# Analyse du panier de consommation (Wordcloud)

# 

## Introduction

# 

L’objectif de ce rapport est de décrire un outil qui utilise des techniques de traitement de texte pour répondre aux besoins de notre client.

Il y a 6 projets, ici sera décrit un des projets.

Dans le secteur de la vente au détail, cette analyse est utilisée pour découvrir des modèles et des relations au sein des transactions et des ensembles d'éléments de transaction (articles fréquemment achetés ensemble).

Dans un premier temps, nous expliquerons les modèles choisis pour trois pays et les méthodes mises en place afin de permettre le regroupement des produits les plus achetés.

Dans un second temps, nous verrons comment généraliser les modèles à plus de pays.

Nous analyserons en détail les problèmes pouvant intervenir avant de conclure sur nos recommandations quant à l’utilisation des différentes méthodes et caractéristiques pour le traitement de texte dans le secteur de la vente.

## L’ensemble de données

Pour effectuer notre analyse, nous utiliserons l'ensemble de données de vente au détail en ligne. Il s'agit d'un ensemble de données contenant des transactions transnationales effectuées sur un magasin de détail en ligne. Il comprend 45 985 lignes, avec 8 features:

* **InvoiceNo :**

Un numéro unique à 6 chiffres est attribué à chaque transaction. Si ce code commence par la lettre "C", la commande a été annulée.

* **StockCode :**

Un numéro unique à 5 chiffres attribué à chaque produit distinct.

* **Description:**

Le nom du produit.

* **Quantité :**

Le nombre de chaque produit (article) acheté par transaction.

* **InvoiceDate:**

La date et l'heure à laquelle chaque transaction a été effectuée.

* **UnitPrice :**

Le prix du produit par unité en livres sterling.

* **CustomerID :**

Un numéro unique à 5 chiffres attribué à chaque client.

* **Country:**

Le nom du pays d'où l'achat a été effectué.

## Exploration des données

### Pays d’où viennent les différentes transactions

Il y a 36 pays avec un total de 2393 transactions et 418 clients.

['France', 'Australia', 'Netherlands', 'Germany', 'Norway', 'EIRE',

'Switzerland', 'Spain', 'Poland', 'Portugal', 'Italy', 'Belgium',

'Lithuania', 'Japan', 'Iceland', 'Channel Islands', 'Denmark',

'Cyprus', 'Sweden', 'Austria', 'Israel', 'Finland', 'Bahrain',

'Greece', 'Hong Kong', 'Singapore', 'Lebanon',

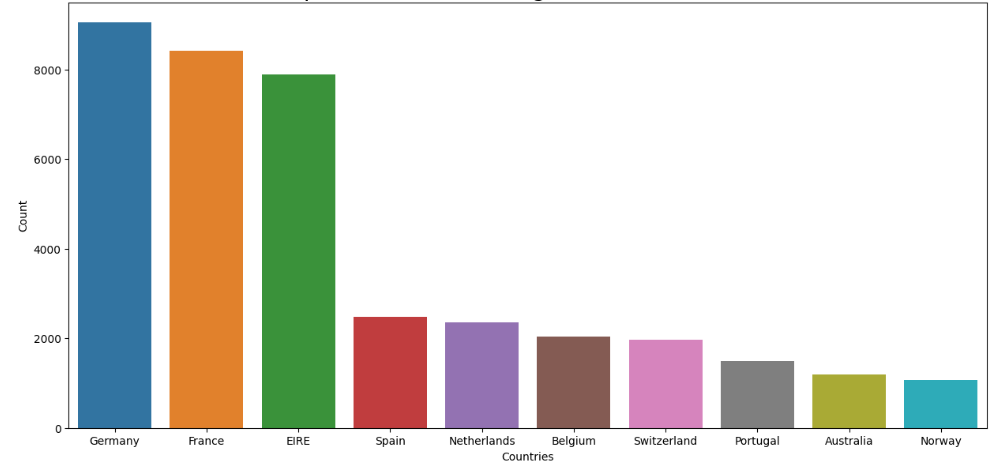
'United Arab Emirates', 'Saudi Arabia', 'Czech Republic', 'Canada',

'Brazil', 'USA', 'European Community', 'Malta', 'RSA']

## Visualisation des données de l’ensemble du dataset



**Figure 1 : Mots les plus représentés du dataset**



**Figure 2 : Top 10 des pays qui ont un site de vente au détail**

### Le nettoyage de données

Il y a des données manquantes dans l'ensemble de données. Plus précisément, la colonne "CustomerID" comporte un total de 1278 valeurs nulles.Comme nous disposons de données suffisantes pour notre analyse, nous n'allons pas essayer d'imputer ces valeurs manquantes et nous allons simplement les supprimer des données.

Après avoir supprimé toutes les valeurs nulles, il nous reste 44707 lignes de données.

Ensuite, nous voulons supprimer toutes les transactions annulées/retournées des données.

Pour rappel, dans la description de l'ensemble de données, les numéros de facture commençant par un "C" sont des transactions annulées/retournées.

Nous les avons supprimées en convertissant tout d'abord la colonne **FactureNo** en un type de données de type string, puis en supprimant toutes les lignes commençant par un "C" à l'aide d'un simple filtre. Toutes les transactions correspondantes ont été supprimées de l'ensemble des données. Cela garantit que seules les transactions valides et pertinentes sont prises en compte dans l'analyse

Enfin, nous avons utilisé la **méthode strip()** pour supprimer les espaces inutiles dans la colonne de description, ce qui permet d'obtenir des données plus cohérentes et facilitera l'analyse ultérieure.

### Séparation des données pour avoir les données par pays

#### France

### 

**Figure 4 : Mots les plus représentés pour la France**

#### 

#### 

#### 

#### Portugal

### 

**Figure 5 : Mots les plus représentés pour la Portugal**

#### 

#### 

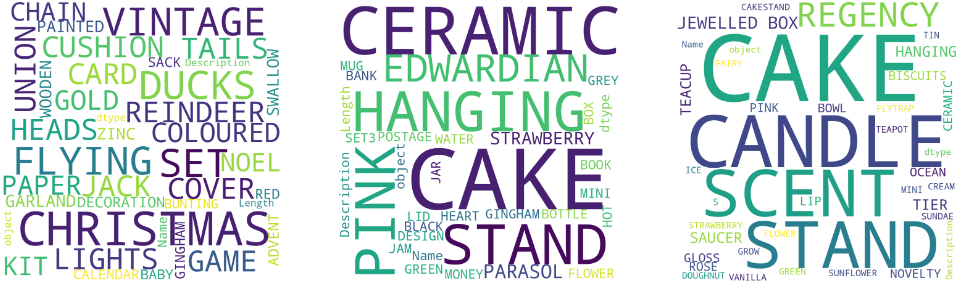
#### 

#### Suède



**Figure 6 : Mots les plus représentés pour la Suède**

### Comparaison des mots choisi pour décrire les produits par pays (france, Portugal, Suède)



**Figure 7 : Comparaison des mots choisi pour décrire les produits par pays (france, Portugal, Suède)**

On constate ici que le mot “cake” est répété pour le portugal et la suède.

## Modèle

### Comment marche l’algorithme Apriori:

L'algorithme Apriori repose sur le principe de l'itemset fréquent. Un itemset est simplement un ensemble d'articles présents dans un panier. Un itemset fréquent est un itemset qui apparaît avec une fréquence supérieure à un seuil prédéfini.

Voici les étapes de base de l'algorithme Apriori :

* Création des itemsets fréquents de taille 1 : Dans cette étape, l'algorithme Apriori identifie les articles individuels qui apparaissent avec une fréquence supérieure au seuil prédéfini. Cela crée les itemsets fréquents de taille 1.
* Génération des candidats : À partir des itemsets fréquents de taille 1, l'algorithme génère des candidats d'itemsets de taille supérieure. Pour cela, il utilise une technique appelée jointure. Il combine les itemsets fréquents de taille k-1 pour produire des candidats de taille k.
* Calcul de la fréquence des candidats : Dans cette étape, l'algorithme parcourt les transactions et compte le nombre d'occurrences des candidats générés. Les candidats dont la fréquence est inférieure au seuil prédéfini sont éliminés.
* Répétition des étapes 3 et 4 : Les étapes 3 et 4 sont répétées jusqu'à ce qu'il n'y ait plus de candidats fréquents à générer. Cela signifie qu'on a atteint la taille maximale des itemsets fréquents.
* Une fois que vous avez obtenu les itemsets fréquents, vous pouvez les utiliser pour analyser les associations entre les articles. Par exemple, si vous avez un itemset fréquent {A, B, C}, cela signifie que les articles A, B et C sont souvent achetés ensemble. Vous pouvez utiliser cette information pour recommander ces articles comme un ensemble ou pour grouper les clients en fonction de leurs achats communs.

### Explication des métriques:

* Antecedents: Cette colonne représente les ensembles d'articles antécédents ou combinaisons d'articles présents dans les transactions. Les antecedents font référence aux articles présents dans la partie "si" de la règle d'association.
* Consequents: Cette colonne représente les ensembles d'articles conséquents ou combinaisons d'articles présents dans les transactions. Les consequents font référence aux articles présents dans la partie "alors" de la règle d'association.
* Antecedent support : Il représente la fréquence d'apparition des ensembles d'articles antecedents dans les transactions.
* Consequent support: Similaire au support des antecedents , il s'agit de la proportion de transactions dans l'ensemble de données qui contiennent les ensembles d'articles consequents. Il représente la fréquence d'apparition des ensembles d'articles consequents dans les transactions.
* Support : Il s'agit de la proportion de transactions dans l'ensemble de données qui contiennent à la fois les ensembles d'articles antecedents et consequents . Le support mesure la fréquence globale d'apparition de la règle complète (à la fois les antecedents et les consequents ) dans les transactions.
* Confidence: Il s'agit de la probabilité conditionnelle de trouver les ensembles d'articles consequents dans une transaction étant donné que les ensembles d'articles antecedents sont déjà présents.

Il indique à quelle fréquence la règle est vraie ou à quelle fréquence les ensembles d'articles consequents apparaissent dans les transactions contenant les ensembles d'articles antécédents.

* Lift : Le lift mesure la force de l'association entre les ensembles d'articles antecedents et conséquents. Il s'agit du rapport entre le support observé et le support attendu si les ensembles d'articles antecedents et consequents étaient indépendants.

Un lift supérieur à 1 indique une association positive, un lift égal à 1 indique une indépendance, et un lift inférieur à 1 indique une association négative.

* Leverage : Il quantifie la différence entre la fréquence d'apparition de la règle ensemble et la fréquence attendue si les ensembles d'articles antecedents et consequents étaient indépendants.

Il mesure la différence absolue entre le support observé et le support attendu.

* Conviction : La conviction est une mesure de la dépendance des ensembles d'articles consequents vis-à-vis des ensembles d'articles antécédents. Elle est calculée comme le rapport entre la confiance attendue (en supposant une indépendance) et la confiance observée.

Une conviction supérieure à 1 indique une forte association, tandis que des valeurs proches de 1 indiquent une association plus faible.

* Zhang's metric : La métrique de Zhang est une mesure qui combine le support, la confiance et le lift pour évaluer la force de l'association entre les ensembles d'articles antecedents et conséquents.

La métrique de Zhang varie de -1 à 1, où des valeurs proches de 1 indiquent une forte association, des valeurs proches de 0 indiquent une indépendance, et des valeurs proches de -1 indiquent une association négative.

Ces métriques permettent de comprendre et d'interpréter les résultats de votre analyse des paniers, en fournissant des informations sur la force, la fréquence et la dépendance des associations entre les ensembles d'articles dans les transactions.

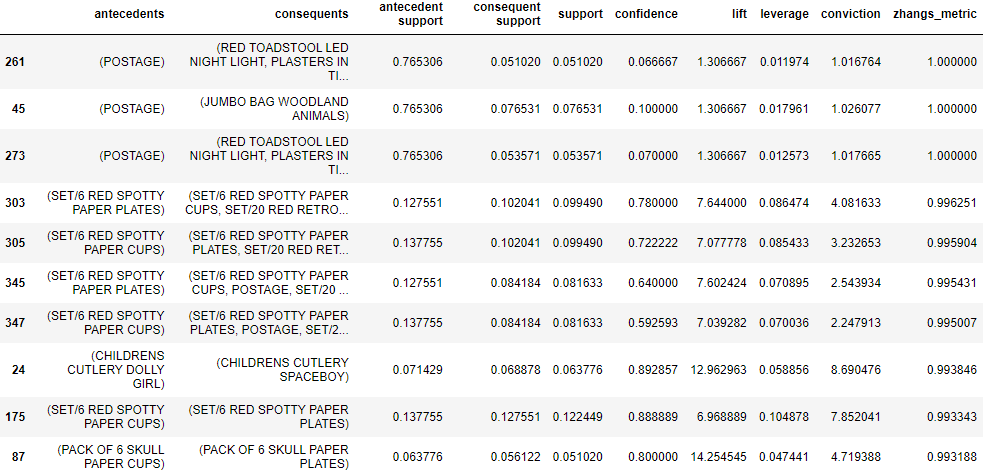
### Exploration pour utiliser le modèle Apriori

* Export des données en json (cela n’a pas fonctionné)
* Transformer les données pour qu’elles soient lisible et compréhensible
* Sortir un dictionnaire de dataframe (pays:dataframe)
* Faire tourner le model sur ce dictionnaire avec une boucle (cela n’a pas fonctionné, pas d’erreur mais pas de possibilité de faire autre chose pour que cela fonctionne)
* Utilisation de signaux (import signal) pour arrêter le programme si il tourne trop longtemps (cela n'a’ pas fonctionné, et l’utilisation n’a pas été bien comprise)
* Utilisation du multiprocessing (utiliser plusieurs coeurs du processeur en même temps), un coeur pour lance le model, l’autre pour compter 20 sec avant d’arrêter le model (cela n’a pas fonctionné)
* Utiliser cette même technique pour taguer les pays avec lequel le modèle peut tourner et en ressortir une liste manuellement (cela n’a pas fonctionné)
* Refaire le dataset avec les pays qui fonctionne dans le modèle
* Et ressortir les résultats du modèle par pays

## 

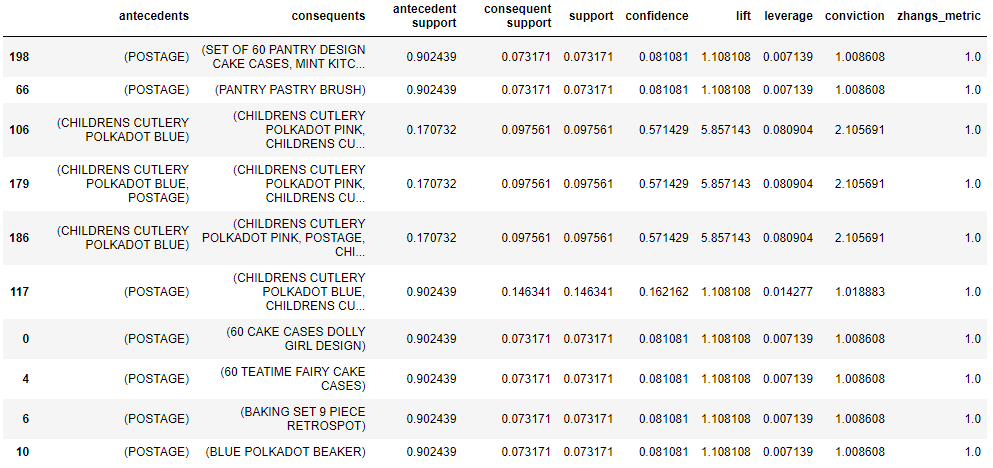
## Résultats

### Résultat France



**Figure 8 : Résultat du modèle pour la france**

### Résultat Finlande



**Figure 9 : Résultat du modèle pour la Finlande**

## 

## Conclusion / Ouverture

Comme constaté précédemment, le modèle Apriori a posé problème, il serait judicieux de tenter d’utiliser le modèle FP-Growth, qui prend moins de données que Apriori, mais est plus rapide, et est plus facilement interprétable.