjust their strategies
  - Web scraping to improve competitive analysis
    - By monitoring competitors' activities through web scraping, companies can gather valuable information about their strategies, performance, and innovations.
    - This data helps identify areas where competitors excel and those where they are vulnerable
  - Web scraping to improve price analysis:
    - Through web scraping, companies can collect data on competitors' prices, as well as price variations based on demand and seasons.
    - This information helps adjust their own product prices competitively and strategically
  - Web scraping to increase revenue:
    - To increase its revenue, organizations can implement the best frameworks based on available data. By analyzing this data, we can identify market opportunities, adjust our offerings, and optimize our processes to maximize sales and profitability
  - Web scraping to retain customers
    - By understanding the specific needs of each segment, we can offer products and services that precisely meet their expectations, thus strengthening customer satisfaction and fostering long-term loyalty

Difficulties to use web scraping in competitive intelligence

- **Difficulties**
  - **Diversity of data and sources (distribution of data across different sources)**
    - This advanced approach involves using multiple servers or scraping instances to collect data on a large scale simultaneously.
  - **Bypassing anti-scraping measures and tools**
    - Companies often face limitations imposed by target websites, such as the use of CAPTCHAs and anti-scraping tools.
  - **Legality of data collection on pages**
    - With the introduction of strict data protection regulations, companies must ensure their scraping practices comply with applicable laws
  - **Ensuring the quality of collected data**
    - Data may be incomplete, inaccurate, or outdated, which can harm the quality of competitive analyses
  - **Costs of implementation and monitoring**
    - Large-scale web scraping can be costly in terms of computing resources and bandwidth, mainly due to the need to manage and process large amounts of data.
  - **Constraints related to personal data protection**
    - The use of web scraping techniques also raises ethical concerns [62]. Companies must ensure their practices respect the rights of content owners and users.

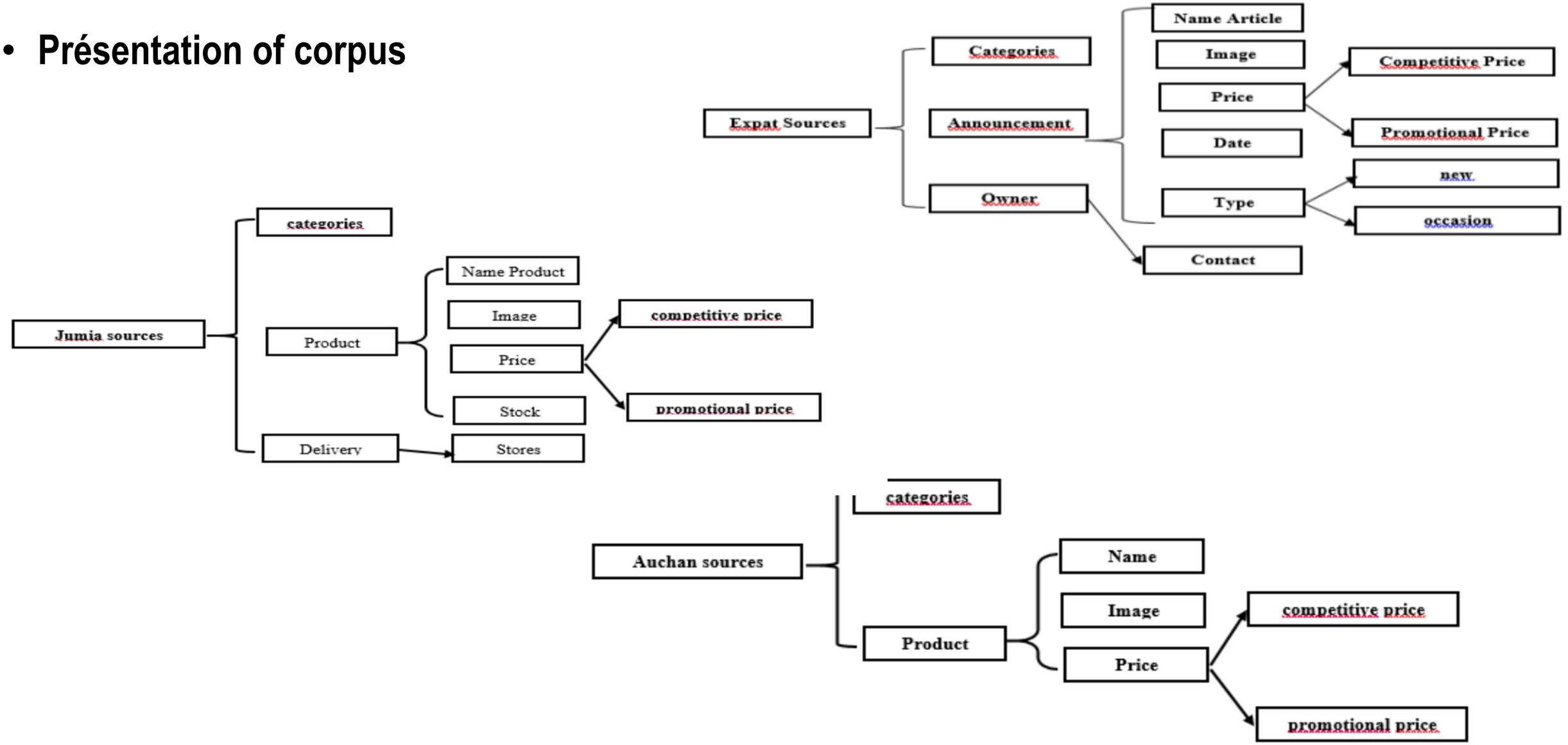
Case of study

- **Global Architecture (03 modules)**
  - **A Data Acquisition Module:** This module, based on intelligent web scraping, synchronously collects data from selected sources
  - **A Preprocessing Module:** This module cleans, standardizes, and merges data into a CSV file. Various Python modules are used to perform the necessary tasks.
  - **An Analysis and Presentation Module**: This module provides real-time decision-making information through graphical representations. Power BI is chosen here as the tool for visualizing decision-making information.



Case of study

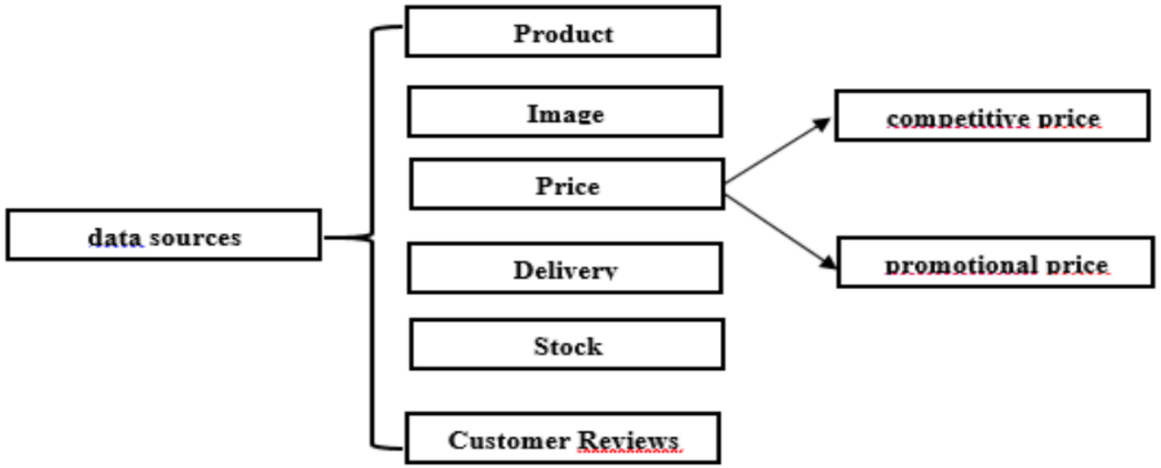
- **Présentation of corpus**



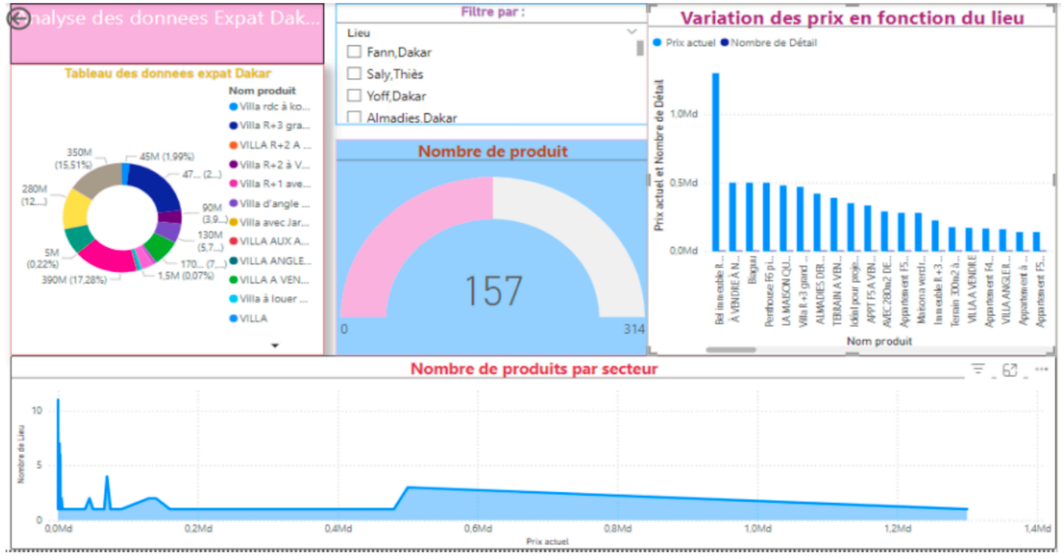
Case of study

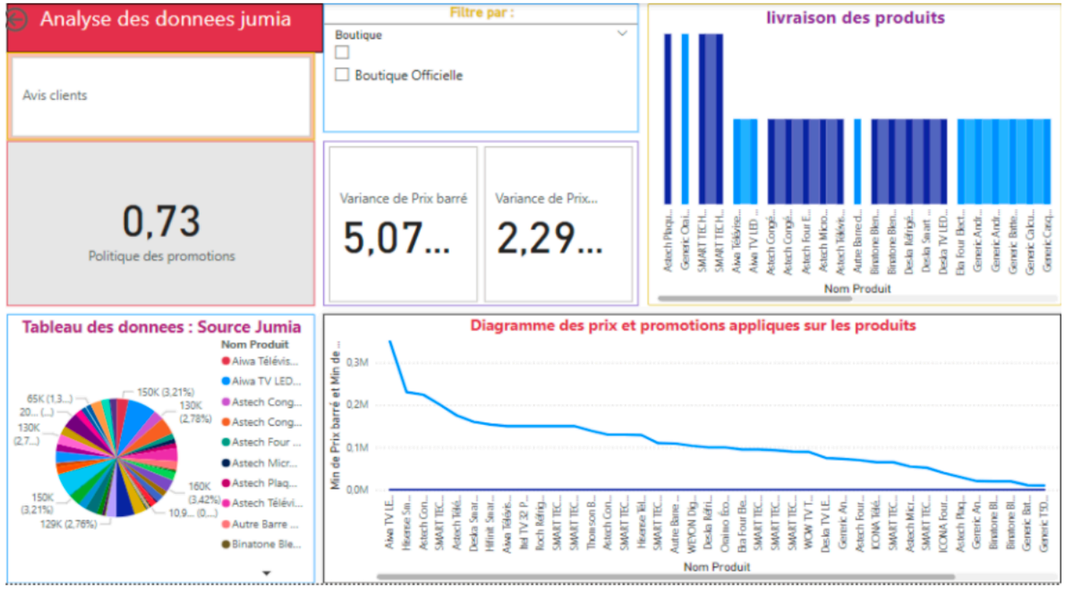
- **Compliance of corpus**

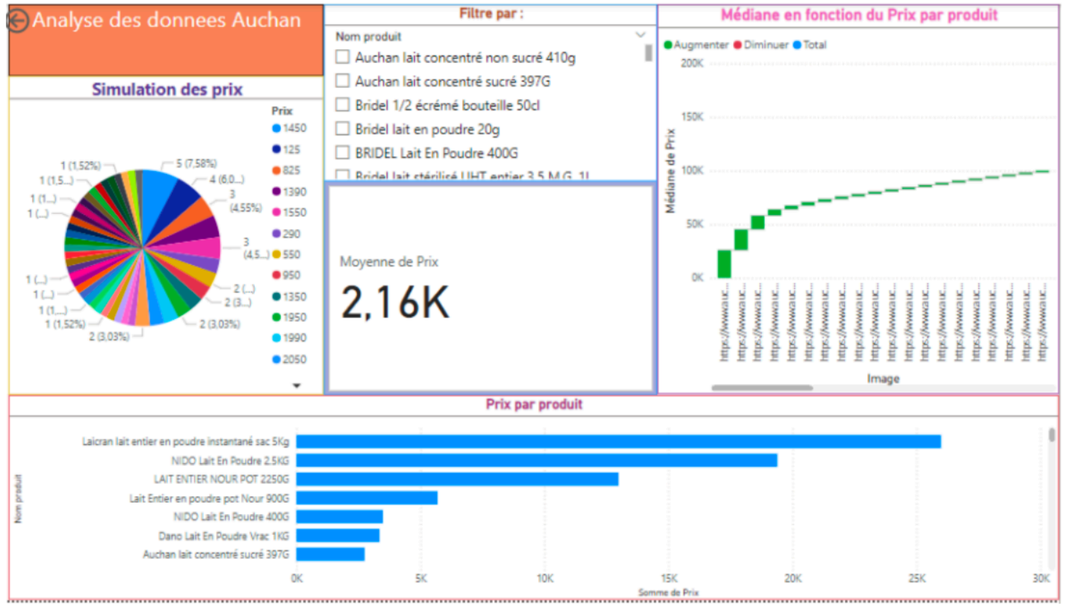
Auchan	Jumia	Expat-Dakar	Final Data
Article	Produit	Announcement	Produit
Nom Article	Nom Produit	Name Article	Nom produit
Prix	Prix	Price	Prix
Flash	Promotions	--	Promotions
Notes et Commande	Notes	--	Avis Client
Livraison	Livraison	Contact	Livraison

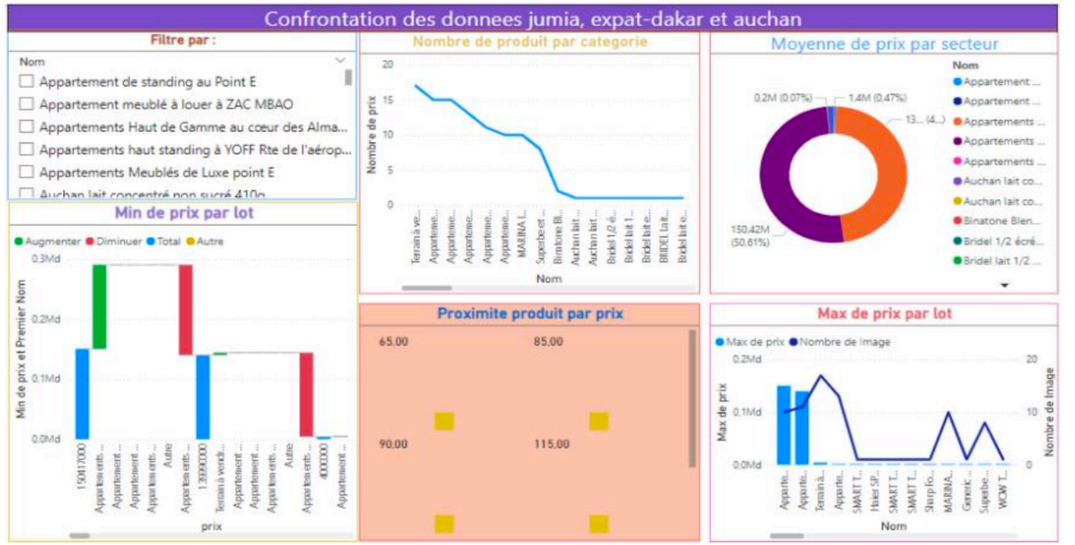


Results and discussions









Results and discussions

- Thus, to improve their position Jumia should strengthen the empowerment of local merchants by providing analytical tools and solid return policies.

- Auchan could adopt a more eco-responsible strategy by promoting local products and enhancing its loyalty program.

- Expat-Dakar should integrate more efficient payment and delivery solutions while offering targeted advertising options to increase the visibility of listings.

Results and discussions

- In this way, companies can collect various types of data, including product specifications, price information, user reviews, and social media mentions.
- This comprehensive data collection provides a holistic view of market dynamics and competitor tactics.
- Therefore, ensuring the accuracy and quality of the collected data is crucial.
- Companies need to implement rigorous data quality control measures, validate the data, and consolidate the information to ensure its reliability.
- To maximize their chances of success, organizations must demonstrate flexibility and innovation in their information gathering by integrating advanced data acquisition systems such as Natural Language Processing (NLP) and web crawlers.
- These tools enable more targeted and efficient data collection, even amid the growing abundance of available data.
- The integration of artificial intelligence and advanced machine learning techniques promises to further enhance these capabilities, allowing for more sophisticated analyses and more accurate market trend predictions. However, this technological evolution comes with significant challenges.
- In this context, web scraping will continue to play a central role in business strategies, helping companies navigate effectively in an increasingly complex and dynamic competitive environment.

Thank you for your attention