MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE RECHERCHE SCIENTIFIQUE

REPUBLIQUE DE COTE D'IVOIRE

Institut Supérieur de Statistique
D'Econométrie et de Data Science

Union-Discipline-Travail

MASTER 1

STATISTIQUE – ECONOMETRIE – DATA SCIENCE

MINI PROJET

STATISTIQUES DESCRIPTIVE

ANALYSE DESCRIPTIVE DE SERIES TEMPORELLES ET PREVISION PAR LISSAGE EXPONETIEL : HOLT WINTER

ANNEE ACADEMIQUE

2024 – 2025

Nom: YOBO

Prenom(s): BAYE GUY ANGE HENOC

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

Avant-Propos

Dans le paysage dynamique de la distribution alimentaire, la précision des prévisions de ventes est un enjeu crucial pour le succès des épiceries. La Corporación Favorita, l'un des principaux détaillants d'alimentation en Équateur, fait face à ce défi avec une large gamme de produits et un réseau étendu de supermarchés. Les enjeux sont multiples : des prévisions excessives peuvent entraîner des surplus de stocks de produits périssables, tandis que des prévisions insuffisantes peuvent mener à des ruptures de stock et à des clients insatisfaits.

Dans ce contexte, la capacité à anticiper avec précision la demande est d'autant plus complexe qu'elle doit tenir compte de nombreux facteurs, tels que la diversité des points de vente, les fluctuations saisonnières des goûts des consommateurs et l'impact des promotions sur les ventes. Corporación Favorita, qui gère plus de 200 000 références, doit constamment affiner ses méthodes de prévision pour s'adapter à un marché en évolution rapide.

Ce projet de prévision des ventes, basé sur des données offre l'opportunité de développer des modèles analytiques robustes pour prédire les ventes de milliers de produits. En exploitant des informations sur les dates, les spécificités des magasins, les promotions et les chiffres de vente, nous visons à améliorer la gestion des stocks et à optimiser la satisfaction des clients.

À travers cette analyse, nous espérons non seulement répondre aux défis logistiques de Corporación Favorita, mais aussi contribuer à l'innovation dans le domaine de la prévision des ventes dans le secteur de l'épicerie.

Table des matières

Liste des figures

Liste des tableaux

INTRODUCTION

Dans un secteur aussi dynamique que celui de l'épicerie, la gestion des ventes repose sur des prévisions précises et réactives. Corporación Favorita, l'un des principaux détaillants d'alimentation en Équateur, fait face à des défis uniques en raison de son vaste réseau de supermarchés et de la diversité de ses produits. Avec plus de 200 000 références, la capacité à anticiper la demande devient cruciale pour éviter à la fois les surplus de stocks de produits périssables et les ruptures de stock qui frustrent les consommateurs.

La nécessité d'une analyse approfondie des ventes s'intensifie dans un environnement où les préférences des consommateurs évoluent rapidement et où les promotions jouent un rôle clé dans la dynamique des ventes. Cette étude vise à comprendre ces dynamiques afin d'optimiser les prévisions et de répondre efficacement aux attentes des clients.

Comment Corporación Favorita peut-elle améliorer ses prévisions de ventes pour ses milliers de produits en tenant compte des spécificités des magasins, des tendances saisonnières et des impacts des promotions ? Cette problématique soulève des enjeux majeurs de gestion des stocks et de satisfaction client.

Nous visons à développer des modèles de prévision des ventes qui permettront d'augmenter la précision des estimations et de réduire les pertes liées aux surstocks et aux ruptures de stock. Ces résultats devraient également contribuer à des recommandations stratégiques pour optimiser les assortiments de produits et la planification des promotions.

Pour atteindre ces objectifs, nous nous appuierons sur nos données, en utilisant des techniques d'analyse de données avancées. Nous examinerons les tendances historiques des ventes, les caractéristiques des magasins et les promotions afin de construire des modèles prédictifs robustes. Ces modèles seront évalués sur leur capacité à fournir des prévisions fiables, contribuant ainsi à une gestion plus efficace des opérations au sein de Corporación Favorita.

Cette étude représente non seulement une opportunité d'amélioration pour Corporación Favorita, mais aussi un pas vers une compréhension plus approfondie des enjeux de la prévision des ventes dans le secteur de l'épicerie.

I) PREPARATION DES DONNEES

La préparation des données consiste en l'ensemble des tâches effectuées pour collecter, importer nettoyer, transformer, combiner, organiser et converties en série temporelle.

1) Dictionnaire de données

Un dictionnaire de données est un document qui présente une description complète de chaque variable utilisée dans une analyse statistique ou économétrique. Il détaille les propriétés, les caractéristiques et le contexte de chaque variable, ainsi que leur signification.

Nom de la Variable	Description	Type de Donnée	Valeurs possibles	Exemples
id	Identifiant unique de chaque enregistrement dans le jeu de données.	Numérique (int)	Identifiant unique	1001, 1002, 1003
date	Date à laquelle les données de vente sont enregistrées.	Date	Format : YYYY-MM-DD	2023-10-01, 2023-10-02
magasin_nbr	Numéro d'identification unique du magasin où les ventes ont été effectuées.	Numérique (int)	Identifiant unique du magasin	101, 102, 103
famille	Catégorie ou famille de produits vendus. Cette variable est un facteur avec 33 niveaux représentant différentes catégories de produits.	Facteur (ou Catégoriel)	33 catégories différentes	"AUTOMOBILE", "PUÉRICULTURE", "ALIMENTATION", ...
ventes	Quantité de ventes réalisées pour chaque enregistrement.	Numérique (int)	Quantité de produits vendus	10, 25, 50
en_promotion	Indicateur de promotion, indiquant si le produit était en promotion lors de la vente.	Binaire (0 ou 1)	0 : Non, 1 : Oui	0, 1

Tableau 1Dictionnaire de données

2) Présentation du jeu de données

Notre dataframe s'intitule **store**

N°	date	store_nb	family	sales	onpromotio
1	01/01/2013	1	AUTOMOTIVE	0	0
2	01/01/2013	1	BABY CARE	0	0
3	01/01/2013	1	BEAUTY	0	0
4	01/01/2013	1	BEVERAGES	0	0
5	01/01/2013	1	BOOKS	0	0
---	----------	---------	----------	----------	--------
3000882	15/08/2017	9	PLAYERS AND ELECTRONICS	6.000	0
3000883	15/08/2017	9	POULTRY	438.133	0
3000884	15/08/2017	9	PREPARED FOODS	154.553	1
3000885	15/08/2017	9	PRODUCE	2419.729	148
3000886	15/08/2017	9	SCHOOL AND OFFICE SUPPLIES	121.000	8

Tableau 2 Extraire du jeu de données

❖ *Structure du jeu de données*

Selon le modèle de notre tableau, le jeu de donnée « store » contient 3000888 observations et 6 variables dont 3 variables qualitatives et 3 variables quantitatives.

'data.frame': 3000888 obs. of 5 variables:
$ date : Factor w/ 1684 levels "2013-01-01","2013-01-02",..: 1 1 1 1 1 1 1 1 1 1 ...
$ store_nbr : int 1 1 1 1 1 1 1 1 1 1 ...
$ family : Factor w/ 33 levels "AUTOMOTIVE","BABY CARE",..: 1 2 3 4 5 6 7 8 9 10 ...
$ sales : num 0 0 0 0 0 0 0 0 0 0 ...
$ onpromotion: int 0 0 0 0 0 0 0 0 0 0 ...

Tableau 3 structure du jeu de données

3) Construction des tables de vente par mois

Après importation de notre jeu de données dans R, nous avons organisé les ventes totales par mois et créé la table des ventes totales par mois qui contient les variables Date et Ventes Totales constituant ainsi la table qui sera utilisée pour la construction de série temporelle des ventes totales par mois que nous étudierons.

janv. 2013	févr. 2013	mars-13	avr. 2013	mai-13	juin-13	juil. 2013	août-13	sept. 2013	oct. 2013	nov. 2013	déc. 2013
10 327 625	9 658 960	11 428 497	10 993 465	11 597 704	11 689 344	11 257 401	11 737 789	11 792 933	11 775 620	12 356 559	15 803 117
janv. 2014	févr. 2014	mars-14	avr. 2014	mai-14	juin-14	juil. 2014	août-14	sept. 2014	oct. 2014	nov. 2014	déc. 2014
18 911 641	12 038 353	20 365 584	12 861 251	13 379 785	13 319 958	19 421 891	13 885 176	20 022 416	20 396 101	20 531 635	24 340 454
janv. 2015	févr. 2015	mars-15	avr. 2015	mai-15	juin-15	juil. 2015	août-15	sept. 2015	oct. 2015	nov. 2015	déc. 2015
14 896 922	13 742 396	15 598 608	14 955 068	17 730 368	21 615 360	22 096 619	22 963 674	23 240 882	23 878 268	22 804 953	27 243 982
janv. 2016	févr. 2016	mars-16	avr. 2016	mai-16	juin-16	juil. 2016	août-16	sept. 2016	oct. 2016	nov. 2016	déc. 2016
23 977 805	21 947 409	23 131 781	25 963 025	24 779 432	22 209 219	23 462 672	22 452 414	22 417 448	24 030 390	24 642 640	29 640 288
janv. 2017	févr. 2017	mars-17	avr. 2017	mai-17	juin-17	juil. 2017	août-17				
26 328 160	23 250 112	26 704 018	25 895 308	26 911 847	25 682 822	27 011 478	12 433 323				

Tableau 4 Présentation de la table de vente par mois

4) Apurement des données dans la table générée

Dans cette partie nous essayerons de déterminer s'il y a des valeurs manquantes et des valeurs extrêmes et ou aberrante.

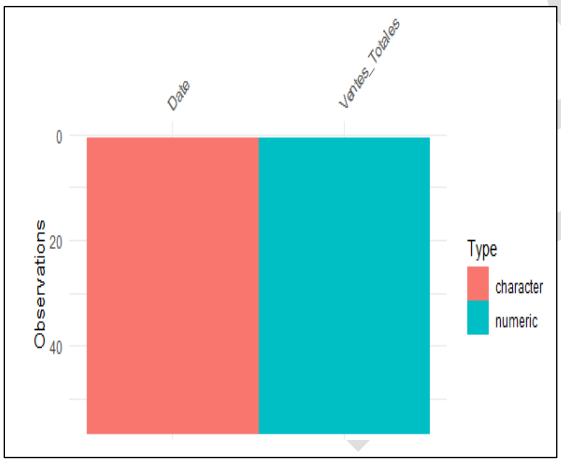


Figure 2 Visualisation des valeurs manquantes

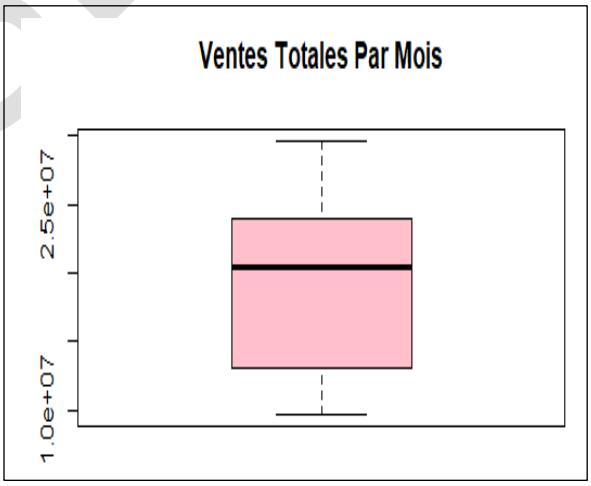


Figure 1 visualisation des valeurs aberrantes

Suite aux analyses nous constatons que la table des ventes par mois ne contient ni de valeurs manquantes.

II) ANALYSE DESCRIPTIVE DE LA SÉRIE TEMPORELLE DES VENTES TOTALES PAR MOIS

L'analyse descriptive de la série temporelle des ventes totales par mois consiste à décrire l'évolution des ventes totales par mois dans le temps (périodicité du phénomène observé, survenue d'événements accidentels).

1) Présentation de la série des ventes totales par mois

anv. 2013	févr. 2013	mars-13	avr. 2013	mai-13	juin-13
10327625	9658960	11428497	10993465	11597704	11689344
uil. 2013	41487	sept. 2013	oct. 2013	nov. 2013	déc. 2013
11257401	11737789	11792933	11775620	12356559	15803117
anv. 2014	févr. 2014	41699	avr. 2014	41760	41791
18911641	12038353	20365584	12861251	13379785	13319958
uil. 2014	41852	sept. 2014	oct. 2014	nov. 2014	déc. 2014
19421891	13885176	20022416	20396101	20531635	24340454
anv. 2015	févr. 2015	42064	avr. 2015	42125	42156
14896922	13742396	15598608	14955068	17730368	21615360
uil. 2015	août-15	sept. 2015	oct. 2015	nov. 2015	déc. 2015
22209619	22963674	23240882	23878268	22804953	27243982
anv. 2016	févr. 2016	mars-16	avr. 2016	mai-16	juin-16
23977805	21947409	23131781	25963025	24779432	22209219
uil. 2016	août-16	sept. 2016	oct. 2016	nov. 2016	déc. 2016
23462672	22452414	22417448	24030390	24642640	29640288
anv. 2017	févr. 2017	mars-17	avr. 2017	mai-17	juin-17
26328160	23250112	26704018	25895308	26911847	25682822
uil. 2017	août-17				
27011478	12433323				

2) Analyse univarié de la variable vente totale par mois

2-1) Tableau statistiques

	Effectif	Eff_cum_crois	Eff_cum_déc	frequence	Freq_cum_crois	Freq_cum_déc
9658959.7774368	1	1	56	0.0179	0.0179	1.0000
10327624.7369095	1	2	55	0.0179	0.0357	0.9821
10993464.7380118	1	3	54	0.0179	0.0536	0.9643
11257400.6076808	1	4	53	0.0179	0.0714	0.9464
11428497.0374875	1	5	52	0.0179	0.0893	0.9286
11597704.0070268	1	6	51	0.0179	0.1071	0.9107
11689344.0622436	1	7	50	0.0179	0.1250	0.8929
11737788.9194727	1	8	49	0.0179	0.1429	0.8750
11775620.3605191	1	9	48	0.0179	0.1607	0.8571
11792933.2318679	1	10	47	0.0179	0.1786	0.8393

Tableau 5extrait du tableau statistique

2-2) Représentation graphiques

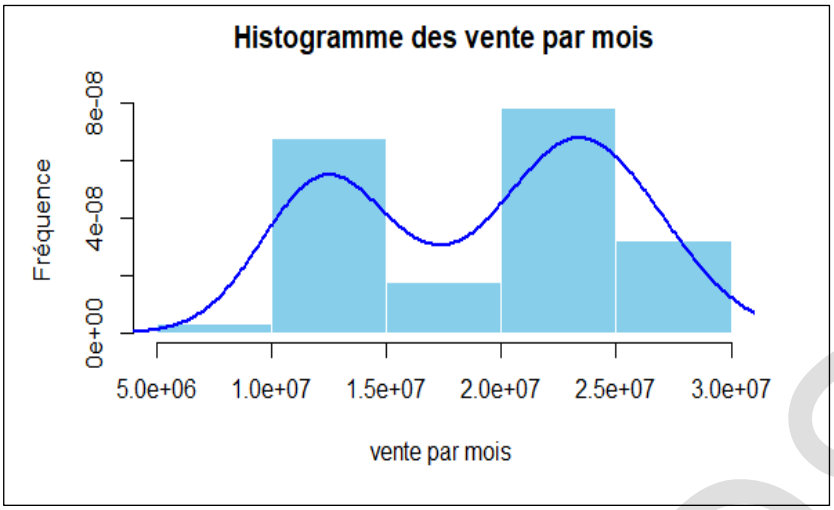


Figure 3: Histogramme des ventes par mois

2-3) Paramètre statistique et interprétation

INDICATEURS	VALEUR	INTERPRÉTATIONS
Indicateurs de tendance centrale		
Minimum	9 658 960	Le minimum du total des ventes par mois de l'épicerie Favorita sur toute la période observée est de 9 658 960
Maximum	29 640 288	Le maximum du total des ventes par mois de l'épicerie Favorita sur toute la période observée est de 29 640 288
Moyenne	19 172 231	Les ventes moyennes totales par mois de l'épicerie Favorita sur toute la période observée est de 19 172 231
Indicateurs de dispersion		
Ecart-type	5 806 956	Sur la période observée, le total des ventes par mois de l'épicerie Favorita variait entre 13 365 275 et 24 979 187
Coefficient de variation	30,28 837	Le coefficient de variation était de 30,29%, ce qui indique une forte hétérogénéité de la distribution
Indicateurs de forme		
Coefficient d'asymétrie skewness	-0,1546965	Le Skewness étant négatif, majoritairement la distribution du total des ventes par mois est à gauche de la moyenne (le total des ventes moyennes par mois) et les valeurs aberrantes encore présentes à droite
Coefficient d'aplatissement (kurtosis)	1,573615	Le Kurtosis étant inférieur à 3, cela indique que par rapport à une distribution normale, la distribution du total des ventes par mois est aplatie et possède des queux fines (la courbe est dite platikurtique)

Tableau 6: parametre statistiques

3) Détermination du model par la méthode analytique

L'objectif est de déterminer si les données suivent un modèle additif ou un modèle multiplicatif. Un modèle additif s'écrit sous la forme suivante : Yt=Tt+St+εtY_t = T_t + S_t + \varepsilon_tYt=Tt+St+εt où YtY_tYt représente la variable observée à la période ttt, TtT_tTt la tendance, StS_tSt la saisonnalité, et εt\varepsilon_tεt l'erreur. En revanche, un modèle multiplicatif s'exprime ainsi : Yt=Tt×St+εtY_t = T_t \times S_t + \varepsilon_tYt=Tt×St+εt

Il existe deux méthodes principales pour déterminer le type de modèle : la méthode graphique et la méthode analytique. Pour cette étude, nous utiliserons la méthode analytique.

Méthode analytique

Avec cette méthode, on commence par calculer les moyennes et les écarts-types pour chaque période considérée. Ensuite, on estime la droite de régression des moindres carrés de l'équation suivante : σ=ay¯+b\sigma = a \bar{y} + bσ=ay¯+b

- Si a=0a = 0a=0, le modèle est additif et peut être représenté par l'équation : μ=a+bt+ε\mu = a + b t + \varepsilonμ=a+bt+ε
- Si a≠0a \neq 0a□=0, le modèle est multiplicatif et prend la forme : Yt=a×bt×εtY_t = a \times b^t \times \varepsilon_tYt=a×bt×εt où μ\muμ est la moyenne, aaa est un coefficient multiplicatif constant, bbb est le facteur multiplicatif (avec btb^tbt représentant la tendance exponentielle au fil du temps), ttt est le temps, et εt\varepsilon_tεt est le terme d'erreur.

Dans notre analyse sous R, nous avons estimé les coefficients aaa et bbb à partir du modèle linéaire (lm) et testé si a=0a = 0a=0 ou non pour déterminer le type de modèle. Les résultats ont montré que a=−7082974404a = -7 082 974 404a=−7082974404 et b=3524128b = 3 524 128b=3524128.

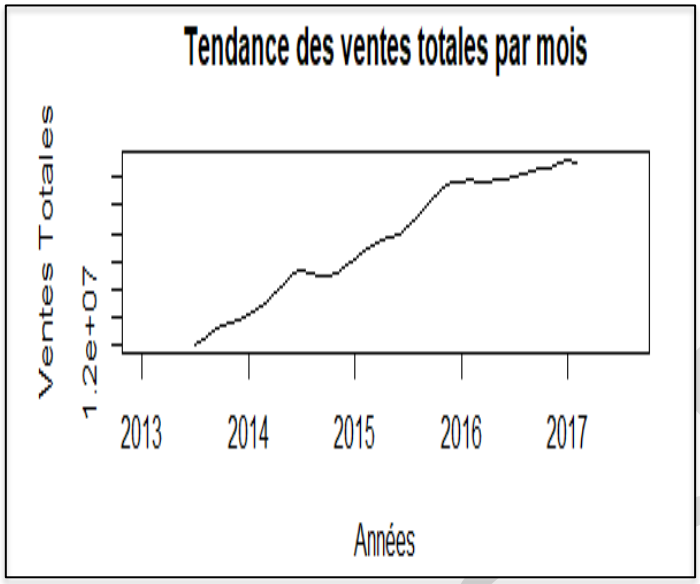
Comme a≠0a \neq 0a□=0, notre modèle est donc **multiplicatif**. L'équation de la droite de régression est ainsi donnée par : μt=−7082974404×(3524128)t\mu_t = -7 082 974 404 \times (3 524 128)^tμt=−7082974404×(3524128)t

Cela confirme que les données suivent une tendance exponentielle, et que le modèle multiplicatif est le plus approprié pour décrire l'évolution des chiffres d'affaires.

4) Décomposition de la série des ventes totales par mois

Étant donné que notre série suit un **modèle multiplicatif**, nous avons choisi de la décomposer afin d'analyser ses différentes **composantes** :

✓ *L'évolution des ventes mensuelles sur la période étudiée*



D'après les données utilisées pour construire la série chronologique, il apparaît une tendance générale à l'augmentation des ventes totales mensuelles au fil des années. Cette évolution de la tendance sera détaillée lors de la description de la série.

Figure 4: Tendance generale de la serie des ventes par mois

✓ *La saisonnalité de la série des ventes totales par mois*

Reproduction d'un phénomène a intervalle de temps régulier. Effet saisonnier clairement visible

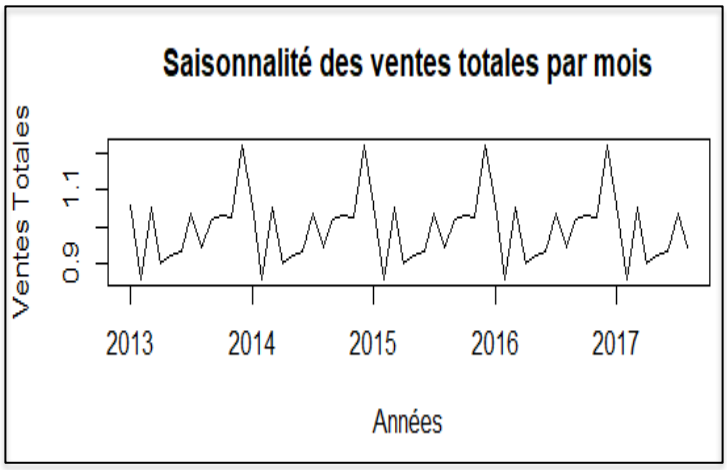


Figure 5: Saisonnalité de la série des ventes totale par mois

✓ ***Le bruit de la série des ventes totales par mois***

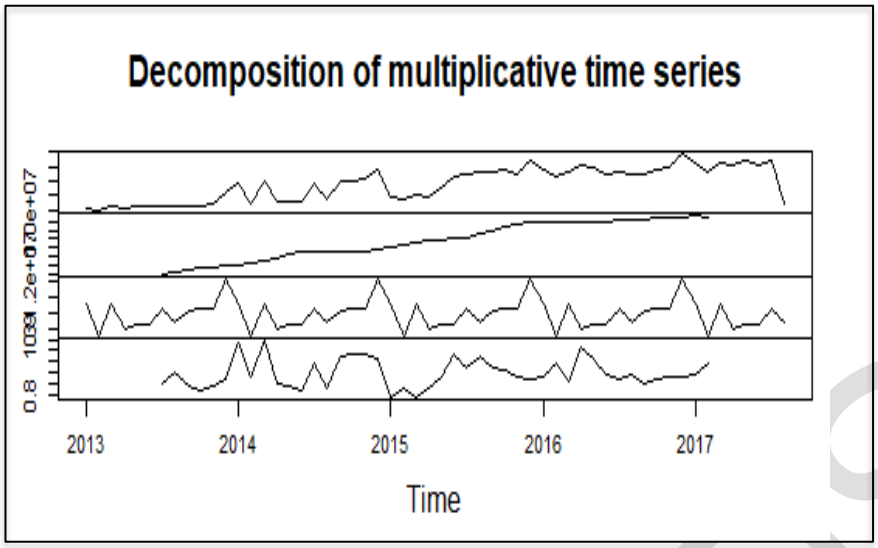


Figure 6: Décomposition de la série des ventes totales par mois (bruits)

La courbe inférieure du graphique représente la composante aléatoire de la série des ventes totales, de janvier 2013 au 31 août 2017

❖ **Indice de dépendance de la série des ventes totales par mois**

La fonction d'autocorrélation est un outil statistique qui mesure la corrélation entre les valeurs d'une série temporelle et les mêmes valeurs à différents décalages temporels. En d'autres termes, elle évalue la similarité entre les observations d'une variable à différents instants dans le temps, permettant ainsi d'identifier des patterns ou des dépendances temporelles dans la série.

Le seuil de significativité sur un graphique d'autocorrélation définit la zone au-delà de laquelle les valeurs de l'autocorrélation sont considérées comme statistiquement significatives. En dessous de ce seuil, les corrélations sont jugées non significatives, ce qui signifie qu'elles peuvent être dues au hasard. Ce seuil est généralement représenté par deux lignes horizontales sur le graphique : une ligne supérieure et une ligne inférieure, formant ainsi une plage autour de zéro. Si les valeurs de l'autocorrélation dépassent ces lignes, elles sont considérées comme significatives, ce qui indique qu'il existe une relation temporelle entre les observations à ces décalages.

✓ **Autocorrelation**

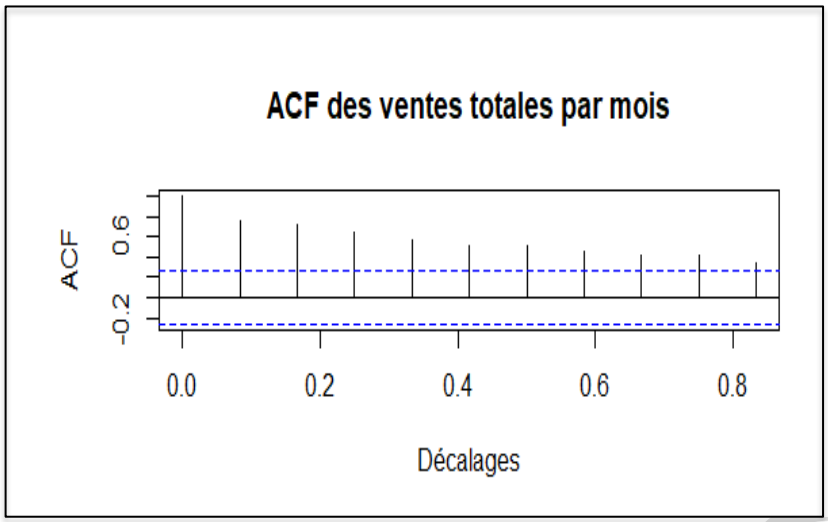


Figure 7: ACF de la series des ventes totales par mois

D'après notre corrélogramme, on observe un seuil critique qui indique une autocorrélation significative. En effet, bien que nous ayons une autocorrélation au décalage zéro, celle-ci n'a pas de signification particulière, car l'autocorrélation de la série elle-même sans décalage est naturellement égale à 1. Cependant, on remarque une autocorrélation relativement forte à chaque comparaison de la série avec des observations de plus en plus éloignées dans le temps, avec des valeurs atteignant des seuils comme 0.75 et 0.634. Cela suggère qu'il existe une certaine relation entre les valeurs de la série à court terme. Toutefois, à mesure que l'on s'éloigne dans le temps avec des décalages plus importants, ces valeurs d'autocorrélation décroissent progressivement, ce qui indique qu'il n'y a pas de dépendance structurelle forte à long terme entre les observations.

✓ **Autocorrelation partielle**

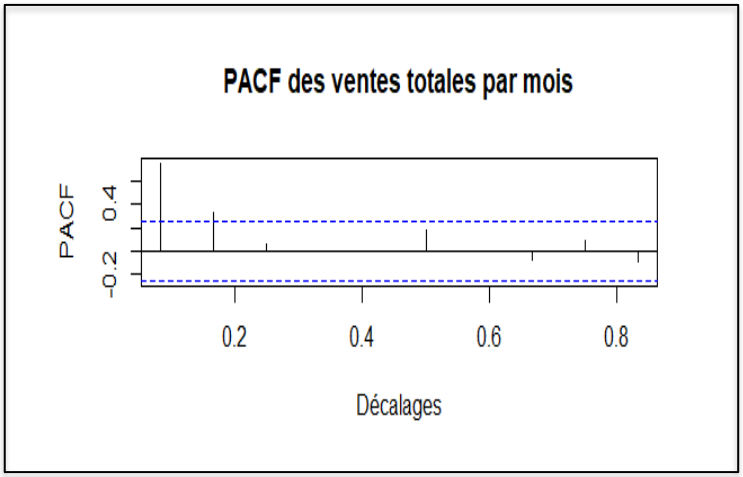


Figure 8: PACF de la série des ventes totales par mois

L'autocorrélation partielle (PACF) permet de quantifier la dépendance linéaire entre deux réalisations successives mais conditionnellement aux réalisations intermédiaires. En d'autres Termes, L'autocorrélation partielle permet de mesurer l'autocorrélation d'un signal pour un Décalage k "indépendamment" des autocorrélations pour les décalages inférieurs.

resultats du pacf :

Partial autocorrelations of series 'vente_mois', by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333
 0.759 0.326 0.057 -0.003 -0.012 0.172 -0.012 -0.079 0.092 -0.098
0.9167 1.0000
 0.024 0.045

D'après nos résultats, on remarque une forte corrélation directe a un décalage de 0.0833 suivie d'une décroissance progressive des valeurs pour des décalages plus éloignés. Notons aussi la présence de valeurs négatives suggérant des relations inverses entre les observations.

En résumé, les autocorrélations observés lors de l'acf n'étaient que les effets résiduels de l'autocorrélation pour le décalage de 0.0833.

❖ Analyse de l'évolution des ventes totales mensuelles

- Illustration visuelle de l'évolution mensuelle des ventes totales.

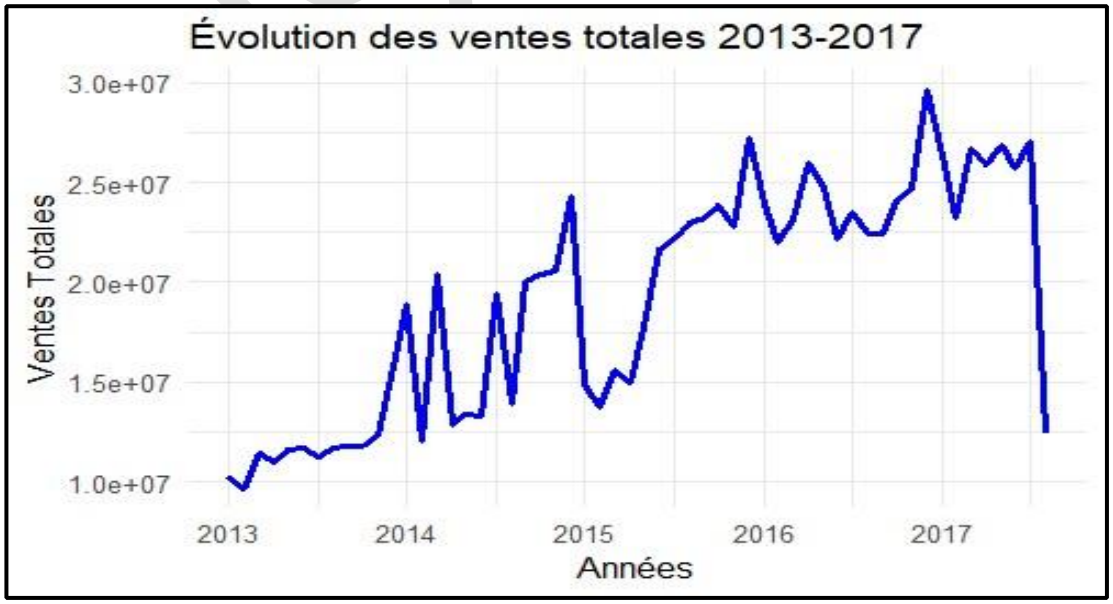


Figure 9: Evolution des ventes par mois

NB :

Sur la période allant de 2013 à 2017, les ventes mensuelles de l'épicerie Favorita ont présenté des variations régulières, ponctuées par des baisses notables :

- **De 2013 à 2014**, les ventes mensuelles ont globalement progressé, malgré quelques légères baisses observées en février, avril et juillet 2013. Elles sont passées de 10 327 625 en janvier 2013 à 18 911 641 en janvier 2014.
- **De janvier à décembre 2014**, les ventes mensuelles ont enregistré des fluctuations importantes et régulières. La moyenne annuelle des ventes sur cette période s'est élevée à 17 456 187.
- **De 2015 à 2017**, les ventes mensuelles ont continué d'augmenter de manière presque constante, bien qu'avec des baisses marquées, notamment en août 2017.

Il est à noter que, de manière générale, les ventes mensuelles de l'épicerie Favorita tendent à diminuer chaque année en février, probablement en raison de facteurs liés à la demande saisonnière.

III) PRÉVISION PAR LISSAGE EXPONENTIEL DE LA SÉRIE TEMPORELLE DES VENTES TOTALES PAR MOIS

La série temporelle des ventes mensuelles commence en janvier 2013 et se termine en août 2017. L'objectif est de prévoir le total des ventes pour les huit premiers mois de 2018, tout en estimant également les quatre derniers mois de 2017, soit une prédiction sur une période de 12 mois.

1) VALIDATION DU MODÈLE PAR L'ANALYSE DES RÉSIDUS DE LA SÉRIE TEMPORELLE DES VENTES TOTALES PAR MOIS

Il s'agit de vérifier si les résidus du modèle choisi suivent un processus bruit blanc. L'analyse des résidus permet de vérifier si le modèle utiliser est bon ou non. Pour se faire, des tests statistiques tels que le ljung-Box ou l'examen des autocorrélations des résidus seront effectués pour détecter la présence de corrélation dans les résidus.

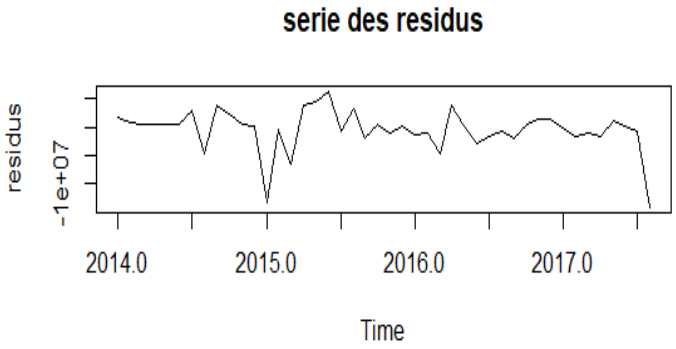


Figure 10: Graphique des residus

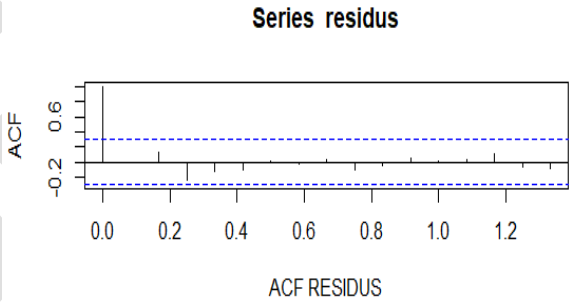


Figure 11: ACF des residus

1-1) Vérification de la blancheur des résidus de la série des quantités de marchandises retournées : graphiques des résidus

Le **graphique des résidus** nous permet de **se faire une idée de la distribution des résidus**. **L'ACF** (autocorrélogramme) **des résidus** est un graphique qui **montre la corrélation entre les résidus à des lags différents**. Nous pouvons voir à partir de l'échantillon du corrélogramme que les résidus ne sont pas corrélés au décalage temporel de la série, ce qui est tout à fait appréciable pour la qualité d'un modèle.

Pour vérifier s'il existe des preuves significatives pour les corrélations non nulles aux décalages 1 à 11, nous pouvons effectuer un test LjungBox.

Résidus du modèle choisi

	Jan	Feb	Mar	Apr
2014	1719432.44	847300.91	424231.00	487398.90
2015	-13161584.45	-325052.59	-6437780.24	3971054.61
2016	-1191121.59	-859488.45	-4750211.77	3807796.52
2017	-24558.68	-1479779.51	-1115684.17	-1726298.33
	May	Jun	Jul	Aug
2014	515970.74	526455.31	2958500.81	-4537737.02
2015	4461222.76	6392867.13	-727705.32	3166799.14
2016	212507.96	-2964608.62	-1584514.35	-685255.26
2017	1252053.27	246013.40	-533475.11	-14125512.68
	Sep	Oct	Nov	Dec
2014	3916480.82	2331447.81	423100.33	289222.01
2015	-1806206.95	480760.48	-871215.77	344930.28
2016	-1867207.79	446189.81	1377018.32	1479824.68
2017				

Tableau 7: Residus du modèle choisi

1-2) Tests de validation du modèle

❖ Test de Ljung

Le test de bruit blanc de la série est très important pour valider un modèle de prévision. Pour cela, nous effectuons le test de Ljung.

- **Hypothèse nulle** : La série est un bruit blanc ;
- **Hypothèse alternative** : la série n'est pas un bruit blanc.

Box-Ljung test
Data : résidus
X-squared = 5,846, df = 11, p-value = 0,8 834

Le résultat du test nous montre que la **statistique du test de Ljung** est de **2,6546** et que la **p-value = 0,89 > 0,05**, on ne peut rejeter H_0, **la série est un bruit blanc**.

Pour être sûr que le modèle prédictif ne peut pas être amélioré, il est judicieux de **vérifier si les erreurs de prévision sont normalement réparties de moyenne zéro et de variance constante**. Pour vérifier si les erreurs de prévision ont une variance constante, nous pouvons établir un **graphique temporel des erreurs de prévision**.

Nous allons pour finir vérifier la normalité des résidus : les erreurs suiventelles un processus gaussien de moyenne nulle (bruit blanc gaussien) ?

❖ Test de normalité des erreurs : test de Shapiro-Wilk

- **Hypothèse nulle** : les résidus suivent une loi normale ;
- **Hypothèse alternative** : les résidus ne suivent pas une loi normale.

Shapiro-Wilk normality test
data: residus
W = 0,82 233, p-value = $9{,}283e^{-06}$

Le résultat du test de Shapiro-Wilk nous montre que la **statistique du test** est de 0,83 et que la **p-value = 9,283e^{-06} < 0,05**, **on rejette H_0, la série ne suit pas la loi normale.**

La **moyenne** des **erreurs** de **prédiction** est de **-424 918,6 et** nous pouvons dire que **le modèle est acceptable car la série est un bruit blanc mais peut être amélioré car il n'est pas un bruit blanc gaussien et centré (les résidus ne suivent pas la loi normale et la moyenne de ces résidus est non nulle).**

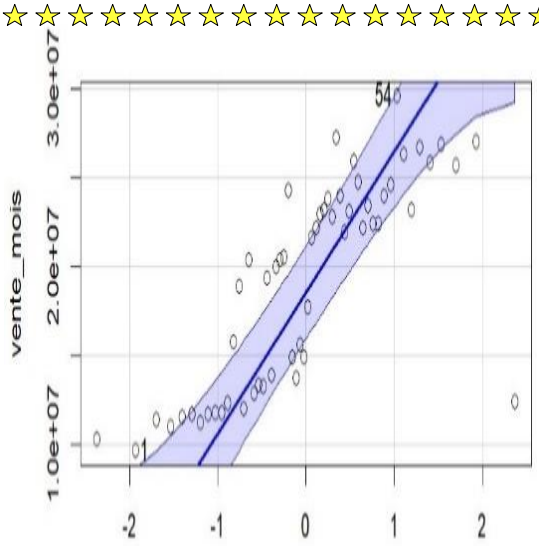


Figure 12 intervale de confiance

2) Modélisation de la série temporelle des ventes mensuelles totales

2-1) Lissage exponentiel multiplicatif (Holt-Winters Multiplicatif) de la série temporelle des ventes totales par mois

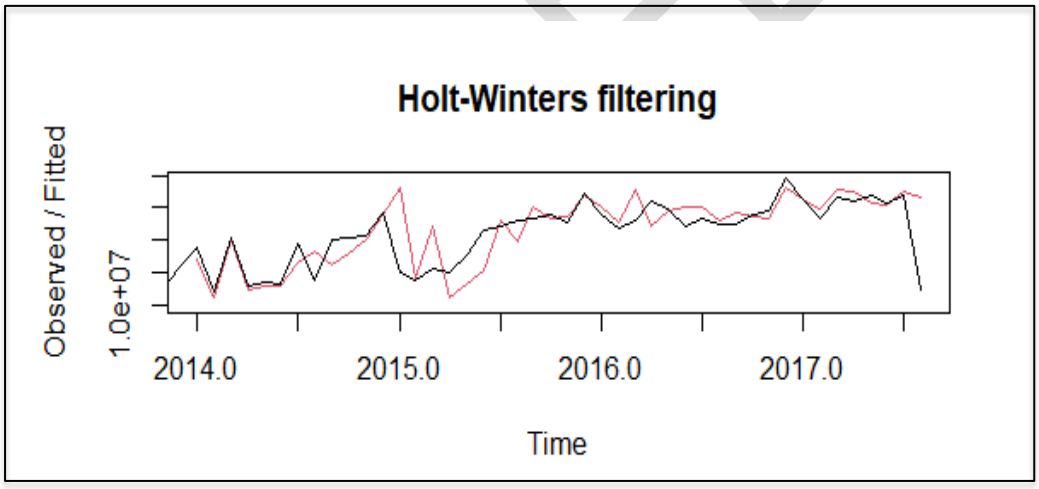


Figure 13Lissage exponentiel par la méthode de Holt-Winters

La méthode de Holt-Winters utilise les caractéristiques spécifiques de la série temporelle pour construire un modèle de prévision adapté à ses dynamiques. En observant les résultats, nous pouvons constater que le modèle de Holt-Winters s'ajuste relativement bien à la série, avec les courbes de prévision et les données réelles superposées l'une sur l'autre.

L'objectif principal ici est de réduire au maximum les erreurs de prédiction. L'écart entre les deux courbes représente l'erreur de prévision : plus cet écart est faible, plus les prévisions sont précises et fiables. À l'inverse, un écart plus important indique que les prévisions sont susceptibles d'être biaisées

2-2) Lissage exponentielle de la série temporelle des ventes totales par mois avec intervalle de confiance

✓ Tableaux statistiques de la prévision

	années	Fit	Upr	Lwr
Sep	2017	20320486	27762741	12878231
oct.	2017	21567336	29927660	13207012
nov.	2017	21419414	30606518	12232311
déc.	2017	25653572	35598959	15708185
janv.	2018	22443581	33093397	11793765
Feb	2018	20209519	31519975	8899062
Mar	2018	24212750	36147333	12278166
apr.	2018	24323908	36851563	11796252
May	2018	24522327	37616219	11428436
Jun	2018	23227290	36863927	9590654
Jul	2018	24875385	39033976	10716794
Aug	2018	18335847	32997824	3673870

Tableau 8: tableau de prevision

Fit : Les valeurs prédites ou estimées pour chaque mois.

Upr : Les bornes supérieures de l'intervalle de confiance pour les prédictions.

Lwr : Les bornes inférieures de l'intervalle de confiance pour les prédictions.

Sept-17 : La valeur prédite des ventes pour septembre 2017 est d'environ 20 320 486, avec un intervalle de confiance sont environ entre 12 878 231 et 27 762 741.

Oct-17 à Juil-18 : Les prédictions et les intervalles de confiance pour ces mois montrent des valeurs de ventes prédites avec des intervalles de confiance variés. Par exemple, les ventes prévues pour juillet 2018 sont d'environ 24 875 385, avec un intervalle de confiance entre environ 10 716 794 et 39 033 976.

Aout-18 : Pour août 2018, la valeur prédite des ventes est d'environ 18 335 847, avec un intervalle de confiance entre environ 3 673 870 et 32 997 824.Représentation graphique du lissage exponentielle

✓ **Représentation graphique de la prévision des ventes totales par mois**

Figure 14: Prévision des ventes totales par mois

Les prévisions réalisées par le modèle, ainsi que l'intervalle de confiance associé à chaque estimation. Les valeurs prédites sont représentées par la courbe verte, entourée de l'intervalle de prévision.

On observe que les ventes totales prévues pour les huit premiers mois de l'année 2018 suivent une tendance fluctuante, avec des hausses et des baisses progressives au fil des mois. Il semble donc probable que les ventes totales diminuent au mois d'août 2018.

NB : suite aux analyse et prédiction sur 1 an, nous avons essayé de pousser l'analyse un peu plus loin en tentant une prévision sur deux ans c'est-à-dire **2018 ; 2019**

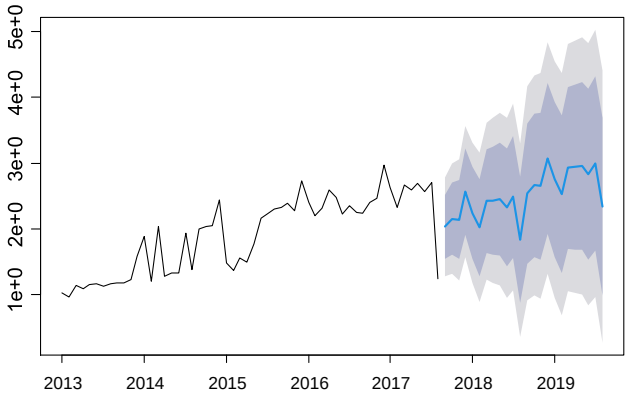


Figure 15: Prévision des ventes totales par mois

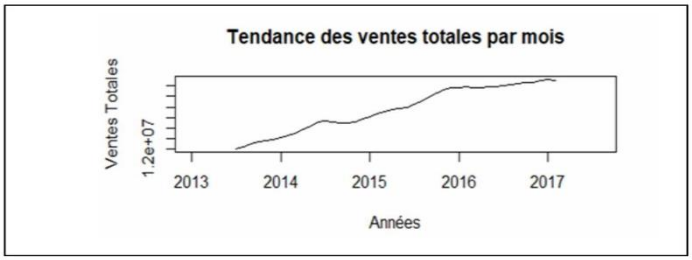
TABLEAU DE BORD

Ce tableau de bord interactif présente une analyse approfondie des indicateurs de vente de différents MAGASIN, offrant une visualisation claire des tendances et facilitant la comparaison des performances à l'échelle commerciale et une prévision pour une bonne prise de décision.

PERFORMANCES DES VENTES ET PREVISION SUR 12 MOIS

MOYENNE VENTE PAR MOIS
19172231

janvier	février	mars	avril	mai	juin
juillet	août	septembre	octobre	novembre	décembre



janv. 2013	févr. 2013	mars 2013	avr. 2013	mai 2013	juin 2013	juil. 2013	août 2013
10327625	9658960	11428497	10993465	11597704	11689344	11257401	11737789
sept. 2013	oct. 2013	nov. 2013	déc. 2013	janv. 2014	févr. 2014	mars 2014	avr. 2014
11792933	11775620	12356559	15803117	18911641	12038353	20365584	12861251
mai 2014	juin 2014	juil. 2014	août 2014	sept. 2014	oct. 2014	nov. 2014	déc. 2014
13379785	13319958	19421891	13885176	20022416	20396101	20531635	24340454
janv. 2015	févr. 2015	mars 2015	avr. 2015	mai 2015	juin 2015	juil. 2015	août 2015
14896922	13742396	15598608	14955068	17730368	21615360	22209619	22963674
sept. 2015	oct. 2015	nov. 2015	déc. 2015	janv. 2016	févr. 2016	mars 2016	avr. 2016
23240882	23878268	22804953	27243982	23977805	21947409	23131781	25963025
mai 2016	juin 2016	juil. 2016	août 2016	sept. 2016	oct. 2016	nov. 2016	déc. 2016
24779432	22209219	23462672	22452414	22417448	24030390	24642640	29640288
janv. 2017	févr. 2017	mars 2017	avr. 2017	mai 2017	juin 2017	juil. 2017	août 2017
26328160	23250112	26704018	25895308	26911847	25682822	27011478	12433323

Conclusion

L'objectif principal de cette analyse était de prédire les ventes mensuelles de l'épicerie "Favorita", une grande chaîne de supermarchés en Équateur, afin de mieux gérer les stocks et optimiser les achats. Ce défi s'inscrit dans le cadre d'une gestion délicate où les détaillants doivent anticiper la demande sans générer de stocks excédentaires ni de ruptures de stock. En utilisant les données historiques des ventes, comprenant des informations sur les produits, les magasins et les promotions, plusieurs techniques ont été explorées pour établir des prévisions fiables sur les ventes futures.

Les résultats obtenus ont montré que les ventes totales de l'épicerie "Favorita" suivent une tendance saisonnière claire, avec des variations mensuelles importantes. Les prévisions, réalisées à l'aide de modèles comme Holt-Winters, ont permis d'identifier des tendances pour l'année suivante, notamment des augmentations et des baisses régulières, telles que des baisses anticipées des ventes en août 2018, cohérentes avec les schémas saisonniers passés. L'analyse des résidus a également révélé une bonne adéquation entre les prévisions et les données réelles, tout en mettant en évidence les marges d'erreur qui restent significatives dans certaines périodes.

À partir de ces résultats, plusieurs recommandations ont été proposées. **Il est crucial d'optimiser la gestion des stocks en ajustant les niveaux de stock selon les prévisions de vente mensuelles. La planification des promotions doit être affinée pour coïncider avec les pics de demande identifiés, et une distribution plus fine des produits entre les différents magasins permettrait de mieux répondre à la demande locale**. Ces actions pourraient réduire les risques liés à la gestion des stocks, notamment l'excédent de produits périssables et les ruptures de stock.

Cependant, cette analyse présente certaines limites. La variabilité des comportements de vente entre les différentes catégories de produits et magasins complique l'élaboration de prévisions précises. De plus, les modèles utilisés ne prennent pas en compte les facteurs externes comme les crises économiques ou les événements mondiaux, qui peuvent influencer de manière significative les comportements d'achat. La complexité des données, notamment les variables catégorielles comme les familles de produits et les magasins, peut également rendre les résultats difficiles à interpréter et à ajuster.

Pour améliorer la précision des prévisions, plusieurs pistes peuvent être explorées. L'intégration de facteurs externes, tels que l'évolution de l'économie, ou des techniques de modélisation avancées, comme les réseaux neuronaux LSTM pour mieux capturer les dépendances temporelles complexes, pourrait enrichir les résultats. Par ailleurs, une analyse plus détaillée des effets des promotions, à l'échelle des différents produits ou familles de produits, permettrait de mieux comprendre leur impact sur les ventes. Enfin, une approche plus granularisée des prévisions, segmentée par magasin, pourrait offrir des insights plus précis et améliorer la gestion locale des stocks.

ANNEXE

Bibliographie

Akposso, **D**. (2024). *Séries chronologiques* : *Support de cours*. Institut Supérieur de Statistique d'Econométrie et de Data Science.

Webographie

https://www.youtube.com/watch?v=RWtnLTFeFwU&list=PLmJWMf9F8euR8QSxkM5mtmUcJnEaxkI5D