REPUBLIQUE DE CÔTE D’IVOIRE



UNION – DISCIPLINE - TRAVAIL



MINISTERE DE L’ECONOMIE DU PLAN ET DU DEVELOPPEMENT



ECOLE NATIONALE SUPERIEURE DE STATISTIQUE ET D’ECONOMIE APPLIQUEE

**Rapport détaillé du projet python**

**ANALYSE DES ATTENTES DES CONSOMMATEURS SUR LES VETEMENTS POUR HOMME : ÉTUDE DES AVIS ET DES COMMENTAIRES SUR JUMIA**

**Sous la supervision de :**

M. KOMENAN Cédric Wilfried

**Rédigé par :**

M. KONATE

Richard

*Elève Ingénieur Statisticien Economiste en première année (option économie)*

*Professeur d’algorithme et programmation*

**Janvier 2025**

Sommaire

[Sommaire 2](#_Toc186910535)

[Résumé 3](#_Toc186910536)

[Abstract 3](#_Toc186910537)

[Introduction 4](#_Toc186910538)

[Contexte et justification 4](#_Toc186910539)

[Problématique 5](#_Toc186910540)

[Questions de recherche 6](#_Toc186910541)

[Objectifs 6](#_Toc186910542)

[Chapitre I : Extraction des données 7](#_Toc186910543)

[I. Méthodologie d'extraction 7](#_Toc186910544)

[II. Technologies utilisées 8](#_Toc186910545)

[III. Résultats 9](#_Toc186910546)

[IV. Défis rencontrés 9](#_Toc186910547)

[Chapitre II : Traitement des données 10](#_Toc186910548)

[I. Méthodologie 10](#_Toc186910549)

[II. Technologies utilisées 11](#_Toc186910550)

[III. Étapes du traitement des données 11](#_Toc186910551)

[IV. Résultats 12](#_Toc186910552)

[V. Défis rencontrés 12](#_Toc186910553)

[Chapitre III : Analyse des données 13](#_Toc186910554)

[I. Méthodologie 13](#_Toc186910555)

[II. Technologies utilisées 14](#_Toc186910556)

[III. Étapes de l’analyse 14](#_Toc186910557)

[IV. Résultats 15](#_Toc186910558)

[V. Défis rencontrés 15](#_Toc186910559)

[Chapitre IV : Visualisation et interprétation 16](#_Toc186910560)

[I. Visualisation des résultats 16](#_Toc186910561)

[II. Interprétation des résultats 18](#_Toc186910562)

[Conclusion et recommandations 19](#_Toc186910563)

[Conclusion 19](#_Toc186910564)

[Recommandations 20](#_Toc186910565)

[Table des matières 21](#_Toc186910566)

Résumé

Cette étude vise à analyser les retours des consommateurs sur les produits disponibles sur Jumia afin de mieux comprendre leurs besoins et leurs attentes. Le processus s'est déroulé en quatre grandes étapes : l'extraction des données, le traitement des données, l'analyse des données, et la visualisation et l'interprétation des résultats.

Les résultats montrent que les Derbies, les Boots hommes, et les Bermudas homme sont les catégories les mieux notées, tandis que la qualité, la taille, et la matière des produits apparaissent comme des aspects essentiels dans les retours des consommateurs. L'analyse des bigrammes a également permis de détecter des motifs récurrents dans les commentaires, offrant des pistes concrètes pour améliorer les produits et leurs descriptions. Enfin, des recommandations sont proposées pour optimiser la satisfaction client et l’expérience d’achat en ligne.

Abstract

This study aims to analyze consumer feedback on products available on Jumia to better understand their needs and expectations. The process was divided into four main steps: data extraction, data processing, data analysis, and results visualization and interpretation.

The results reveal that Derbies, Boots hommes, and Bermudas homme are the highest-rated categories, while quality, size, and material are critical aspects mentioned in consumer feedback. The analysis of bigrams also identified recurring patterns in the comments, providing concrete directions to enhance products and their descriptions. Finally, recommendations are provided to optimize customer satisfaction and the online shopping experience.

Introduction

Contexte et justification

Dans un monde de plus en plus tourné vers le numérique, les plateformes de commerce électronique jouent un rôle central dans les décisions d'achat des consommateurs. En Afrique, des acteurs majeurs comme Jumia ont révolutionné la manière dont les consommateurs accèdent aux produits, notamment les vêtements. Les achats en ligne offrent une commodité indéniable, mais ils s'accompagnent également de défis spécifiques, notamment en matière de perception de la qualité, d'ajustement des tailles et de satisfaction globale.

Les avis en ligne laissés par les consommateurs sont devenus une source essentielle pour guider les décisions d'achat. Ces commentaires ne reflètent pas seulement les expériences des clients, mais contiennent également des informations précieuses sur leurs attentes, leurs besoins et leurs frustrations. Ces données, lorsqu'elles sont exploitées correctement, peuvent offrir des insights stratégiques pour les vendeurs et les plateformes e-commerce.

Le choix de se concentrer sur les vêtements pour homme sur Jumia repose sur plusieurs raisons :

* ***Importance stratégique :*** Les vêtements constituent une catégorie très concurrentielle, où la qualité perçue et la satisfaction des clients déterminent le succès des vendeurs.
* ***Accessibilité des données :*** Les avis et notes laissés par les clients sur Jumia offrent un corpus riche à analyser pour extraire des tendances.
* ***Valeur ajoutée :*** Les résultats de cette analyse peuvent être utilisés pour améliorer la qualité des produits, renforcer la satisfaction client et optimiser l'expérience utilisateur sur Jumia.

En combinant le scraping des avis, l’analyse des sentiments et la visualisation des données, ce projet vise à répondre à ces enjeux en fournissant une compréhension approfondie des perceptions et préférences des consommateurs.

Problématique

Dans un environnement d’achat en ligne en constante évolution, les consommateurs s’appuient de plus en plus sur les avis et évaluations pour guider leurs décisions. Les plateformes comme Jumia, qui proposent une large gamme de vêtements pour homme, offrent une mine d’informations sous la forme d’avis clients, mais ces données restent largement sous-exploitées.

Dans la catégorie des **vêtements pour homme**, plusieurs questions cruciales émergent :

* Quelles catégories de vêtements sont les plus appréciées ?
* Quels aspects des produits retiennent le plus l’attention des consommateurs dans leurs commentaires ?

Ces interrogations sont essentielles pour comprendre les attentes des consommateurs, identifier les facteurs qui influencent leur satisfaction, et orienter les décisions des vendeurs et de la plateforme. Cependant, l’absence d’une analyse structurée des avis rend difficile l’extraction de tendances claires, que ce soit par type de vêtement.

Problème clé :

Comment analyser les avis des consommateurs pour identifier :

* Les catégories les plus appréciées par les consommateurs sur Jumia ?
* Les aspects des produits qui retiennent le plus l’attention des consommateurs dans leurs commentaires ?

Ce projet vise à répondre à cette problématique en combinant des techniques de scraping, d’analyse de sentiments et de visualisation des données.

Questions de recherche

Les questions de recherche sont les suivantes :

***Question principale :***

* Quelles sont les catégories les plus appréciées par les consommateurs sur Jumia, ainsi que les aspects des produits qui retiennent le plus l’attention des consommateurs dans leurs commentaires ?

***Questions secondaires :***

* Quelles catégories de vêtements pour homme (chemises, pantalons, vestes, etc.) suscitent les notes moyennes les plus élevées ?
* Quels termes ou expressions reviennent fréquemment dans les avis positifs et négatifs (par exemple, "bonne qualité", "taille incorrecte") ?

Objectifs

***Objectif général***

* Connaitre les catégories les plus appréciées par les consommateurs sur Jumia, ainsi que les aspects des produits qui retiennent le plus l’attention des consommateurs dans leurs commentaires.

***Objectifs spécifiques***

* Déterminer les notes moyennes des vêtements par catégories de vêtements pour homme (chemises, pantalons, vestes, etc.), et les classer par ordre décroissant.
* Déterminer les termes les plus récurrents dans les commentaires afin de savoir aspects des produits qui retiennent le plus l’attention des consommateurs dans leurs commentaires.

# Extraction des données

L'extraction des données constitue une étape essentielle dans le processus d'analyse. Dans ce projet, les données ont été collectées à partir du site Jumia, une plateforme de commerce en ligne, pour analyser les avis clients et les préférences liées aux vêtements pour homme. Cette étape a été réalisée en deux phases distinctes : la collecte des liens des produits par catégorie et l'extraction des détails des produits.

## Méthodologie d'extraction

### Phase 1 : Collecte des liens des produits par catégorie

Lors de cette phase, l'objectif était de collecter les liens des produits disponibles sur Jumia, organisés par catégorie (chemises, pantalons, vestes, etc.). Pour ce faire, un script Python utilisant les bibliothèques Selenium et BeautifulSoup a été développé.

* Configuration du WebDriver :
  + Selenium a été configuré avec Chrome WebDriver pour interagir avec les pages dynamiques de Jumia.
  + Un user-agent spécifique a été utilisé pour éviter les restrictions liées au scraping.
* Récupération des liens des catégories : À partir de la page principale, les liens des différentes catégories ont été extraits. Le script gérait les pop-ups contextuels pour garantir un flux de navigation fluide.
* Extraction des produits par catégorie : Chaque catégorie a été explorée pour collecter les liens des produits et leurs informations de base (nom, prix, nombre d'avis).
* Sauvegarde des résultats : Les données collectées ont été sauvegardées dans un fichier JSON (product\_links.json), organisé par catégorie pour faciliter les étapes ultérieures.

### Phase 2 : Extraction des détails des produits

Cette phase visait à enrichir les données des produits en collectant les avis clients, les notes moyennes et le nombre total d'avis.

* Chargement des liens des produits : Les liens des produits extraits à l'étape précédente ont été chargés à partir du fichier JSON.
* Accès aux pages des produits : Chaque lien de produit a été visité pour collecter des informations supplémentaires, notamment :
  + Les commentaires clients.
  + Les notes moyennes.
  + Le nombre d'avis.
* Gestion des commentaires multi-pages : Un mécanisme a été implémenté pour naviguer à travers les pages suivantes des commentaires lorsque celles-ci étaient disponibles.
* Sauvegarde des données : Les commentaires et métadonnées collectés ont été sauvegardés dans des fichiers JSON spécifiques à chaque catégorie (comments\_<categorie>.json).

## Technologies utilisées

Les bibliothèques et outils suivants ont été utilisés pour l'extraction des données :

* Selenium : Pour interagir avec les pages dynamiques et naviguer sur le site.
* BeautifulSoup : Pour analyser le code HTML des pages et extraire les informations pertinentes.
* Python : Pour écrire et exécuter les scripts d'extraction.
* JSON : Pour organiser et stocker les données collectées.

## Résultats

A l’issue de la phase 1 :

* Nombre total de catégories explorées : 12.
* Nombre de produits collectés : environ 6 000.
* Résultat stocké dans product\_links.json.

A l’issue de la phase 2 :

* Nombre total de commentaires collectés : plus de 10 000.
* Données enrichies avec des notes moyennes et des métadonnées.
* Résultat stocké dans des fichiers JSON spécifiques par catégorie.

## Défis rencontrés

**Fenêtres contextuelles (pop-ups) :**

* Ces fenêtres interrompaient le flux d'extraction.
* Solution : un script a été développé pour détecter et interagir avec ces éléments automatiquement.

**Structure dynamique des pages web :**

* Les changements fréquents dans la structure HTML rendaient l'extraction délicate.
* Solution : des sélecteurs CSS robustes ont été utilisés et régulièrement mis à jour.

**Temps d'attente pour le chargement des pages :**

* Les pages nécessitaient un temps de chargement variable.
* Solution : utilisation de time.sleep et des attentes explicites de Selenium.

En somme, l'extraction des données a permis de collecter une quantité substantielle d'informations pertinentes pour les étapes suivantes de traitement et d'analyse. Cette approche méthodique garantit des données structurées et exploitables, alignées avec les objectifs du projet.

# Traitement des données

Après l'extraction des données, il est crucial de les nettoyer et de les structurer pour garantir leur qualité et leur pertinence dans les phases ultérieures d'analyse. Ce chapitre présente les différentes étapes du traitement des données, depuis leur fusion jusqu'à leur nettoyage, tout en prenant en compte des aspects tels que la gestion des champs numériques, le traitement des textes et la prise en compte des négations dans les commentaires.

## Méthodologie

### Fusion des données

Les données provenant des produits (liens, prix, avis) et des commentaires (contenus, titres, métadonnées) ont été fusionnées pour obtenir un ensemble cohérent et structuré. Cette fusion a permis d'associer chaque produit à ses commentaires respectifs.

### Nettoyage des champs numériques

Deux champs importants, average\_rating (note moyenne) et total\_reviews (nombre total d'avis), ont été convertis dans des formats numériques exploitables :

average\_rating a été converti en float pour permettre des calculs précis.

total\_reviews a été converti en int en extrayant uniquement les valeurs numériques des textes (exemple : 71 avis vérifiés).

### Traitement des textes

Les commentaires ont été traités avec spaCy, une bibliothèque de traitement de langage naturel, pour :

Nettoyer les textes : suppression des caractères spéciaux, conversion en minuscules.

Prendre en compte les négations : association des mots de négation avec les termes qu’ils modifient (exemple : "ne pas content" devient "content\_ne").

Supprimer les mots non pertinents (stop-words) à l’aide d’une liste personnalisée.

## Technologies utilisées

Les outils suivants ont été mobilisés :

Python : pour la manipulation des données et l'automatisation des tâches.

spaCy : pour le traitement des textes et l'analyse linguistique, utilisé car plus adaptée à la langue françaises qui est la langue dans laquelle les commentaires sont écrits.

JSON : pour structurer et stocker les données finales.

## Étapes du traitement des données

### Chargement des données

Les données des produits et des commentaires ont été chargées à partir de fichiers JSON. Les liens des produits ont été organisés par catégorie, et les commentaires des clients ont été regroupés par produit.

### Conversion et nettoyage des données

Conversion des champs numériques :

* Les notes moyennes (average\_rating) ont été converties en nombres flottants pour les calculs statistiques.
* Les nombres d'avis (total\_reviews) ont été extraits des chaînes de caractères et convertis en entiers.

Nettoyage des textes :

* Les titres (title) et contenus (content) des commentaires ont été analysés avec spaCy.
* Les mots de négation ont été associés aux termes qu’ils modifient pour préserver le sens du texte.
* Les mots non pertinents (stop-words personnalisés) ont été exclus.

### Fusion des données

Les données des produits et des commentaires ont été fusionnées pour obtenir un fichier unique contenant toutes les informations nécessaires.

## Résultats

Volume de données traitées :

* Nombre total de produits : environ 6 000.
* Nombre total de commentaires : plus de 10 000.

Fichier final : Les données fusionnées et nettoyées ont été sauvegardées dans un fichier JSON (data\_fusion\_cleaned.json), prêtes pour l’analyse.

## Défis rencontrés

**Gestion des données incomplètes :** Certains produits n’avaient pas de commentaires ou de notes. Une logique a été implémentée pour ignorer ces cas sans interrompre le traitement.

**Prise en compte des négations :** Les commentaires contenant des négations nécessitaient une approche spécifique pour conserver le sens des phrases.

**Performance :** Le traitement d’un grand volume de données a nécessité l’optimisation des boucles et la gestion efficace de la mémoire.

En conclusion, il convient de retenir que le processus de traitement a permis de structurer et d’enrichir les données collectées pour répondre aux exigences de l’analyse. Les données finales sont prêtes pour les étapes d’analyse des préférences et des sentiments des consommateurs.

# Analyse des données

L’analyse des données vise à extraire des informations pertinentes pour répondre aux objectifs de l’étude. Cette étape exploite les données traitées pour :

* Déterminer les catégories les plus appréciées par les consommateurs à travers le calcul des notes moyennes pondérées.
* Identifier les termes et expressions récurrents dans les commentaires afin de mieux comprendre les aspects des produits les plus importants aux yeux des consommateurs.

## Méthodologie

### Calcul des notes moyennes pondérées

Les notes moyennes pondérées par catégorie sont calculées en tenant compte du nombre d’avis pour chaque produit. Cela permet de valoriser les produits ayant reçu un grand nombre de commentaires et de réduire l’influence des produits avec peu d’avis.

### Analyse des termes fréquents (unigrammes et bigrammes)

* Unigrammes : Identifie les mots isolés les plus fréquents dans les commentaires des consommateurs. Cela donne un aperçu des thèmes principaux.
* Bigrammes : Analyse les paires de mots consécutifs les plus fréquents. Cela permet de capturer des expressions clés (par exemple : "bonne qualité", "taille correcte").

### Visualisation des résultats

Les résultats sont présentés sous forme de graphiques :

* Un nuage de mots pour chaque catégorie, montrant les termes fréquents (unigrammes).
* Un graphique à barres horizontales pour visualiser les bigrammes les plus courants.
* Un graphique des notes pondérées par catégorie.

## Technologies utilisées

**Python :** Pour le traitement et l’analyse des données.

**spaCy :** Pour l’analyse linguistique avancée.

**matplotlib :** Pour la création de graphiques.

**WordCloud :** Pour générer des nuages de mots.

## Étapes de l’analyse

**Chargement des données :** Les données des catégories et des produits nettoyés ont été chargées à partir des fichiers JSON générés lors des étapes précédentes.

**Calcul des notes moyennes pondérées :** Les notes moyennes pondérées par catégorie ont été calculées en fonction des données suivantes :

* Note moyenne (average\_rating) : La moyenne des évaluations des utilisateurs pour un produit.
* Nombre d’avis (total\_reviews) : Le nombre total de commentaires pour un produit.

**Analyse des termes fréquents :**

* Unigrammes : Les commentaires des produits sont analysés pour extraire les mots les plus fréquents. Les stop-words (mots non pertinents) sont filtrés à l’aide d’une liste personnalisée.
* Bigrammes : Les paires de mots consécutifs sont identifiées dans les commentaires pour capturer des expressions significatives. Les stop-words sont également exclus.

**Visualisation des résultats :**

* Un graphique des notes pondérées par catégorie met en évidence les catégories les mieux notées.
* Un nuage de mots par catégorie illustre les termes les plus fréquents.
* Un graphique des bigrammes fréquents montre les expressions clés.

## Résultats

**Notes moyennes pondérées par catégorie :**

* Les catégories ayant les meilleures notes moyennes pondérées reflètent les préférences des consommateurs.
* Exemple : Les chemises casuals et les pantalons ont une note moyenne élevée, indiquant leur popularité.

**Termes fréquents (unigrammes et bigrammes) :**

* Les unigrammes révèlent des thèmes récurrents comme la "taille", la "qualité", et le "tissu".
* Les bigrammes mettent en évidence des expressions comme "bonne qualité" et "taille parfaite".

**Visualisation :**

* Les nuages de mots montrent les termes dominants pour chaque catégorie.
* Les graphiques des bigrammes offrent une vue plus détaillée des préoccupations ou des appréciations des consommateurs.

## Défis rencontrés

**Précision de l’analyse linguistique :**

* Certains commentaires qui comportaient des fautes d’orthographes, rendant plus complexe le traitement des commentaires.
* La gestion des négations dans les commentaires a nécessité une configuration spécifique pour capturer le sens exact des phrases.
* Les bigrammes peuvent inclure des combinaisons moins pertinentes. Des ajustements supplémentaires des stop-words peuvent être nécessaires.

**Volume des données :** Le traitement de plusieurs milliers de commentaires a demandé une gestion efficace des ressources.

Au terme de ce chapitre, retenons que l’analyse des données a permis d’obtenir des informations précieuses sur les catégories les mieux notées et les aspects des produits les plus importants pour les consommateurs. Ces résultats seront exploités dans la phase de visualisation et d’interprétation pour guider les recommandations finales.

# Visualisation et interprétation

La visualisation des données a pour objectif de transformer les résultats analytiques en représentations graphiques claires et accessibles, permettant d’interpréter facilement les préférences des consommateurs et les aspects des produits les plus importants. Dans ce chapitre, nous analyserons les données à travers les visualisations suivantes :

* Les notes moyennes pondérées par catégorie pour identifier les catégories les plus appréciées.
* Les termes fréquents (unigrammes) et expressions fréquentes (bigrammes) pour comprendre les préoccupations des consommateurs.

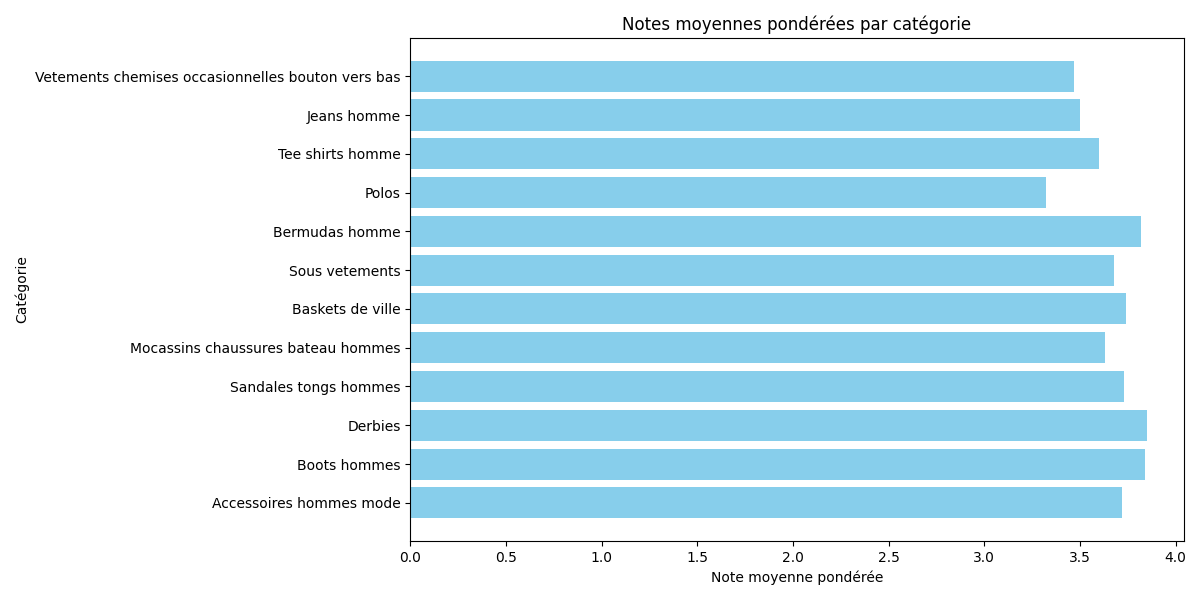
## Visualisation des résultats

### Notes moyennes pondérées par catégorie

Les notes moyennes pondérées reflètent les préférences globales des consommateurs pour chaque catégorie. Les résultats montrent que les catégories Derbies et Boots hommes sont les mieux notées, avec des notes moyennes respectives de 3.85 et 3.84, tandis que la catégorie Polos se trouve en bas de classement avec une note moyenne de 3.32.

Voici les catégories triées par ordre décroissant de leur note moyenne pondérée :

* Derbies : 3.85
* Boots hommes : 3.84
* Bermudas homme : 3.82
* Baskets de ville : 3.74
* Sandales tongs hommes : 3.73
* Accessoires hommes mode : 3.72
* Sous-vêtements : 3.68
* Mocassins chaussures bateau hommes : 3.63
* Tee-shirts homme : 3.6
* Jeans homme : 3.5
* Chemises occasionnelles : 3.47
* Polos : 3.32



Les notes moyennes pondérées sont représentées dans un graphique à barres. Ce graphique permet d’identifier rapidement les catégories les plus populaires auprès des consommateurs.

### Analyse des termes fréquents par catégorie

Les unigrammes (termes uniques) permettent de repérer les thèmes récurrents dans les commentaires. Voici quelques observations :

* Dans la catégorie Chemises occasionnelles, les termes "taille", "chemise", et "tissu" apparaissent fréquemment, ce qui indique que ces aspects sont cruciaux pour les consommateurs.
* Pour les Tee-shirts homme, des termes comme "qualité", "taille", et "coton" sont les plus fréquents, mettant en avant les préoccupations concernant la matière et la taille.
* Dans la catégorie Accessoires hommes mode, le mot "ceinture" ressort fortement, ce qui reflète l’intérêt pour cet article spécifique.

Les nuages de mots générés pour chaque catégorie permettent de visualiser ces résultats de manière claire et esthétique. Les termes les plus fréquents y sont représentés avec une taille proportionnelle à leur fréquence d’apparition.

### Analyse des bigrammes fréquents par catégorie

Les bigrammes (paires de mots consécutifs) révèlent des expressions clés dans les commentaires. Voici quelques exemples intéressants :

* Dans la catégorie Chemises occasionnelles, des expressions comme "bel chemise" et "taille petit" indiquent à la fois des points positifs et des problèmes récurrents.
* Pour les Baskets de ville, des bigrammes tels que "bel chaussure" et "chaussure confortable" montrent que l’esthétique et le confort sont des préoccupations majeures.
* Dans la catégorie Accessoires hommes mode, des bigrammes comme "bel ceintur" et "ceinture cuir" mettent en avant la qualité et le style des produits.

Ces résultats sont visualisés à l’aide de graphiques à barres horizontales, où chaque bigramme est représenté avec sa fréquence d’occurrence.

## Interprétation des résultats

### Préférences des consommateurs

Les résultats montrent que les consommateurs apprécient particulièrement les catégories Derbies, Boots, et Bermudas, qui combinent style, confort et qualité. Les catégories avec des notes plus faibles, comme les Polos, pourraient nécessiter des améliorations sur certains aspects tels que la qualité du tissu ou la coupe.

### Préoccupations majeures

L’analyse des unigrammes et des bigrammes montre que les consommateurs se préoccupent principalement de la qualité, de la taille, et du confort. Ces aspects apparaissent dans presque toutes les catégories, avec des variations selon les produits :

* Pour les vêtements, les termes liés à la taille et au tissu sont récurrents.
* Pour les chaussures, des termes comme "confortable", "léger", et "cuir" reviennent souvent.
* Pour les accessoires, les consommateurs mentionnent fréquemment la "qualité" et la "matière" des produits.

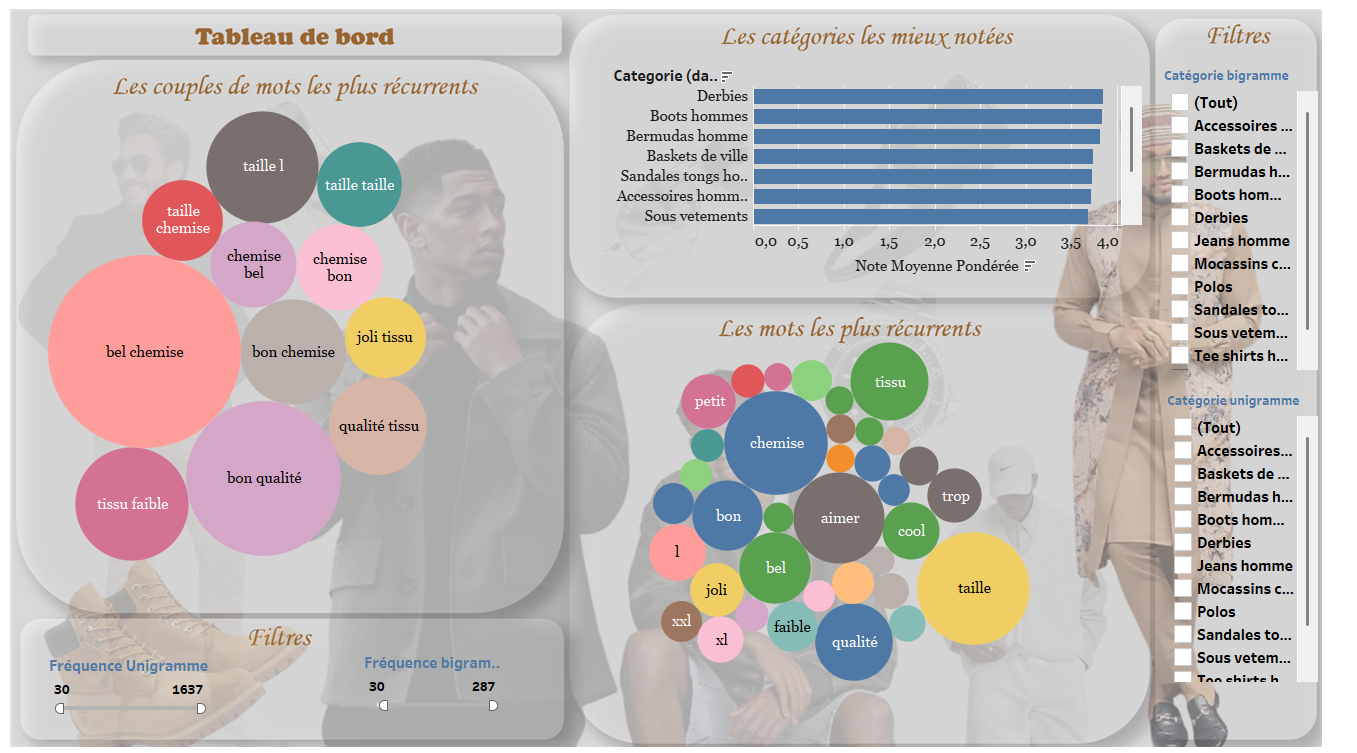
### Points d’amélioration

Les expressions négatives identifiées dans les bigrammes, comme "taille petit" ou "tissu faible", signalent des problèmes spécifiques à résoudre pour améliorer la satisfaction des consommateurs.

Au terme de ce chapitre, il semble prépondérant de savoir que les visualisations et les analyses effectuées dans ce chapitre offrent des insights précieux sur les préférences et les préoccupations des consommateurs. Ces informations peuvent guider les décisions stratégiques pour améliorer les produits, affiner les descriptions en ligne, et renforcer la satisfaction des clients.

Le lien vers le Dashboard est le suivant :

<https://public.tableau.com/views/Dashboard_17360088295270/Overview?:language=fr-FR&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>



Conclusion et recommandations

Conclusion

L'analyse des données issues de l'extraction, du traitement, et de l'analyse a permis de mettre en évidence des éléments clés sur les préférences et les préoccupations des consommateurs de produits sur Jumia. Les points principaux de cette étude sont :

* Les notes moyennes pondérées par catégorie montrent que les catégories les mieux appréciées sont les Derbies, les Boots hommes, et les Bermudas homme, indiquant une satisfaction élevée pour ces produits. En revanche, les catégories comme les Polos présentent des opportunités d'amélioration.
* L'analyse des termes fréquents (unigrammes) a révélé que des aspects comme la taille, la qualité, et la matière sont les plus souvent mentionnés, ce qui reflète les attentes majeures des consommateurs.
* L'analyse des bigrammes fréquents a permis d'identifier des expressions spécifiques, telles que "taille petit" ou "qualité prix", qui mettent en avant les points positifs et négatifs des produits.

Ces résultats montrent l'importance d'un suivi attentif des retours des consommateurs pour mieux comprendre leurs besoins et adapter les produits ou les services en conséquence.

Recommandations

Sur la base des résultats obtenus, voici des recommandations pour améliorer la satisfaction des consommateurs et maximiser les performances des produits sur la plateforme :

### Améliorations des produits

* Pour les chemises :
  + Mieux ajuster les tailles pour éviter les plaintes concernant des produits "trop petits" ou "pas conformes".
  + Mettre l'accent sur la qualité des tissus, en évitant les matières perçues comme "faibles" ou "fragiles".
* Pour les chaussures :
  + Prioriser des modèles offrant confort et légèreté, tout en maintenant un bon rapport qualité/prix.
  + Inclure des informations détaillées sur la matière (ex. cuir véritable) pour répondre aux attentes de durabilité.
* Pour les accessoires : Garantir des matériaux de qualité pour les ceintures et autres articles. Les mots "cuir" et "conforme" ressortent fortement comme critères essentiels.

### Optimisation des descriptions de produits

* Mettre en avant dans les descriptions les points positifs identifiés dans les commentaires, comme la qualité, le style, ou la matière.
* Fournir des informations précises sur les tailles, en proposant par exemple un guide des tailles pour chaque produit.
* Inclure les détails sur les matériaux et la fabrication, qui sont des préoccupations majeures des consommateurs.

### Gestion des retours et de la satisfaction client

* Analyser les retours contenant des expressions négatives comme "tissu faible" ou "taille petit" pour identifier les causes récurrentes d'insatisfaction.
* Encourager les consommateurs satisfaits à laisser des commentaires positifs, ce qui peut renforcer la crédibilité des produits auprès des futurs acheteurs.

### Suivi et mise à jour

Mettre à jour les catégories et descriptions des produits en fonction des retours obtenus pour rester aligné avec les besoins des consommateurs.

Réaliser des analyses périodiques des données pour suivre les tendances et anticiper les évolutions des attentes des clients.

Table des matières

[Sommaire 2](#_Toc186910370)

[Résumé 3](#_Toc186910371)

[Abstract 3](#_Toc186910372)

[Introduction 4](#_Toc186910373)

[Contexte et justification 4](#_Toc186910374)

[Problématique 4](#_Toc186910375)

[Questions de recherche 5](#_Toc186910376)

[Objectifs 6](#_Toc186910377)

[Chapitre I : Extraction des données 7](#_Toc186910378)

[I. Méthodologie d'extraction 7](#_Toc186910379)

[1. Phase 1 : Collecte des liens des produits par catégorie 7](#_Toc186910380)

[2. Phase 2 : Extraction des détails des produits 7](#_Toc186910381)

[II. Technologies utilisées 8](#_Toc186910382)

[III. Résultats 8](#_Toc186910383)

[IV. Défis rencontrés 8](#_Toc186910384)

[Chapitre II : Traitement des données 10](#_Toc186910385)

[I. Méthodologie 10](#_Toc186910386)

[1. Fusion des données 10](#_Toc186910387)

[2. Nettoyage des champs numériques 10](#_Toc186910388)

[3. Traitement des textes 10](#_Toc186910389)

[II. Technologies utilisées 11](#_Toc186910390)

[III. Étapes du traitement des données 11](#_Toc186910391)

[1. Chargement des données 11](#_Toc186910392)

[2. Conversion et nettoyage des données 11](#_Toc186910393)

[3. Fusion des données 12](#_Toc186910394)

[IV. Résultats 12](#_Toc186910395)

[V. Défis rencontrés 12](#_Toc186910396)

[Chapitre III : Analyse des données 13](#_Toc186910397)

[I. Méthodologie 13](#_Toc186910398)

[1. Calcul des notes moyennes pondérées 13](#_Toc186910399)

[2. Analyse des termes fréquents (unigrammes et bigrammes) 13](#_Toc186910400)

[3. Visualisation des résultats 13](#_Toc186910401)

[II. Technologies utilisées 13](#_Toc186910402)

[III. Étapes de l’analyse 14](#_Toc186910403)

[IV. Résultats 14](#_Toc186910404)

[V. Défis rencontrés 15](#_Toc186910405)

[Chapitre IV : Visualisation et interprétation 16](#_Toc186910406)

[I. Visualisation des résultats 16](#_Toc186910407)

[1. Notes moyennes pondérées par catégorie 16](#_Toc186910408)

[2. Analyse des termes fréquents par catégorie 17](#_Toc186910409)

[3. Analyse des bigrammes fréquents par catégorie 17](#_Toc186910410)

[II. Interprétation des résultats 18](#_Toc186910411)

[1. Préférences des consommateurs 18](#_Toc186910412)

[2. Préoccupations majeures 18](#_Toc186910413)

[3. Points d’amélioration 18](#_Toc186910414)

[Conclusion et recommandations 19](#_Toc186910415)

[Conclusion 19](#_Toc186910416)

[Recommandations 19](#_Toc186910417)

[1. Améliorations des produits 20](#_Toc186910418)

[2. Optimisation des descriptions de produits 20](#_Toc186910419)

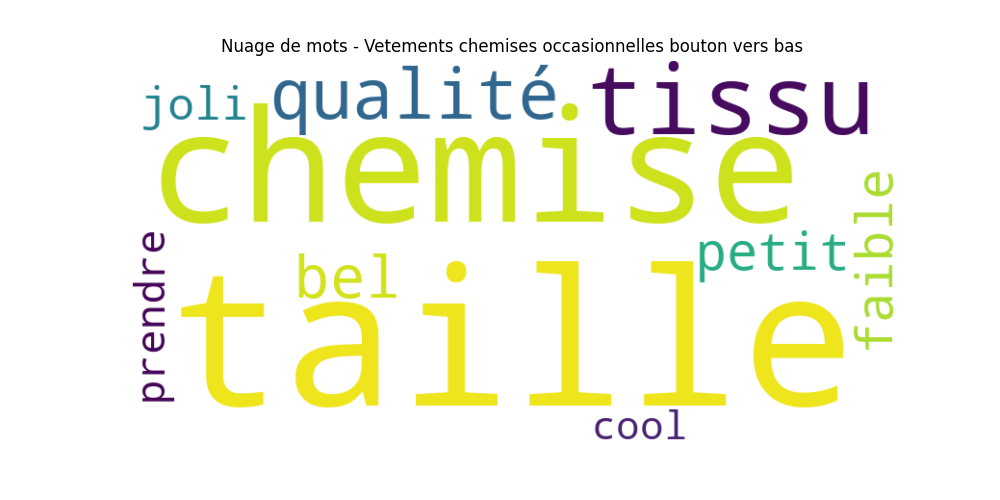
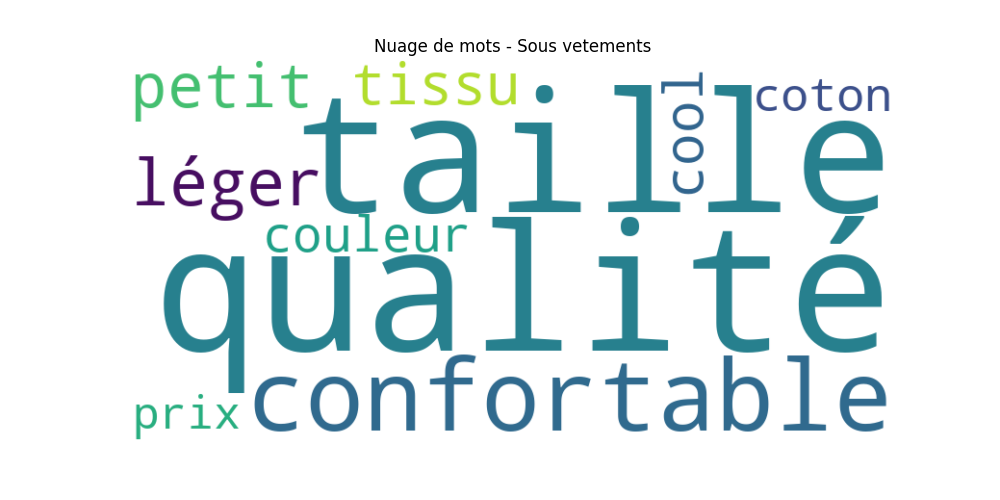
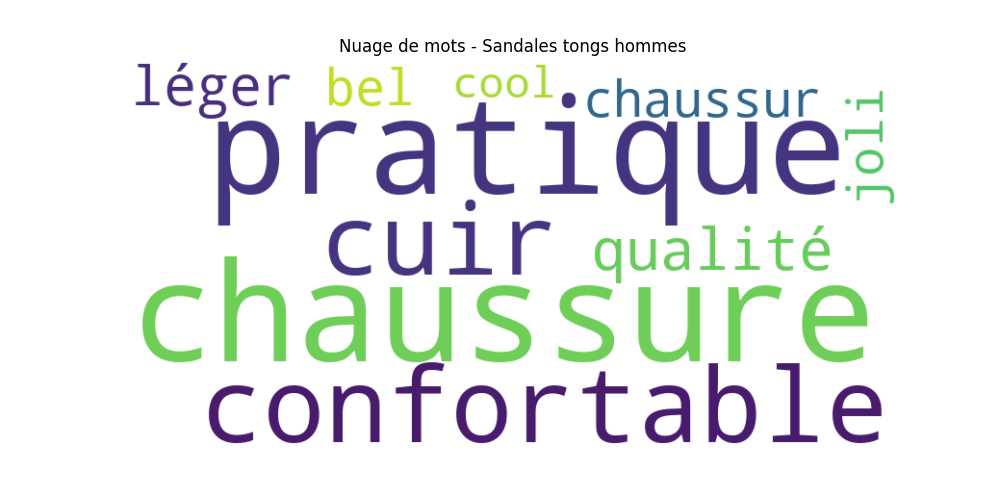
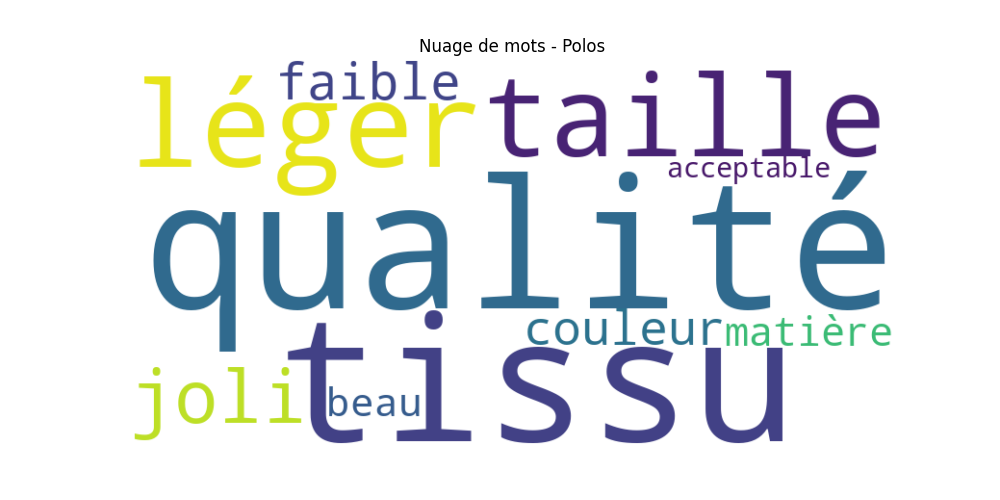
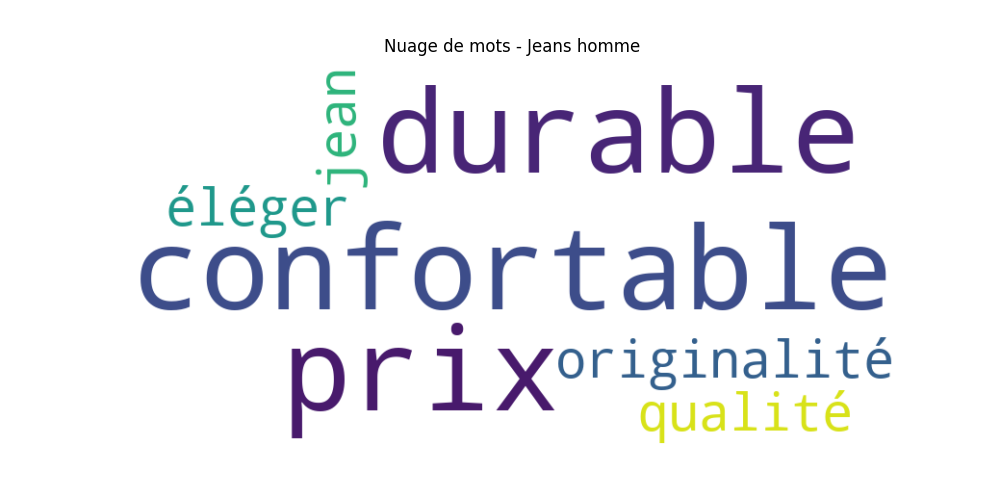
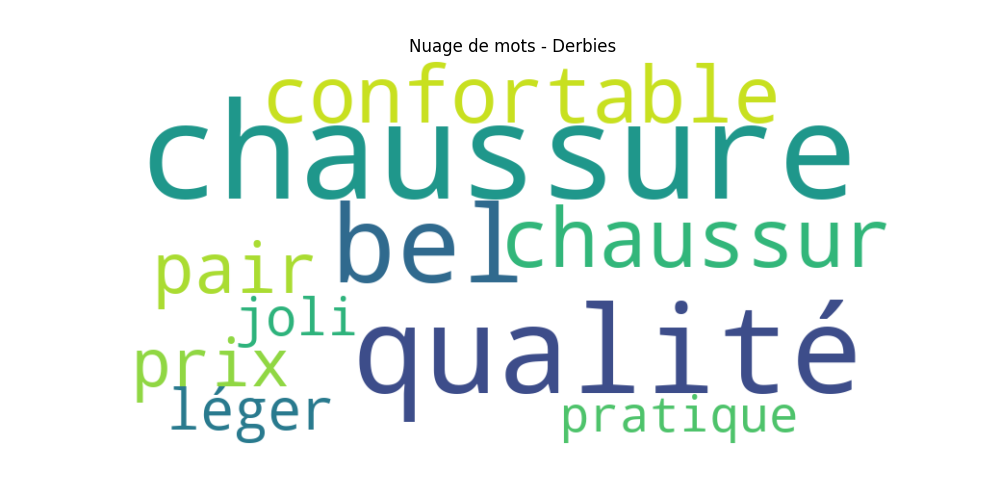
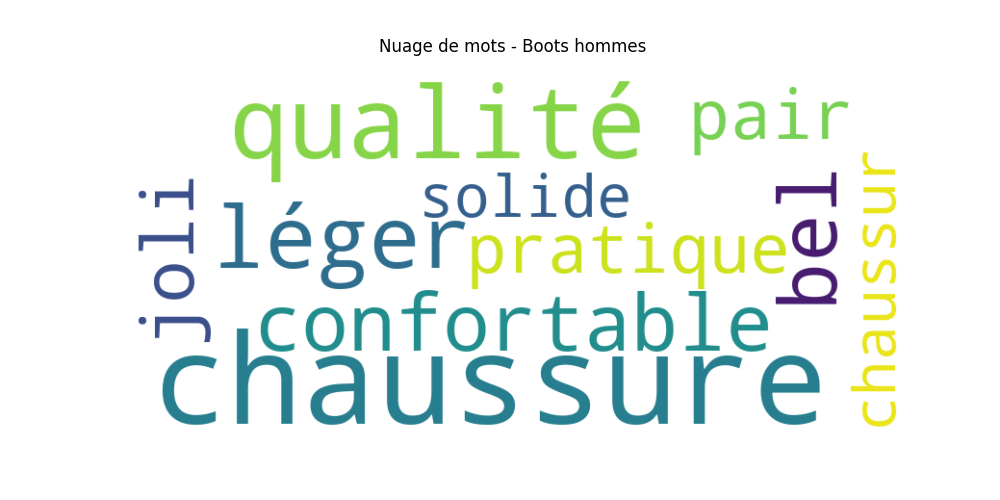
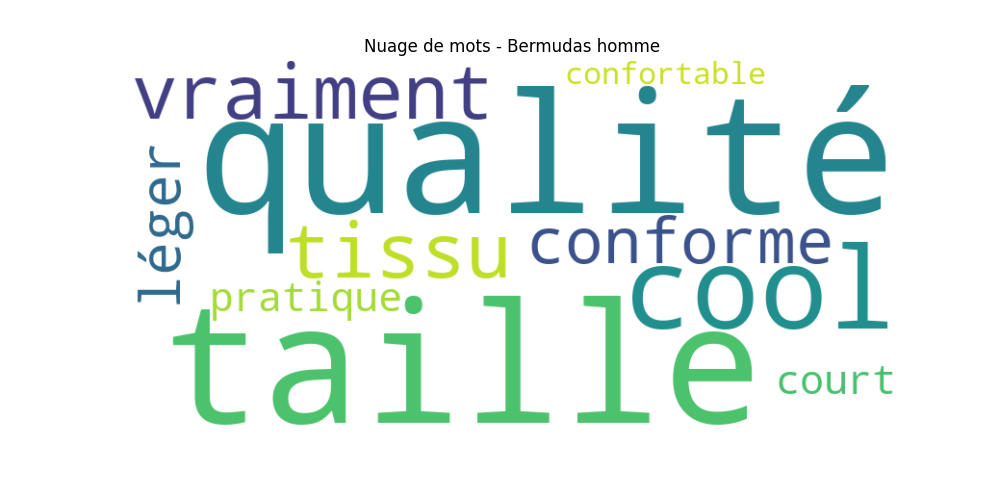
[3. Gestion des retours et de la satisfaction client 20](#_Toc186910420)

[4. Suivi et mise à jour 20](#_Toc186910421)

[Table des matières 21](#_Toc186910422)

Annexe

Annexe 1 : Unigrammes



Annexe 2 : Bigrammes

