UNIVERSITE CLERMONT AUVERGNE

IUT d'Aurillac

Sciences de données

Projet pour la SAE 5.EMS.01

Analyse des ventes de produits dans un réseau de magasins

Membres du groupe:

Dosse Inza **COULIBALY**

Jeanne **MEDENOU**

Lore **IGOUWE**

Mame Diarra Bousso **NOBA**

Encadrant :

Mme Emilie **SOHIER**

Table des matières

Introduction

Dans le cadre de l'étude de la performance des 06 magasins d'un groupe de distribution alimentaire qui requiert notre expertise en modélisation statistique, deux bases de données sur les informations utiles à nos analyses sont mises à notre disposition.

La première base relative aux ventes, contient les informations sur les quantités vendues, les prix et les promotions puis la seconde base, regroupe les principaux postes de charges et le chiffre d'affaires de chaque magasin.

L'objectif principal de cette étude est d'identifier les facteurs clés influençant les ventes et d'analyser les différences de performance entre les magasins. Cette analyse doit permettre de formuler des recommandations opérationnelles visant à optimiser les ventes au sein du réseau.

Pour répondre à cette problématique, l'étude s'articulera autour de plusieurs étapes : le nettoyage et la préparation des données, l'analyse statistique descriptive, la construction et la comparaison de modèles statistiques, puis l'interprétation des résultats dans une perspective décisionnelle.

I. Exploration et nettoyage des données

1. Traitement des valeurs manquantes

Un diagnostic initial des bases de données a permis d'identifier la présence de valeurs manquantes. La base financière ne comporte aucune donnée manquante, tandis que la base des ventes présente des valeurs manquantes uniquement sur la variable **« Quantité Vendue »**. Au total, **2 040 observations** sont concernées par ces valeurs manquantes.

Nous avons eu l'idée de les imputer avec les données du Chiffre d'affaires (CA) de la base finances. En effet, avec les quantités vendues connues et les prix associés, on obtient le chiffre d'affaires connu qui aurait tout simplement été soustrait du CA total. Cette quantité restante du CA aurait permis de répartir les quantités vendues manquantes à condition qu'il n'y ait qu'un seul produit manquant par jour par magasin.

Malheureusement, après vérification, cette méthode s'avérait impossible puisque plusieurs produits manquaient par jour et nous n'étions pas en mesure de répartir les quantités proportionnellement à celles attendues.

On constate aussi qu'en calculant le chiffre d'affaires connu et en la comparant avec le chiffre d'affaires du fichier finances, nous avons des valeurs anormalement divergentes. Pour cela nous traiterons les fichiers différemment en dénonçant l'écart flagrant entre les valeurs théoriques et empiriques du chiffre d'affaires des magasins.

Afin de conserver l'ensemble des observations et d'éviter une perte d'information, les quantités manquantes ont été imputées à l'aide d'une méthode par régression, mise en œuvre via le package "***mice***". Cette approche permet d'estimer les valeurs manquantes à partir des autres variables disponibles. Les quantités imputées ont ensuite été converties en valeurs entières afin de respecter la nature discrète de la variable. Enfin, le chiffre d'affaires journalier a été recalculé sur la base des données complétées, garantissant ainsi des indicateurs cohérents pour la suite de l'analyse.

2. Traitement des valeurs aberrantes

Après correction des valeurs manquantes et visualisation des valeurs aberrantes, nous décidons de ne pas supprimer les valeurs aberrantes ici mais de les corriger car dans ce contexte, les promotions appliquées ou d'autres facteurs peuvent expliquer pourquoi nous avons des valeurs plus importantes que la moyenne. Nous avons donc utilisé une fonction de R qui permet de les corriger.

II. Analyse descriptive

Après la phase de nettoyage et de préparation des données, une analyse statistique descriptive a été réalisée afin de mieux comprendre la structure et les principales caractéristiques des données. Elle s'est appuyée à la fois sur des indicateurs statistiques (moyennes, médianes, dispersions) et sur des représentations graphiques. Elle a permis d'examiner l'évolution des ventes dans le temps, l'impact des promotions, les différences entre les jours de semaine et les week-ends, ainsi que les écarts de performance entre les magasins.

Par ailleurs, une étude des données financières a été menée afin de comparer les niveaux de charges, de chiffre d'affaires et de bénéfices entre les magasins. Cette approche permet d'obtenir une première vision globale de la rentabilité et de l'efficacité de gestion des différents points de vente.

Les données issues des deux bases, telles qu'illustrées ci-dessous, couvrent une période allant du 17 avril 2022 au 26 février 2024. Par ailleurs, chaque magasin est représenté de manière équilibrée dans les deux bases de données. En effet, on observe le même nombre d'observations pour chacune des variables et pour chacun des magasins, ce qui garantit une structure homogène des données. Cette répartition équilibrée permet de comparer les performances des magasins sans biais lié à une sur- ou sous-représentation de chacun des magasins.

- **Aperçu des données de la base des ventes**

```
      Date                 Produit          Magasin      Promotion   Prix_Unitaire
 Min.   :2022-04-17   Beurre      : 4086   Magasin_1:6810   Non:36780   Min.   : 0.490
 1st Qu.:2022-10-04   Bœuf        : 4086   Magasin_2:6810   Oui: 4080   1st Qu.: 0.990
 Median :2023-03-23   Café        : 4086   Magasin_3:6810               Median : 1.490
 Mean   :2023-03-23   Jus d'orange: 4086   Magasin_4:6810               Mean   : 3.195
 3rd Qu.:2023-09-09   Lait        : 4086   Magasin_5:6810               3rd Qu.: 3.390
 Max.   :2024-02-26   Pain        : 4086   Magasin_6:6810               Max.   :12.600
                      (Other)     :16344
 Quantité_Vendue
 Min.   :  3
 1st Qu.: 38
 Median : 58
 Mean   : 63
 3rd Qu.: 84
 Max.   :124
```

- **Aperçu des données de la base finance**

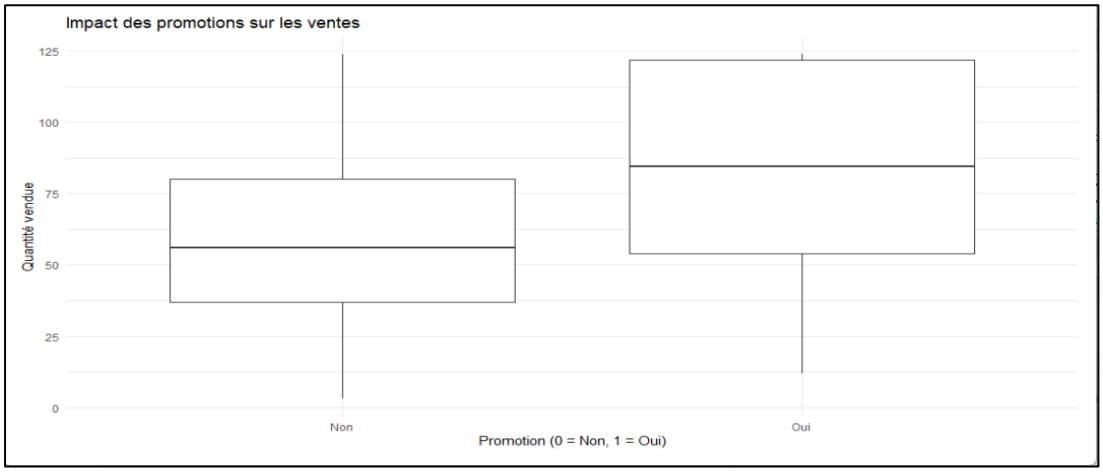
```
      date                magasin    masse_salariale depenses_fonctionnement
 Min.   :2022-04-17   Magasin_1:681   Min.   :1601    Min.   :1000
 1st Qu.:2022-10-04   Magasin_2:681   1st Qu.:1761    1st Qu.:1487
 Median :2023-03-23   Magasin_3:681   Median :1876    Median :1988
 Mean   :2023-03-23   Magasin_4:681   Mean   :1875    Mean   :1990
 3rd Qu.:2023-09-09   Magasin_5:681   3rd Qu.:1984    3rd Qu.:2495
 Max.   :2024-02-26   Magasin_6:681   Max.   :2252    Max.   :3000
 chiffre_affaires budget_publicite
 Min.   : 2293    Min.   : 0.08
 1st Qu.: 4129    1st Qu.:24.94
 Median : 5007    Median :50.16
 Mean   : 4999    Mean   :50.04
 3rd Qu.: 5780    3rd Qu.:75.51
 Max.   :10133    Max.   :99.98
```

L’analyse descriptive faite a notamment permis d’identifier les tendances générales, les écarts entre produits et entre magasins, ainsi que l’impact potentiel de certains facteurs tels que les promotions ou les saisons (on considère que les magasins sont dans un pays qui compte les O4 saisons-été, automne, hiver, printemps).

Au terme de celle-ci, on retient que les magasins appliquent approximativement la même politique des prix. Bien qu’on suppose que les prix ont un effet significatif sur les quantités vendues dans chaque magasin, ils n’expliquent pas la différence des quantités vendues par magasin

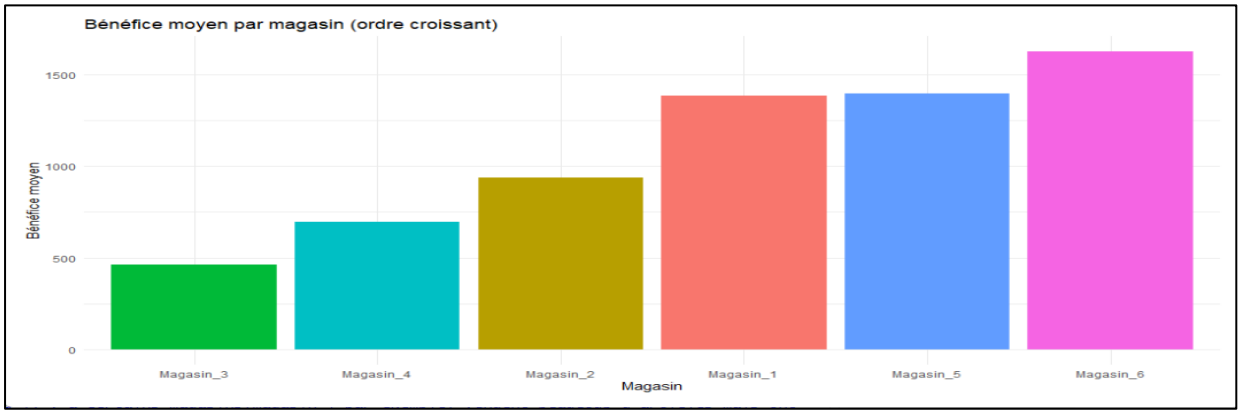
On retient également que la demande est relativement stable pour la plupart des produits mais que certains facteurs comme les promotions entraînent des pics de ventes notables.

En effet, le boxplot ci-dessous relève l’impact globalement positif des promotions sur les quantités vendues. On observe que la médiane des ventes est plus élevée lorsque les produits en promotion, ce qui suggère une hausse générale de la demande durant ces périodes. Par ailleurs, l’étendue des valeurs (dispersion) est également plus importante avec promotion : les volumes vendus varient davantage, ce qui indique que l’effet promotionnel n’est pas strictement uniforme et peut dépendre du contexte (produit, magasin, période). Ainsi, les promotions constituent un levier efficace pour augmenter les ventes, tout en entraînant une variabilité plus forte des performances.



En ce qui concerne le bénéfice moyen par magasin, calculé comme la différence entre le chiffre d’affaires et la somme des charges (masse salariale, dépenses de fonctionnement et budget publicitaire), on observe que le magasin 6 présente le bénéfice moyen le plus élevé, tandis que le magasin 3 affiche le bénéfice moyen le plus faible. Ce résultat indique que la performance financière ne dépend pas uniquement des quantités vendues, mais également de la structure des coûts. En particulier, certains magasins peuvent présenter de bons volumes de vente tout en générant un bénéfice plus faible, ce qui suggère des charges

relativement plus importantes. Ainsi, les charges apparaissent comme un facteur déterminant dans la rentabilité globale des magasins.



III. Modélisation statistique

Nous avons testé deux modèles de régression adaptés aux variables positives : la régression de Poisson et la régression Gamma. Le modèle Gamma présente un AIC plus faible, ce qui traduit un meilleur ajustement numérique. Toutefois, ce modèle n'est pas adapté à la nature de la variable étudiée, qui est la quantité vendue. En effet, la quantité vendue correspond à une variable de comptage, discrète et non bornée, tandis que la régression Gamma suppose une variable continue strictement positive. Ainsi, malgré un meilleur ajustement en termes d'AIC, l'utilisation du modèle Gamma serait théoriquement incohérent avec les données. Nous retenons donc la régression de Poisson, plus appropriée pour modéliser une variable de comptage.

Le modèle de Poisson (m0) met en évidence un effet positif significatif de la promotion sur la quantité vendue. En moyenne, la présence d'une promotion augmente la quantité vendue d'environ 37 % par rapport à une situation sans promotion.

Dans un second temps, l'ajout des variables Produit et Magasin (m1) améliore fortement la qualité du modèle. Les résultats montrent que l'effet de la promotion reste stable et significatif : une promotion augmente toujours la quantité vendue d'environ 37 %, même après prise en compte des différences entre produits et entre magasins. Les ventes varient également significativement selon le type de produit et le magasin, ce qui explique une part importante de la variable observée.

L'introduction du prix unitaire (m2) conduit à une légère amélioration du modèle (AIC plus faible). L'effet de la promotion demeure significatif, avec une augmentation d'environ 38 % des ventes en période promotionnelle. Le prix unitaire présente un effet positif, mais faible :

une hausse d'une unité du prix est associée à une augmentation d'environ 0,6 % de la quantité vendue.

L'ajout d'une variable temporelle sous forme de tendance linéaire (Date, modèle m3) améliore l'ajustement, ce qui met en évidence une évolution significative des ventes au cours du temps.

Afin de mieux modéliser la saisonnalité, nous avons ensuite testé deux formulations plus adaptées : la saison (M4) et le mois (M5). Le modèle M4 montre que la saison a un effet très marqué sur les ventes : par rapport à l'été, les ventes diminuent fortement en automne (-29 %), augmentent en hiver (+14 %) et augmentent fortement au printemps (+34 %). L'ajout de la saison améliore considérablement la performance du modèle (forte baisse de l'AIC).

Le modèle M5 met en évidence un effet encore plus significatif du mois sur la quantité vendue. Comparé à janvier, les ventes augmentent au printemps (février à mai), tandis qu'elles diminuent fortement entre juillet et novembre, avec une baisse maximale en octobre (-42 %). Le mois permet donc de capturer des effets calendaires précis (vacances, rentrée, fêtes), ce qui explique une grande partie de la variabilité des ventes.

Nous avons également étudié l'effet des jours de la semaine. Le modèle M6 montre que les ventes augmentent d'environ 15 % le week-end par rapport aux jours de semaine. Le modèle M7, plus détaillé, confirme que les ventes sont environ 13 à 14 % plus faibles du lundi au vendredi par rapport au dimanche, tandis que le samedi n'est pas significativement différent du dimanche. Ces résultats confirment un comportement d'achat concentré sur le week-end.

Enfin, la comparaison globale des AIC montre que le modèle M5 est le plus performant parmi l'ensemble des modèles testés. Il offre le meilleur compromis entre qualité d'ajustement et complexité, tout en intégrant explicitement la saisonnalité via le mois. Ainsi, nous retenons le modèle M5 comme modèle final. Il se présente comme suit :

```
Call:
glm(formula = Quantité_Vendue ~ Promotion + Produit + Magasin +
    Prix_Unitaire + Mois, family = poisson(link = "log"), data = ventes1)

Coefficients:
                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)          4.159967   0.003883 1071.350  < 2e-16 ***
PromotionOui         0.322466   0.002170  148.595  < 2e-16 ***
ProduitBœuf         -0.032367   0.010915   -2.965  0.00302 **
ProduitCafé         -0.169099   0.003118  -54.225  < 2e-16 ***
ProduitJus d'orange -0.202107   0.003025  -66.803  < 2e-16 ***
ProduitLait          0.109387   0.002965   36.899  < 2e-16 ***
ProduitPain          0.162026   0.002920   55.479  < 2e-16 ***
ProduitPâtes         0.090094   0.002912   30.944  < 2e-16 ***
ProduitPoulet       -0.034221   0.006993   -4.894 9.90e-07 ***
ProduitRiz          -0.061968   0.003009  -20.597  < 2e-16 ***
ProduitYaourt       -0.167416   0.003077  -54.403  < 2e-16 ***
MagasinMagasin_2     0.129797   0.002160   60.092  < 2e-16 ***
MagasinMagasin_3    -0.078082   0.002274  -34.339  < 2e-16 ***
MagasinMagasin_4     0.085184   0.002183   39.018  < 2e-16 ***
MagasinMagasin_5     0.059251   0.002198   26.962  < 2e-16 ***
MagasinMagasin_6     0.171774   0.002139   80.317  < 2e-16 ***
Prix_Unitaire        0.009418   0.001135    8.301  < 2e-16 ***
MoisFévrier          0.133167   0.002749   48.448  < 2e-16 ***
MoisMars             0.174199   0.003177   54.823  < 2e-16 ***
MoisAvril            0.155349   0.002882   53.901  < 2e-16 ***
MoisMai              0.104743   0.002676   39.136  < 2e-16 ***
MoisJuin             0.013013   0.002757    4.720 2.36e-06 ***
MoisJuillet         -0.143737   0.002848  -50.463  < 2e-16 ***
MoisAoût            -0.355339   0.003026 -117.441  < 2e-16 ***
MoisSeptembre       -0.519752   0.003213 -161.753  < 2e-16 ***
MoisOctobre         -0.539304   0.003198 -168.660  < 2e-16 ***
MoisNovembre        -0.414851   0.003107 -133.528  < 2e-16 ***
MoisDécembre        -0.200984   0.002894  -69.459  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pour étudier l'impact des charges sur la performance financière, nous avons modélisé le chiffre d'affaires via une régression de gamma car c'est une variable continue positive.

Le modèle Gamma met en évidence des différences significatives de chiffre d'affaires entre magasins : le Magasin_6 présente un chiffre d'affaires significativement plus élevé que les autres magasins. Il représente notre référence. Parmi les variables de charges, seule la masse salariale est associée significativement au chiffre d'affaires (effet positif sur le CA), tandis que les dépenses de fonctionnement et le budget publicitaire ne présentent pas d'effet statistiquement significatif dans ce modèle. Ces résultats suggèrent que la performance est davantage liée à des facteurs structurels (taille, fréquentation, organisation) et à la capacité opérationnelle (personnel) qu'au niveau de dépenses de fonctionnement ou de publicité.

IV. Interprétation des résultats

1. Différences entre les magasins

Le magasin 6 apparaît comme le plus performant, suivi des magasins 2 et 4, tandis que le magasin 3 enregistre le volume de ventes le plus faible. Les magasins 1 et 5 se situent dans une position intermédiaire. En effet, le magasin 6 a plus de ventes, le chiffre d'affaires le plus élevé et une grande rentabilité. Le magasin 2 et 4 présentent également des quantités de ventes élevées, mais le magasin 4 semble être pénalisé par un bénéfice plus faible probablement lié à des charges plus importantes que le chiffre d'affaires. Alors que le magasin

3 a le plus faible bénéfice car il cumule une faible quantité de ventes et une part de charges très élevée (plus élevée que le chiffre d'affaires). Enfin, les prix étant très similaires entre magasins, les écarts observés s'expliquent principalement par des facteurs comme l'emplacement du magasin et la taille du magasin au plutôt que par une stratégie des prix et des promotions.

2. Les leviers d'action potentiels pour optimiser les ventes

- Les promotions ont un effet significatif sur les quantités vendues. Elles constituent ainsi un levier efficace pour stimuler la demande. Il est donc recommandé de renforcer les actions promotionnelles durant les périodes creuses, caractérisées par de faibles volumes de ventes, en particulier dans les magasins dont les performances sont inférieures à la moyenne. Cette stratégie permettrait de dynamiser les ventes tout en contribuant à réduire les écarts de performance entre les magasins.

- Les charges des magasins, et en particulier la masse salariale, sont associées de manière significative à leur performance financière. Une part élevée de charges par rapport au chiffre d'affaires peut entraîner une diminution du bénéfice. Dans cette optique, il pourrait être pertinent d'ajuster la masse salariale en fonction de l'activité réelle de chaque magasin, afin d'optimiser la gestion des coûts tout en maintenant un niveau de service adapté. Afin d'améliorer la rentabilité sans réduire les effectifs, il est recommandé d'optimiser l'organisation du travail. Cela peut passer par une adaptation des plannings aux périodes d'affluence, le développement de la polyvalence des employés, la réduction des heures supplémentaires ou encore l'automatisation de certaines tâches. Ces mesures permettent de mieux aligner la masse salariale avec le niveau réel d'activité, tout en maintenant la qualité de service.

V. Conclusion

Cette étude a permis d'identifier les principaux facteurs influençant la quantité vendue dans un réseau de six magasins, à partir des données de ventes et de finances sur la période 2022–2024. Les résultats montrent que les promotions ont un effet fortement positif sur les ventes, et que la saisonnalité joue un rôle majeur, avec des variations importantes selon les mois. Les ventes sont également plus élevées le week-end, confirmant un comportement d'achat concentré sur la fin de semaine.

Les modèles statistiques indiquent aussi des différences structurelles entre magasins : le Magasin_6 apparaît comme le plus performant, tandis que le Magasin_3 est le plus fragile. Enfin, les écarts de performance ne s'expliquent pas par les prix (très similaires), mais plutôt par des facteurs structurels et organisationnels. Les leviers d'action principaux sont donc le renforcement des promotions sur les périodes creuses et l'optimisation des charges, notamment via une meilleure organisation de la masse salariale.