République algérienne démocratique et populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université 20 août 1955 Skikda

Faculté des sciences Département

d'informatique

Mémoire de fin d'études en vue de l'obtention du diplôme

Master Académique

Option : Réseaux et Systèmes Distribués

Thème

L'analyse des données des clients en gros

Réalisé par :

- **BOUHEZZA ROFAIDA**
- **HADJI CHAIMA**

Encadré par :

Dr. Ramdane Chafika

Session : Juin 2024

Dédicace

Je dédie ce modeste travail

Au meilleur des pères ***Rabeh Bouhezza*** *et à ma très chère maman* ***ghania louhichi*** *qu'ils trouvent en moi la source de leur fierté qui ne cessent de me donner avec amour le nécessaire pour que je puisse arriver à ce que je suis aujourd'hui.*

Que dieu les protège et que la réussite soit toujours à ma portée pour que je puisse vous combler de bonheur.

À mes sœurs et mes frères ***chaima – SaifAlisslam – khairelddine - Amir***

À toute ma famille

Et A toutes mes amies,

A tous les gens qui me connaissent et que je connais

À tous ceux qui me sont chers, aux personnes qui m'ont aidé et encouragé de près ou de loin, qui étaient toujours à mes côtés et qui m'ont accompagné durant mon chemin d'études

Dédicace

Je dédie ce mémoire à mes chers parents Mon père ***Abed Aziz Hadji*** *et mon maman* ***Boulahia Fahima*** *qui ont ététoujours à mes côtés et m'ont toujours soutenu tout au long de ces longues années d'études. En signe de reconnaissance, qu'ils trouvent ici, l'expression de ma profonde gratitude pour tout ce qu'ils ont consenti d'efforts et de moyens pour me voir réussir dans mes études.*

À mes sœurs et frère ***Chams ,Amel , Lamiss , Mohamed Saleh Eddin***

A toute ma famille

Et A toutes mes amies,

A tous les gens qui me connaissent et que je connais en particulier

Et à tous ceux qui aiment le bon travail et ne reculent pas devant les obstacles de la vie.

REMERCIEMENT

Tout d'abord, nous exprimons notre gratitude envers Allah le Tout-Puissant, qui nous a accordé la force, la volonté et le courage nécessaires pour mener à bien ce modeste projet. Nous avons accompli un travail qui a été le fruit d'un parcours pédagogiqueapprofondi qui a duré tout au long de notre parcours éducatif dans le domaine de l'enseignement supérieur. Nous souhaitons exprimer notre profonde gratitude envers le Dr. Ramdane pour nous avoir offert son soutien et son encadrement. Son regard critique, sa patience et son dévouement ont joué un rôle essentiel dans la réalisation de ce travail modeste.

Je tiens à exprimer ma gratitude spéciale envers nos très chers parents, frères, sœurs, collègues et amies qui nous ont soutenus et encouragés tout au long de notre cheminement.

Résumé :

Ce travail de master se concentre sur l'analyse des données des clients dans le domaine de la vente en gros. L'objectif principal était de convertir les données brutes en informations exploitables pour mieux comprendre le comportement des clients.

Nous avons utilisé une variété de méthodes d'analyse, incluant l'analyse unidimensionnelle, bidimensionnelle et multidimensionnelle. Les techniques telles que le clustering et l'analyse factorielle ont été appliquées pour identifier des segments de clients et explorer les relations entre les variables.

Une étape cruciale de notre méthodologie était l'interprétation des résultats graphiques obtenus à chaque étape de l'analyse. Cette approche visuelle a permis de tirer des conclusions significatives et de fournir des recommandations pertinentes pour les stratégies commerciales.

Mots-clés : analyse des données des clients, segmentation des clients, clustering, analyse en composantes principales, analyse unidimensionnelle, analyse bidimensionnelle, analyse multidimensionnelle, interprétation graphique.

Abstract:

This master's thesis addresses the critical issue of customer data analysis in the wholesale sector. Data analysis plays a pivotal role in transforming raw data into actionable insights to better understand customer behavior.

We employed various analytical techniques, including unidimensional, bidimensional, and multidimensional analysis. These methods encompassed clustering techniques and factor analysis to identify customer segments and explore relationships among variables.

A key focus was placed on interpreting the graphical results obtained at each stage ofanalysis. This graphical interpretation was instrumental in deriving meaningful conclusions and providing relevant recommendations for business strategies.

Keywords: customer data analysis, customer segmentation, clustering, principal component analysis (PCA), unidimensional analysis, bidimensional analysis, multidimensional analysis, graphical interpretation.

Liste des matières

Chapitre 2 : Le clustering des Données

Chapitre 3 : Description de l'application d'analyse des données des clients de vente en gros.

Chapitre 4 : Analyse et interprétation des résultats.

Liste des figures

Liste des tables

Introduction Générale

Introduction générale :

À notre époque où les données sont devenues la nouvelle devise du succès organisationnel, l'analyse de données est devenue un pilier fondamental pour les organisations cherchant à tirer profit de l'avalanche croissante de données disponibles. Elle désigne le processus méthodique de collecte, de nettoyage, d'analyse et d'interprétation de données pour en extraire des informations pertinentes et actionnables. Cette discipline joue un rôle crucial dans la transformation des données brutes en informations stratégiques, facilitant ainsi la prise de décisions éclairées et la formulation de stratégies efficaces.

L'importance de l'analyse de données réside dans sa capacité à transformer des données en informations exploitables, permettant aux entreprises de mieux comprendre leur marché, leurs clients et leurs opérations. En particulier, l'analyse de données pour l'analyse des clients permet de segmenter efficacement les consommateurs, d'évaluer leurs comportements d'achat et de personnaliser les interactions pour répondre aux besoins spécifiques de chaque segment de marché.

Dans cette optique, nous allons viser dans ce travail de master, d'analyser des données de clients de la vente en gros. Nous allons utiliser différents types d'analyses pour explorer ces données. L'analyse unidimensionnelle se concentre sur l'examen de variables individuelles, fournissant des informations simples mais cruciales. L'analyse bidimensionnelle explore les relations entre deux variables, offrant une compréhension plus approfondie des interactions. Enfin, l'analyse multidimensionnelle intègre plusieurs variables simultanément, permettant une vision complète et complexe des données et de leurs relations.

En analyse multidimensionnelle, nous explorerons différents types d'approches, notamment le clustering et l'analyse factorielle, ainsi qu'une hybridation entre ces types. Le clustering est essentiel car il permet de regrouper les données similaires en clusters distincts, facilitant ainsi l'identification de catégories de clients ayant des comportements ou des caractéristiques communes. En parallèle, l'analyse factorielle, spécifiquement l'analyse en composantes principales, joue un rôle crucial en réduisant la dimensionnalité des données. Cette technique identifie les principales sources de variation parmi un ensemble de variables, ce qui simplifie l'analyse des relations complexes et permet une meilleure interprétation des structures sous-jacentes des données clients.

Ce mémoire est divisé en quatre chapitres distincts, chacun explorant des aspects spécifiques de l'analyse de données et de son application pratique.

Le premier chapitre introduit de manière détaillée les concepts fondamentaux de l'analyse de données, mettant en avant son rôle crucial dans la prise de décisions éclairées et stratégiques. Dans le deuxième chapitre, l'accent est mis sur le clustering des données, où différentes techniques sont examinées en profondeur pour leur capacité à regrouper efficacement des données similaires.

Le troisième chapitre se concentre sur la description et l'implémentation pratique d'une application concrète, illustrant comment les principes théoriques sont mis en œuvre dans un contexte réel

.
Enfin, le quatrième chapitre analyse et interprète rigoureusement les résultats obtenus à travers l'application des méthodes d'analyse de données.
Le mémoire se termine par une conclusion et différentes extensions de ce travail.

Chapitre 01 :

Principes de l'Analyse des données

1.1. Introduction :

Pour développer votre entreprise, et parfois même votre vie personnelle, il suffit parfois d'une analyse ! Si votre entreprise ne se développe pas, vous devez revenir en arrière, reconnaître vos erreurs et établir un nouveau plan sans les répéter. Et même si votre entreprise croît, vous devez continuer à analyser pour favoriser une croissance encore plus grande. L'analyse consiste à examiner attentivement les données et les processus de votre entreprise]. L'analyse des données est le processus de classement et d'organisation des données brutes de manière à en extraire des informations utiles. Organiser et réfléchir aux données est essentiel pour comprendre ce qu'elles contiennent ou non. Il existe de nombreuses méthodes pour aborder l'analyse des données, et il est remarquablement facile de manipuler les données durant la phase d'analyse pour soutenir certaines conclusions ou agendas. C'est pourquoi il est crucial de prêter attention à la présentation de l'analyse des données et de réfléchir de manière critique aux données et aux conclusions obtenues [1].

1.2. Définition de l'Analyse de Données :

L'analyse des données est une méthode complète d'inspection, de nettoyage, de transformation et de modélisation des données pour découvrir des informations utiles, tirer des conclusions et soutenir la prise de décision. C'est un processus multi facette qui implique diverses techniques et méthodologies pour interpréter des données provenant de diverses sources sous différents formats, structurés et non structurés[2].

L'analyse des données n'est pas seulement un processus; c'est un outil qui permet aux organisations de prendre des décisions éclairées, de prédire les tendances et d'améliorer l'efficacité opérationnelle. C'est l'épine dorsale de la planification stratégique dans les entreprises, les gouvernements et d'autres organisations.

Prenons l'exemple d'une entreprise leader dans le commerce électronique. Grâce à l'analyse des données, elle peut comprendre le comportement d'achat de ses clients, leurs préférences etleurs habitudes. Elle peut ensuite utiliser ces informations pour personnaliser les expériences des clients, prévoir les ventes et optimiser les stratégies marketing, ce qui entraîne finalement la croissance de l'entreprise et la satisfaction des clients.

1.3. Types de données :

Chaque type de données a la rare qualité de décrire les choses après leur avoir attribué une valeur spécifique. Pour l'analyse, il faut organiser ces valeurs, les traiter et les présenter dans un contexte donné, afin de les rendre utiles. Les données peuvent se présenter sous différentes formes ; voici les principaux types de données [3].

- **Données qualitatives :**

  Lorsque les données présentées sont accompagnées de mots et de descriptions, on parle alors de données qualitatives . Bien que vous puissiez observer ces données, ellessont subjectives et plus difficiles à analyser dans le cadre d'une recherche, en particulier à des fins de comparaison. Exemple : Les données de qualité représentent tout ce qui décrit le goût, l'expérience, la texture ou une opinion qui est considérée comme une donnée de qualité. Ce type de données est généralement collecté par le biais de groupes de discussion, d'entretiens qualitatifs personnels, d'observations qualitatives ou de questions ouvertes dans le cadre d'enquêtes.

- **Données quantitatives :**

  Toute donnée exprimée en nombre de chiffres est appelée données quantitatives . Ce type de données peut être classé en catégories, regroupé, mesuré, calculé ou classé. Exemple : des questions telles que l'âge, le rang, le coût, la longueur, le poids, les scores, etc. relèvent de ce type de données. Vous pouvez présenter ces données sous forme de graphiques, de tableaux ou appliquer des méthodes d'analyse statistique à ces données. Les questionnaires OMS (Outcomes Measurem en Systems) des enquêtes constituent une source importante de collecte de données numériques.

- **Données catégorielles :**

  Il s'agit de données présentés en groupes. Toutefois, un élément inclus dans les données catégorielles ne peut appartenir à plus d'un groupe. Exemple : Une personne qui répond à une enquête en indiquant son mode de vie, sa situation matrimoniale, son habitude de fumer ou de boire fait partie des données catégorielles. Le test du chi-carréest une méthode standard utilisée pour analyser ces données.

1.4. Processus d'analyse de données :

Pour la plupart des entreprises et des agences gouvernementales, le manque de données n'est pas un problème. En fait, c'est le contraire : il y a souvent trop d'informations disponibles pour prendre une décision claire. Le processus d'analyse de données se compose de plusieurs parties. Il existe quelques parties importantes du traitement des données telles que la collecte de données, le traitement des données, le nettoyage des données, l'analyse des données, la communication. Tout d'abord, nous devons être clairs sur un concept. Pourquoi avons-nous besoin d'analyser les données et que ferons-nous avec elles ? Après avoir compris cela, nous pouvons passer à la deuxième étape. Il s'agit de la collecte de données. La collecte de données est un processus de collecte d'informations auprès de toutes les sources pertinentes pour trouver des réponses au problème de recherche, tester l'hypothèse et évaluer les résultats [4].

Les méthodes de collecte de données peuvent être divisées en deux catégories : les méthodes de collecte de données secondaires et les méthodes de collecte de données primaires. Les données secondaires sont un type de données qui ont déjà été publiées dans des livres, des journaux, des magazines, des revues, des portails en ligne, etc. Il y a une abondance de données disponibles dans ces sources sur votre domaine de recherche en études commerciales,presque indépendamment de la nature du domaine de recherche. Par conséquent, l'application du bon ensemble de critères pour sélectionner les données secondaires à utiliser dans l'étude joue un rôle important en termes d'augmentation des niveaux de validité et de fiabilité de la recherche. Les méthodes de collecte de données primaires peuvent être divisées en deux groupes : quantitatives et qualitatives. Les méthodes de collecte de données quantitatives sont basées sur des calculs mathématiques sous divers formats. Les méthodes de collecte et d'analyse de données quantitatives comprennent des questionnaires avec des questions à choixmultiples, des méthodes de corrélation et de régression, la moyenne, le mode et la médiane, entre autres. Les méthodes quantitatives sont moins coûteuses à appliquer et peuvent être appliquées dans un laps de temps plus court par rapport aux méthodes qualitatives. De plus,en raison d'un haut niveau de standardisation des méthodes quantitatives, il est facile de comparer les résultats. Les méthodes de recherche qualitative, en revanche, n'impliquent pas de nombres ou de calculs mathématiques. La recherche qualitative est étroitement associée aux mots, aux sons, aux sentiments, aux émotions, aux couleurs et à d'autres éléments qui ne sont pas quantifiables [5].

Figure 1.2. Les étapes de l'analyse des données

Le nettoyage des données est une partie cruciale de l'analyse des données, en particulier lorsque vous collectez vos propres données quantitatives. Après avoir collecté les données, vous devez les saisir dans un programme informatique tel que SAS, SPSS ou Excel. Pendant ce processus, qu'il soit effectué à la main ou par un scanner informatique, il y aura des erreurs.Peu importe à quel point les données ont été saisies avec soin, les erreurs sont inévitables. Cela pourrait signifier un codage incorrect, une lecture incorrecte de codes écrits, une détection incorrecte de marques noircies, des données manquantes, etc. Le nettoyage des données est le processus de détection et de correction de ces erreurs de codage. Il existe deux types de nettoyage des données qui doivent être effectués sur les ensembles de données. Il s'agit du nettoyage de code possible et du nettoyage de contingence. Les deux sont cruciaux

pour le processus d'analyse des données car s'ils sont ignorés, vous produirez presque toujours des résultats de recherche trompeurs. Après avoir nettoyé les données, nous pouvons passer à l'analyse des données [6].

1.5. Types de l'Analyse des Données :

Il existe plusieurs types de techniques d'analyse des données basées sur le domaine des affaires et la technologie. Les principaux types d'analyse des données sont :

1.5.1. Analyse de Texte :

L'analyse de texte est également appelée fouille de données. C'est une méthode pour découvrir des modèles dans de grands ensembles de données en utilisant des bases de données ou des outils de fouille de données. Elle est utilisée pour transformer des données brutes en informations commerciales. Des outils d'intelligence d'affaires sont présents sur le marché et sont utilisés pour prendre des décisions stratégiques pour les entreprises. Globalement, elle offre un moyen d'extraire et d'examiner les données, de dériver des modèles et enfin d'interpréter les données.[7].

1.5.2. Analyse Statistique :

L'analyse statistique montre "ce qui s'est passé" en utilisant des données passées sous forme de tableaux de bord. L'analyse statistique inclut la collecte, l'analyse, l'interprétation, la présentation et la modélisation des données. Elle analyse un ensemble de données ou un échantillon de données. Il existe deux catégories pour ce type d'analyse - l'analyse descriptive et l'analyse inférentielle.

1.5.3. Analyse Descriptive :

L'analyse descriptive traite des données complètes ou d'un échantillon de données numériques résumées. Elle présente la moyenne et l'écart type pour les données continues, tandis qu'elle affiche les pourcentages et les fréquences pour les données catégorielles [8].

1.5.4. Analyse Prédictive :

L'analyse prédictive répond à la question « Que va-t-il probablement se passer ? » en utilisant des données antérieures. Par exemple, si l'année dernière j'ai acheté deux robes en fonction de mes économies et que cette année mon salaire a doublé, alors je pourrais acheter quatre robes. Cependant, ce n'est pas aussi simple, car il faut prendre en compte d'autres circonstances, comme la possibilité d'une augmentation des prix des vêtements cette année ou le fait que je préfère acheter une nouvelle moto ou une maison. Cette analyse fait donc des prédictions sur les résultats futurs en se basant sur des données actuelles ou passées. La prévision n'est qu'une estimation dont la précision dépend de la quantité et de la qualité des informations disponibles.[9].

1.5.5. Analyse Prescriptive :

L'analyse prescriptive combine les informations de toutes les analyses précédentes pour déterminer quelle action entreprendre face à un problème ou une décision

actuelle. La plupart des entreprises axées sur les données utilisent l'analyse prescriptive, car les analyses prédictives et descriptives ne suffisent pas à améliorer les performances des données. En fonction des situations et des problèmes actuels, elles analysent les données et prennent des décisions. [10].

1.5.6. L'analyse de données unidimensionnelle :

L'analyse de données unidimensionnelle, également connue sous le nom d'analyse uni variée, est une méthode statistique qui se concentre sur l'examen et la compréhension d'une seule variable à la fois. Cette approche est souvent utilisée pour explorer les caractéristiques d'une variable particulière dans un ensemble de données sans tenir compte des relations avecd'autres variables [11].

Les techniques couramment utilisées dans l'analyse univariée comprennent la description des statistiques de base telles que la moyenne, la médiane, l'écart-type, la variance, le mode, etc. Elle peut également impliquer des méthodes de visualisation telles que les histogrammes, les diagrammes à barres, les diagrammes

en secteurs, les boîtes à moustaches, etc. Ces méthodes permettent de comprendre la distribution et les tendances des données d'une seule variable.

L'analyse univariée est souvent utilisée comme première étape dans l'exploration des données avant d'aller vers des analyses plus complexes impliquant plusieurs variables (analyse bidimensionnelle, analyse multi variée, etc.).

1.5.7. L'analyse de données bidimensionnelle :

L'analyse de données bidimensionnelle, également connue sous le nom d'analyse bi-variée, est une méthode statistique qui examine les relations entre deux variables simultanément. Contrairement à l'analyse unidimensionnelle qui se concentre sur une seule variable à la fois, l'analyse bidimensionnelle explore la relation entre deux variables en même temps [12].

Cette approche permet de déterminer s'il existe une corrélation, une association ou une relation causale entre deux variables. Elle est utile pour comprendre comment les changements dans une variable sont associés à des changements dans une autre variable. Par exemple, dans une étude sur la relation entre le temps de travail et le niveau de stress, une analyse bidimensionnelle pourrait être utilisée pour examiner comment le niveau de stress varie en fonction du nombre d'heures travaillées par semaine.

Les techniques couramment utilisées dans l'analyse bidimensionnelle comprennent les coefficients de corrélation, les tests d'indépendance (comme le test du chi-carré), les diagrammes de dispersion (scatter plots), les analyses de régression, etc. Ces méthodes permettent de quantifier et de visualiser la relation entre les deux variables étudiées.

L'analyse bidimensionnelle est souvent utilisée pour explorer les relations préliminaires entre les variables avant d'effectuer des analyses plus complexes impliquant plusieurs variables (analyse multivariée).

1.5.8. analyse de données multidimensionnelle :

L'analyse de données multidimensionnelle est une méthode d'analyse statistique qui examine des ensembles de données complexes contenant plusieurs variables. Plutôt que de se limiter à l'examen de relations entre deux variables (comme dans l'analyse bidimensionnelle), l'analyse multidimensionnelle prend en compte plusieurs variables simultanément.[13].

Elle est souvent utilisée dans des domaines tels que la recherche marketing, la finance, la biologie, la psychologie et d'autres sciences sociales pour identifier des tendances, des schémas ou des relations cachées dans les données qui pourraient ne pas être évidentes lorsqu'on les examine de manière traditionnelle. Cette méthode peut être réalisée à l'aide de techniques telles que l'analyse des composantes principales (PCA), l'analyse factorielle, l'analyse des correspondances multiples (MCA) et d'autres méthodes statistiques avancées. En résumé, l'analyse de

données multidimensionnelle permet de mieux comprendre la structure et les relations complexes des données en tenant compte de plusieurs variables à la fois.

il ya des techniques plusieurs dans analyse multidimensionnelle parmi cette technique ACP(Analyse en composantes principales) les plus couramment utilisé dans cette analyse :

1.6. Analyse factorielle des données:

L'objectif de l'analyse factorielle est de fournir des représentations dans un espace de faible dimension des informations consignées dans de volumin eux tableau. En général, ces représentations sortent des visualisations graphiques qui mettenten évidence les traits essentiels caractérisant les données.
Plusieurs méthodes factorielles ont été proposées, elles ont le même principe de base mais chacune d'elle s'adapte à un contexte particulier [14] .

1. Méthode factorielles classiques; Spearman(1904)et Tu stone(1947).
2. Analyse en composantes principales; Pearson(1904)et Hotte ling(1953).
3. Analyse des correspondances; Hirsfeld(1935)et Beng(1964).
4. Analyse des proximités.

1.6.1. Principe de base des méthodes factorielles:

Le principe général est de résumer au mieux un tableau de données représenté par une matrice X à n lignes et p colonnes de terme général x_{ij}.
Résumer au mieux signifie ici la possibilité de reconstituer les (np)valeurs numériques initiales de X par un nombre de valeurs numériques nettement inférieurs à np et tel que la reconstitution soit bonne.

- dans se cas on a opté le type 'analyse de composant principal ' Également connu 'ACP':

1.6.2. Analyse en composantes principales(ACP):

Les données mises en jeu par l'ACP sont relatives à des variables quantitatives, continues, homogènes ou non, à priori corrélées entre elles deux à deux, ces données sont consignées

dans un tableau individu-variable.

1.6.3. L'ACP normée:

En général, les variables du tableau de données sont hétérogènes de points de vue de leur moyenne, de leur dispersion et de leur nature (les unités de mesure sont exprimées en quantités non comparable).
Il faut donc transformer les variables X_j de telle sorte que ses variables soient de variance unité et moyenne nulle [14].
Le tableau X de départ est transformé comme suit:

$$X = \left\{ x_{ij} = \frac{x_{ij} - \overline{X}_j}{\sigma_j \sqrt{n}} \right\}$$

Où σ_j désigne l'écart type de la variable X_j.

1.6.3.1. L'analyse des points individus dans l'espase R^p :

La transformation du tableau initial aura comme signification la translation des axes initi aux au centre de gravité des points.

- Calcul de la matrice X^tX à diagonaliser:

$$X^t \cdot X = \left\{ \sigma_{jk} = \sum_{i=1}^{n} \frac{(x_{ij} - \overline{X}_j)(x_{ik} - \overline{X}_k)}{\sigma_j \sigma_k n} \right\}$$

X^tX est la matrice des corrélations linéaires entre les variables. L'analyse consiste donc à chercher les valeurs propres λ_α et les vecteurs propres u_α de la matrice des corrélations.
Les cordonnées des point si sur l'axe de rang sont les produits scalaires $\overrightarrow{OM_i} \cdot \overrightarrow{u_\alpha}$ qui constituent les lignes du vecteur X u_α $\left(X \cdot u_\alpha = \sqrt{\lambda_\alpha} \cdot v_\alpha\right)$.

- Distance entre deux points individus h et i:

$$d^2(h,i) = \sum_{j=1}^{p} \left(\frac{x_{hj} - \overline{X}_j}{\sigma_j \sqrt{n}} - \frac{x_{ij} - \overline{X}_j}{\sigma_j \sqrt{n}} \right)^2 = \boxed{\sum_{j=1}^{p} \frac{(x_{hj} - x_{ij})^2}{{\sigma_j}^2 n}}$$

1.6.3.2. L'analyse des points variables dans l'espace R^n:

Distance d'un point j par rapport a l'origine O:

$$d^2(O,j) = \sum_{i=1}^{n} \left(0 - \frac{x_{ij} - \overline{X}_j}{\sigma_j \sqrt{n}} \right)^2 = \sum_{i=1}^{n} \frac{(x_{ij} - \overline{X}_j)^2}{{\sigma_j}^2 n} = \frac{1}{{\sigma_j}^2} \times \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{X}_j)^2 = \frac{1}{{\sigma_j}^2} \times {\sigma_j}^2 = \boxed{1}$$

$$d^2(O,j) = 1$$

- Distance entre deux points variables j et k:

$$d^2(j,k) = \sum_{i=1}^{n} \left(\frac{x_{ij} - \overline{X}_j}{\sigma_j \sqrt{n}} - \frac{x_{ik} - \overline{X}_k}{\sigma_k \sqrt{n}} \right)^2 = \sum_{i=1}^{n} \left(\frac{x_{ij} - \overline{X}_j}{\sigma_j \sqrt{n}} \right)^2 + \sum_{i=1}^{n} \left(\frac{x_{ik} - \overline{X}_k}{\sigma_k \sqrt{n}} \right)^2 - 2 \sum_{i=1}^{n} \left(\frac{x_{ij} - \overline{X}_j}{\sigma_j \sqrt{n}} \right) \left(\frac{x_{ik} - \overline{X}_k}{\sigma_k \sqrt{n}} \right) =$$

$$1 + 1 - 2 \sum_{i=1}^{n} \frac{(x_{ij} - \overline{X}_j)(x_{ik} - \overline{X}_k)}{\sigma_j \sigma_k n} = 2 - 2r_{jk} = \boxed{2(1 - r_{jk})}$$

r_{jk} est le coefficient de corrélation linéaire entre les variables j et k.

- **Interprétation des résultats:**

- Deux variables fortement corrélées entre elles(positivement)sont très proche l'une de l'autre:

$$r_{jk} \approx 1 \implies d(j,k) \approx 0$$

- Deux variables fortement corrélées entre elles(négativement)sont très éloignées l'une de l'autre:

$$r_{jk} \approx -1 \implies d(j,k) \approx 2$$

- Deux variables très faiblement corrélées entre elles sont orthogonales:

$$r_{jk} \approx 0 \implies d(j,k) \approx \sqrt{2}$$

1.6.4. les objectifs d'une analyse en composante principale :

Mêlant approche géométrique (représentation le lien entre variables et individus dans un espace rectangulaire) et approche statistique (recherche sur des axes indépendants décrivant la variance), l'ACP a trois grands objectifs : comprendre la structure d'un ensemble de variables, créer des instruments pour analyser des éléments impossibles à mesurer directement, et condenser les informations issues d'un grand nombre de variables dans un ensemble restreint en garantissant une perte minime.[15]

1.6.5. les principales étapes de l'analyse en composantes principales :

Il existe quatre principales étapes lors d'une analyse en composantes principales[16] :

1. Définir les objectifs de l'analyse et l'approche (exploratoire ou confirmatoire) adaptée au type de problème, selon l'existence ou non d'a priori théoriques.
2. Préparer l'analyse en déterminant le nombre de variables conservées, le type de variables (continues ou dichotomiques), et la taille de l'échantillon.
3. S'assurer de l'existence de corrélations minimales entre les variables analysées, en recourant à une matrice de corrélation, puis mesurer l'adéquation de l'échantillonnage et réaliser un test de sphéricité dit de Bartlett.
4. Choisir le nombre de facteurs à extraire grâce à l'ACP en se fiant à deux critères distincts (la valeur de Eigen et le coude de Cattell).

1.7. l'Importance de l'analyse des données pour les entreprises :

L'importance croissante de l'analyse des données pour les entreprises a véritablement changé le monde, mais la personne moyenne reste souvent inconsciente de l'impact de l'analyse des données dans le domaine des affaires. Certaines des façons dont cela a affecté les entreprises incluent les suivantes:

1.7.1. Amélioration de l'efficacité :

Toutes les données collectées par l'entreprise ne sont pas uniquement liées aux individus externes à l'organisation. La plupart des données collectées par les

entreprises sont analysées en interne. Avec les progrès technologiques, il est devenu très pratique de collecter des données. Ces données aident à connaître la performance des employés ainsi que celle de l'entreprise.

1.7.2. Compréhension du marché :

Avec le développement d'algorithmes de nos jours, de vastes ensembles de données peuvent être rassemblés et analysés. Ce processus d'analyse est appelé "Mining". Contrairement aux autres types de ressources physiques, la collecte de données se fait sous forme brute et est ensuite affinée. Cela permet de collecter des données auprès d'une grande variété de personnes, ce qui s'avère ensuite fructueux pour une meilleure stratégie marketing.

1.7.3. Réduction des coûts :

Les technologies du big data comme l'analyse basée sur le Cloud et Hadoop peuvent apporter d'énormes avantages en termes de coûts s'il s'agit de stocker de grandes quantités de données. Elles peuvent également identifier les moyens efficaces de mener les affaires. Vous économisez non seulement de l'argent en termes d'infrastructure, mais aussi sur le coût du développement d'un produit qui aurait un ajustement parfait sur le marché.

1.7.4. Prise de décision plus rapide et meilleure :

Grâce à l'analyse en mémoire haute vitesse et à Hadoop combinés à la capacité d'analyser de nouvelles sources de données, les entreprises peuvent analyser l'information presque instantanément. Cela se révèle être un grand gain de temps, car vous pouvez désormais livrer de manière plus efficace et gérer vos délais avec aisance.

1.7.5. Nouveaux produits/services :

Avec le pouvoir de l'analyse des données, les besoins et la satisfaction des clients sont mieux pris en compte. Cela permet de s'assurer que le produit/service est aligné sur les valeurs du public cible.

1.7.6. Connaissance de l'industrie :

La connaissance de l'industrie peut être comprise et elle peut montrer comment une entreprise peut fonctionner dans un avenir proche. De plus, elle peut vous indiquer quel type d'économie est déjà disponible à des fins d'expansion commerciale. Cela ouvre non seulement de nouvelles avenues pour la croissance des entreprises, mais cela aide également à construire unécosystème solide autour de la marque.

1.8. L'analyse marketing :

L'analyse des données marketing joue un rôle essentiel dans la compréhension du marché. Peu d'entreprises peuvent aujourd'hui se passer de cet outil primordial dans la poursuite de leurs objectifs commerciaux. Car connaître rapidement le succès – ou l'échec – de sesopérations marketing, c'est réagir vite et se démarquer face aux concurrents !

L'étude marketing fait partie intégrante de toute stratégie . Elle permet en effet de :

- Comprendre les besoins des clients : des besoins qui changent vite et auxquels les produits de l'entreprise doivent apporter une réponse rapide.
- Prendre de meilleures décisions : l'étude de marché permet de cibler ses actions marketing dès le départ, et donc de gagner temps, énergie et efficacité.
- Évaluer les performances de l'entreprise : avec l'analyse marketing, l'entreprise prend du recul sur sa manière de procéder et améliore ses méthodes en permanence.

Mais une étude marketing peut également renforcer significativement l'image de l'entreprise aux yeux des clients, notamment face aux concurrents de son marché ! Avec à sa disposition des données clients précises fournies par sa solution CRM, la marque peut personnaliser davantage son offre, faire évoluer ses produits et s'adapter en permanence aux attentes des consommateurs. Des atouts imparables pour suivre sans faillir les rapides changements de marché [17].

1.9. les principes de l'analyse marketing :

L'analyse marketing ne se limite pas à un simple croisement de données. Le processus nécessite en effet une bonne compréhension du contexte commercial de l'entreprise

L'analyse de marché permettra alors de :

- Comprendre quelles sont les forces et faiblesses de l'entreprise sur son marché ;
- Déterminer le profil des clients, leur comportement et leurs attentes par rapport aux produits ;
- Connaître les concurrents, leurs avantages et leurs spécificités sur le marché ;
- Identifier les opportunités pouvant être exploitées rapidement et optimiser sa gestion des ventes ;
- Cibler les étapes de sa stratégie marketing qui sont à améliorer, les actions qui performent le plus, etc.

L'analyse marketing donne ainsi à l'entreprise une vision claire sur ses concurrents et sur les attentes de son public cible : en découlent une stratégie plus efficace, des prises de décision plus efficaces et des actions marketing qui vont droit au but.

On pourrait croire que la mise en œuvre de l'analyse marketing se fait à certains moments précis de la vie de l'entreprise. Mais il s'agit en réalité d'un véritable processus de fond qui suit une stratégie claire et définie, qu'il est nécessaire de renouveler constamment et d'étendreà tous les services. De la production au service après-vente, de nombreux acteurs de l'organisation ont en effet leur rôle à jouer dans la satisfaction des clients : grâce à l'étude de marché, c'est ainsi tout l'environnement commercial de l'entreprise qui est mieux maîtrisé.

Afin de toujours pouvoir suivre les évolutions de l'entreprise, de sa stratégie marketing, des clients, des concurrents, etc ; il est primordial de mettre en place un logiciel CRM adapté et évolutif. L'étape de la rédaction du cahier des charges CRM est donc fondamentale pour exprimer clairement les besoins de l'entreprise en matière de fonctionnalités, et offrir à ses équipes, un outil qui saura répondre à leurs enjeux [18].

1.10. L'importance de l'analyse de marketing :

L'analyse marketing, en se basant sur les sources fournies, est un processus vital pour comprendre ce qui fonctionne dans votre stratégie marketing et ce qui peut être amélioré. Elle

fournit des informations précieuses pour prendre des décisions basées sur des données réelles.[19]

Dans le monde du marketing et de l'analyse de données, l'analyse marketing joue un rôle crucial. Elle aide les entreprises à plusieurs niveaux :

1. Comprendre les besoins et le comportement des clients : En analysant les données, les entreprises peuvent mieux comprendre ce que recherchent leurs clients, comment ils interagissent avec la marque et quels sont leurs besoins non satisfaits.

2. Prendre de meilleures décisions marketing et commerciales : En utilisant les informations collectées, les entreprises peuvent prendre des décisions plus éclairées en ce qui concerne leurs efforts de marketing et leurs activités commerciales, en se basant sur des données concrètes plutôt que sur des suppositions.

3. Évaluer les performances de l'entreprise et de ses campagnes : L'analyse marketing permet de mesurer l'efficacité des différentes initiatives marketing et commerciales, en identifiant ce qui fonctionne bien et ce qui peut être amélioré.

4. Personnaliser l'offre et s'adapter aux attentes des consommateurs : En comprenant mieux les besoins et les comportements des clients, les entreprises peuvent personnaliser leurs offres et leurs messages pour répondre aux attentes spécifiques des consommateurs.

5. Identifier les forces, faiblesses, opportunités et menaces sur le marché : L'analyse marketing permet également d'identifier les tendances du marché, les opportunités

de croissance, ainsi que les défis potentiels auxquels l'entreprise pourrait être confrontée

En résumé, l'analyse marketing permet aux entreprises de mieux comprendre leur marché, d'optimiser leurs stratégies marketing et de prendre des décisions éclairées pour améliorer leur performance globale [20].

1.11. Conclusion :

Le processus d'évaluation des données en utilisant le raisonnement analytique et logique pour examiner chaque composant des données fournies. Cette forme d'analyse n'est qu'une des nombreuses étapes qui doivent être franchies lors de la conduite d'une expérience de recherche. Les données de diverses sources sont rassemblées, examinées, puis analysées pour former une sorte de constat ou de conclusion. Il existe une variété de méthodes d'analyse de données spécifiques, dont certaines incluent l'exploration de données, l'analyse de texte, l'intelligence d'affaires et les visualisations de données. L'importance de l'analyse de données change vraiment le monde. Que ce soit dans le domaine du sport, du commerce ou simplement dans les activités quotidiennes de la vie humaine, l'analyse de données a changéla façon dont les gens agissaient auparavant. Elle joue maintenant un rôle majeur dans les affaires et est également utilisée dans le développement de l'intelligence artificielle, le suivi

des maladies, la compréhension du comportement des consommateurs et la détection des faiblesses des adversaires dans le sport ou la politique. C'est le nouvel âge des données et il a un potentiel illimité. Chaque organisation tente de rassembler des données, par exemple, en surveillant les performances de ses concurrents, les chiffres de vente et les tendances d'achat, etc., dans le but d'être plus compétitive. Cependant, personne ne peut comprendre les comportements des clients et les performances des concurrents sans les compétences pour analyser toutes ces données. L'analyse des données est donc une nécessité pour prendre des décisions bien informées et efficaces. L'analyse des données est ce qui aide les organisations à déterminer leur position sur le marché par rapport aux concurrents. C'est ce qui nous aide à identifier les risques potentiels à éviter et les opportunités à saisir pour croître. En fait, c'est l'analyse des données qui nous permet d'évaluer le niveau de satisfaction des clients et leurs besoins afin de proposer de nouveaux produits et services qui leur apportent une plus grande satisfaction. Par conséquent, il est peu dire que l'analyse des données est importante pour le succès des entreprises.

Chapitre 02 :

Le Clustering des données

2.1. Introduction :

Le clustering est une tâche que les humains pratiquent depuis des milliers d'années et qui a été entièrement automatisée ces dernières décennies grâce aux avancées des technologies de calcul. Le clustering est généralement défini comme la tâche consistant à identifier des groupes naturels dans un ensemble de données multidimensionnelles. Le regroupement est effectué de manière à ce que les données d'un même cluster soient plus similaires entre elles que celles appartenant à des clusters différents. Bien que ce concept soit intuitif, il est difficileà mettre en œuvre dans la pratique. Cette difficulté est due à l'absence de définition unique et précise d'un cluster, en raison du manque d'informations préalables sur les distributions des données.

Dans la littérature, il existe plusieurs définitions d'un cluster, de ce qui constitue un cluster et de la relation entre les clusters. Sur cette base, une multitude de techniques de clustering ont été développées. Ces techniques peuvent différer dans leurs principes, leurs propriétés, leurs paramètres et les formes générales des partitions générées. La catégorisation de ces techniques peut être réalisée selon plusieurs aspects : la mesure de proximité utilisée entre les données, la théorie ou les concepts fondamentaux sur lesquels reposent les techniques, la nature des données manipulées et bien d'autres critères. En réalité, la catégorisation de ces techniques n'est ni simple ni canonique, car les catégories se chevauchent.

2.2. Définition de clustering de donnée :

Le clusteringest généralement défini comme la tâche qui consiste à trouver des groupes naturels dans un ensemble de données multidimensionnelles. Le regroupement est fait tel que les données dans le même cluster ou groupe sont plus similaires que celles dans des clusters différents. Ainsi qu'il est une tâche d'apprentissage "non supervisée" car on ne dispose d'aucune autre information préalable que la description des données ce qui implique que les classes possibles ne sont pas connues à l'avance [21].

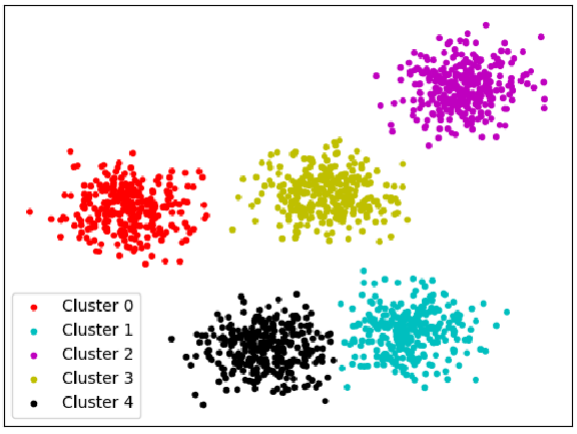

Figure 2.1. Le clustering de données .

2.3. Définition d'un cluster :

La définition de cheque constitue un cluster n'est pas bien défini et le terme, cluster n'a pas de définition précise Cependant, plusieurs définitions d'un cluster sont généralement utilisées[22][23].

2.3.1. Définition d'un cluster bien séparé :

Est un cluster dans lequel les objets ont peu de similarité avec les objets des autres clusters. Autrement dit, les objets d'un cluster bien séparé sont fortement groupés ensemble et sont clairement distincts des objets dans les autres clusters. Ce concept

algorithme doit être capable de produire des clusters bien séparé qui reflètent les structures cachées dans les données.

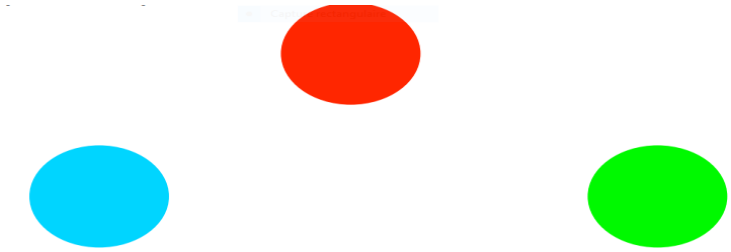

Figure 2.2. Clusters bien séparé.

2.3.2. Définition d'un cluster basé sur un centre :

Est un cluster dans lequel chaque cluster est représenté par un objet central appelé center de cluster. Les autres objets du cluster sont associés au centre en fonction de leur similarité ou de Les clustering des données leur proximité. Les algorithmes de clustering baséssur le centre utilisent généralement une métrique de distance pour mesurer la proximité entre les objets et le centre de cluster. Le modèle final dépend de la position des centres et des associations entre. Les algorithmes de clustering basés sur le centre incluent k-menas et le modèle de mélange gaussien.

Figure 2.3. Quatre clusters basés sur le centre.

2.3.3. Définition de Cluster contiguë :

Un cluster contigus désigne un cluster ou les objets sont groupés de maniéré a former une forme continue ou une région contiguë dans l'espace des données. Autrement dit, les objets d'un cluster contiguës sont proches les uns des autres et forment une région cohérente est souvent utilisé dans les algorithmes de clustering qui utilisent des techniques de partitionnement de l'espace pour diviser les données en clusters .Les algorithmes qui produisent des clusters contigus sont utiles pour les applications qui impliquent la segmentation de l'image ou la géographie.

Figure 2.4 Huit clusters contigus.

2.3.4. Définition basée sur la densité :

Est une approche de clustering qui regroupe les objets qui sont proches les uns des autres dans l'espace des données en formant des régions de haute densité. Cette approche considère que les clusters sont des régions ou la densité de rapport à leur environnement immédiat. Les algorithmes basés sur la densité, tels que DBSCAN et OPTICS, utilisent des méthodes pour mesurer la densité dans l'espace des données et trouver les régions de haute densité pour former les clusters. Cette approche est particulièrement utile pour les données de grande dimension etles données complexes, car elle est capable de détecter des structures complexes dans les données.

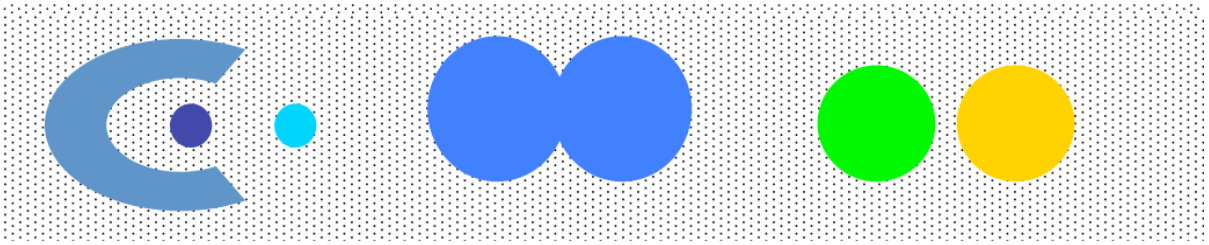

Figure 2.5. Six clusters denses.

2.4. Les concepts de clustering :

- **La matrice de données**:
  Les objets (échantillons, mesures, modèles, événements) sont habituellement représentés comme des points (vecteurs) dans un espace multidimensionnel, où chaque dimension représente un attribut distinct (variable, mesure) décrivant l'objet. Ainsi, un ensemble d'objets est représenté comme une matrice m×n, avec m lignes, une pour chaque objet et n colonnes, une pour chaque attribut. Cette matrice est appelée matrice de données ou jeu de données

- **La matrice de proximité**:
  Plusieurs algorithmes de clustering utilisent la matrice de données originale et beaucoup D'autres emploient une matrice de similarité, ou une matrice de dis similarité. Pour la convenance, les deux matrices sont généralement mentionnées comme une matrice de proximité, P. Une matrice de proximité, P, est une matrice m×m contenant toutes les dis similarités ou les similarités entre les objets considérés. Si p i et p j sont le i émet et le j émet objets, respectivement, alors l'entrée à la i Emme ligne et la j Emme colonne de la matrice de proximité est la similarité, ou la dis similarité, entre pi et p j.

2.5. Types et échelles de données :

Le mesure d'approximation et le type de clustering utilisé dépend

du type et de l'échelle d'Attributs de données Les trois types d'attributs sont montrés dans le tableau 2.1.

les différentes échelles de données sont présentées dans le tableau 2.2. [24].

Binaire	Deux valeurs possibles, vraies ou fausses
Discret	Nombre de valeur finie ou d'entiers
Continu	Nombre de valeurs infinies ou de réels

Tableau 2.1. Les différents types d'attributs.

Qualitative	Nomina	les valeurs sont juste des noms différents. Par exemple : les codes postaux, les couleurs, le sexe.
Qualitative	Ordinal	les valeurs reflètent un ordre, rien plus. Par exemple : bon, meilleur, mieux ou couleurs ordonnées par le spectre.
Quantitative	Intervalle	la différence entre les valeurs est significative par exemple, l'intervalle de température.
Quantitative	Ratio	rapport entre deux grandeurs. par exemple : les quantités monétaires, comme le salaire et le bénéfice et beaucoup de quantités physiques comme courant électrique, pression, etc.

Tableau 2.2. Les différentes échelles de données.

2.6. Les différentes techniques de clustering:

2.6.1. Clustering par partitionnement :

- **K-menas *:***

C'est l'un des algorithmes de clustering les plus répandus. Il permet d'analyser un jeu de données caractérisées par un ensemble de descripteurs, afin de regrouper les données "similaires" en groupes (ou clusters) [25].

La similarité entre deux données peut être inférée grâce à la "distance" séparant leurs descripteurs ; ainsi deux données très similaires sont deux données dont les descripteurs sont très proches. Cette définition permet deformuler le problème de partitionnement des données comme la recherche de ***K*** "données prototypes", autour desquelles peuvent être regroupées les autres données.

Ces données prototypes sont appelés *centrioles ;* en pratique l'algorithme associe chaque donnée à son centroïde le plus proche, afin de créer des *clusters*. D'autre part, les moyennes des descripteurs des données d'un *cluster*, définissent la position de leur centroïde dans l'espace des descripteurs : ceci est à l'origine du nom de cet algorithme (K- moyennes ou *K-menas* en anglais).
Après avoir initialisé ses centroïdes en prenant des données au hasard dans le jeu de données, *K-means* alterne plusieurs fois ces deux étapes pour optimiser les centroïdes et leurs groupes :
1. Regrouper chaque objet autour du centroïde le plus proche.
2. Replacer chaque centroïde selon la moyenne des descripteurs de son groupe.

Après quelques itérations, l'algorithme trouve un découpage stable du jeu de données : on dit que l'algorithme a convergé.

- **Le fonctionne le K-menas:**

L'algorithme peut être décomposé en 4 à 5 étapes :

1- Choisir le nombre de clusters :

Choix du nombre de clusters La première étape consiste à définir le nombre K de clusters dans lesquels nous allons regrouper les données. Choisissons K=3.

2- Initialisation des centroïdes :

Le centroïde est le centre d'un cluster, mais initialement, le centre exact des points de données sera inconnu. Donc, nous sélectionnons des points de données aléatoires et les définissons comme centroïdes pour chaque cluster. Nous initialiserons 3 centroïdes dans l'ensemble de données.

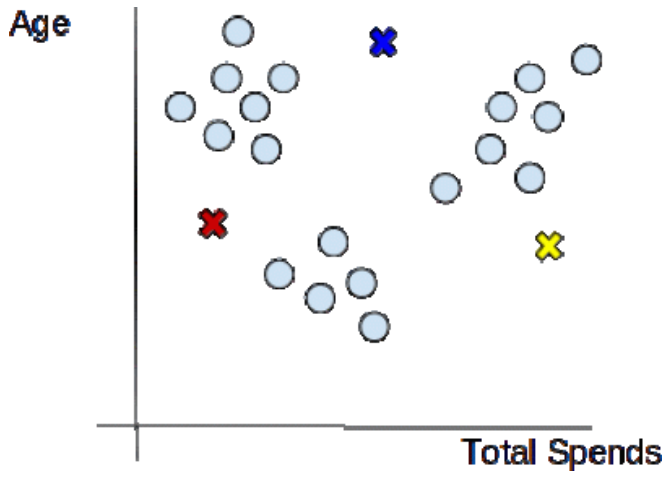

Figure 2.6. Deuxième étape de kmenas**.**

3-Attribuer les points de données au centre de regroupement le plus proche :

Maintenant que les centroïdes sont initialisés, l'étape suivante consiste à attribuer les points de données X n à leur centroïde de regroupement le plus proche KC

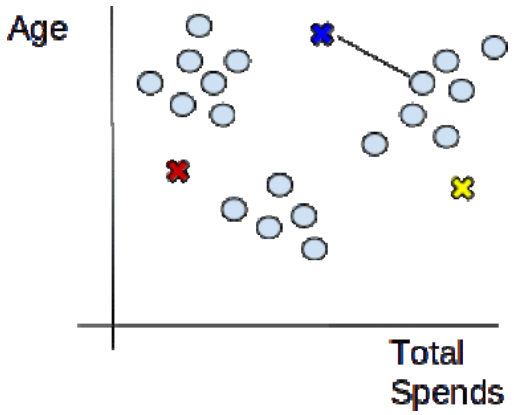

Figure 2.7. troisième étape de kmenas.

Dans cette étape, nous allons d'abord calculer la distance entre le point de données X et le centroïde C en utilisant la métrique de distance euclidienne

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Et ensuite, choisir le cluster pour les points de données où la distance entre le point de données et le centroïde est minimale.

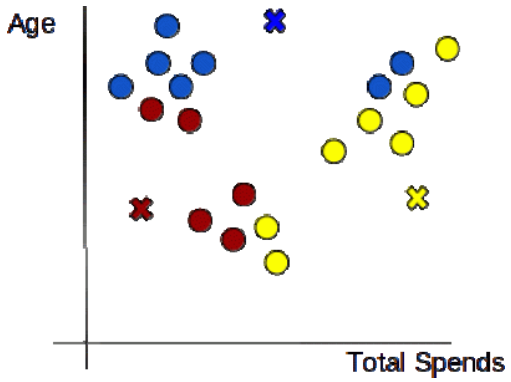

Figure 2.8. troisième étape suite de kmenas

4-Réinitialiser les centroïdes :

Ensuite, nous allons réinitialiser les centroïdes en calculant la moyenne de tous les points de données de ce cluster.

$$Ci = \frac{1}{|Ni|} \sum xi$$

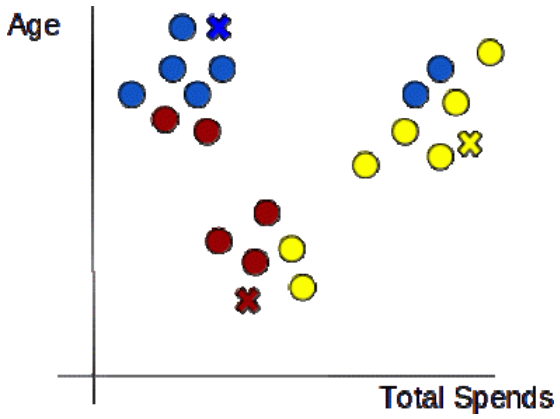

Figure 2.9. Quatrième étape de k-menas

5-Répétez les étapes 3 et 4 :

Nous continuerons de répéter les étapes 3 et 4 jusqu'à ce que nous ayons des centroïdes optimaux et que les attributions des points de données aux bons clusters ne changent plus.

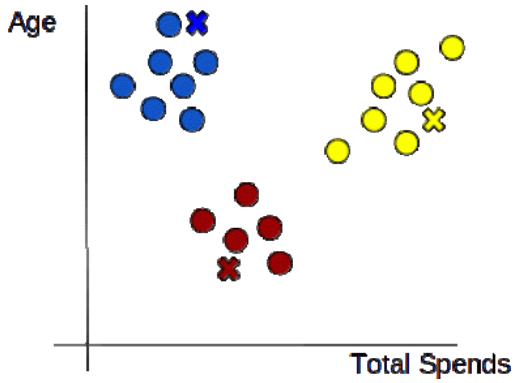

Figure 2.10. Cinquième étape suite de k-menas

- **K-Medoids :**

Le clustering K-Medoids est une technique de clustering par partitionnement similaire à l'approche KMenas, mais avec une différence clé dans le choix des centres des clusters. Alors que KMenas cherche à minimiser la variance au sein des clusters, K-Medoids minimise la somme des dis similarités entre les points de données et le point désigné comme le centre (ou medoid) de leur cluster. Cette méthode est souvent préférée dans des situations où la moyenne (utilisée dans KMeans) n'est pas une représentation significative du centre d'un cluster, ou lorsque la mesure de distance n'est pas euclidienne ,les étapes de K-medoid sont décrites comme suit :

1. Choisir k points initiaux. Ces points sont les médoïdes candidats qui sont destinés à être les points les plus centraux de leurs clusters.

2. Considérer l'effet de remplacer un des points choisis (médoïdes) avec un des points non choisis. Conceptuellement, ceci est fait de la façon suivante :

On calcule la distance entre chaque point non choisi et le médoïde candidat le plus proche et on calcule la somme de toutes les distances, cette somme représente le "coût" de la configuration actuelle. Tous les échanges possibles d'un point non choisi par un autre choisi sont considérés, et le coût de chaque configuration est calculé.

3. Choisir la configuration avec le coût le plus bas. Si c'est une nouvelle configuration, alors répéter l'étape 2.

4. Sinon, associer chaque point non choisi au point choisi le plus proche (médoïde) et arrêter.

L'un des grands avantages de cette méthode est sa robustesse. L'utilisation des médoïdes pour définir des clusters rend cette méthode très résistante contre les bruits de données, elle est un peu moins sensible au bruit que k-Menas, mais elle a un certain nombre d'inconvénients
Elle ne doit pas stocker une vaste quantité de l'information en plus des données originales dans la mémoire;

- En utilisant des médoïdes, cette méthode ne fournit aucune manière de décrire les clusters, autre que le détail d'adhésion.

- La recherche de médoïdes est plus coûteuse que le simple calcul de centroïdes [20]

2.6.2. Classification hiérarchique :

La classification hiérarchique est une technique complexe et polyvalente dans le domaine de l'apprentissage automatique et de l'analyse des données, visant à organiser et à regrouper des points de données similaires selon des critères spécifiques en groupes ou clusters. Cette approche se caractérise par la création d'une structure hiérarchique ou arborescente qui reflète les relations entre les points de données ou les sous-groupes de manière graduelle, que ce soit en fusionnant les éléments individuels en groupes plus grands étape par étape (classification hiérarchique agglomérat ive), ou en divisant un grand groupe en groupes plus petits progressivement (classification hiérarchique divisiez) [26].

Figure2.11. La classification hiérarchique de clustering

A-Classification hiérarchique agglomérative :

- Début : Chaque point de donnée est considéré comme un groupe individuel.
- Fusion : À chaque étape, la paire de groupes la plus proche est fusionnée ensemble, la proximité étant calculée selon des critères spécifiques (comme la distance la plus courte, la distance la plus longue, la distance moyenne, ou la distance de Ward).
- Répétition : Ce processus est répété jusqu'à ce que tous les points de données soient fusionnés en un seul grand groupe.

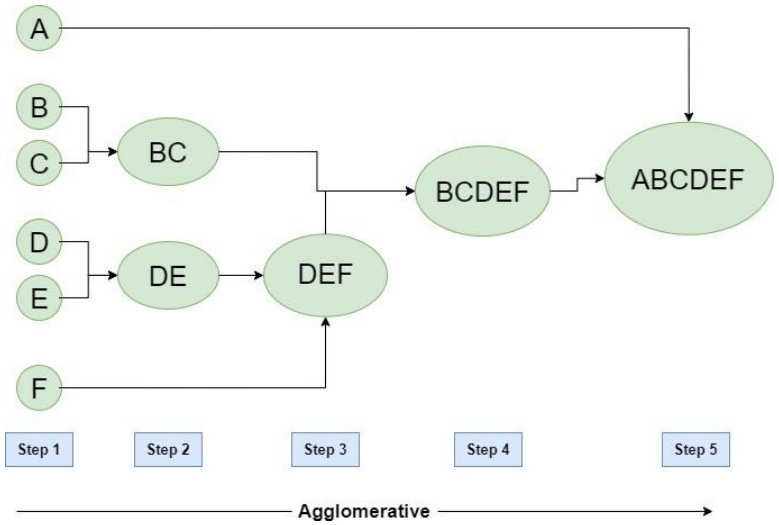

Figure2.12. Classification hiérarchique agglomérative.

B-Classification hiérarchique divisive :

- Début : Tous les points de données sont considérés comme un seul grand groupe.
- Division : À chaque étape, le groupe actuel est divisé en deux groupes selon des critères spécifiques, afin d'atteindre le plus haut degré de différence entre les groupeset le plus bas degré de différence à l'intérieur de chaque groupe.
- Répétition : Ce processus de division est répété jusqu'à obtenir un nombre spécifié de groupes ou jusqu'à ce que chaque groupe corresponde à un point de données individuel.

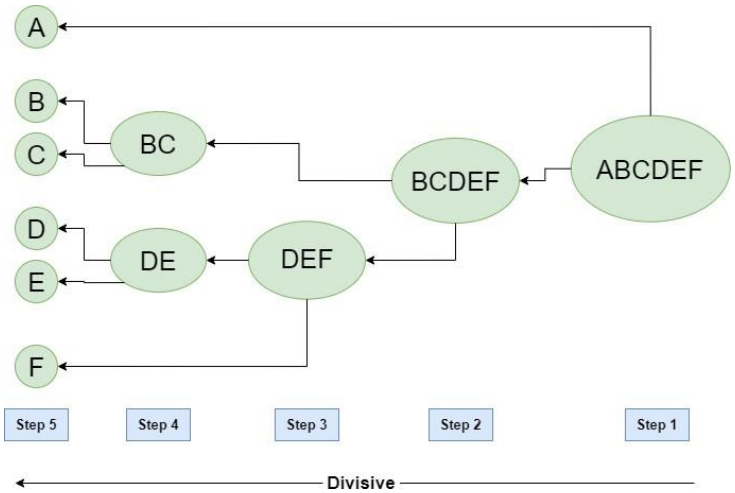

Figure 2.13. Classification hiérarchique divisive.

2.7. Les liens entre les groupes données :

Lors de la détermination de la manière de fusionner ou de diviser les groupes, une méthode de mesure de la proximité ou de la distance entre les groupes doit être adoptée. Les types principaux de liens comprennent[27] :

A-Single Link :

Cette méthode évalue la proximité entre deux clusters en utilisant la distance la plus courte entre n'importe quel membre d'un cluster et n'importe quel membre de l'autre cluster. Elle tendà favoriser la formation de clusters longs et minces car elle ne prend en compte que la paire la plus proche de points.

La distance entre les clusters A et B est définie comme la distance la plus courte entre n'importe quel membre d'un cluster et n'importe quel membre de l'autre cluster : DSingle $(A,B)=\min\{d(x,y):x\in A,y\in B\}$.

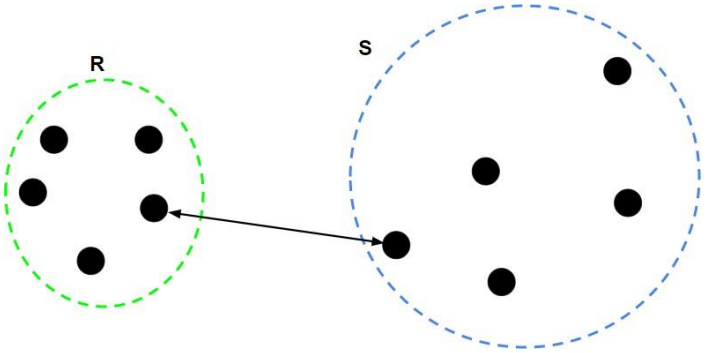

Figure2.14.graphe de Single Linke.

B-Complete Link:

À l'opposé du lien le plus proche, cette méthode utilise la distance la plus longue entre les membres de deux clusters différents pour évaluer leur distance. Cela conduit généralement à des clusters plus compacts et plus uniformément distribués, car elle considère la paire la plus éloignée de points.

La distance entre les clusters A et Best définie comme la distance la plus longue entre les membres de deux clusters différents :

DComplete$(A,B)=\max\{d(x,y):x\in A,y\in B\}$

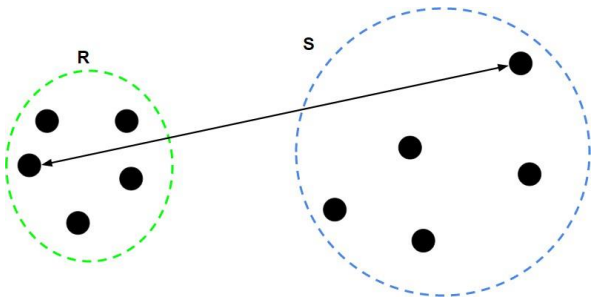

Figure2.15.graphe de Complete Linke.

C-Average Link:

Cette méthode calcule la distance moyenne entre tous les paires de points des deux clusters. Elle offre un équilibre entre les approches du lien le plus proche et du lien le plus éloigné, résultant en des clusters qui ne sont ni trop étirés ni trop compacts.

La distance entre les clusters *A*et*B*est définie comme la distance moyenne entre toutes les paires de points des deux clusters :

$$DAverage(A,B)=|A|\times|B|1\sum x\in A\sum y\in Bd(x,y)$$

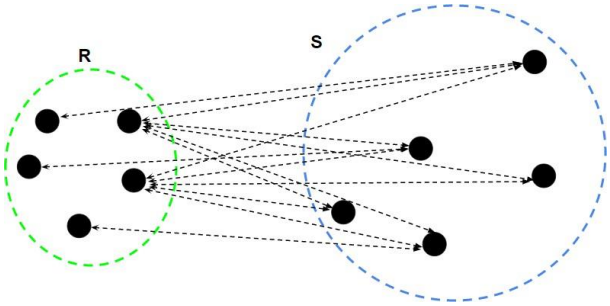

Figure2.16. graphe de Average Linke

2.8. Algorithme de recherche de coucou :

Les coucous pondent leurs œufs dans les nids des autres oiseaux hôtes. La première motivation fondamentale pour développer un nouvel algorithme d'optimisation est la ponte et la reproduction des œufs de coucou. Si l'oiseau hôte reconnaît que les œufs ne sont pas les siens, il les jettera hors de son nid ou videra simplement le nid et en construira un nouveau[28].

Chaque œuf dans un nid représente une solution. L'œuf de coucou représente une nouvelle et bonne option. La réponse obtenue est une nouvelle option basée sur celle existante avec certaines caractéristiques modifiées. Dans sa forme la plus élémentaire, chaque nid contient un œuf de coucou, et chaque nid contenant plusieurs œufs représente un ensemble d'options. La recherche de coucous a idéalisé ce comportement de reproduction. L'algorithme de recherche de coucou peut être appliqué à un large éventail de problèmes d'optimisation. Cet algorithme d'optimisation améliore l'efficacité, la précision et le taux de convergence.

Pour commencer, chaque coucou ne peut pondre qu'un œuf à la fois. Déposez-le ensuite dans un nid choisi au hasard. Deuxièmement, les meilleurs nids contenant des œufs de haute qualité seront transmis aux générations futures. Troisièmement, le nombre de nids d'hôtes disponibles est fixe.

2.8.1. Étapes de l'algorithme de recherche de coucou :

Découvrons maintenant les étapes de l'algorithme de recherche de coucou.

A. Initialisation :

Les coucous préfèrent pondre leurs œufs dans les nids d'autres oiseaux.

B. Vol de prélèvement :

C'est un vol ou une marche aléatoire. Les étapes sont définies en termes de longueurs de pas qui ont une certaine distribution de probabilité avec des directions aléatoires. Ce type de vol est observé chez différents animaux et insectes. Le mouvement suivant est déterminé par la position actuelle [29].

C.Calcul de la condition physique :

Le calcul de l'aptitude est réalisé en utilisant la fonction d'aptitude pour trouver la meilleure solution. Le nid est choisi au hasard. La condition physique de l'œuf de coucou (nouvelle solution) est ensuite comparée à celle des œufs hôtes (solutions) dans le nid. Si la valeur de la fonction fitness de l'œuf de coucou est inférieure ou égale à la valeur de la fonction fitness du nid choisi au hasard, le nid choisi au hasard est remplacé par la nouvelle solution.

D.Résiliation :

La fonction de fitness compare les solutions dans l'itération en cours et seule la meilleure solution est transmise plus loin. Si le nombre d'itérations est inférieur au maximum, le meilleur nid est retenu. Tous les coucous sont prêts pour leurs prochaines actions après avoir terminé les processus d'initialisation, de vol de prélèvement et de calcul de condition physique. L'algorithme de recherche de coucou se terminera une fois que le nombre maximumd'itérations sera atteint. Ces étapes sont applicables à tout problème d'optimisation. Dans de tels cas, chaque œuf de coucou et chaque nid de coucou jouent un rôle important.

2.9. L'objectif de l'analyse de clusters dans le marketing :

Analyse des clusters dans l'étau marketing pour regrouper des clients ou prospects Et des segments homogènes avec de nouveaux comprendre leur portement, leurs préférences et leurs besoins. L'objectif principal est de diviser une population de clients en groupes distincts, applications et autres services spécialisés. Ces activités sont basées sur l'évolution démographique. services, habitudes de consommation, spécifications des produits et valeurs. Tous nos critères sont permanents dans l'entreprise [30].

2.10. Les avantages de l'analyse de clusters :

Les avantages de l'analyse de clusters dans le marketing ont les suivants :

1. **Segmentation du marché** : L'analyse des clusters permet la segmentation du marché en groupes de groupes, ainsi que la capacité de suivre l'évolution du marketing. Ce cluster peut avoir une caractéristique distincte avec des stratégies marketing spécifiques qui s'adaptent à ses caractéristiques uniques.
2. **Personnalisation des offres** : En comparant les produits existants et les particularités du réseau, les entreprises peuvent proposer aux gens des offres de produits, de services et de communication. Cela permet à l'utilisateur de continuer les messages marketing et d'accéder à l'engagement client.
3. **Optimisation des ressources** : L'analyse des clusters permet decommercialiser toutes les ressources sans avoir besoin d'achats ou de locations supplémentaires. Les entreprises speuvent concentrer leurs efforts et leurs investissements sur les clusters les plus prometteurs, ce peut améliorer l'efficacité des activités marketing.
4. **Prévision des comportements futurs** : Etudiant les clusters et en analysant les modèles de comportement passés, les entreprises speuventirer des conclusions sur les tendances et les comportements futurs. Cela peut vous aider à protéger vos clients, à développer de nouvelles offres et à prendre des décisions stratégiques prédéterminées.

Dans le rapport, l'analyse des clusters dans le marketing est une méthode utilisée pour segmenter le marché, personnaliser les stratégies marketing, optimiser les ressources et se préparer.

Les activités futures. Il permet aux entreprises d'accompagner leurs clients et de développer des initiatives marketing plus efficaces.

2.11. Conclusion :

Le clustering de données est une pierre angulaire dans le domaine de l'analyse des données et de l'apprentissage automatique, offrant une voie puissante pour l'exploration et la compréhension des ensembles de données complexes. Que ce soit à travers la visualisation de données, la création de prototypes, l'échantillonnage ou l'amélioration des modèles prédictifs, les applications du clustering sont vastes et variées, reflétant son importance et sa flexibilité. Bien que les algorithmes de clustering comme les K-menas et la classification hiérarchique soient largement utilisés, le choix de l'algorithme approprié dépend des spécificités du jeu de données et des objectifs de l'analyse. En dépit des défis tels que la détermination du nombre de clusters ou la sensibilité au bruit et aux outliers, le clustering reste un outil indispensable pour les data scientists cherchant à extraire des connaissances précieuses des données brutes. En somme, le clustering enrichit significativement notre capacité à analyser et interpréter les données, ouvrant la voie à des découvertes et innovations dans de nombreux domaines.

Chapitre 03 :

Description de l'application d'analyse des données des clients de vente en gros

3.1. Introduction :

Dans ce chapitre, notre principal objectif est de fournir une description générale des composants de base de notre système d'analyse de données.

Nous allons présenter les composants de l'environnement de travail ainsi que les différentes fenêtres de l'application.

3.2. Architecture général de notre système :

Notre système est composé des étapes suivantes :

- Le choix du jeu de données à analyser.
- L'analyse unidimensionnelle.
- L'analyse bidimensionnelle.
- L'analyse multidimensionnelle.
- L'analyse de représentations graphiques.

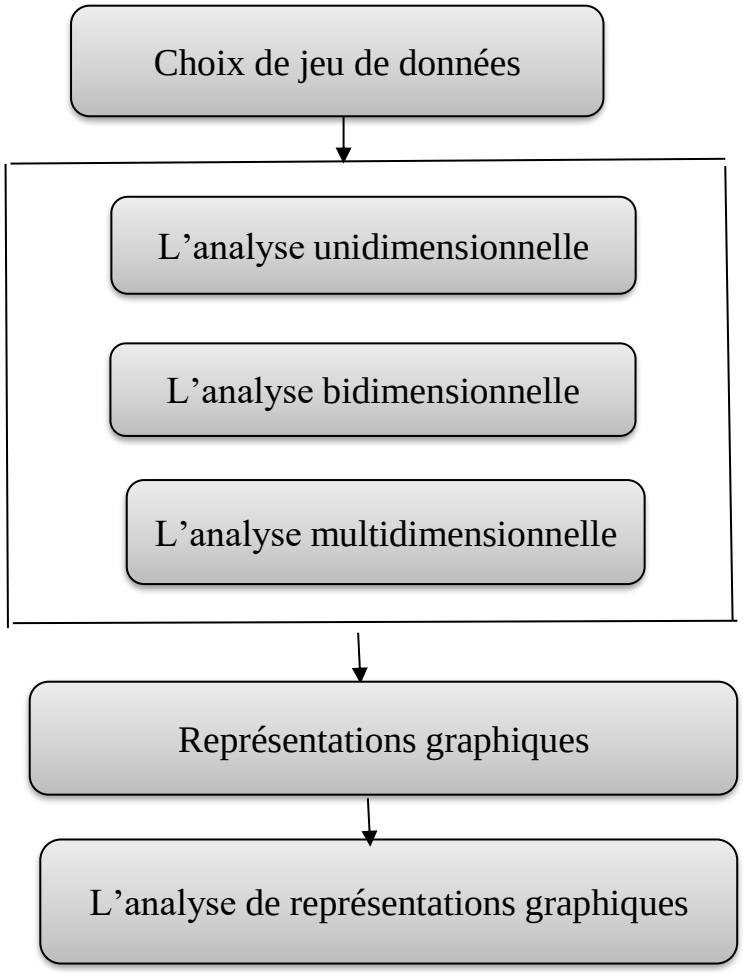

Figure 3.1. Architecture général de notre système.

3.3.Les étapes de notre système :

3.3.1. Choix du jeu du données :

dans cette étape , on peut choisir n'importe quel jeu de données , dans notre cas , on opté pour le jeu de données Wholesale Customer data .Le jeu de données utilisé ici est le jeu de données appelé WholesaleCustomer data extrait du site[31] .

Il contient les propriétés de 440 consommateurs .de. Ce jeu de donnes inclut 8 attributs : Fresh , Milk, Grocery, Frozen, Detergents_Paper, Delicassen, Channel , Région.

1) FRESH : dépenses annuelles (m.u.) sur les produits frais (Continu).

2) MILK : dépenses annuelles (m.u.) sur les produits laitiers (Continu).

3) GROCERY : dépenses annuelles (m.u.) sur les produits d'épicerie (Continu).

4) FROZEN : dépenses annuelles (m.u.) sur les produits surgelés (Continu).

5) DETERGENTS_PAPER : dépenses annuelles (m.u.) sur les détergents et les produits en papier (Continu).

6) DELICATESSEN : dépenses annuelles (m.u.) sur les produits de charcuterie (prêts à manger) (Continu).

CHANNEL : canal des clients - Horeca (Hôtel/Restaurant/Café) ou canal de détail (Nominal).

REGION : région des clients - Lisbonne, Porto ou Autre (Nominal).

3.3.2. L'Analyse Unidimensionnelle : Dans cette phase, chaque dimension du jeu de données est analysée de manière individuelle. Pour ce faire, nous avons choisi de présenter les histogrammes et les boxplots de chaque dimension.

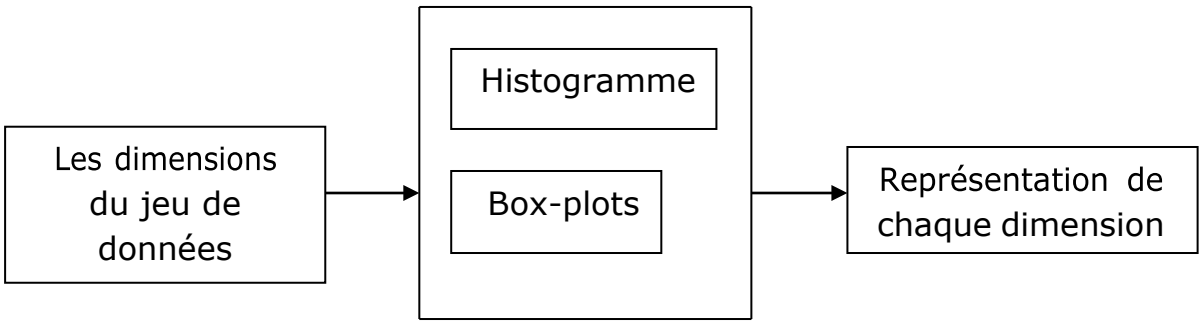

Figure 3.2. Analyse Unidimensionnel.

3.3.3. L'Analyse Bidimensionnelle: Dans cette étape, nous analysons les relations potentielles entre chaque paire de variables. Pour ce faire, nous avons opté pour le calcul de la matrice des corrélations et la création des box-plots pour chaque paire de dimensions.

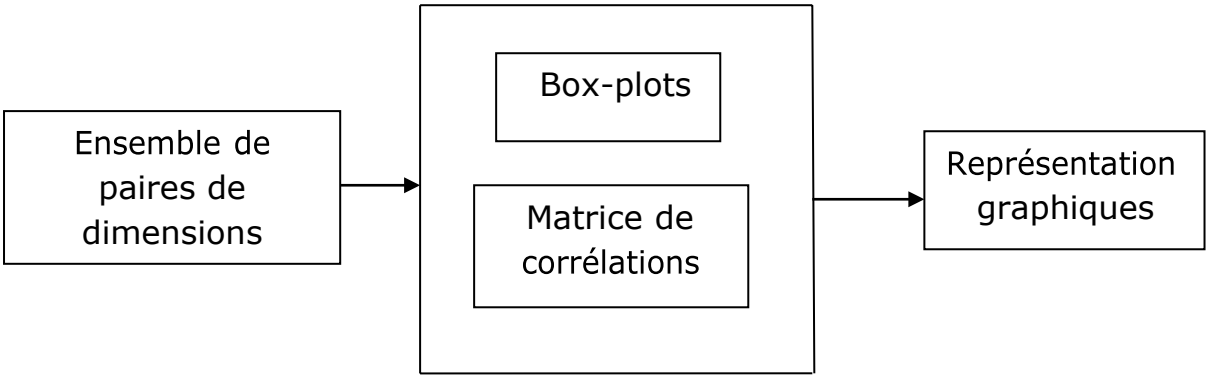

Figure 3.3. Analyse Bidimensionnelle.

3.3.4. L'Analyse Multidimensionnelle : Dans cette étape, nous avons décidé d'effectuer deux types d'analyses différents : le clustering et l'analyse factorielle, ainsi qu'une combinaison des deux.

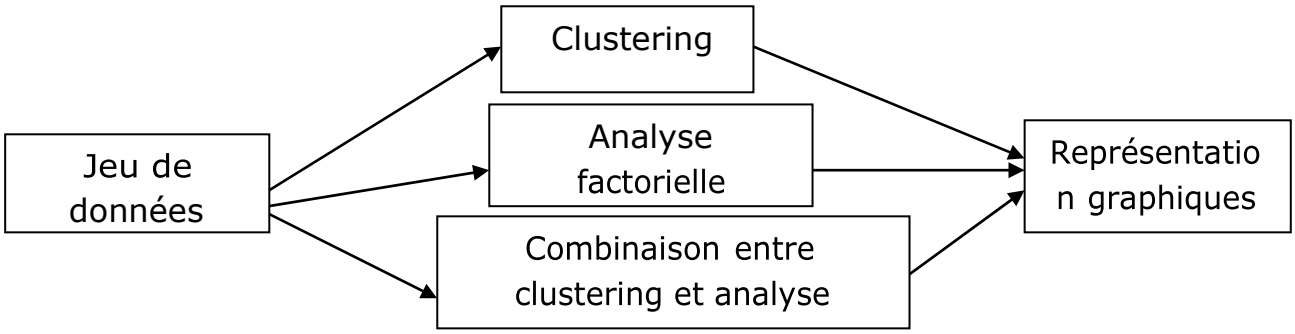

Figure 3.4. Analyse Multidimensionnelle.

3.3.4.1. Le CLustering :

Le clustering vise à regrouper les données en formant des groupes homogènes afin de réduire leur taille. À cet effet, nous avons choisi d'utiliser les algorithmes de clustering suivants : Kmeans, Single-Link et CuckooSearch (CS).

Ces algorithmes se distinguent par leur nature :

- Kmeans est un algorithme de clustering par partitionnement ;
- Single-Link est un algorithme de clustering hiérarchique ;
- CuckooSearch (CS) est un algorithme de clustering basé sur uneméta heuristique.

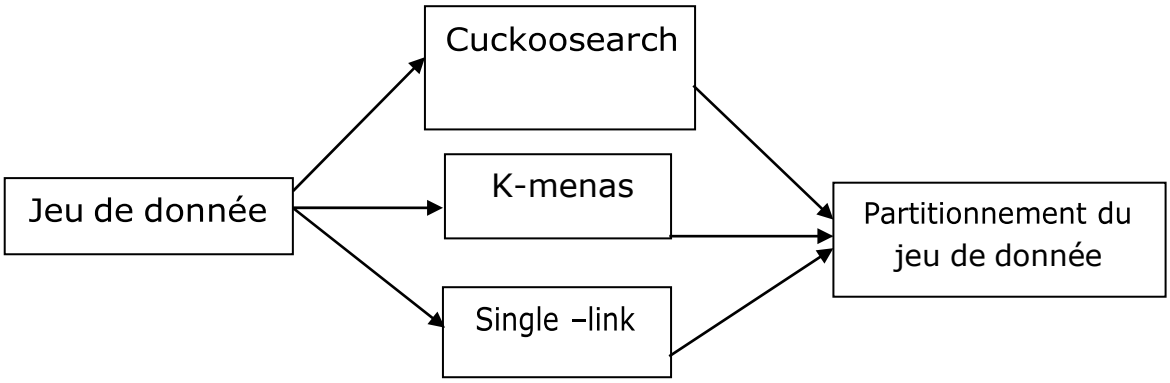

Figure 3.5.leclustering.

3.3.4.2. L'analyse factorielle

L'analyse factorielle vise à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques. Dans notre système, nous avons opté pour l'analyse en composantes principales (ACP), qui est une méthode courante parmi les méthodes factorielles disponibles.

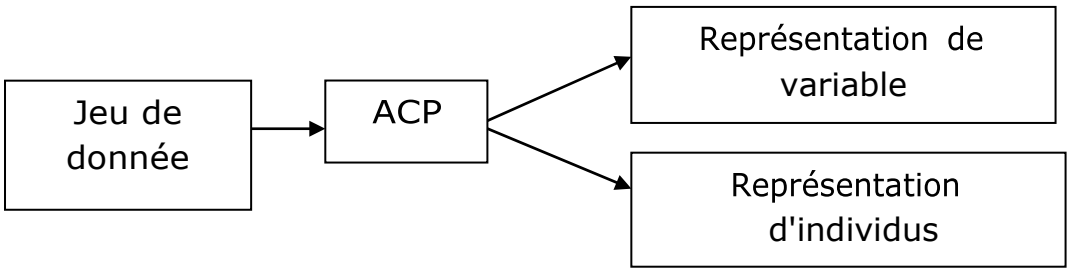

Figure 3.6. Analyse en composant principale .

3.3.4.3. Combinaison entre le clustering et l'analyse factorielle :

La combinaison est réalisée de la manière suivante : nous appliquons d'abord l'ACP pour réduire le nombre de variables du jeu de données, puis nous appliquons l'algorithme k-means sur le jeu de données réduit.

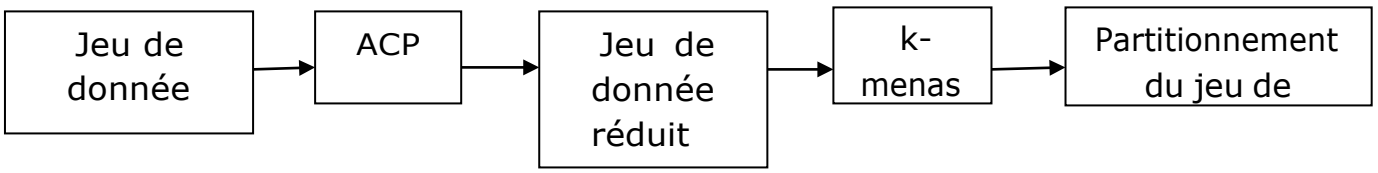

Figure 3.7 combinaison entre le clustering et ACP.

3.4. Implémentation:

Les composants utilisés dans ce travail est :

- Le système d'exploitation : Windows 7 professionnel.
- Le langage de programmation : MATLAB R2018a.

3.4.1. Représentation de MATLAB :

MATLAB est un langage de programmation de haut niveau conçu pour les ingénieurs et les scientifiques, qui permet d'exprimer directement les opérations mathématiques sur des matrices. Il est utilisé pour une gamme variée de tâches, allant de l'exécution de simples commandes interactives au développement d'applications complexes à grande échelle.

3.4.2.Déroulement d'une session de travail :

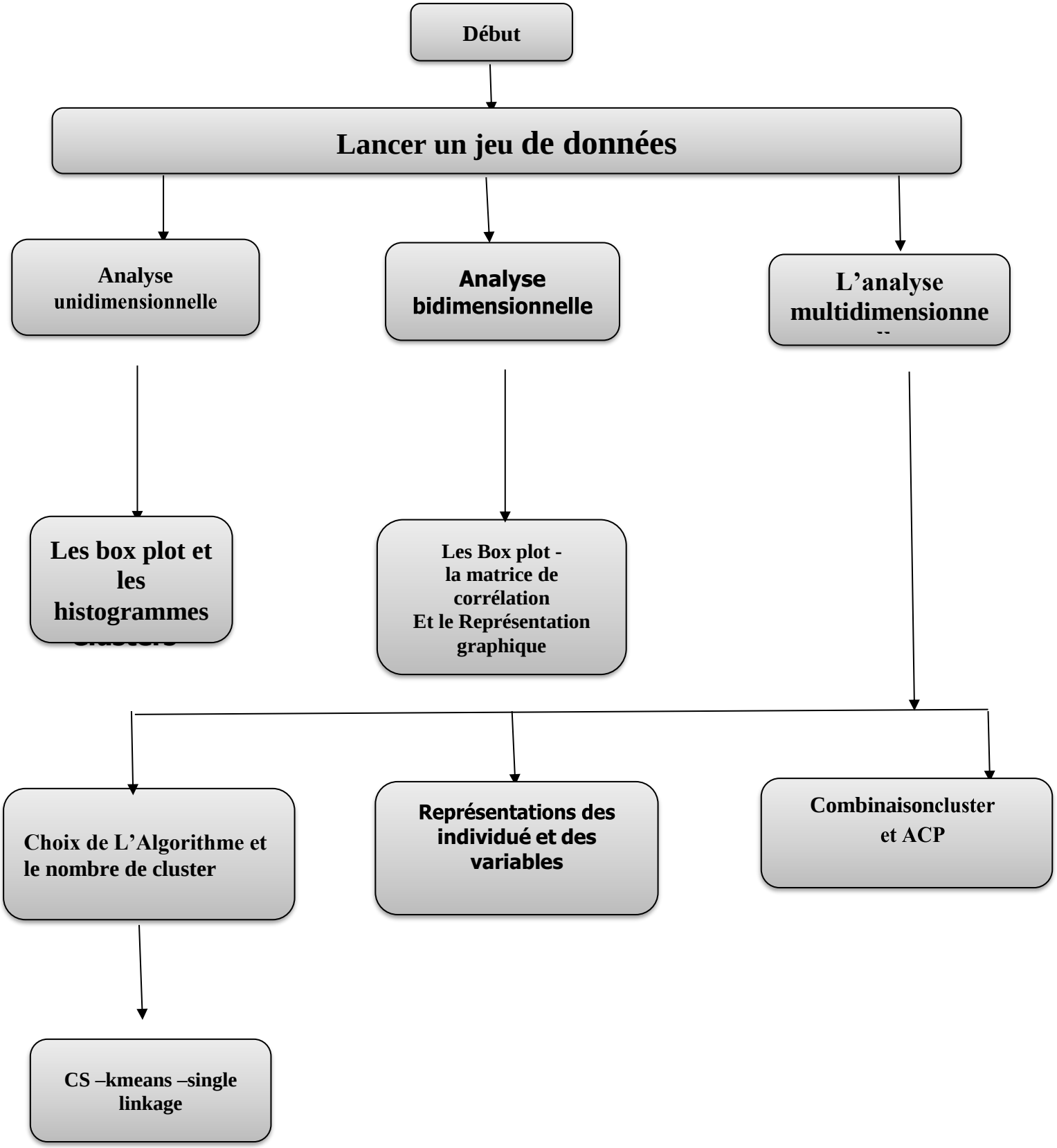

Figure 3.8 Déroulement d’une session du travail.

3.5. Présentation de l'application :

A-La fenêtre principale :

est présentée dans la figure, elle est divisée en 3 zones.

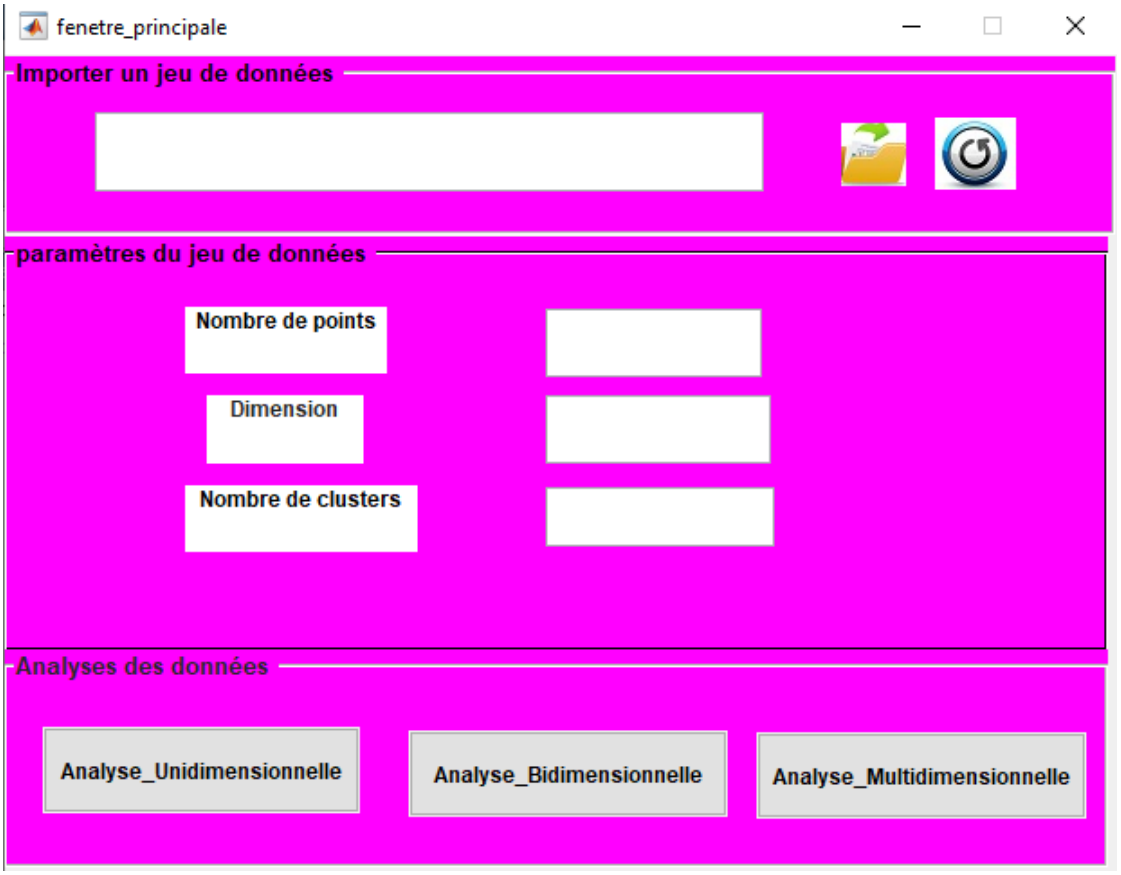

Figure 3.9. Fenêtre principal.

1)La première zone « Importer un jeu de données » permet à l'utilisateur de lancer le jeu de données, lorsqu'on clique sur le bouton  , une fenêtre s'affiche sur l'écran pour sélectionner le fichier du jeu de données.

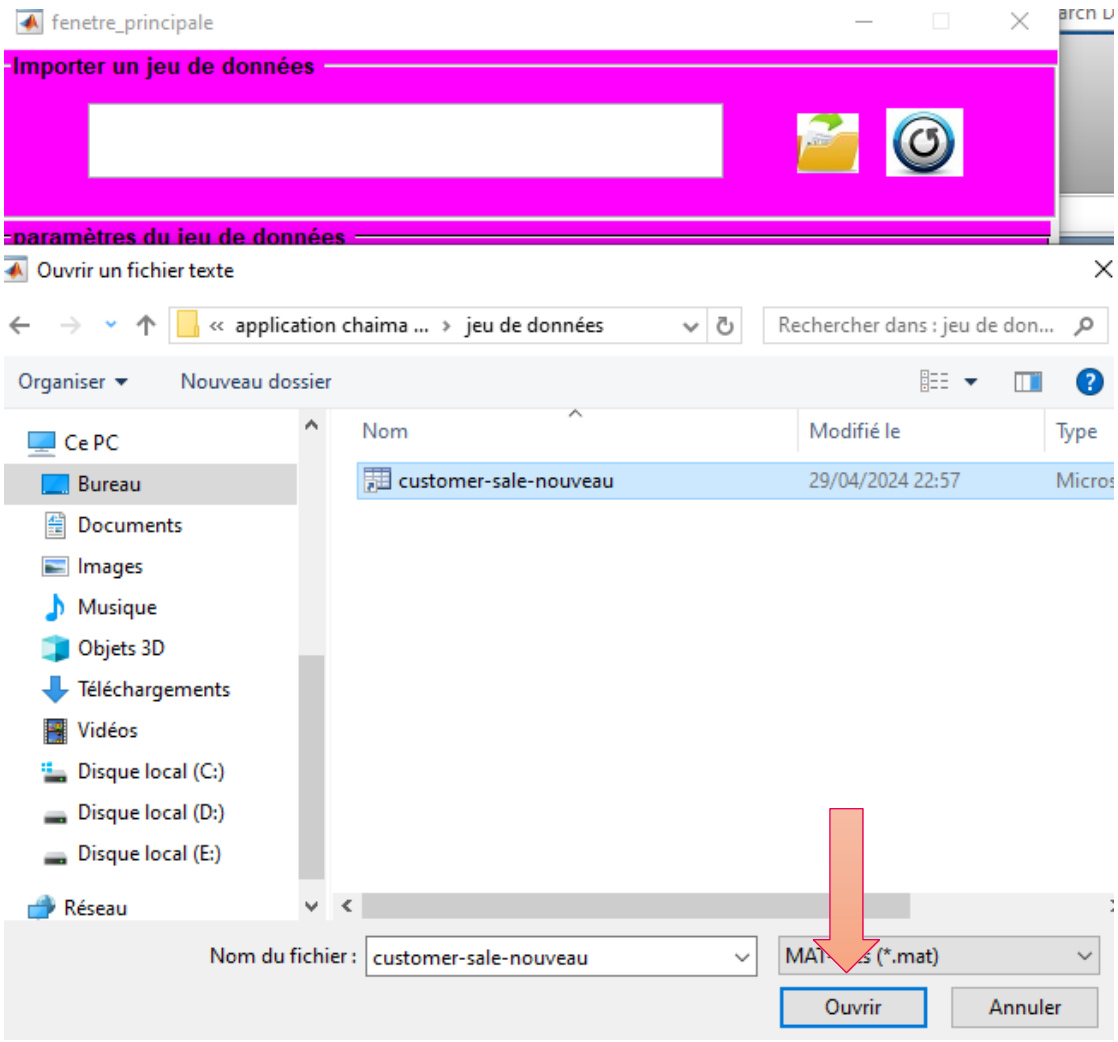

Figure 3.10. Fenêtre de la sélection du jeu de donnée.

2) Dans la deuxième zone « Paramètres du jeu de données » En cliquant sur le bouton « ouvrir» les paramètres du jeu de données qui sont : Le nombre de clusters K, le nombre de points nbr et nombre de dimensions nd. Qui peut être modifiée par l'utilisateur.

Le bouton  permet de sélectionner un autre jeu de données.

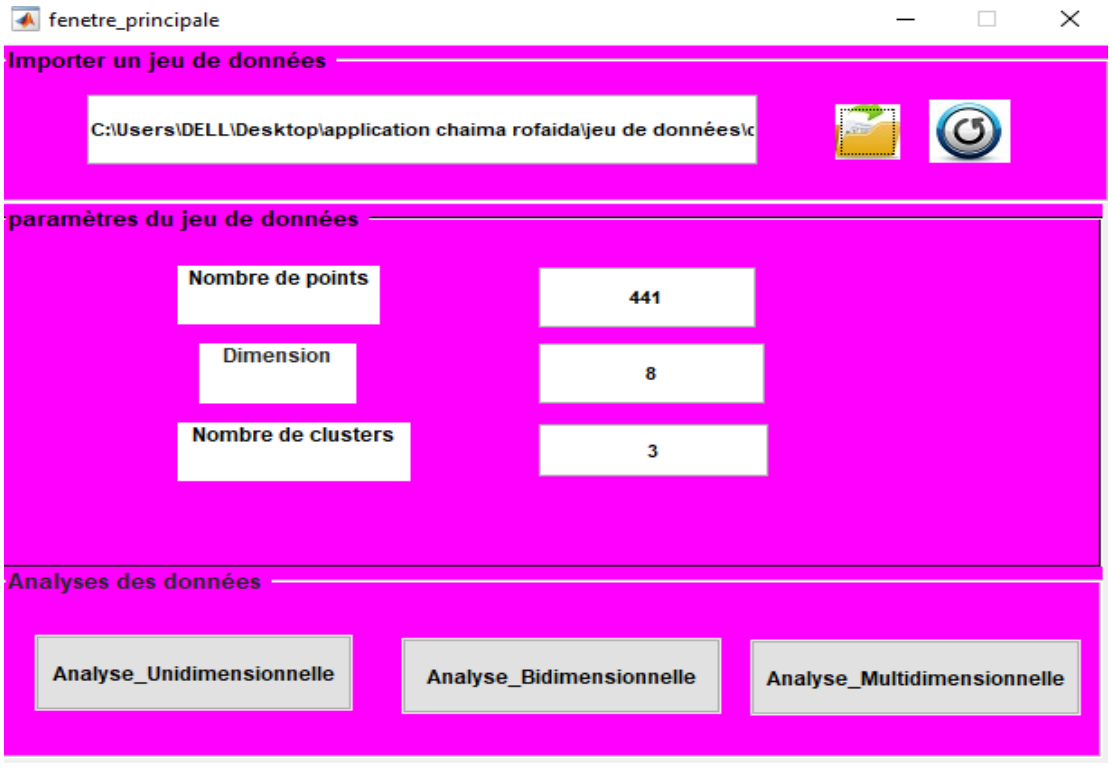

Figure 3.11. Paramètres du jeu de données.

3) La troisième zone « Analyse des données» : l'utilisateur peut choisir un type de l`analyse des données :

B. Fenêtre d'Analyse Unidimensionnelle :

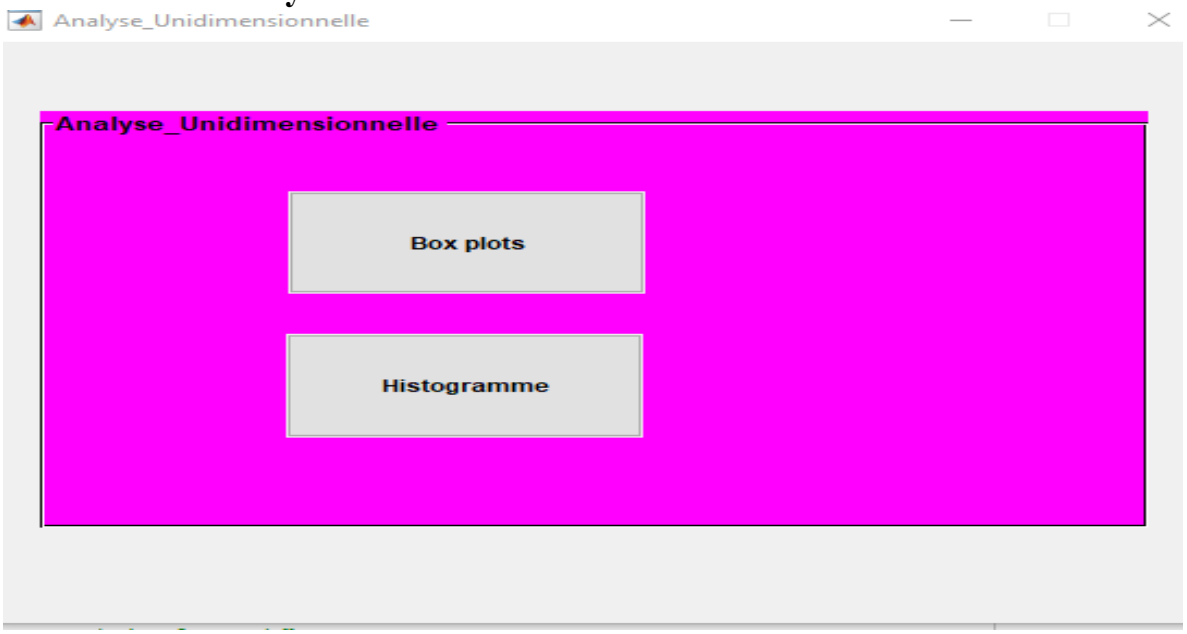

Figure 3.12. Représentation de l'analyse Unidimensionnelle.

Dans l'analyse Unidimensionnelle, l'utilisateur peut choisir la représentation graphique Box plots ou histogramme de chaque dimension.

C.Fenêtre d'Analyse Bidimensionnelle :

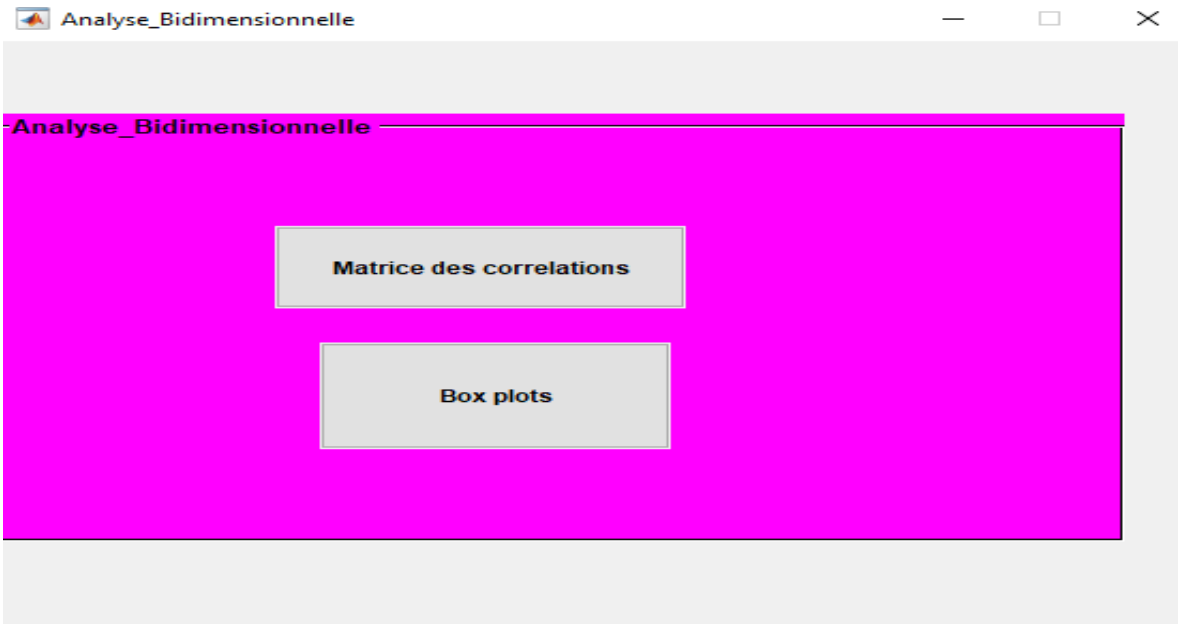

Figure 3.13. Représentation de l'analyse Bidimensionnelle.

Dans l'analyse Bidimensionnelle, l'utilisateur peut choisir les deux outils Box plots ou Matrice des corrélations de chaque paires de dimensions.

D.Fenêtre d'Analyse multidimensionnelle :

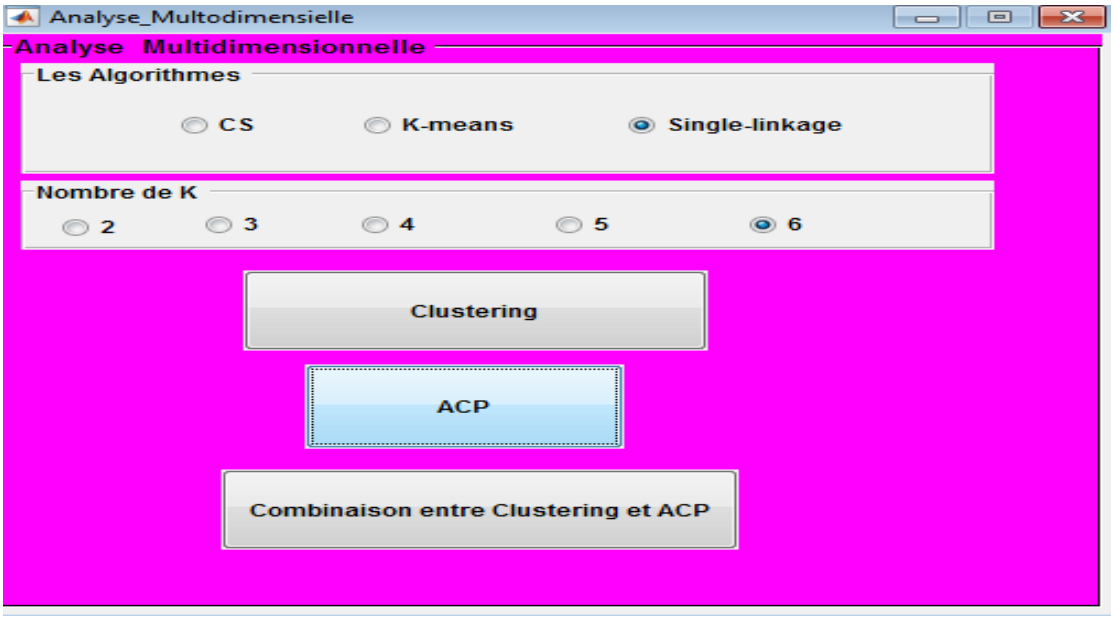

Figure 3.14. Représentation de l'analyse Multidimensionnelle.

Dans cette fenêtre, nous avons deux sections. Dans la première, nous sélectionnons un algorithme et un nombre de clusters K. Dans la seconde, nous lançons les différentes méthodes disponibles (clustering, ACP, combinaison de clustering et ACP).

3.6. Conclusion :

Dans ce chapitre, nous avons présenté la conception de notre application, nous avons montré le déroulement de notre application, ainsi que la présentation détaillée des fenêtres de applications.

Chapitre 04 :

Analyse et interprétation des résultats

4.1. Introduction :

Dans ce chapitre, nous présenterons les résultats de notre système d'analyse du jeu de données Wholesale Customer Data, accompagnés de leurs visualisations graphiques utilisant les boxplots, les histogrammes, la matrice des corrélations, l'ACP et les algorithmes de clustering. Nous examinerons également les analyses des représentations graphiques obtenues afin d'extraire des connaissances à partir des données.

4.2. Analyse des Résultats:

4.2.1. Analyse unidimensionnelle :

A-Analyse du jeu de données basée sur les box-plots:

La figure 4.1 montre les box plots de chaque variable du jeu de données. Conformément aux boxplots de la figure 4.1, nous avons des valeurs aberrantes dans notre jeu de données, mais la suppression de ces valeurs entraînera une perte de données. Ces valeurs aberrantes représentent les consommateurs qui effectuent des achats nettement supérieurs aux autres pour chaque catégorie de produit. Ces clients sont particulièrement importants pour le grossiste. Les box plots de la figure 4.1 fournissent un classement décroissant des montants des ventes pour les produits Fresh, Grocery, Milk, Detergents_paper, Frozen, et Delicassen.

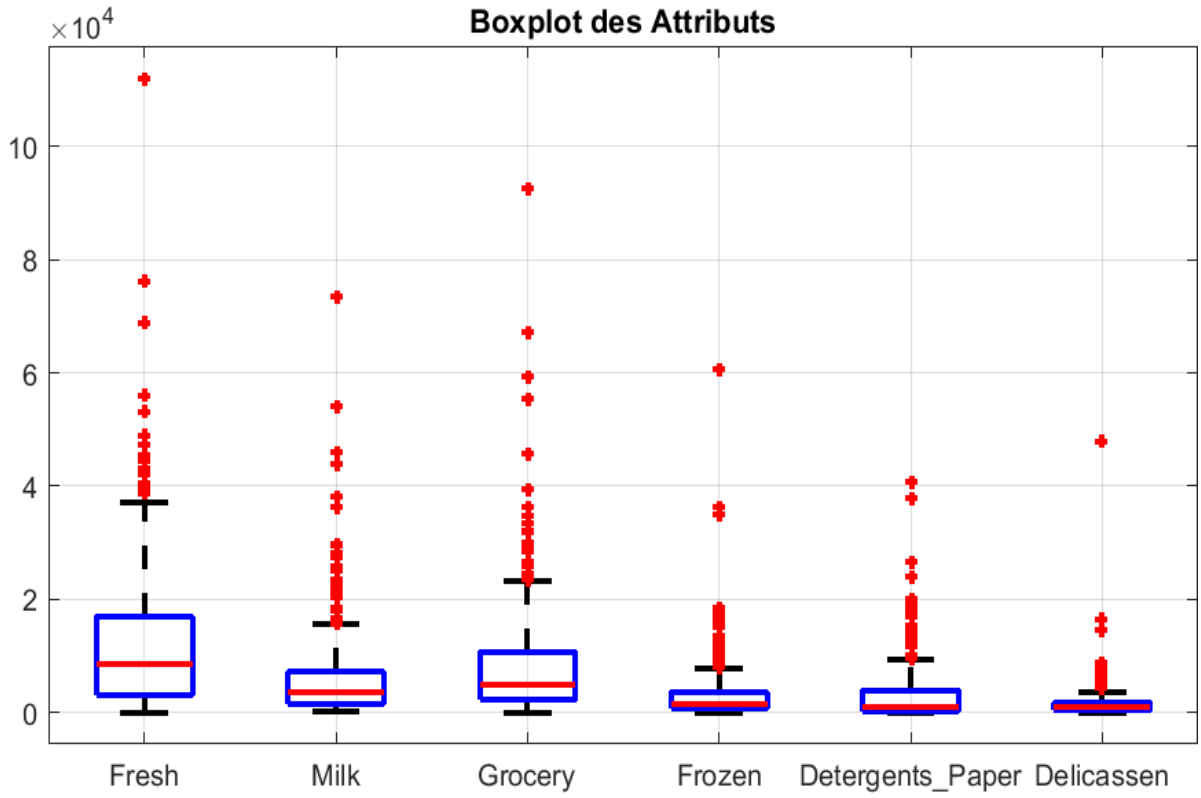

Figure4.1 Box plots de chaque dimension du jeu de données.

B-Analyse du jeu de données basée sur les Histogrammes:

La Figure 4.2 présente les histogrammes de chaque dimension du jeu de données. Pour l'attribut Channel, il existe deux valeurs possibles : Horeca et Retail. Horeca représente les consommateurs Hotel/Restaurant/Cafe, tandis que Retail représente les consommateurs détaillants. Il est évident sur les Figures 4.2 et 4.3 que les deux tiers des consommateurs sont des Hotel/Restaurant/Cafe.
Quant à l'attribut région, il comporte trois valeurs possibles : région 1, région 2, région 3. La majorité des consommateurs proviennent de la région 3.

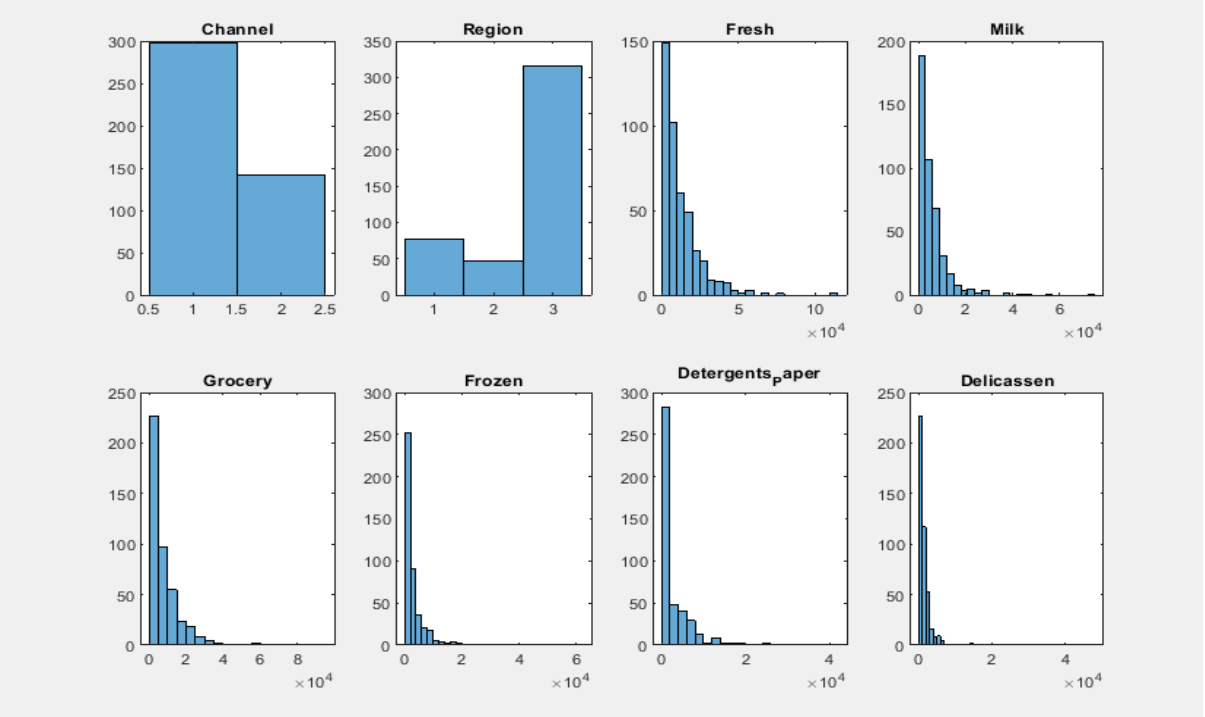

Figure 4.2 Histogrammes de chaque dimension du jeu de données.

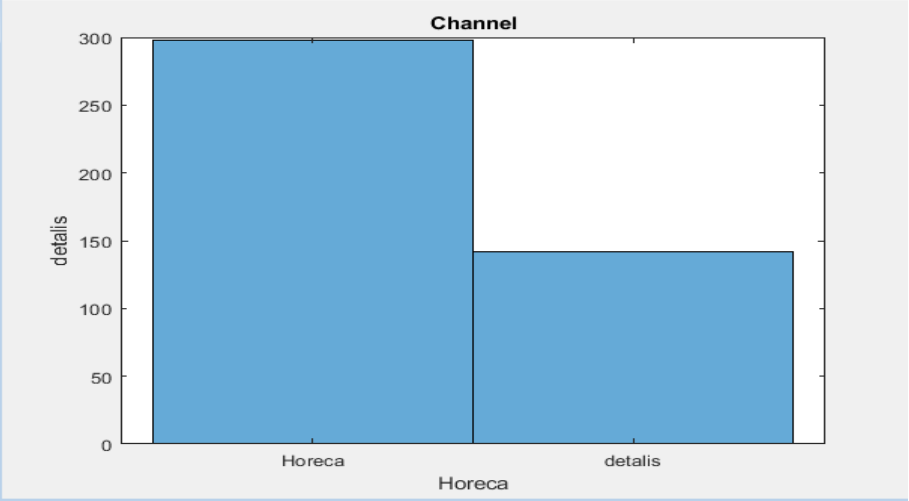

Figure4.3 Histogramme de l'attribut Channel.

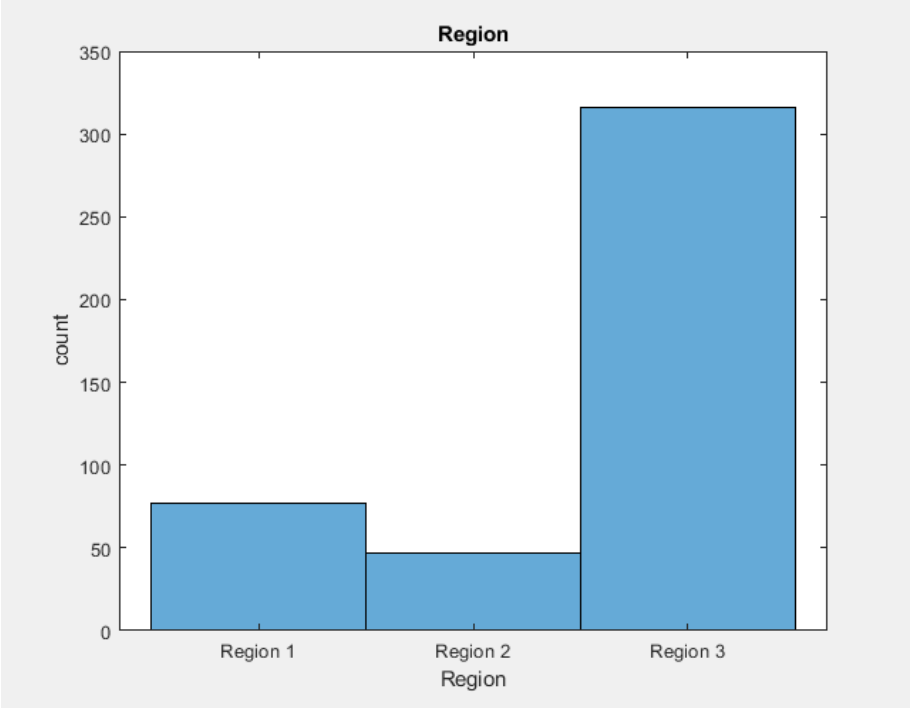

Figure4.4 Histogramme de l'attribut région.

4.2.2. Analyse bidimensionnelle :

A-L'analyse des box-plots :

La figure 4.5 représente la relation entre l'attribut Channel et chacun des attributs du jeu de données. Pour les produits Milk, Grocery et detergent_paper, delicassen les detaillants consomment plus que les Horeca. Pour les produit fresh et frosen, les consommateurs Horeca consomment plus que les détaillants.

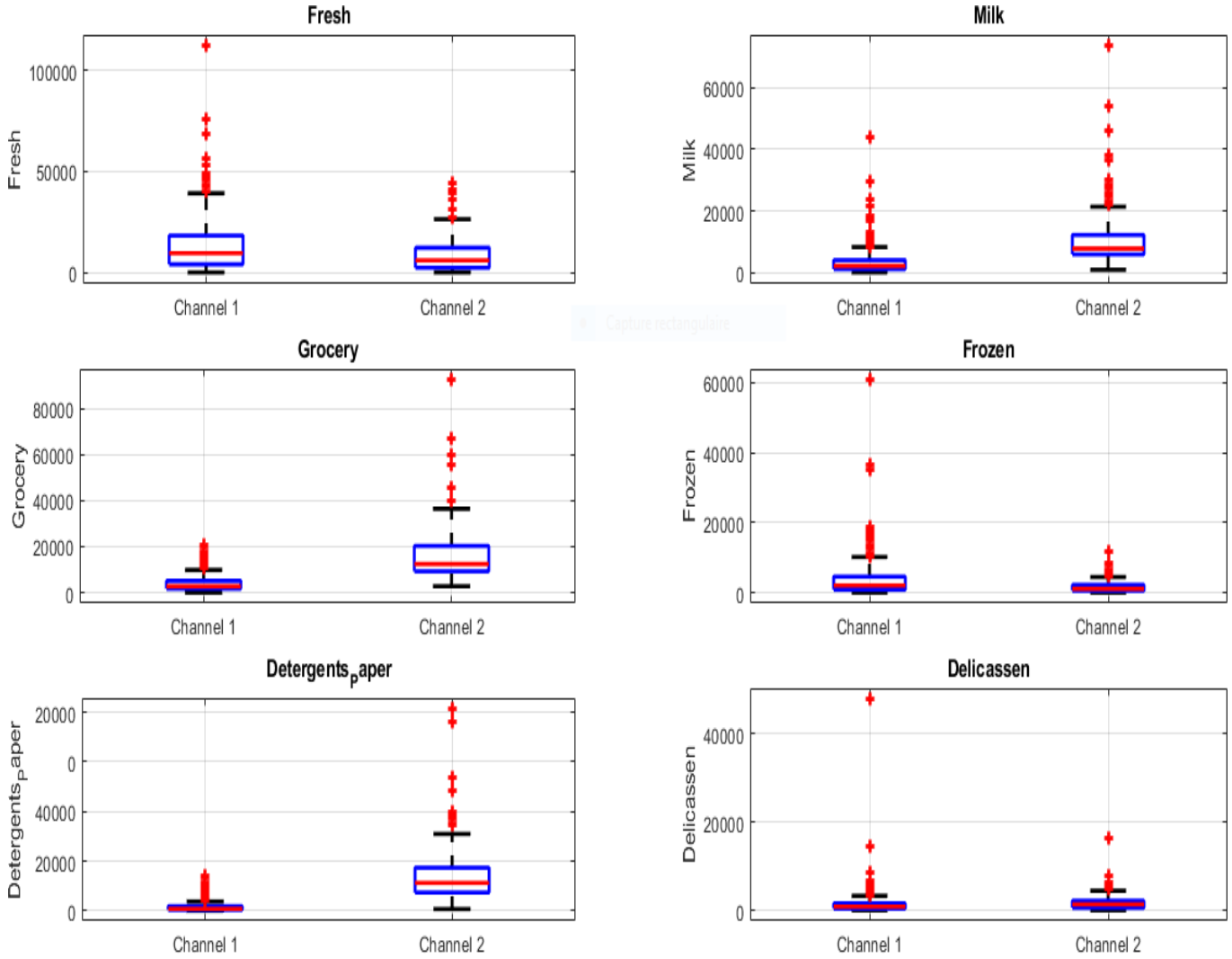

Figure4.5 Relation entre l'attribut Channel et les autres attributs.

B.la matrice de corrélation et la représentation graphique :

B.1. La matrice de corrélation :

Une matrice de corrélation est une table qui montre les coefficients de corrélation entre un ensemble de variables. Les coefficients de corrélation indiquent la force et la direction de la relation linéaire entre chaque paire de variables.

La figure 4.6 représente la matrice des corrélations du jeu de données.

Il existe une corrélation linéaire très forte et positive entre les produits Grocery et Detergents_paper (corrélation = 0.9246). Une corrélation assez forte et positive est également observée entre Milk et Grocery (corrélation = 0.7283), ainsi qu'entre Milk et Detergents_paper (corrélation = 0.6618), Grocery et Channel (corrélation = 0.6088), et Detergents_paper et Channel (corrélation = 0.636).

La signification de ces corrélations est illustrée dans la Figure 4.7. Une augmentation de la consommation de produits Grocery est associée à une augmentation de la consommation de Detergents_paper, et vice versa. De même, une augmentation de la consommation de Milk est corrélée à une augmentation de la consommation de Grocery, et vice versa.

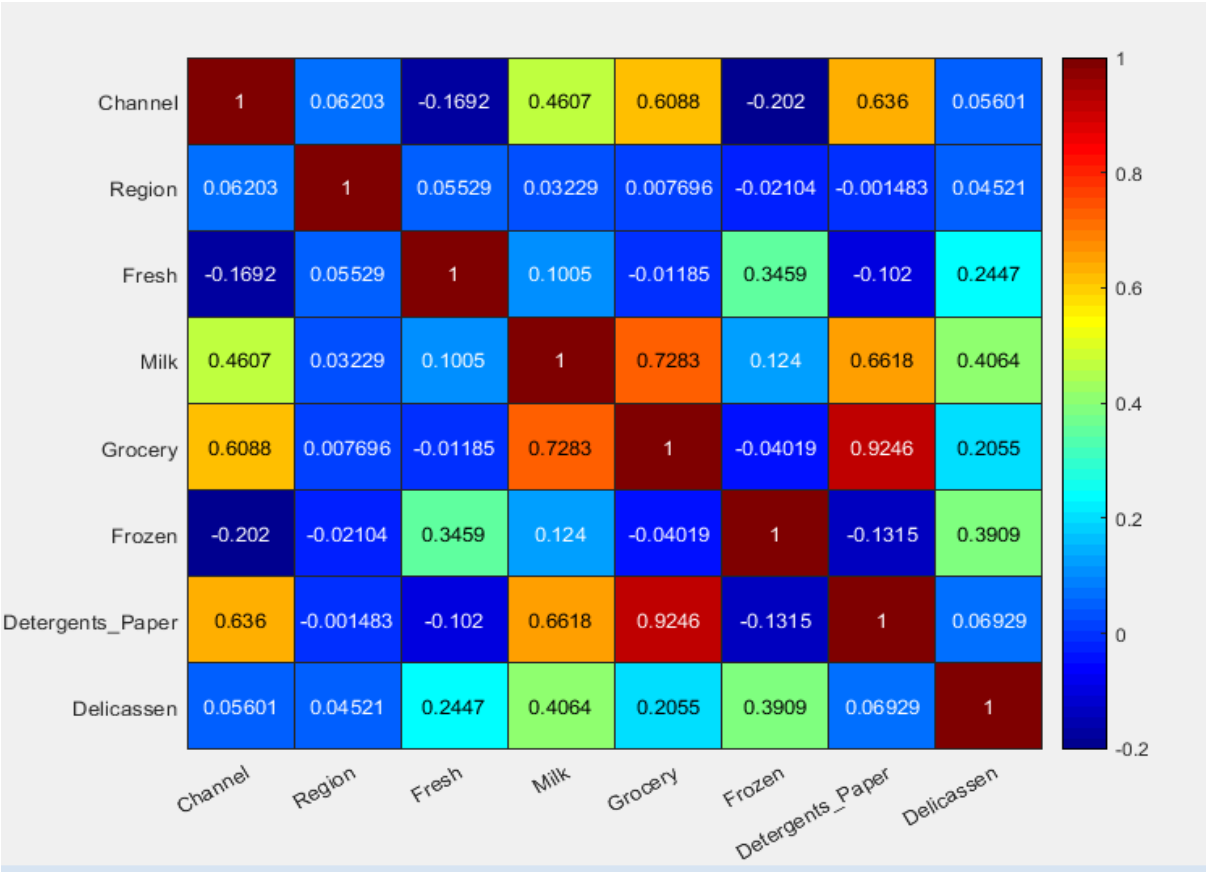

Figure4.6 Représentation graphique de jeux de donnée basée sur Matrice des corrélations .

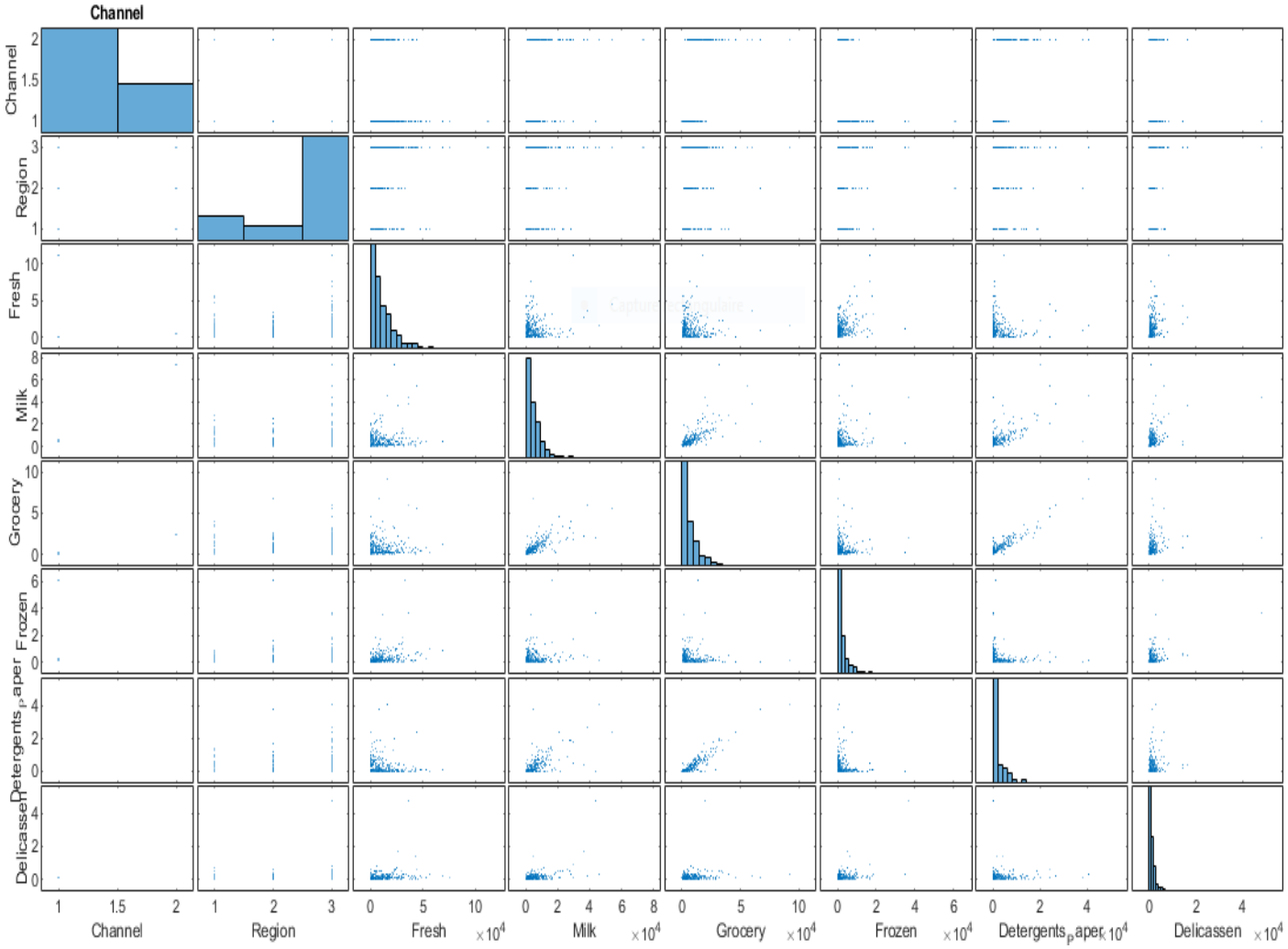

Figure 4.7. la représentation graphique de La matrice de corrélation.

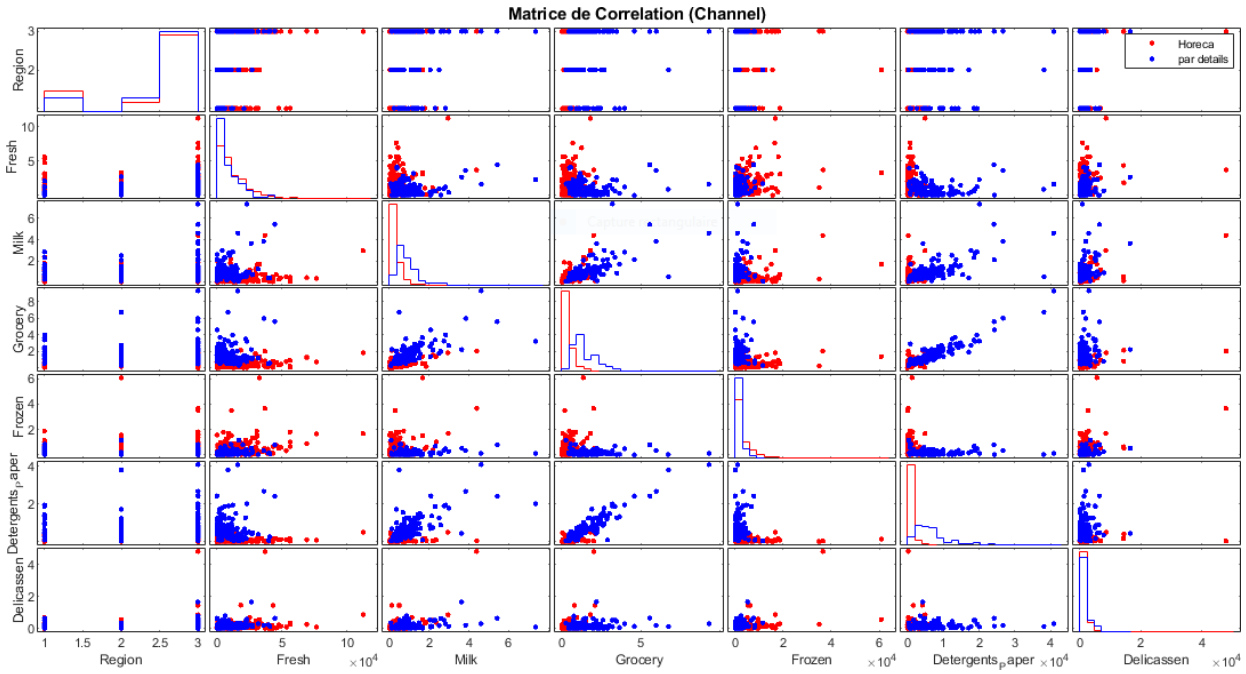

Figure 4.8. Représentation graphique de jeux de donnée basée sur la matrice de corrélation(CHANNEL).

La figure 4.8 montre la distribution des clients de deux types Horeca et détaillants.
Sur la figure 4.9, Il semble que les clients de chaque région présentent une tendance similaire lorsqu'ils achètent des produits en gros. Bien qu'il existe une différence d'ampleur, cela peut s'expliquer par le fait que la « Région 3 » fait référence à toutes les régions à l'exception de la région 1 et la région 2.

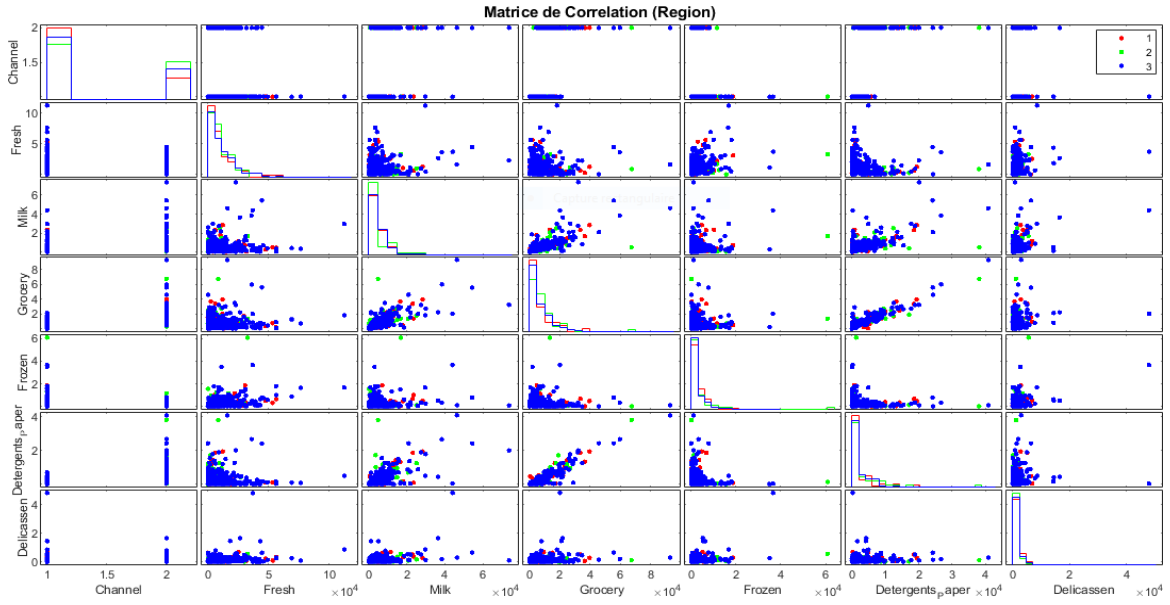

Figure4.9. Représentation graphique de jeux de donnée basée sur la matrice de corrélation (Régions).

4.2.3. Analyse multidimensionnelle :

4.2.3.1. L'analyse factorielle(ACP) :

L'application de l'analyse en composantes principales (ACP) fournit deux représentations graphiques : la représentation de variables et la représentation des individus.

Dans la Figure 4.10, une forte corrélation positive est observée entre les trois attributs (Milk, Grocery et Detergents_paper) et la première composante principale (dimension 1). Cela signifie que plus les consommateurs sont situés du côté positif de la première dimension, plus leur consommation de ces trois produits augmente (voir Figure 4.11).

Dans la Figure 4.10, une corrélation positive marquée est observée entre les trois attributs (Delicassen, Frozen et Fresh) et la deuxième composante principale (Dimension 2). Cela indique que les consommateurs positionnés du côté positif de la deuxième dimension ont tendance à consommer une quantité plus importante de ces trois produits par rapport aux consommateurs situés du côté négatif de la Dimension 2 (voir Figure 4.11).

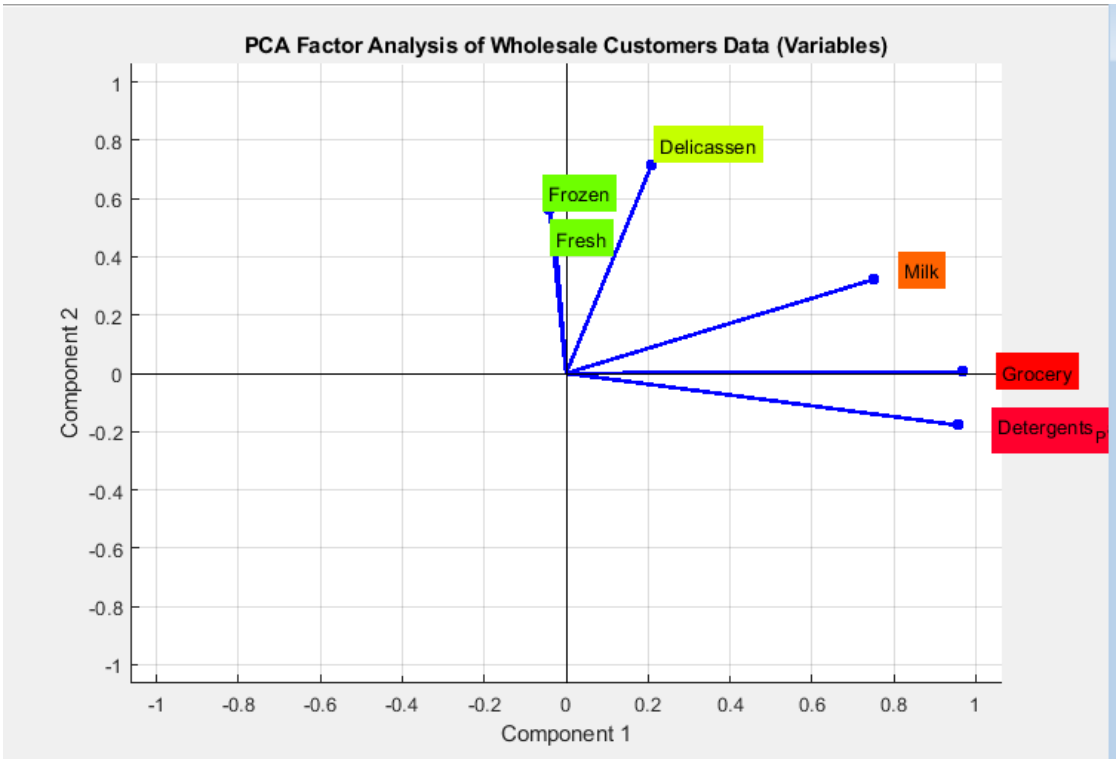

Figure 4.10 Représentation des variables.

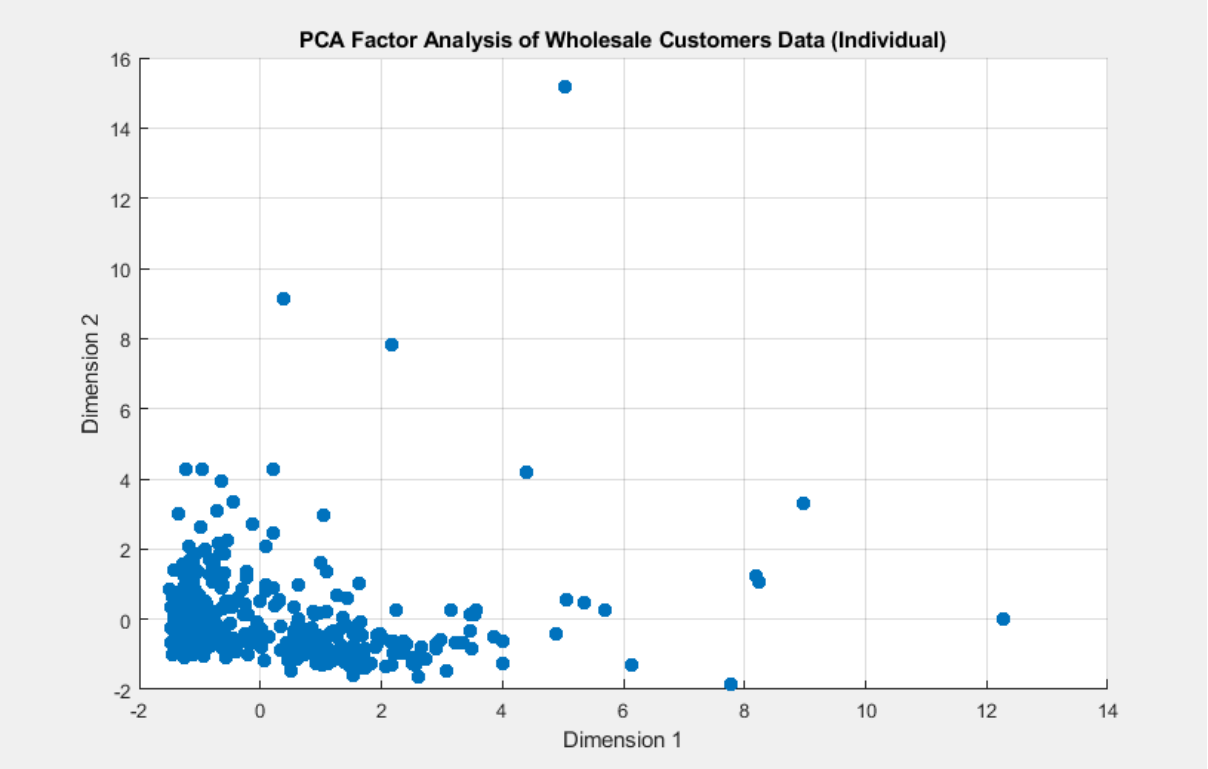

Figure 4.11 Représentation des individus.

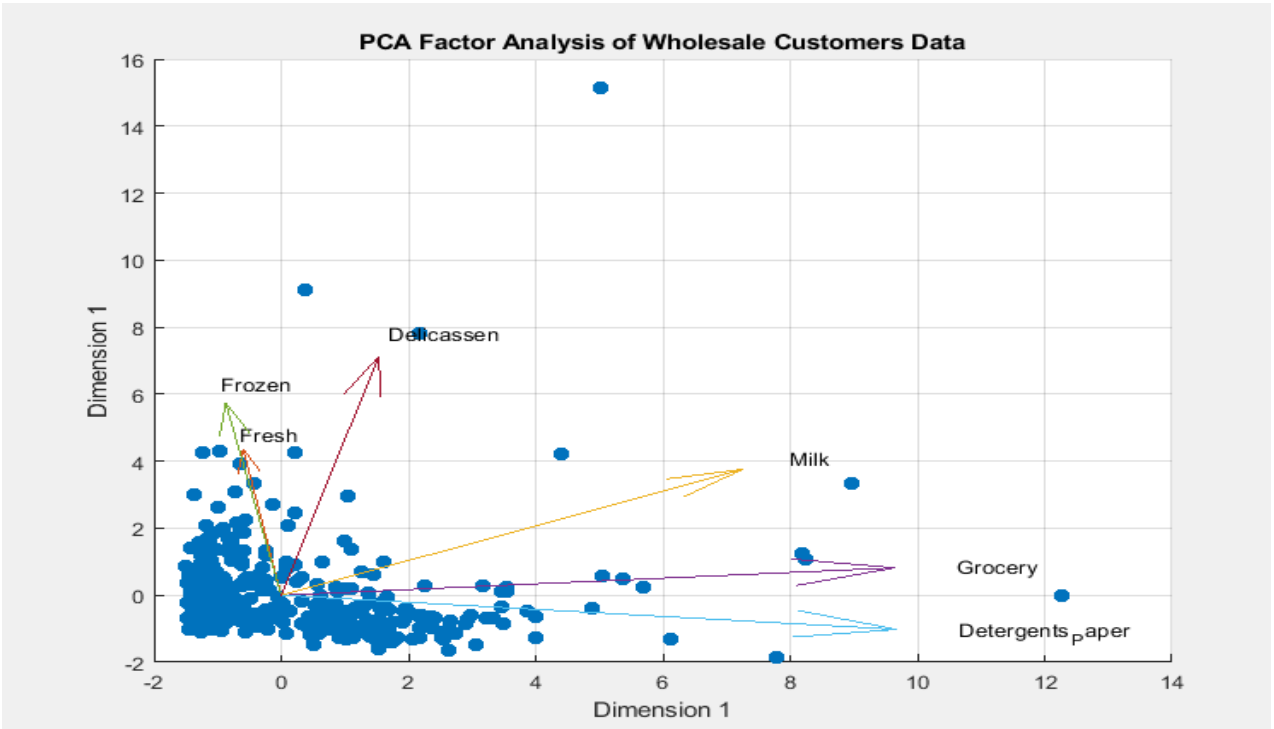

Figure 4.12 Représentation entre les individus et les variables.

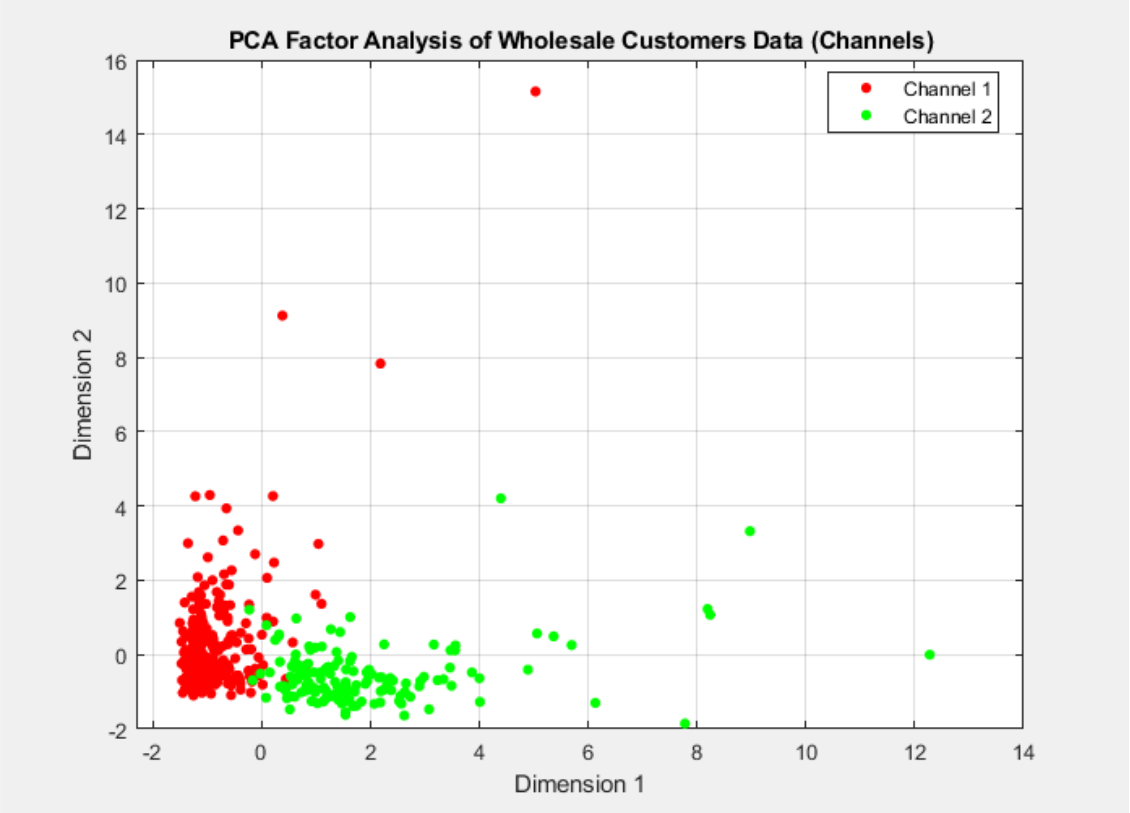

Figure 4.13 Représentation des individus libellés par Channel.

Dans la Figure 4.13, chaque consommateur est étiqueté selon la valeur de son attribut Channel, représenté par la couleur. On observe que les détaillants consomment davantage de produits (Milk, Grocery et Detergents_paper), tandis que les consommateurs HORECA ont une consommation plus élevée de produits (Delicassen, Frozen et Fresh).

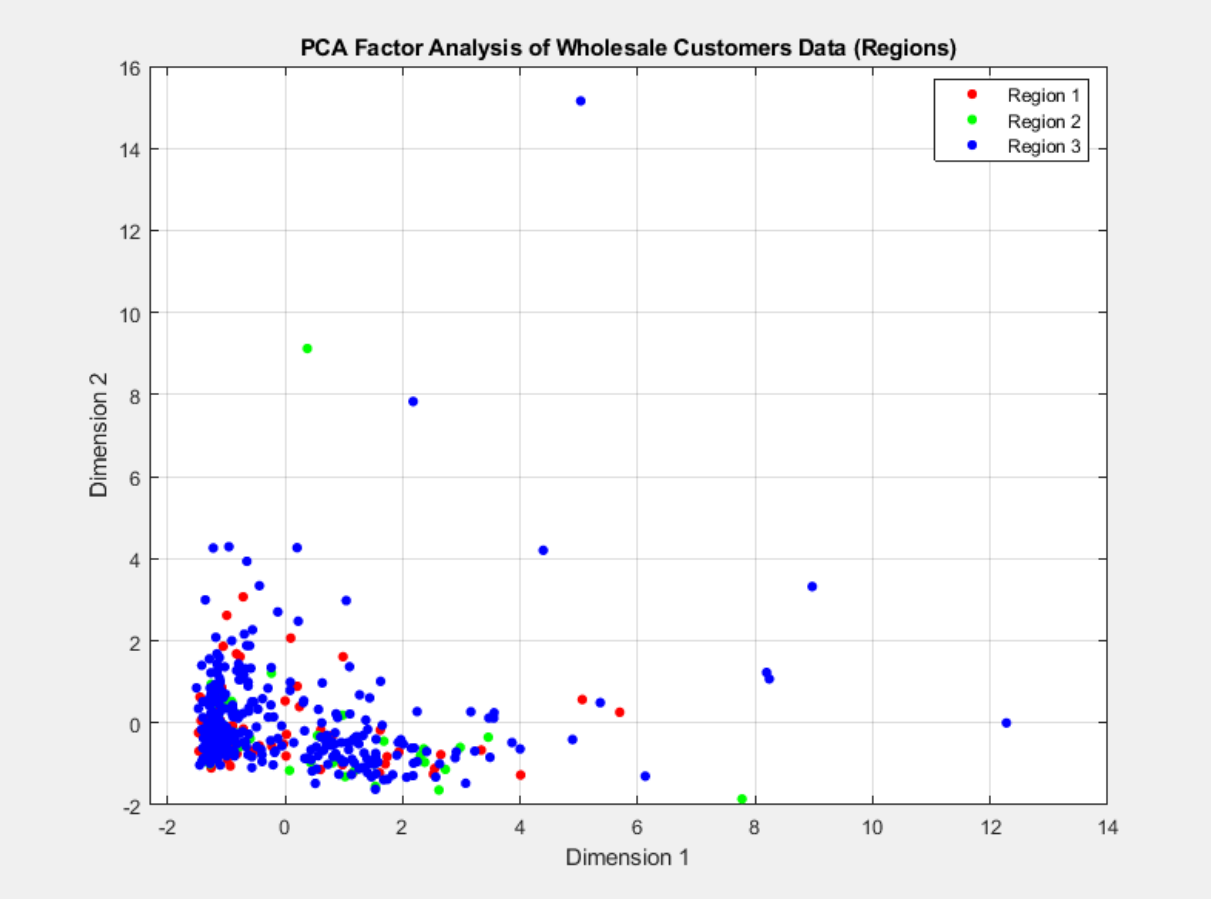

Figure 4.14 Représentation des individus libellés par région.

Pour la Figure 4.14, où les consommateurs sont étiquetés par l'attribut région, on peut observer que la consommation des produits est similaire dans les trois régions.

4.2.3.2. Le clustering :

A. Les résultats de l'application de Kmeans pour k=3 :

Les figure 4.15 à 4.22 donnent le partitionnement du jeu de données en utilisant l'algorithme kmeans pour un nombre de clusters égale à 3. Les figures montrent trois clusters différents où chaque cluster représente une catégorie de consommateurs.

Le clusters vert regroupe les consommateurs ayant une grande consommation de produits (Detergent_paper, Grocery, Milk).

Le cluster rouge regroupe les consommateurs ayant une consommation modérée de produits (Detergent_paper, Grocery, Milk).

Le cluster bleu regroupe les consommateurs ayant grande une consommation de produits (Fresh).

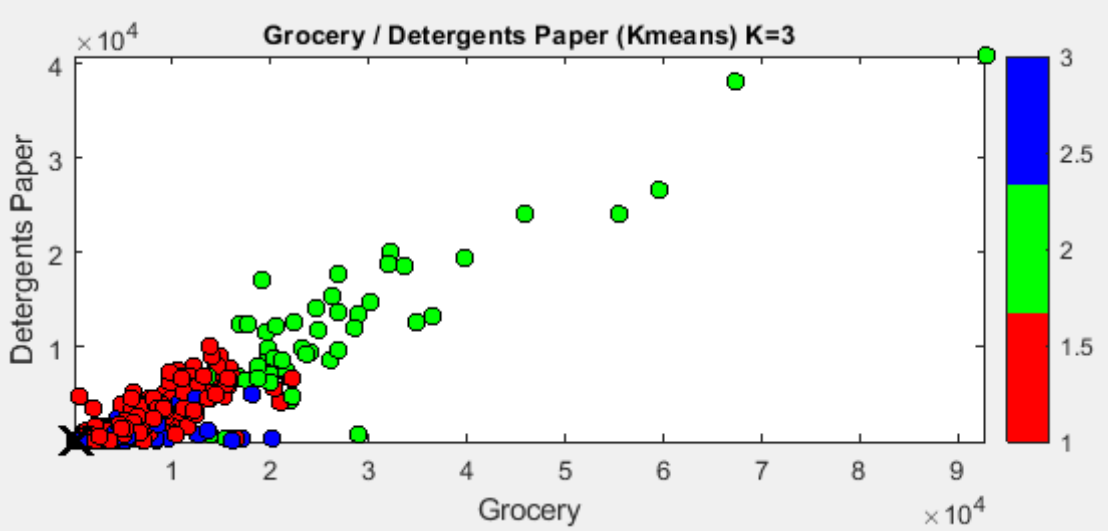

Figure 4.15 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions detergentpaper et grocery.

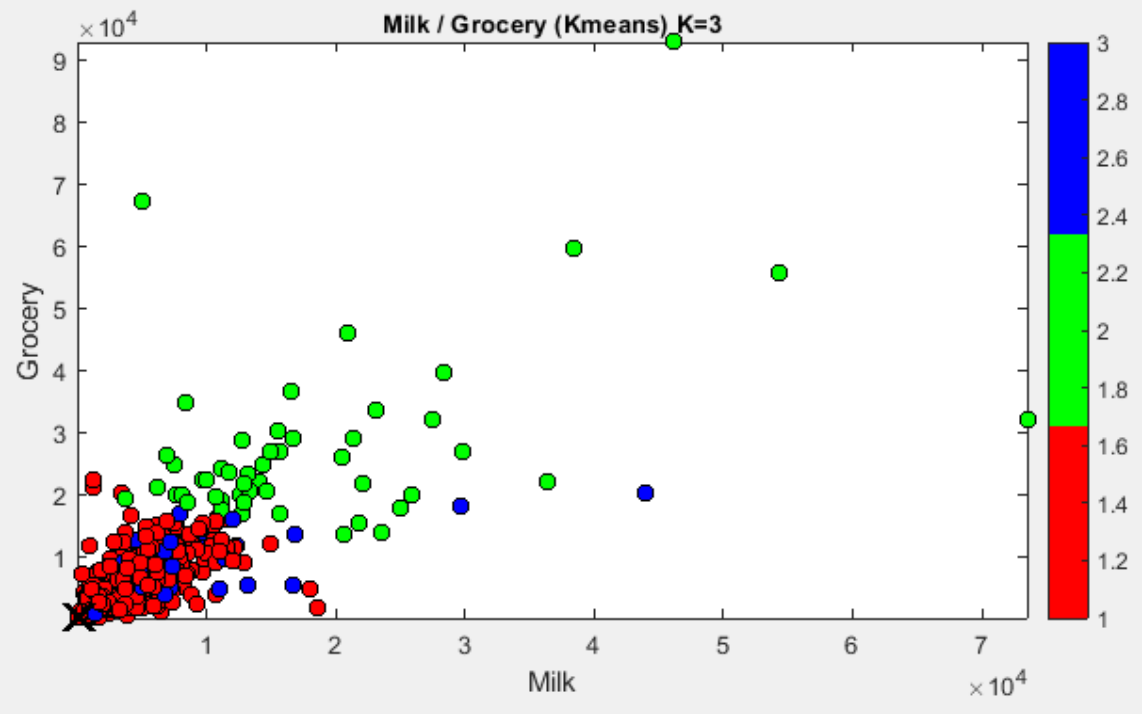

Figure 4.16 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Milk et grocery.

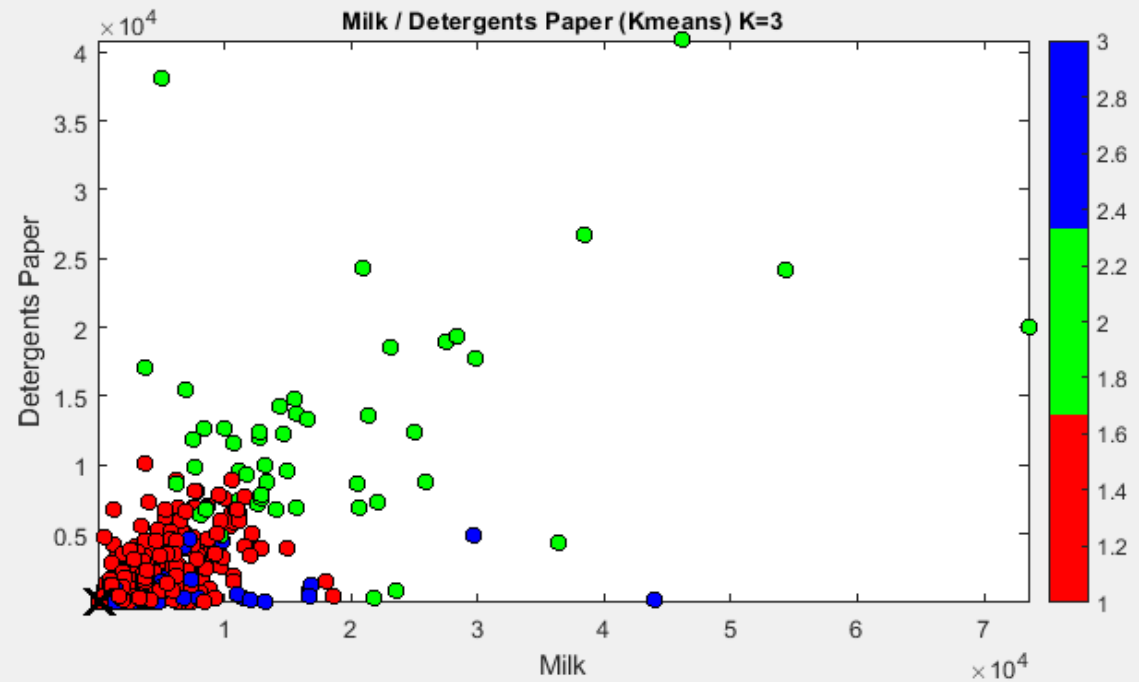

Figure 4.17 Projection du partitionnement du jeu de données par k-menas sur les 2 dimensions Milk et Detergents_paper.

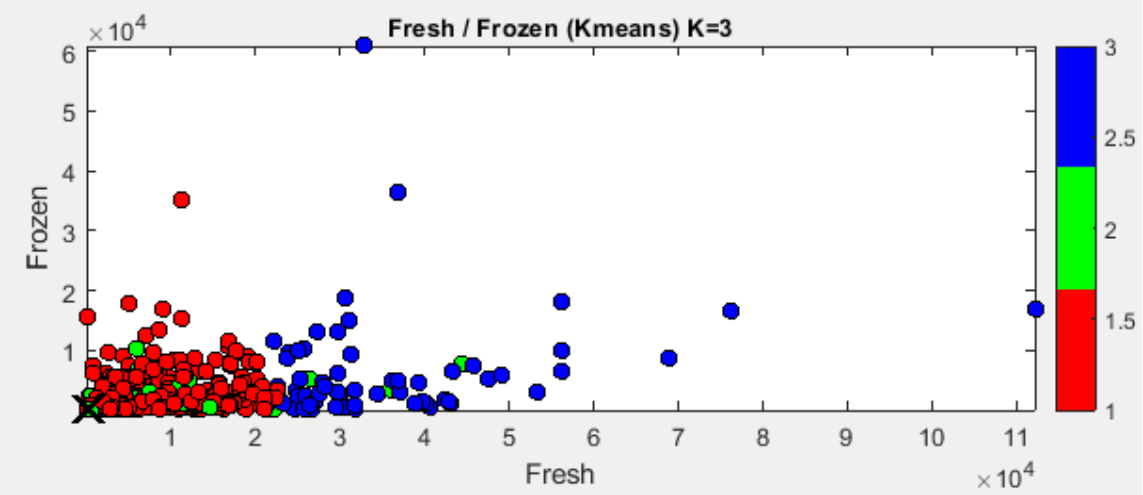

Figure 4.18 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Fresh et Frozen.

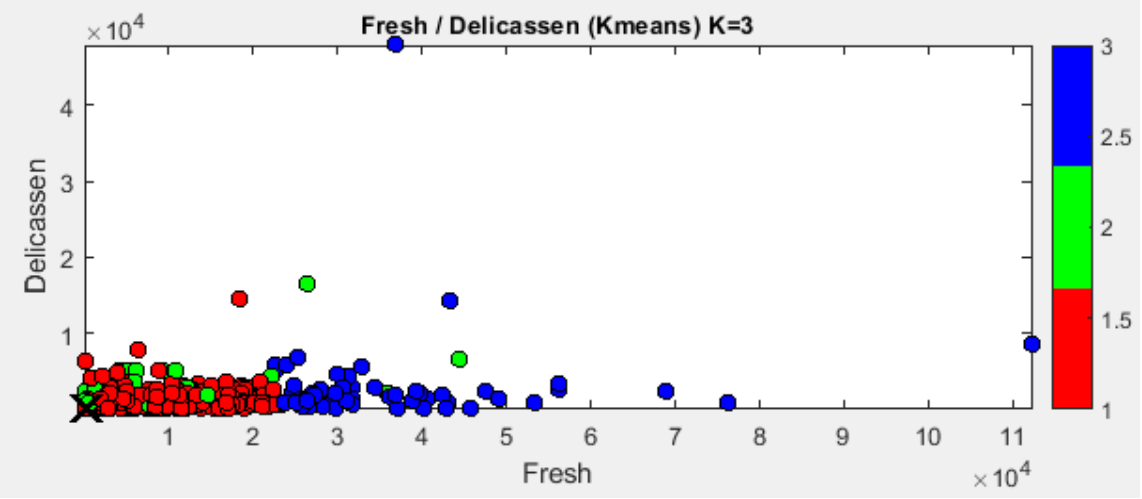

Figure 4.19 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Fresh et Delicassen.

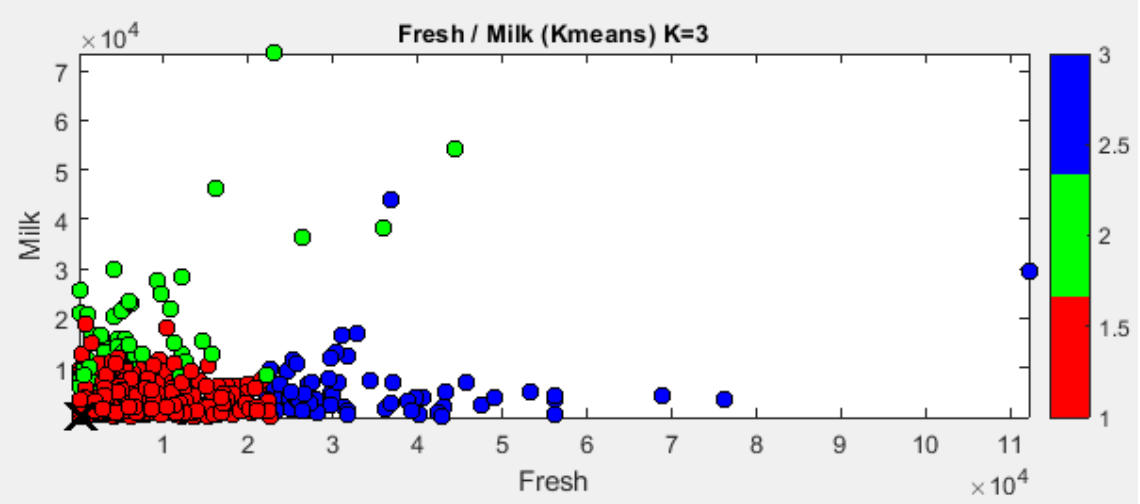

Figure 4.20 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Fresh et Milk.

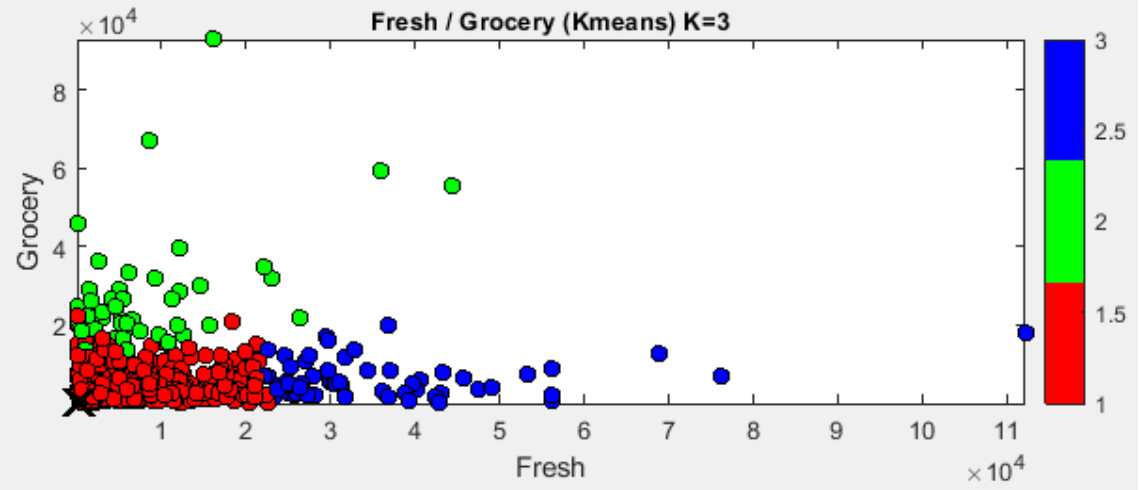

Figure 4.21 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Fresh et Grocery.

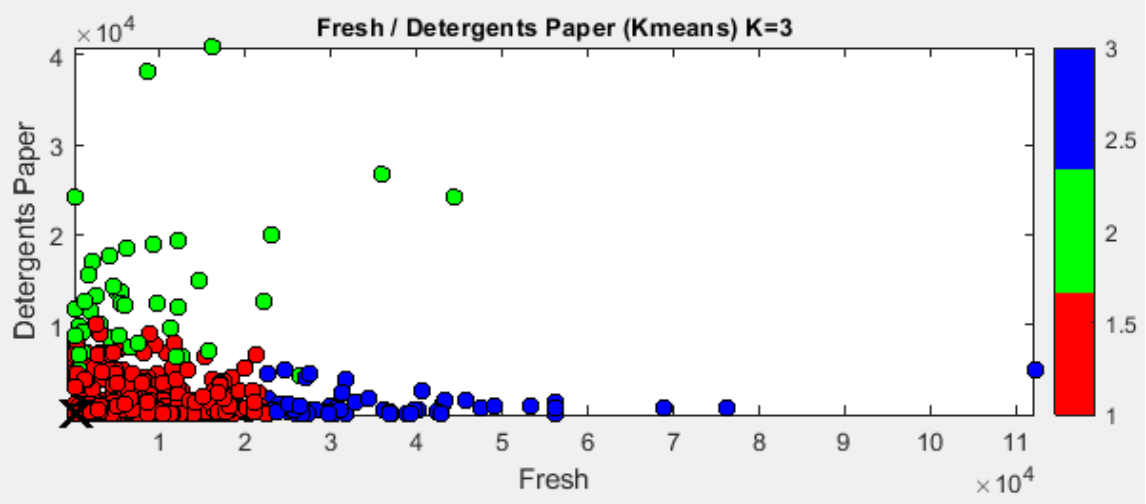

Figure 4.22 Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Fresh et Detergents_paper.

B. Les résultats de l'application de Single link pour k=4 :

Les figure 4.23 à 4.30 donnent le partitionnement du jeu de données en utilisant l'algorithme Single link pour un nombre de clusters égal à 4. Les figures montrent quatre clusters différents où chaque cluster représente une catégorie de consommateurs.

Les résultats sont similaires à celui de k-menas , la différence réside dans le 3 eme cluster qui a été divisé selon la consommation en produit (Fresh).

Le clusters violet regroupe les consommateurs ayant une grande consommation de produits (Detergent_paper, Grocery, Milk).

Le cluster cyan regroupe les consommateurs ayant une consommation modérée de produits (Detergent_paper, Grocery, Milk) et une faible consommation de produits(Detergent_paper, Grocery, Milk).

Le cluster rouge regroupe les consommateurs ayant une consommation moyenne de produits (Fresh).

Le cluster vert regroupe les consommateurs ayant une grande consommation de produits (Fresh).

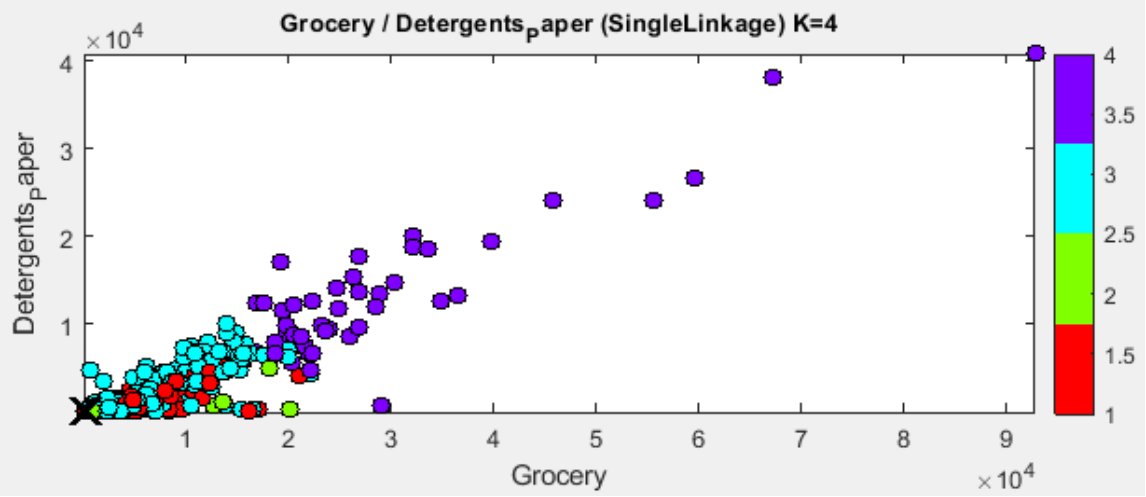

Figure 4.23 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Grocery et Detergents_paper.

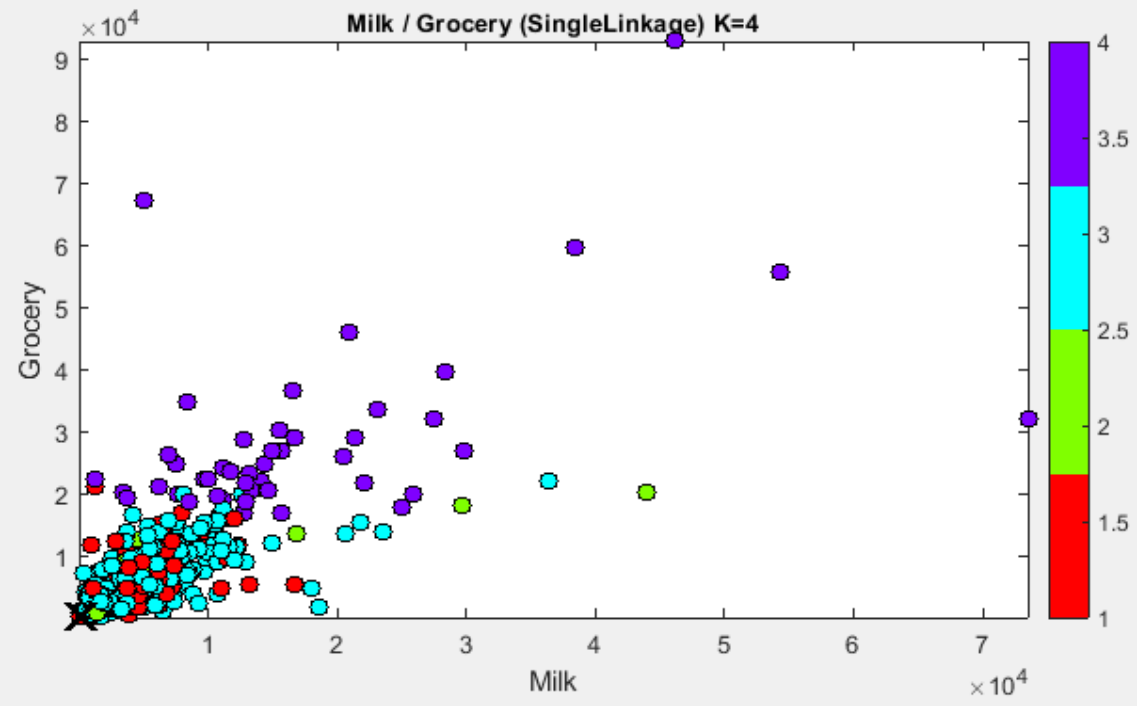

Figure 4.24 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Milk et Grocery.

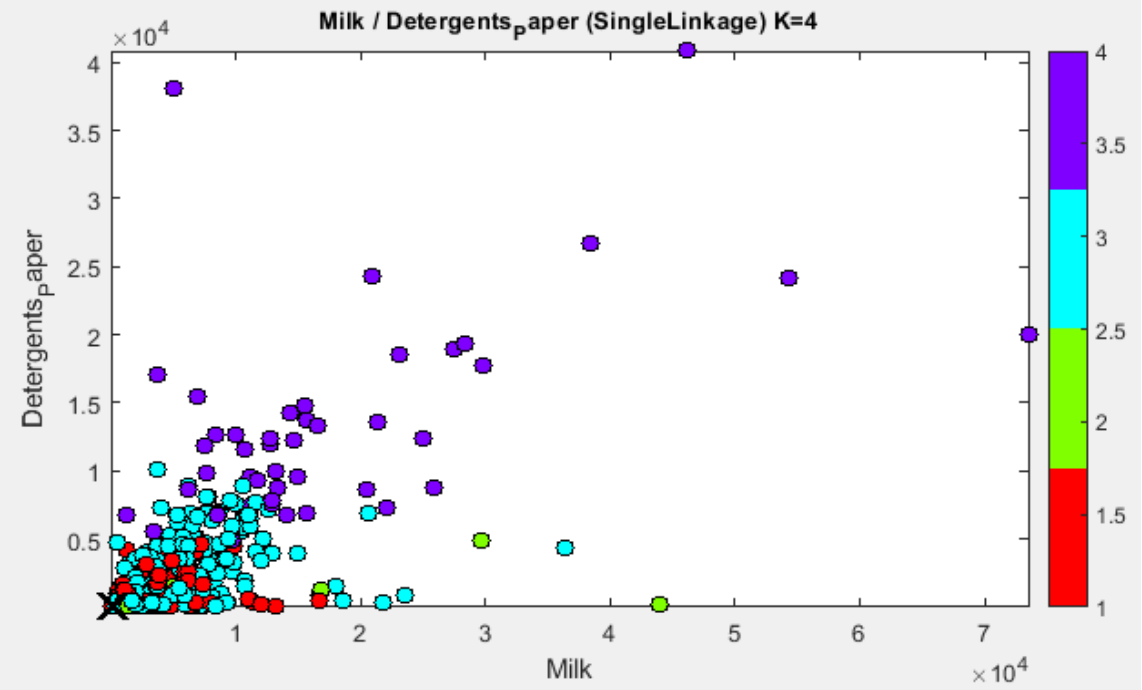

Figure 4.25 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Milk et Detergents_paper.

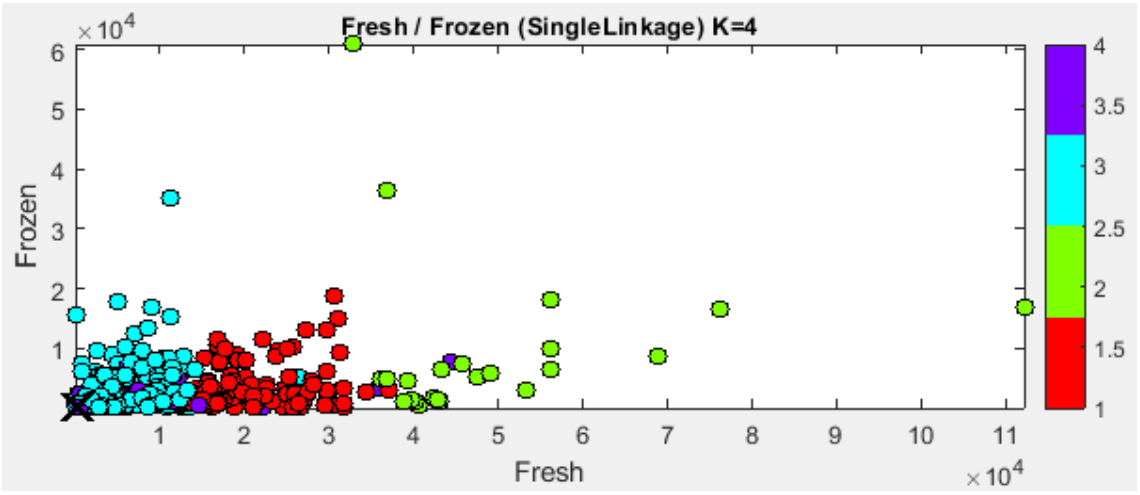

Figure 4.26 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Fresh et Frozen.

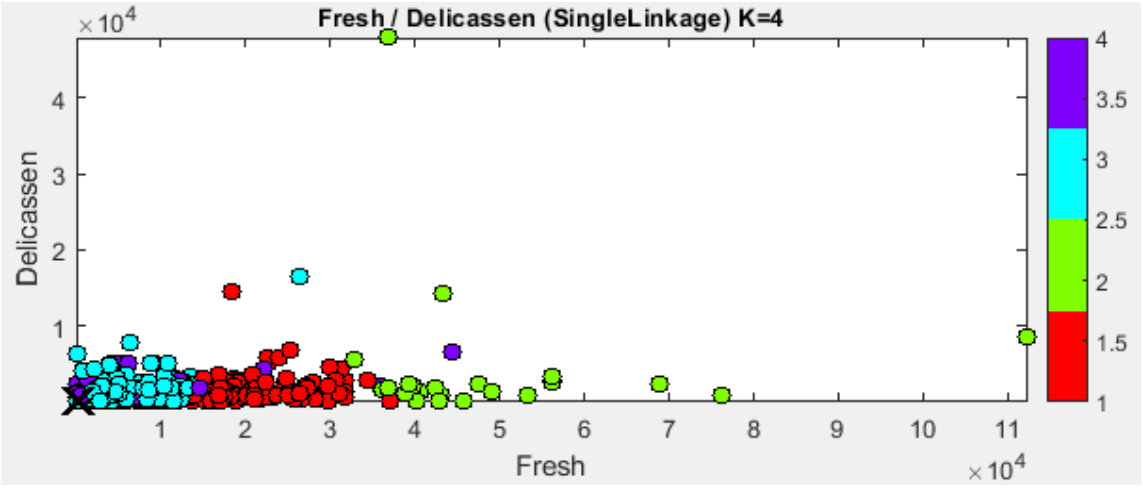

Figure 4.27 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Fresh et Delicassen.

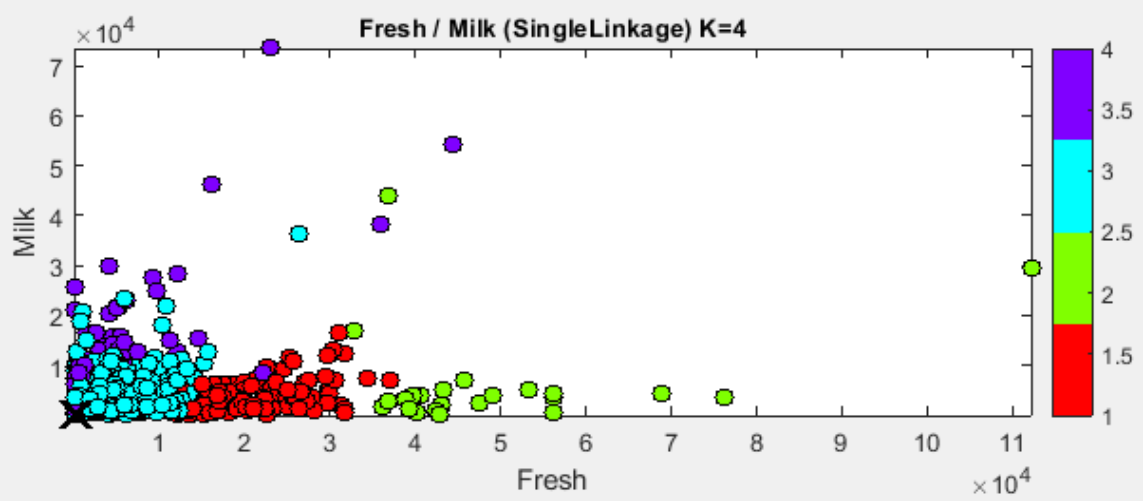

Figure 4.28 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Fresh et Milk.

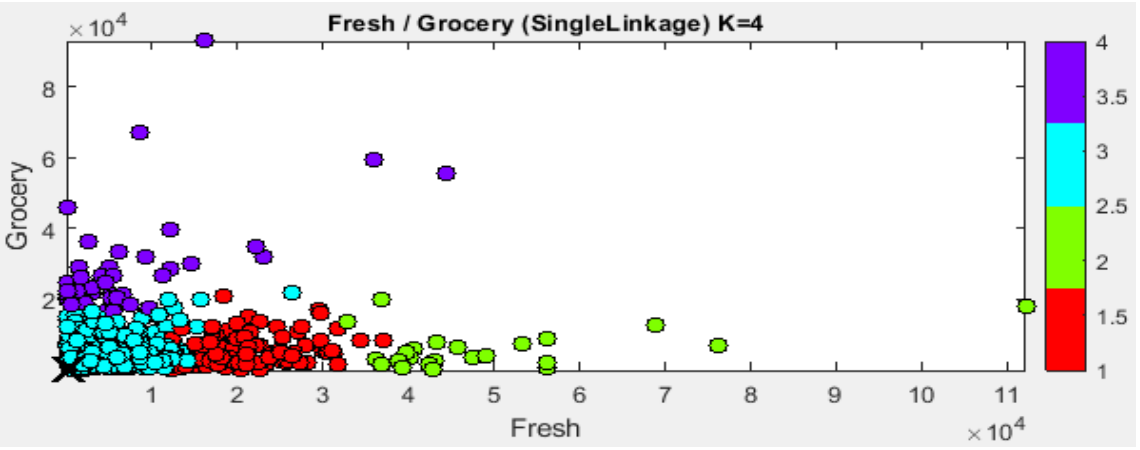

Figure 4.29 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Grocery et Fresh .

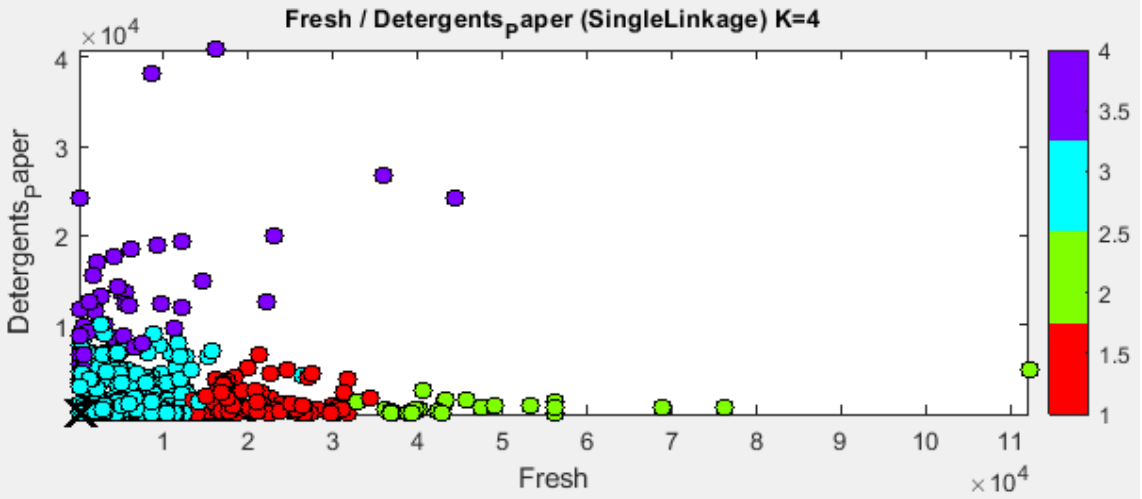

Figure 4.30 Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Fresh et Detergents_paper .

C. Les résultats de l'application de CuckooSearch (CS) pour k=3 :

Les figure 4.31….. 4.33 donnent le partitionnement du jeu de données en utilisant l'algorithme CuckooSearch (CS) pour un nombre de clusters égal à 3.

L'algorithme semble séparer les consommateurs aberrants.

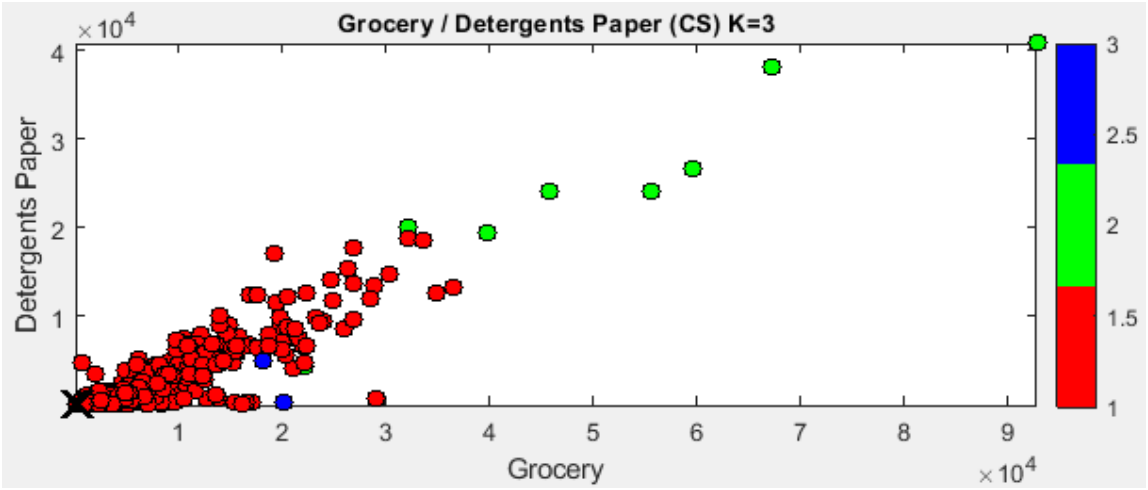

Figure 4.31 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Grocery et Detergents_paper .

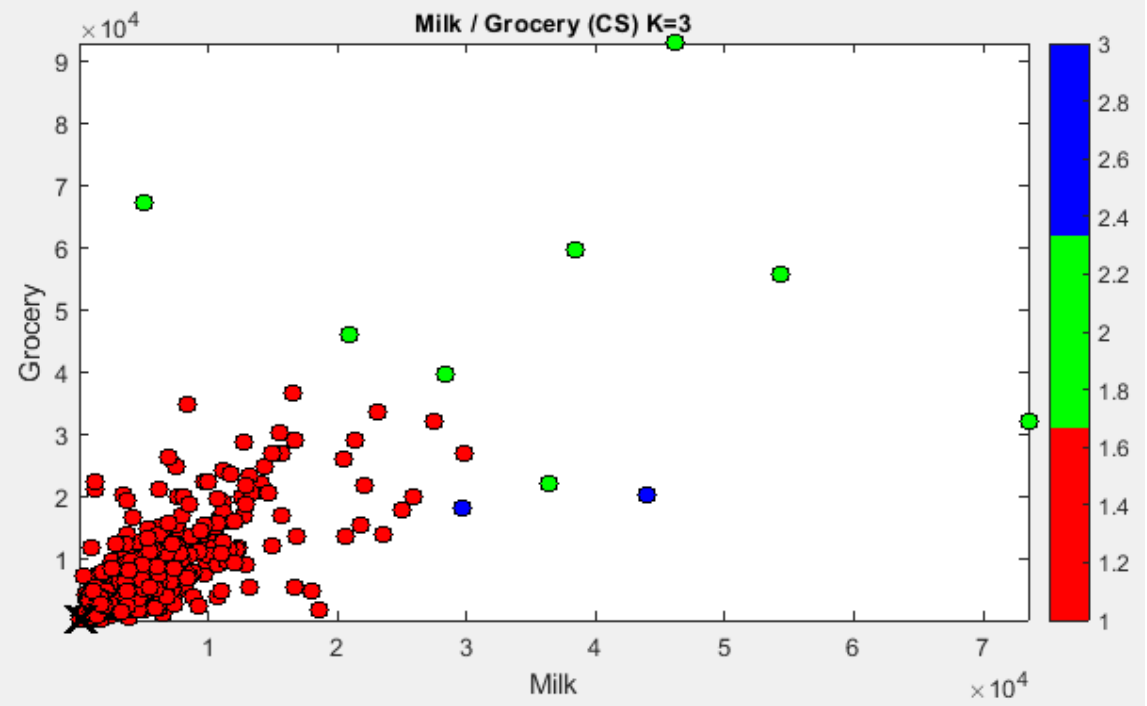

Figure 4.32 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions MIlk et Grocery .

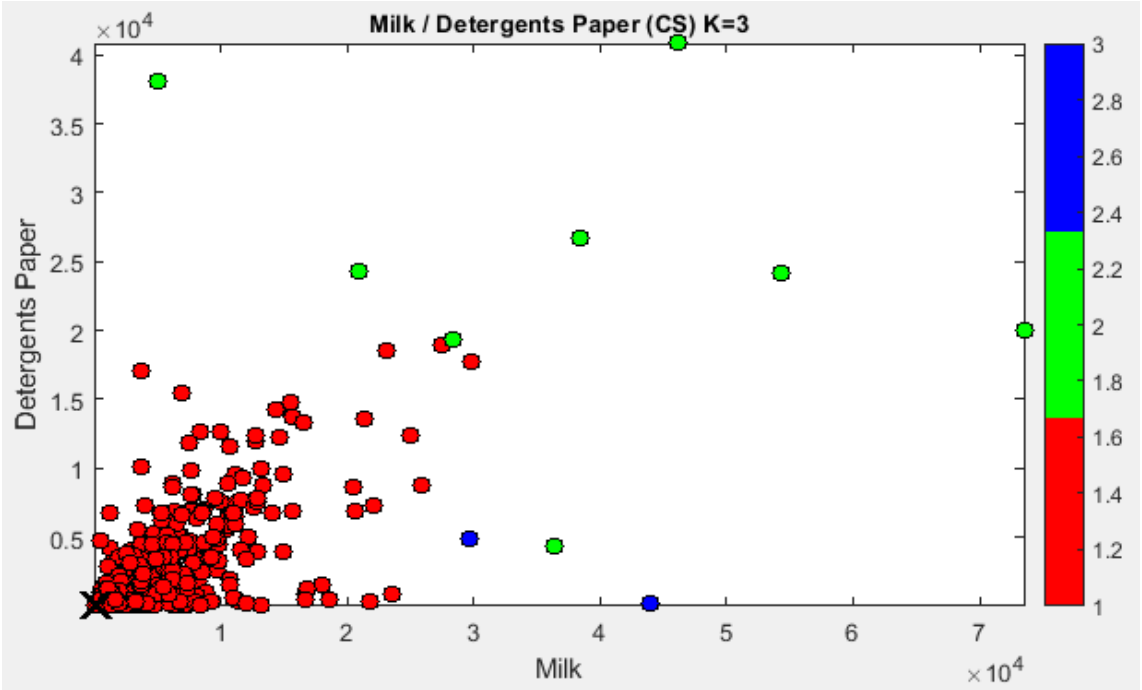

Figure 4.33 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions MIlk et Detergents_paper ..

4.2.3.3. Combinaison entre le clustering et l'analyse en composantes principales :

Les figures 4.34 montre le résultat de l'application de l'algorithme Kmeans sur le jeu de donnée réduit par l'ACP. La figure 4.34 montre trois clusters, le cluster en vert regroupe les clients ayant une grande consommation de produits (MILK, GROCERY, DETERGENT_PAPER). Pour le distributeur, les clients n° 86, 48 ,334,87 ,62 sont les meilleurs de cette catégorie.

Le cluster en bleu regroupe les clients ayant une grande consommation de produit (FRESH, FROSEN, DELICASSEN). Les meilleurs clients de cette catégorie sont ceux ayant le numéros : 184,326,182 ,88,126,104.

Le cluster en rouge regroupé les clients de consommation modérée.

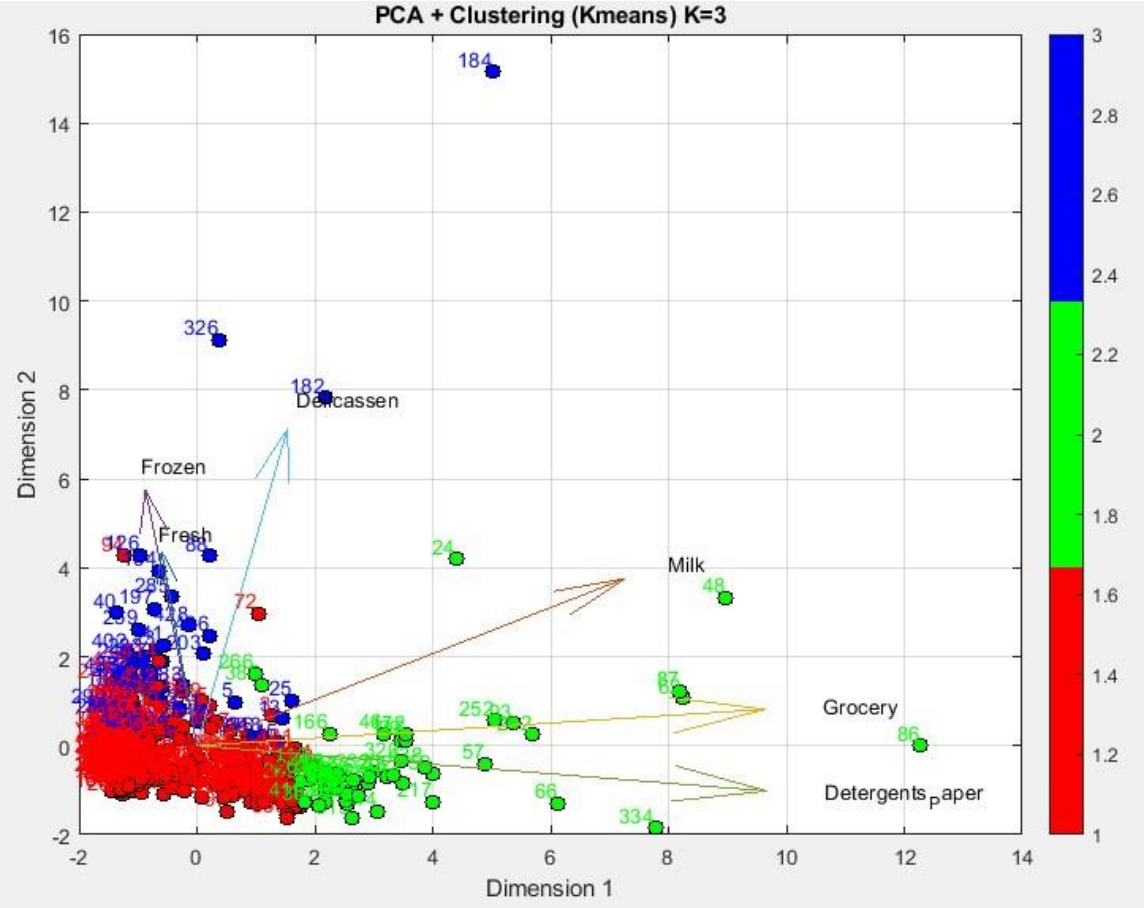

Figure 4.34 Clustering du jeu de données obtenu pas l'ACP (kMeans K=3).

4.3.Conclusion :

Dans ce chapitre, nous avons examiné, analysé et interprété les résultats graphiques issus de diverses analyses. Les informations extraites peuvent être précieuses pour le distributeur afin d'éclairer ses décisions futures.

Conclusion Générale :

Au cours de ce travail de master, nous nous sommes intéressés à l'exploration approfondie et pratique de l'analyse de données appliquée au contexte spécifique de la vente en gros et de l'analyse des clients. En mettant l'accent sur l'importance stratégique de transformer les données en informations exploitables, nous avons examiné différentes méthodologies allant de l'analyse unidimensionnelle à l'analyse bidimensionnelle et multidimensionnelle, en incluant des techniques telles que le clustering et l'analyse factorielle.

Dans ce mémoire, nous avons commencé par présenter les notions et les concepts de base du domaine de l'analyse de données, son importance et ces différents types. Nous avons passé ensuite à la description de notre application. À la fin, nous avons focalisé nos efforts sur l'analyse approfondie des représentations graphiques résultant de l'application des différentes méthodes d'analyse aux données.

Pour enrichir ce travail, nous envisageons d'inclure une méthode de prédiction fondée sur les règles d'association.

Bibliographie

[1] M. JAMBU. (1999). Méthode de base de l'analyse de données.

[2] T. Yves. Cours de statistique descriptive.

[3] G. SAPORTA (2006). Probabilités analyse des données et statistique. EditionTECHNIP, paris. France.

[4] J.P.BENZECRI (1980). L'analyse de données (Tome1) la taxonomie. Dunod

[5] J. DUDA (2002). Pattern recognition. MIT.

[6] P.DEMARTINES et J. HeRAULT (1997). Curvilinear component analysis : a self organizing neural network for non linear of mappinf of data set. IEEE transactions on neural networks, 8(1) :148-54

[7] L. LEBART , A. MORINEAU et M. PIRON (1995). Statistique exploratoire multidimensionnele. Dunod

[8] J. ZHANG (2012). Kernel principal compenent analysis. Expert systems with application. Vol10 :11- 25

[9] F. CHEVALIER et J. LE BALLAC (2013). Rapport sur la classification. Université de RENNES1. [10] G. CELEUX, E. DIDAY, G. GOVAERT (1989). Classification automatique des données. Dunod.

[11] G. BROSSIER (2003). Les élémentsfondamentaux de la classification. Hermes Sciences publication.
[12] N. WICKER (2001). Cours d'analyse de données. North-western European Journal of mathematics

[13] F.DAZY et J-F LE BARZIC (1996). L'analyse des données évolutives « méthodes et applications ». Edition TECHNIP.

[14] C. E. Lawrence and A. A. Reilly (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," Proteins: Structure, Function, and Bioinformatics, vol. 7, pp. 41-51.

[15] J. De Lagarde (1995). Initiation à l'analyse des données. Dunod.

[16] A. LAGNOUX (2018). Série chronologique. Université de TOULOUS.

[17] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59–63, 21-24 Oct. 2012. [13] Wie, Jiang, Ravi V. T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[18] K, Chitharanjan, and Kala Karun A. "A review on hadoop - HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.

[19] F. C. P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, p. 32–37, 2013.

[20] Castelino, C., Gandhi, D., Narula, H. G., &Chokshi, N. H. (2014). Integration of Big Data and Cloud Computing. International Journal of Engineering Trends and Technology (IJETT), 100-102.

[21] : Laurent Candillier, rapport technique : "La classification non supervisée", Septembre 2004.

[22] : V. Kumar, rapport technique: "An Introduction to Cluster Analysis for Data Mining", C.S. Dept. Univ. Minnesota, 2000.

[23] : P. Berkhin, Rapport techniques: "Survey of Clustering Data Mining Techniques", Accrue software, San Jose, California, 2002.

[4] : A. K. Jain et R.C. Dubes: "Algorithms for Clustering Data", Prentice Hall series de reference avancée, 1988.

[25] : Laetitia Jourdan, thèse de doctorat : "Métaheuristiques pour l'extraction de connaissances application à la génomique", Novembre 2003.

[26] : RamdaneChafika, mémoire de magistère :"Le clustering des données : une nouvelle approche évolutionnaire quantique ", Université Mentouri de Constantine, Juin 2006.

[27] :D.P.Mercer, linacre college : " Clustering large Datasets ", Octobre 2003

[28] :Krishnapuram R. et Keller, J.M. " A possibilistic approach to clustering", IEEE Trans.Fuzzy

[29] : Yang, M.S. et Wu, K.L. 'Unsupervised possibilistic clustering', Pattern Recognition,Vol.

[30] : RamdaneChafika, thése de doctorat :" Apports du Calcul Quantique et des Concepts Possibilistes à la Classification non supervisée Evolutionnaire ", Université Mentouri de Constantine, Novembre 2013.

[31]https://archive.ics.uci.edu/dataset/292/wholesale%2Bcustomers

Annexe

Résultats du clustering du jeu de données

1. Les résultats de l'application de Kmeans pour k=3 :

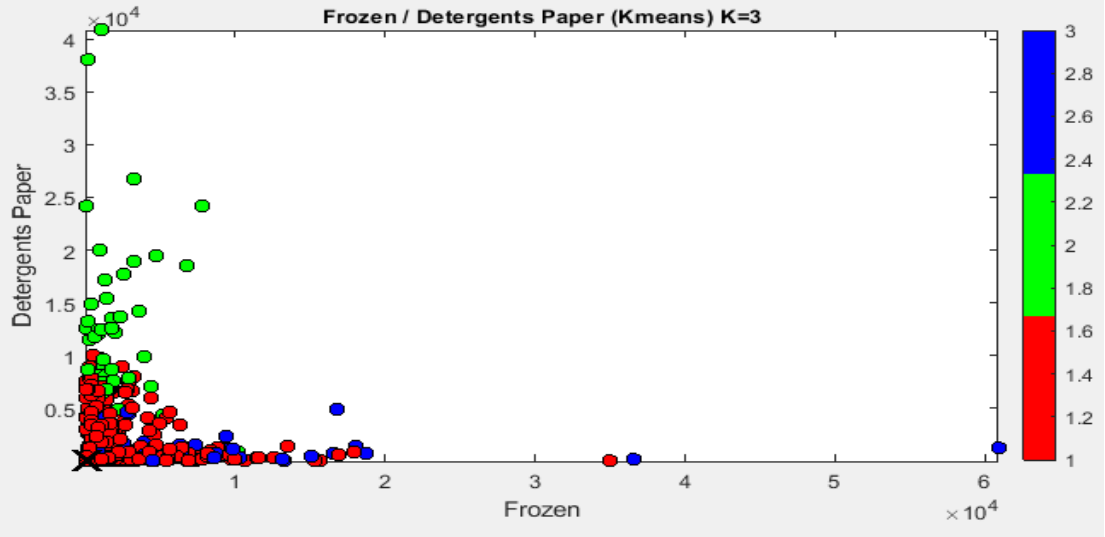

Figure 1. Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions detergent-paper et frozen.

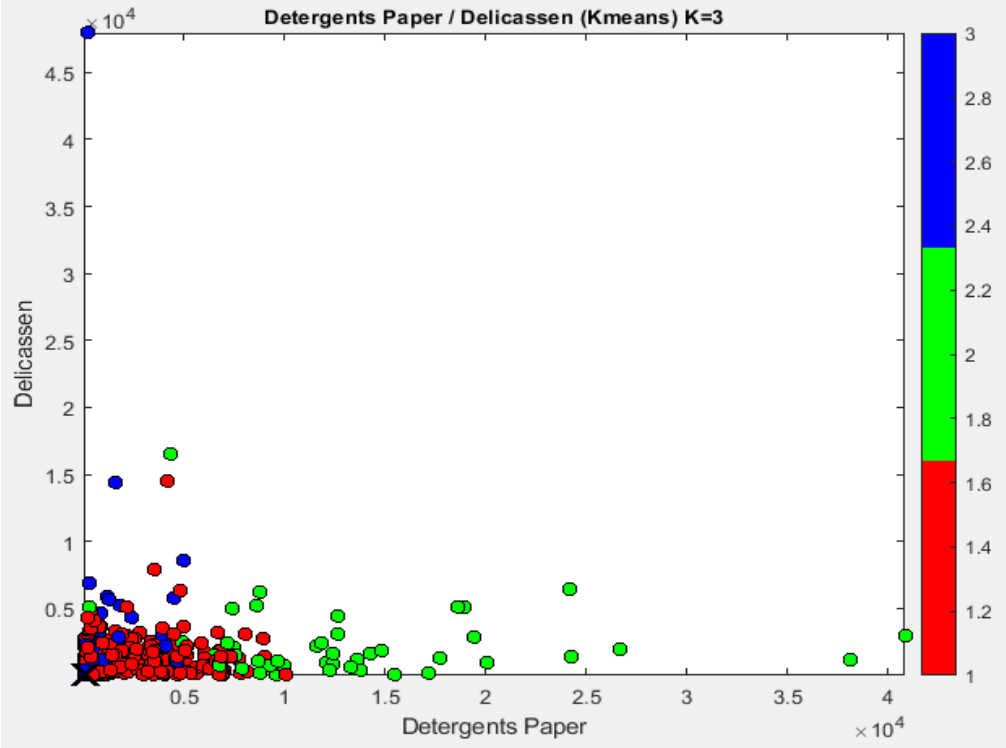

Figure 2. Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Delicassen et detergent-paper.

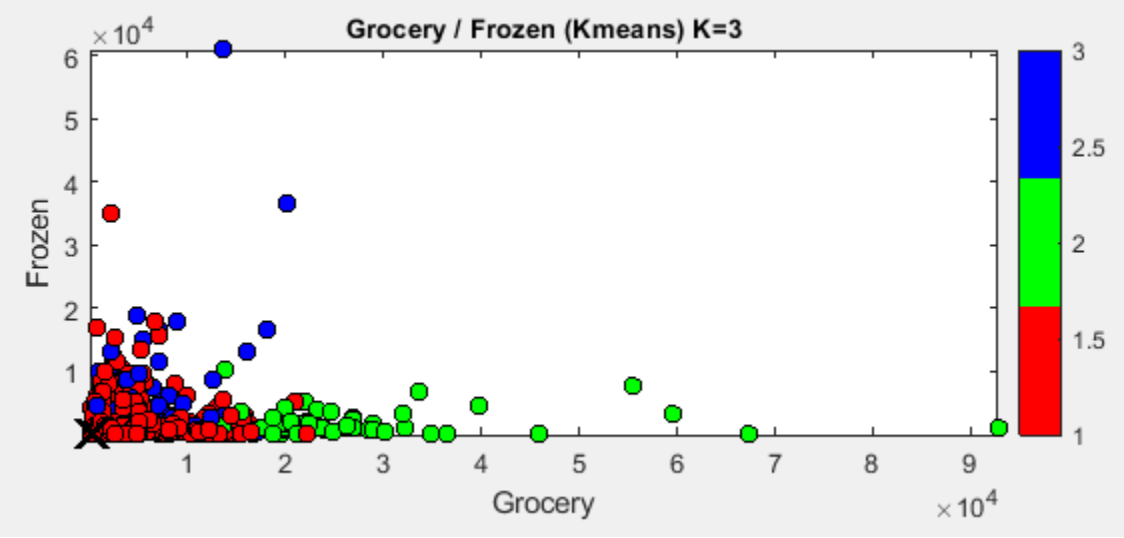

Figure 3. Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions frozen et grocery.

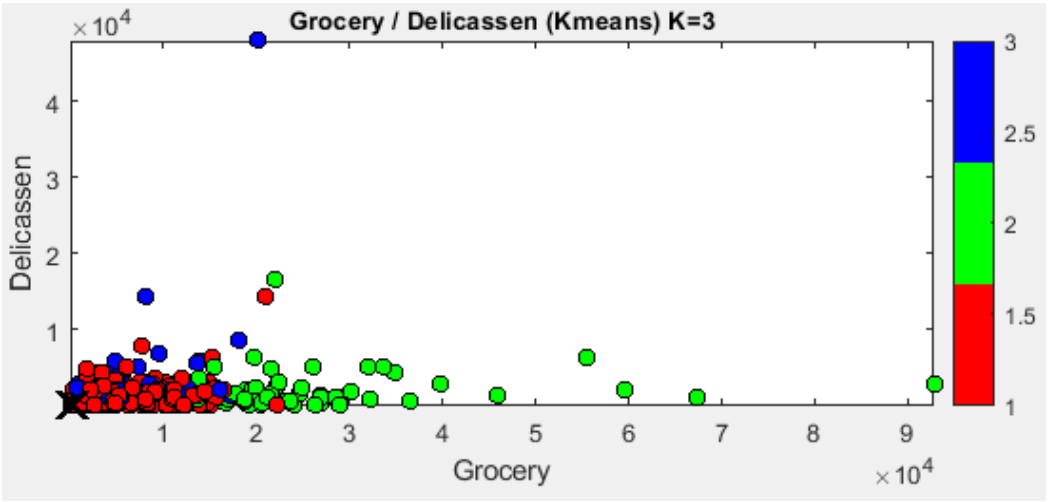

Figure 4. Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Delicassen et Grocery.

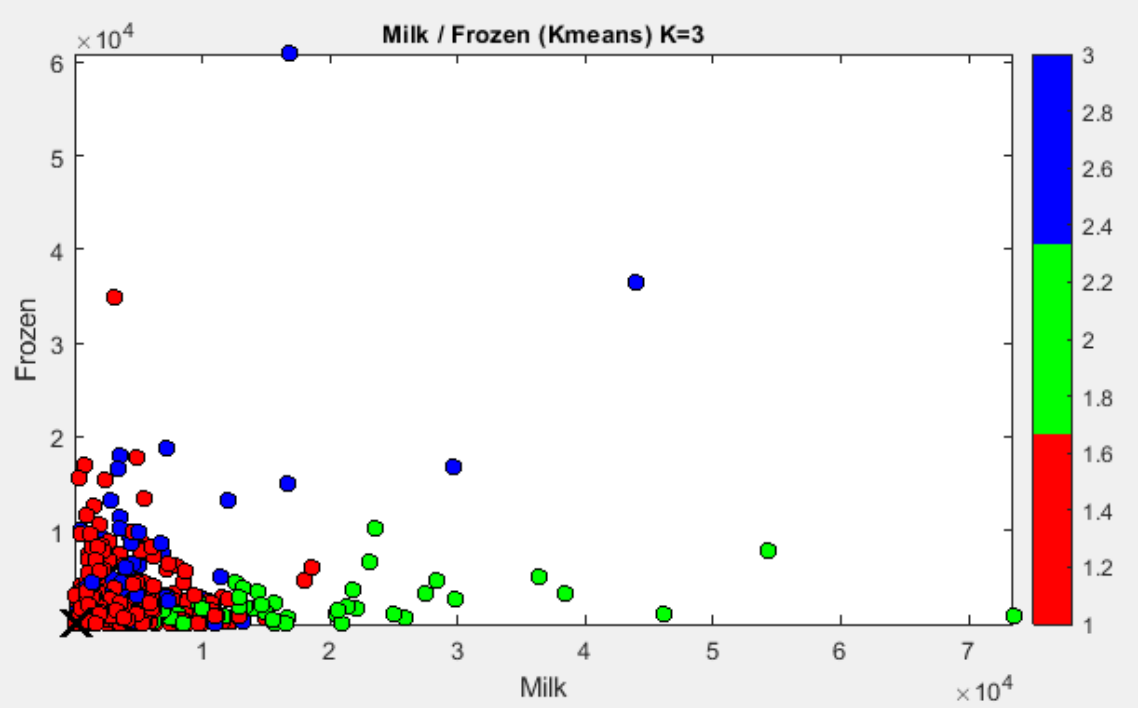

Figure 5. Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions frozen et Milk.

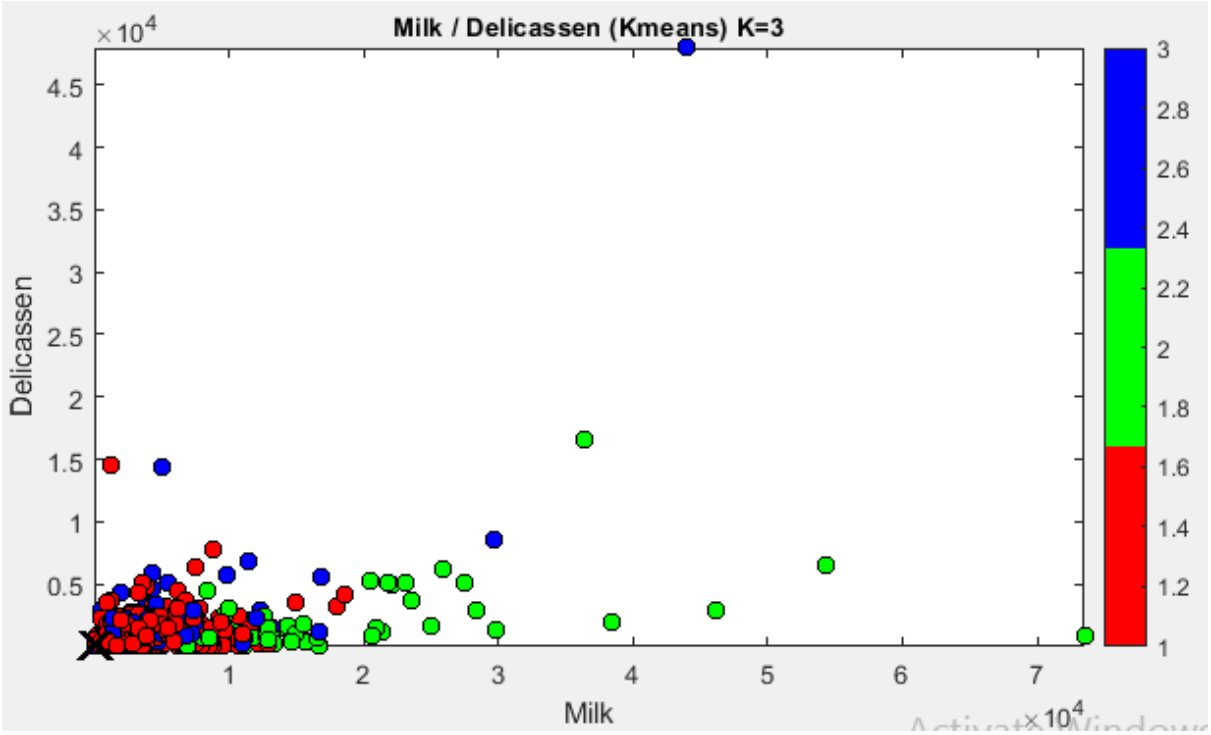

Figure 6. Projection du partitionnement du jeu de données par kmeans sur les 2 dimensions Delicassen et Milk.

2. Les résultats de l'application de Single link pour k=4 :

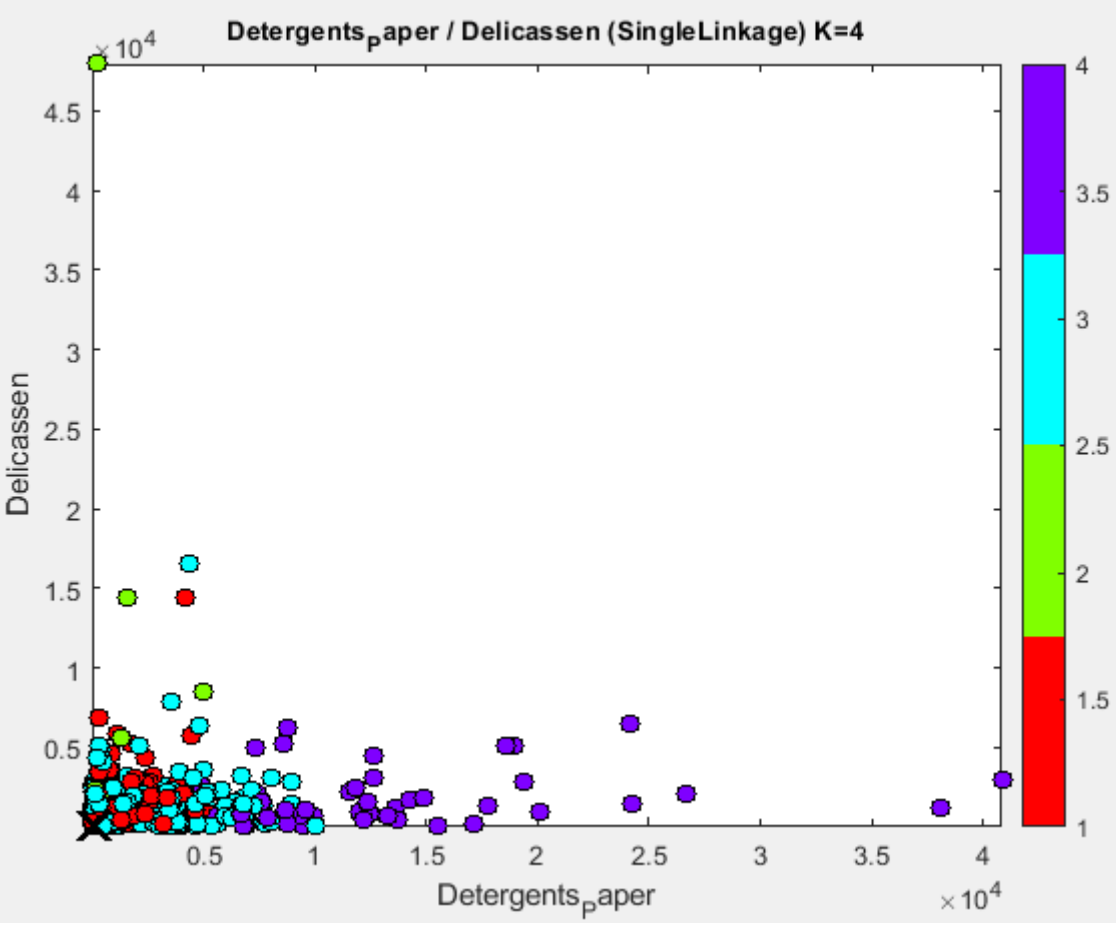

Figure 7. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Delicassen et Detergents-paper.

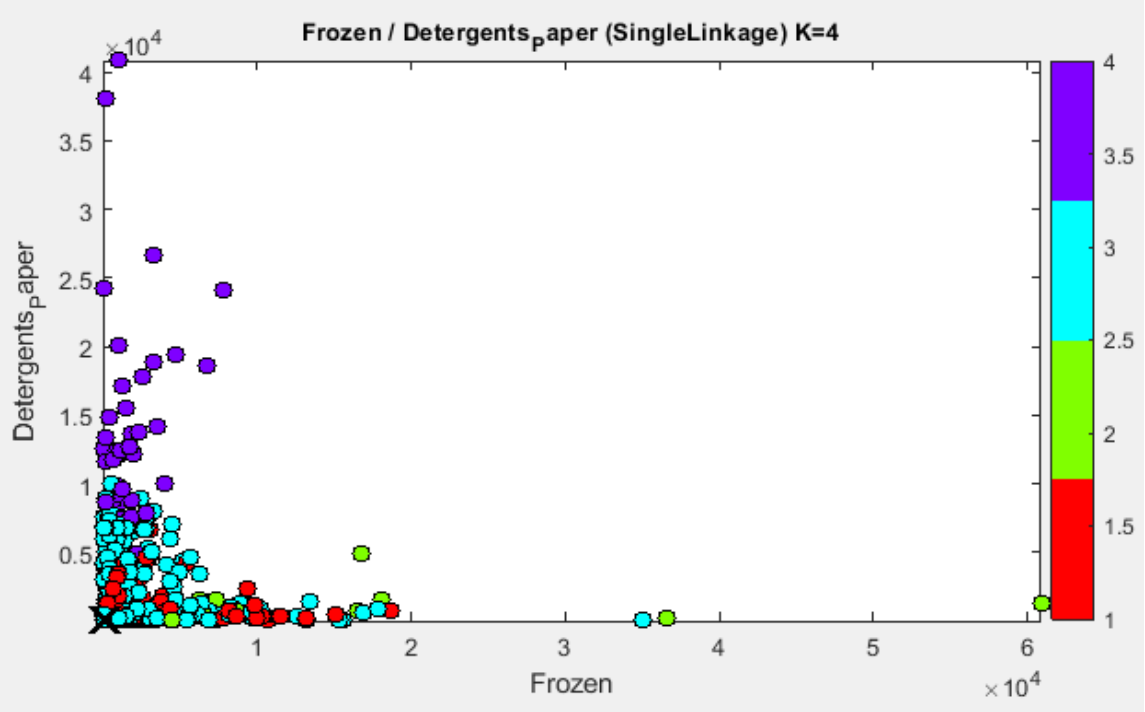

Figure 8. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Detergents-paper et Frozen.

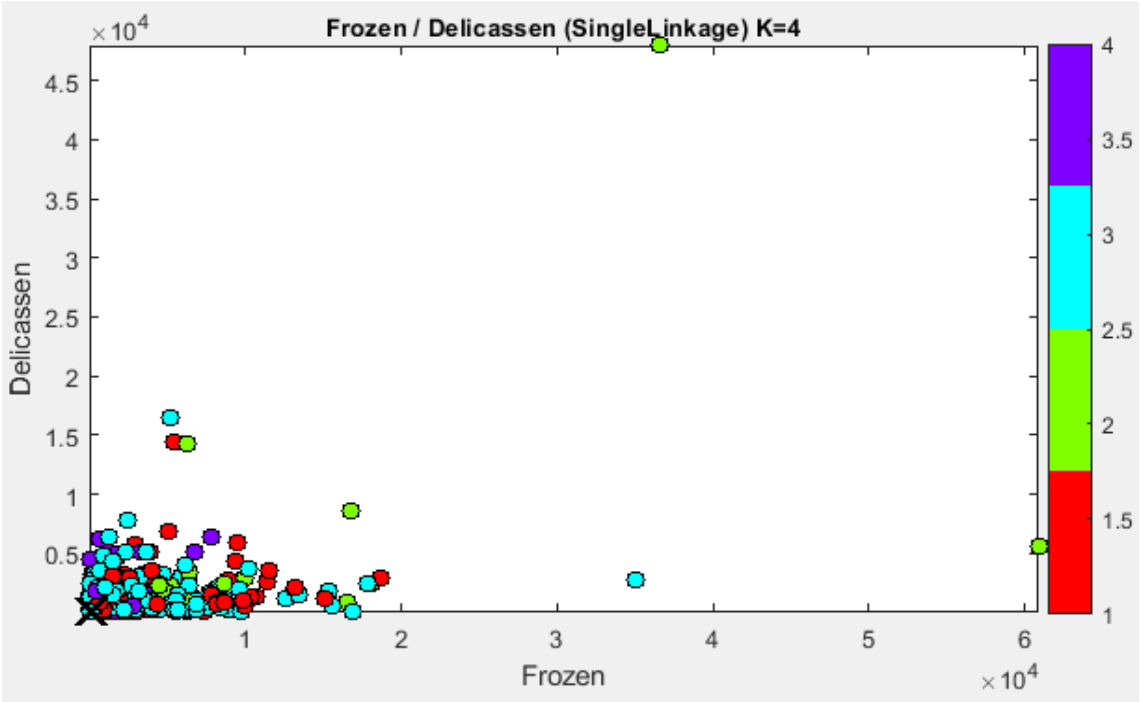

Figure 9. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Delicassen et Frozen.

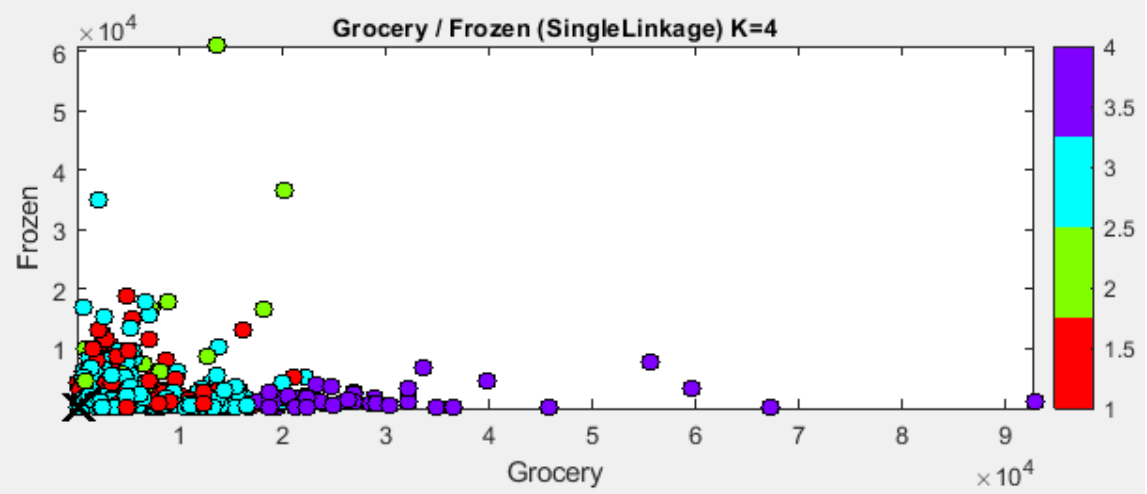

Figure 10. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Frozen et Grocery.

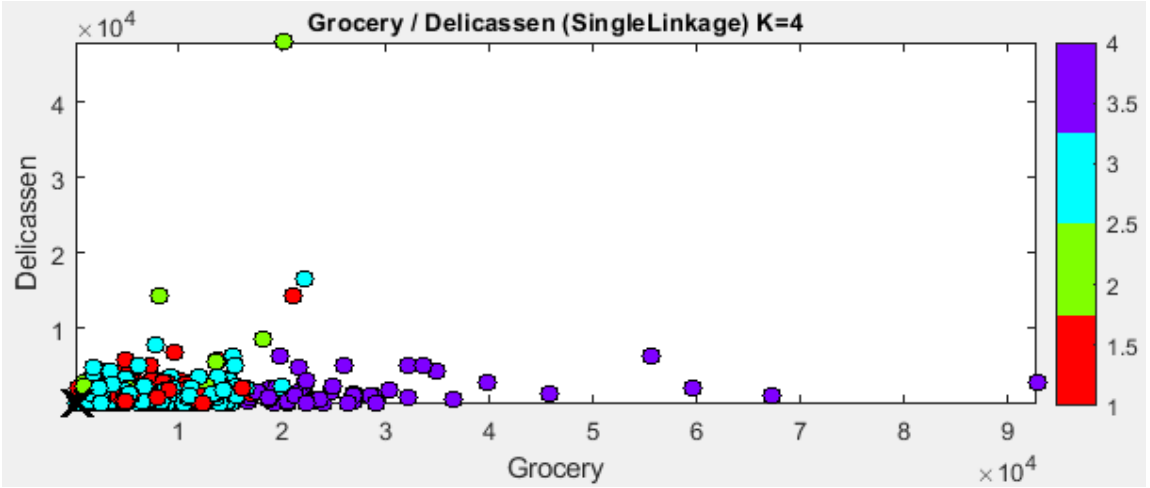

Figure 11. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Delicassen et Grocery.

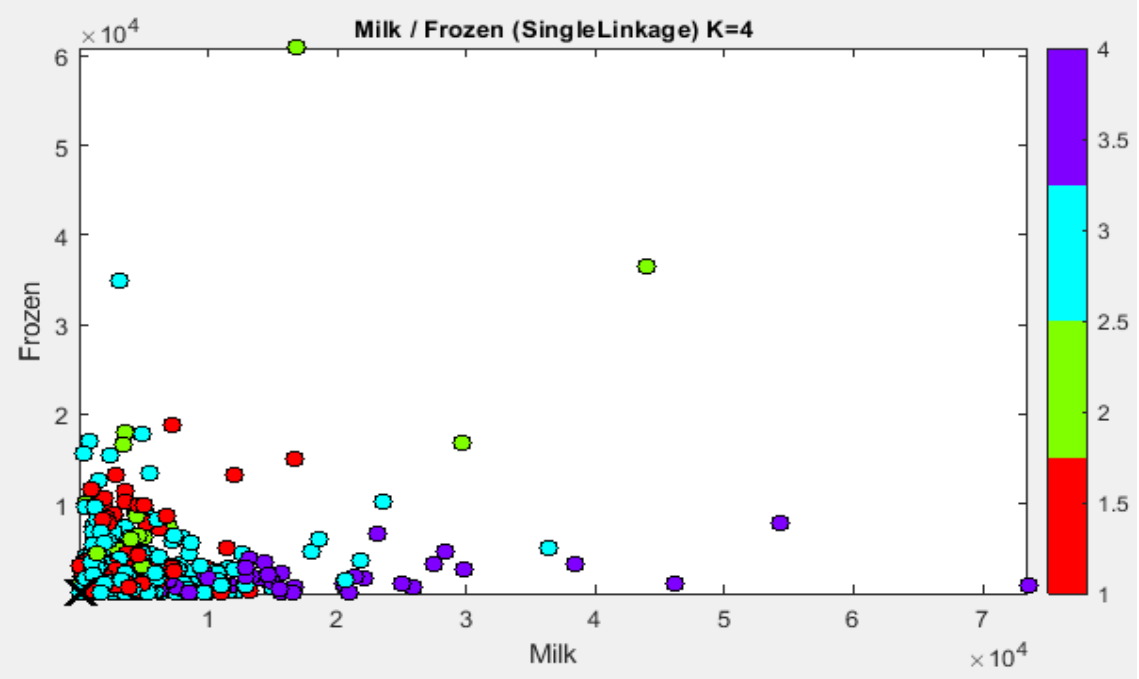

Figure 12. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Frozen et Milk.

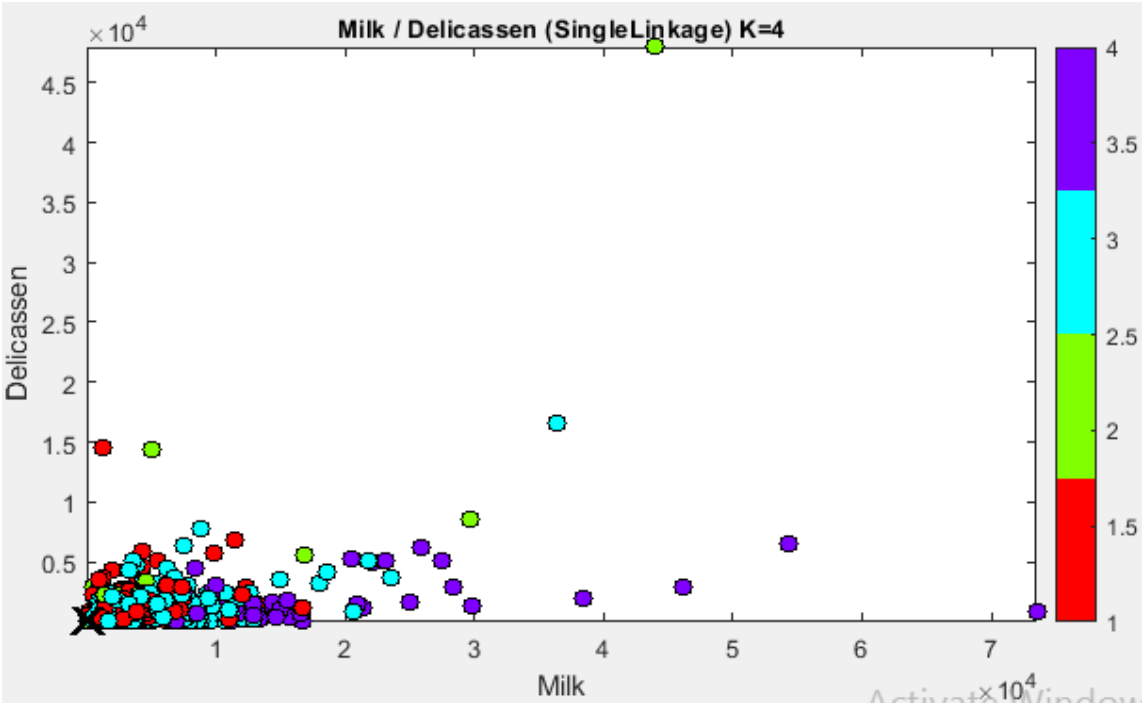

Figure 13. Projection du partitionnement du jeu de données par Single_linkage sur les 2 dimensions Delicassen et Milk.

3. Les résultats de l'application de Cuckoo Search CS pour k=3 :

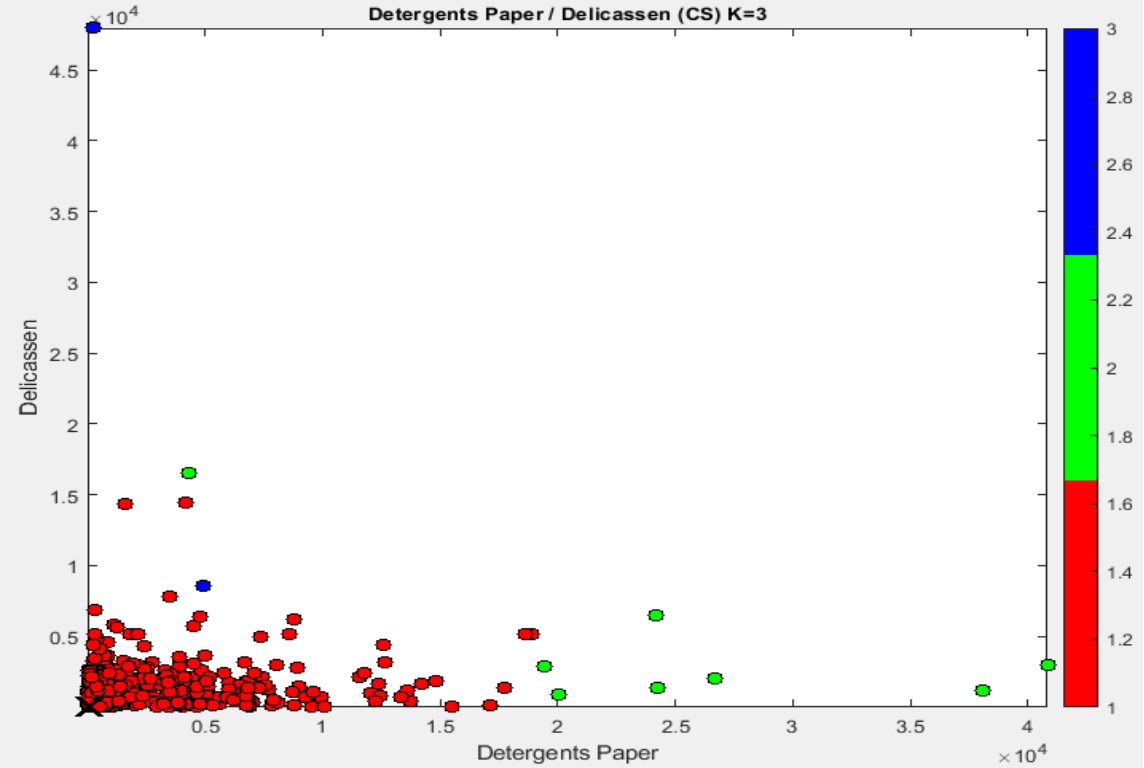

Figure 14 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Delicassen et Detergents-paper .

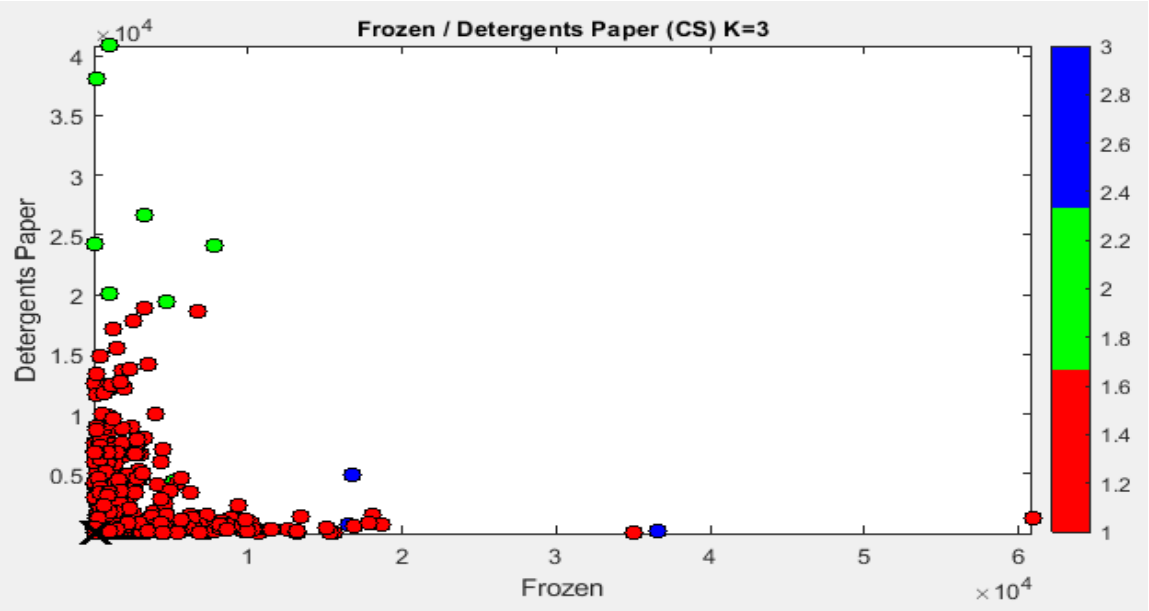

Figure 15 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Detergents-paper et Frozen .

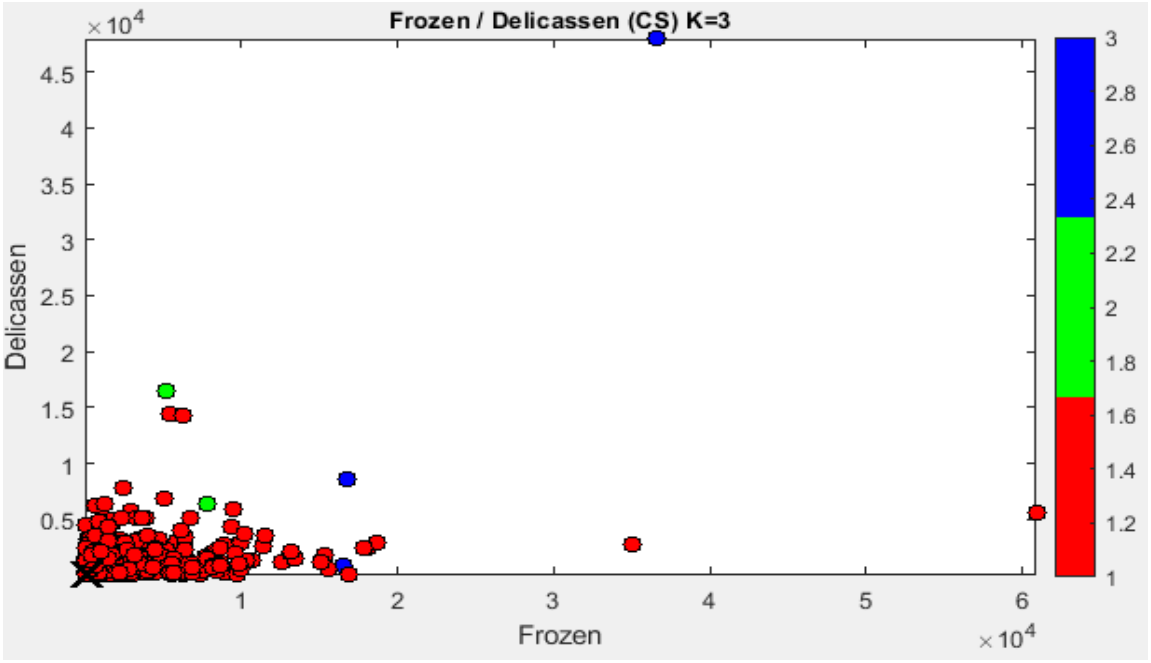

Figure 16 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Delicassen et Frozen .

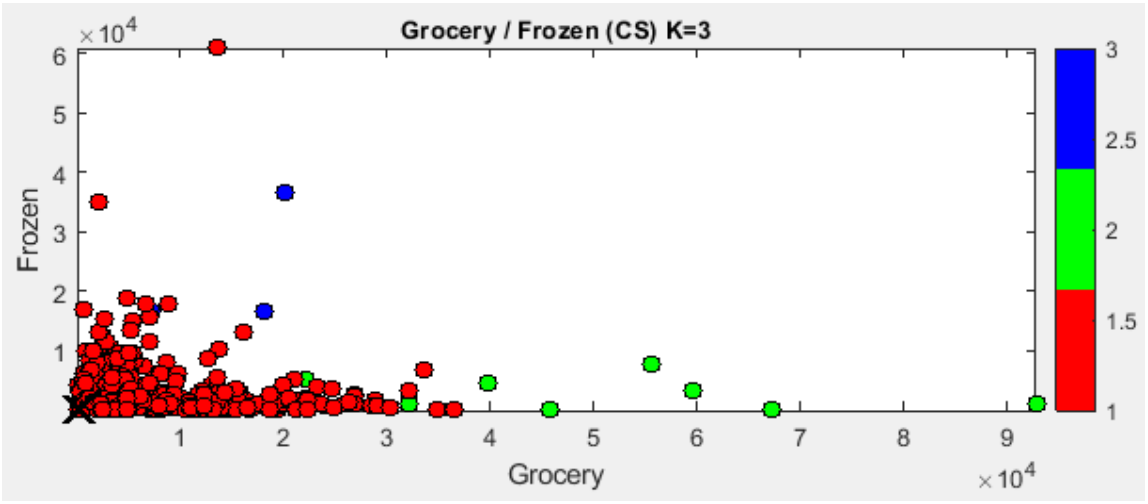

Figure 17 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Frozen et Grocery .

Figure 18 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Delicassen et Grocery .

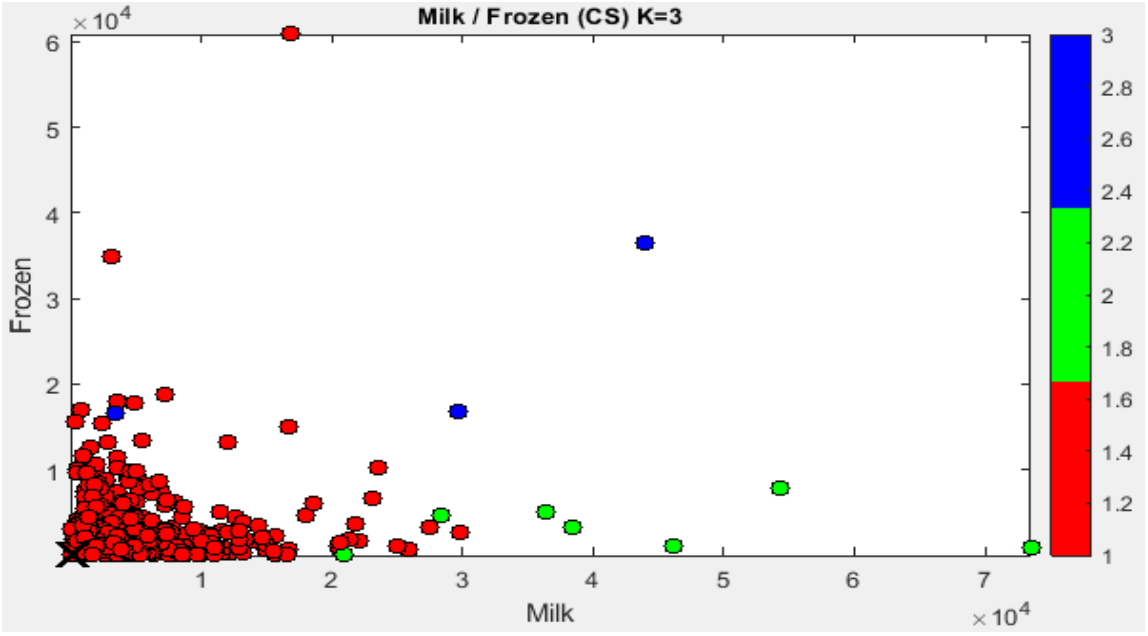

Figure 19 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Frozen et Milk .

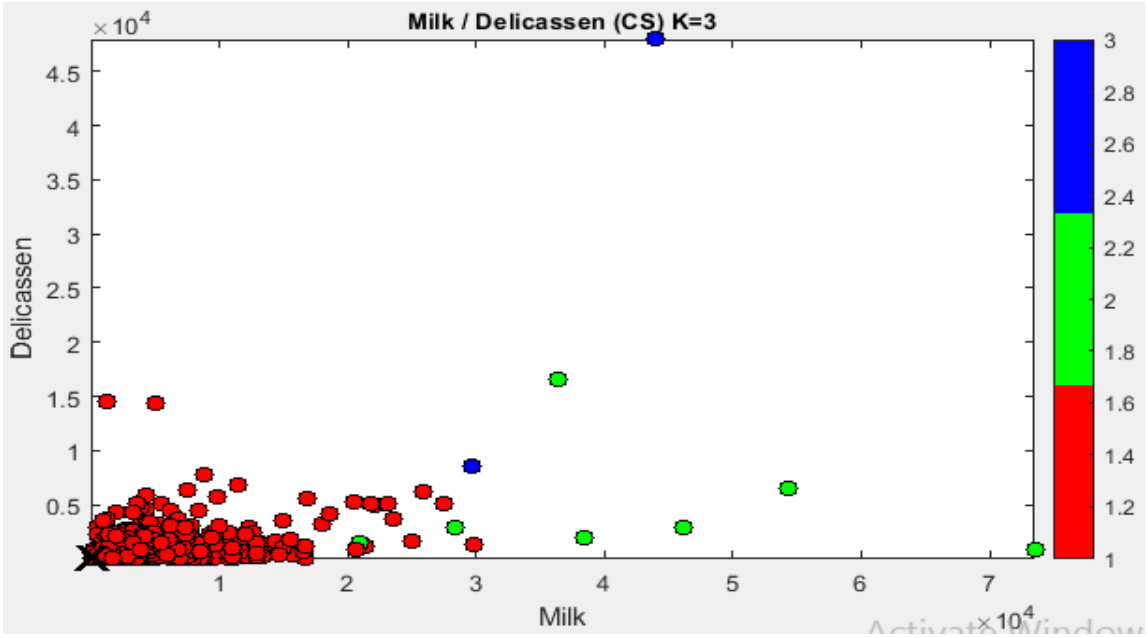

Figure 20 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Delicassen et Milk .

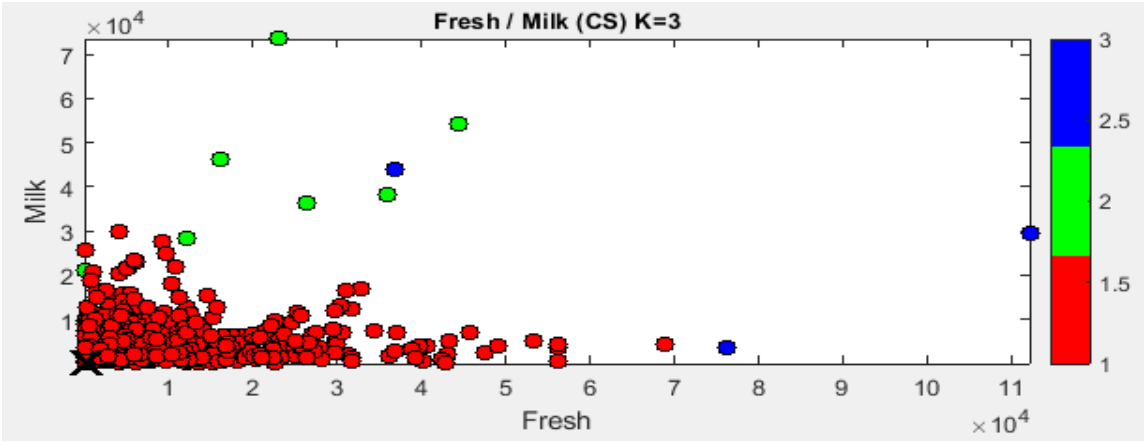

Figure 21 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Milk et Fresh .

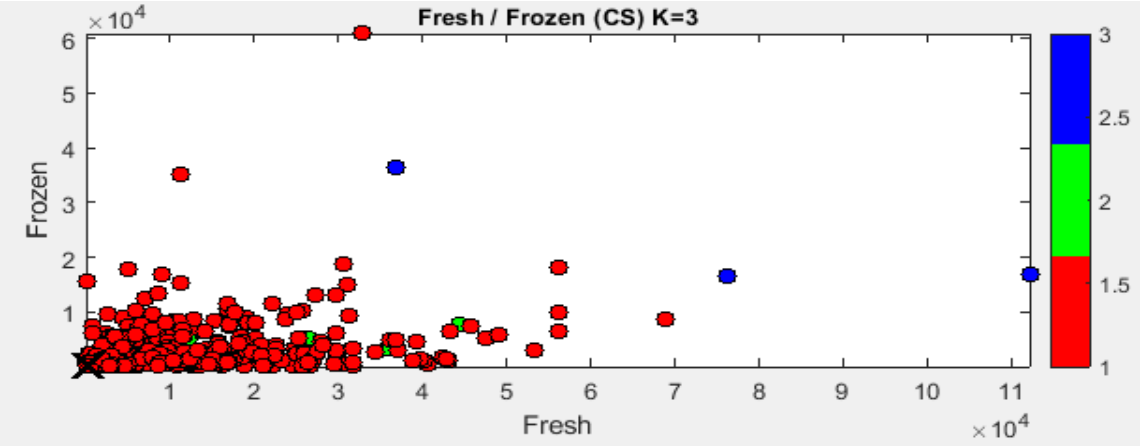

Figure 22 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Frozen et Fresh .

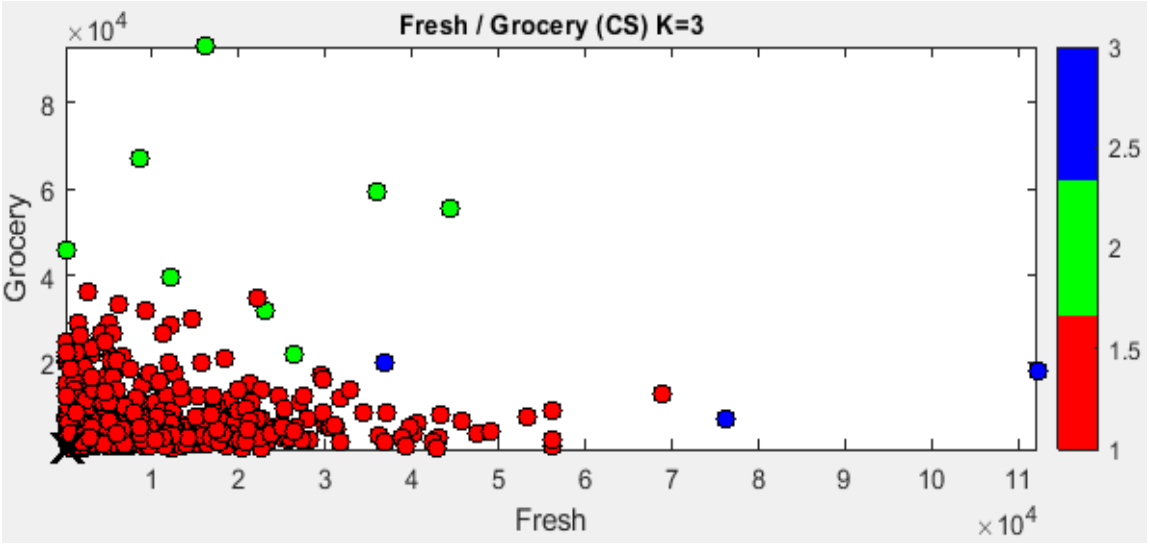

Figure 23 Projection du partitionnement du jeu de données par CuckooSearch sur les 2 dimensions Fresh et Grocery .

Combinaison entre le clustering et l'analyse en composantes principales

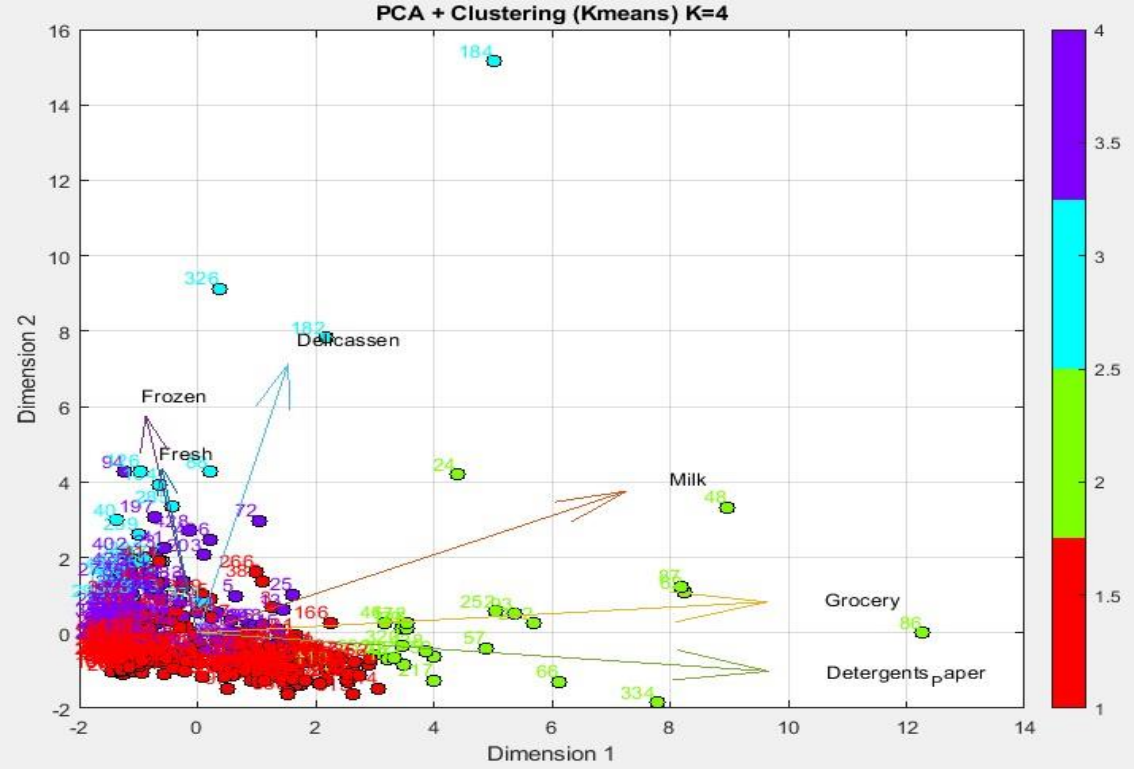

Figure 24 Clustering du jeu de données obtenu pas l'ACP (kMeans K=4).

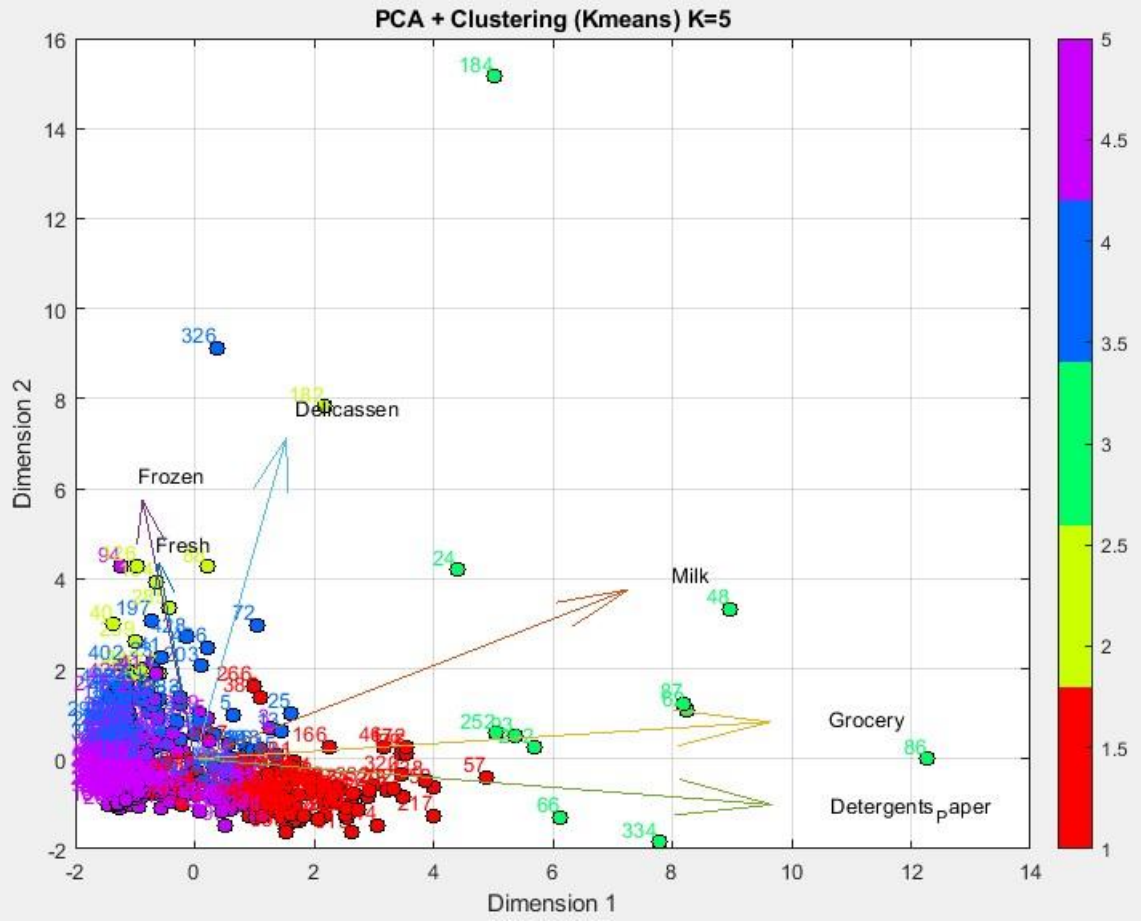

Figure 23 Clustering du jeu de données obtenu pas l'ACP (kMeans K=5).

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليـــــــم العالـــــــي و البحث العلمـــــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Août 1955- skikda-

Faculté des Sciences

Département d'Informatique

جـامعة 20 أوت 1955- سكيكدة

كـــــلية العلـــــوم

قسم الاعلام الالي

الرقم : ..../ ق/ ا/ أ/ ك.ع/2024

Autorisation de Dépôt de Mémoire de Master

Je soussigné: ...RAMDANE...CHAFIKA.....

Certifie que l'étudiant(e) :...Bouhenza...Roufaida.....

Spécialité : ....M2.RSD.....

Ayant soutenu le projet intitulé :...Analyse...des...données....
.....des...clients...en...gros.....
.....

A apporté les corrections nécessaires sur son manuscrit de Master

Signature de l'encadreur

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليـــــــم العالـــــــي و البحث العلمـــــــي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـامعة 20 أوت 1955- سكيكدة
كـــــلية العلـــــوم
مكتبـة الكليـة

Université 20 Août 1955 - Skikda
Faculté des Sciences
Bibliothèque de la Faculté

سكيكدة في 2024

بطاقة معلومات خاصة بمذكرة التخرج

اسم و لقب الطالب : رقم التسجيل :

Bouhezza Rafaista * 360054487 *

Hadji chaima * 360039644 *

* *

* *

اسم و لقب المشرف على المذكرة : Ramdane chaFika

عنوان المذكرة : Analyse des données des clients en gros

القسم : informatique

المستوى : Master 2

التخصص : Réseau et Système distribués (RSD)

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Août 1955- skikda-

Faculté des Sciences

Département d'Informatique

جامعة 20 أوت 1955- سكيكدة

كلية العلوم

قسم الاعلام الالي

الرقم : ..../ ق/ ا/ ا/ ك.ع/2024

Autorisation de Dépôt de Mémoire de Master

Je soussigné: .....RAMDANE...CHAFIKA.....

Certifie que l'étudiant(e) :...Hadji..Chaima.....

Spécialité : .....M2.RSD.....

Ayant soutenu le projet intitulé :..Analyse...des..données.....
.....des...clients..en gros.....
.....

A apporté les corrections nécessaires sur son manuscrit de Master

Signature de l'encadreur