MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE RECHERCHE SCIENTIFIQUE

REPUBLIQUE DE COTE D'IVOIRE

UNION-DISCIPLINE-TRAVAIL

UNION - DISCIPLINE - TRAVAIL

MASTER 2

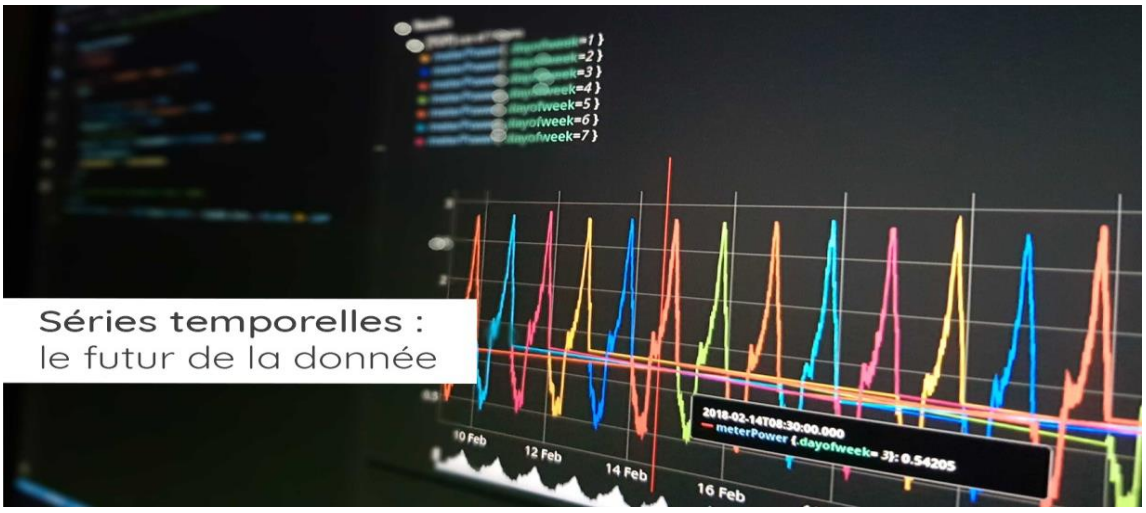
STATISTIQUE-ECONOMETRIE-DATA SCIENCE

MINI-PROJET

ANALYSE DESCRIPTIVE DE SERIE TEMPORELLE ET PREVISION PAR LISSAGE EXPONENTIEL

PREDIRE LES VENTES DES MILLIERS DE FFAMILLES DE PRODUITS VENDUS DANS LES MAGASINS DE L'EPICERIE "FAVORITA" SITUES EN EQUATEUR

2023-2024



ETUDIANTE :

WAWA LAURALIE MARIE-LICHELLE

ENSEIGNANT ENCADREUR :

AKPOSSO DIDIER MARTIAL

AVANT PROPOS

Cher lecteur,

C’est avec un grand enthousiasme que je vous présente ce mini-projet, fruit de Plusieurs semaines de réflexion, de recherche et de travail acharné. Ce mini-Projet est le résultat de mon engagement envers l’apprentissage, la créativité et la mise en pratique des connaissances acquises.

Ce mini-projet est de démontrer ma capacité à appliquer les compétences et les concepts que j’ai acquis dans le cadre de l’obtention du diplôme de Master Professionnel en Statistique-Econométrie- Machine Learning à l’institut Supérieur de Statistique d’Econométrie et de Data Science (INSSEDS). Le sujet abordé dans le cadre de ce mini-projet me semble à la fois stimulant et pertinent.

L'épicerie est un type de commerce de détail spécialisé dans la vente au détail de produits alimentaires et d'autres articles ménagers de base. L'objectif principal de ce projet est de développer des modèles prédictifs robustes pour estimer les ventes futures des différentes familles de produits. Ces modèles pourraient aider les gestionnaires de magasins à prendre des décisions éclairées en matière de gestion des stocks, de planification de promotions et de prévisions de revenus.

Je tiens à remercier M. AKPOSSO Didier Martial, Directeur des études et Encadreur de ladit formation, les enseignants de l’INSSEDS, ma famille et mes amis qui m’ont soutenu tout au long de la rédaction de ce mini-projet en m’apportant leur expertise, leurs conseils et leur encouragement.

Enfin, je vous invite à parcourir ce document avec attention, en espérant qu’il vous apportera une vision claire et détaillé de mon travail, tout en suscitant votre intérêt et votre réflexion.

Merci de prendre le temps de découvrir ce mini-projet. Je vous souhaite une agréable lecture.

Cordialement

Contenu

INTRODUCTION GENERALE

L'industrie de la vente au détail, en particulier dans le secteur de l'épicerie, est un domaine dynamique et complexe où la compréhension des tendances de consommation et des facteurs influençant les ventes est cruciale. Les données fournies ici présentent une opportunité passionnante d'explorer les ventes de milliers de familles de produits dans les magasins d'épicerie de la chaîne "Favorita", situés en Équateur.

Ce jeu de données, extrait du site Kaggle, offre un aperçu détaillé des transactions de vente, comprenant des informations sur la date, le magasin, le type de produit et même la promotion éventuelle sur les produits. Les données sur les ventes peuvent être fractionnaires, reflétant la réalité du commerce où les produits sont souvent vendus en quantités variables.

L'objectif principal de cette analyse est de prédire les ventes futures en fonction de diverses variables, telles que la date, le magasin, le type de produit et la promotion. En comprenant ces modèles de vente, les détaillants peuvent mieux anticiper la demande, ajuster leurs stocks et leurs stratégies de marketing, et ainsi optimiser leurs performances commerciales.

Dans cette étude, nous explorerons ces données pour identifier les tendances, les corrélations et les modèles prédictifs qui pourraient aider "Favorita" à améliorer ses opérations et à mieux servir sa clientèle. En combinant des techniques d'analyse de données avancées avec une compréhension approfondie du marché de détail, nous chercherons à fournir des recommandations précieuses pour une prise de décision stratégique éclairée.

Le présent document s’articulera autour de trois grandes parties :

➢ **PREMIERE PARTIE: IMPORTATION ET ORGANISATION DES VENTES TOTALES PAR MOIS**

➢ **DEUXIEME PARTIE: ANALYSE DESCRIPTIVE DES VENTES**

➢ **TROISIEME PARTIE: PREVISION POUR LES 12 PROCHAINS MOIS PAR LA METHODE DE HOLT WINTER**

CONTEXTE ET JUSTIFICATION DE L'ETUDE

L'étude des ventes des familles de produits dans les magasins de l'épicerie "Favorita" en Équateur permet d'optimiser les opérations, de prévoir la demande, de maximiser l'efficacité des promotions, de mieux comprendre la clientèle et de prendre des décisions stratégiques éclairées. Elle constitue ainsi un outil précieux pour améliorer la performance globale de l'entreprise et répondre aux besoins changeants du marché.

PROBLEMATIQUE

Comment optimiser les ventes et les opérations des magasins d'épicerie "Favorita" en Équateur ?

PRINCIPAUX RESULTATS ATTENDUS

Les principaux résultats attendus de cette étude est une compréhension approfondie pour fournir des informations précieuses pour aider les magasins d'épicerie "Favorita" à améliorer leurs performances commerciales, à mieux répondre aux besoins de leurs clients et à rester compétitifs sur le marché de la vente au détail en Équateur.

PREMIERE PARTIE : PRETAITEMENT DES DONNEES ET ORGANISATION DES VENTES TOTAALES PAR MOIS

I-PRETAITEMENT DES DONNEES

Le prétraitement des données est une étape cruciale dans le processus d'analyse des données, visant à préparer les données brutes en vue de les rendre appropriées pour une analyse ultérieure. Cette partie consiste, dans un premier temps, à importer le jeu de données dans un logiciel statistique. Ensuite, on procèdera à la visualisation du jeu de données afin d'en dégager la structure. Enfin, la dernière étape consistera à apurer le jeu de données. L'apurement de données, souvent appelé "data cleansing" en anglais, est le processus de nettoyage et de correction des données stockées dans une base de données ou un ensemble de données. L'objectif de l'apurement de données est d'identifier et de corriger les incohérences, les erreurs, les doublons et les données obsolètes ou incorrectes afin d'assurer la qualité et la précision des données.

1-DESCRIPTION DU JEU DE DONNEE : DICTIONNAIRE DES DONNEES

Le dictionnaire de données est une documentation détaillée qui répertorie et décrit chacune des variables du jeu de données. Il sert de référence pour comprendre la signification, la structure et les propriétés des données stockées, ce qui facilite la gestion, la maintenance, l'analyse et l'utilisation des données. Le jeu de données de notre étude se décrit comme suit :

Tableau 1: Dictionnaire de données

VARIABLE	NATURE	DESCRIPTION	MODALITES
Id	Quantitative	Identifiants	Numérique
Date	Quantitative	Date à laquelle les ventes ont été enregistrées	Date
Store_nbr	Quantitative	Identification du magasin dans lequel les produits sont vendus.	Numérique
Family	Qualitative	Identification du type de produit vendu.	Chaine de caractère
Sales	Quantitative	Le total des ventes d'une famille de produits dans un magasin particulier à une date donnée.	Numérique
onpromotion	Quantitative	Le nombre total d'articles d'une famille de produits qui étaient en promotion dans un magasin	Numérique

2-Structure des données

Notre jeu de données est constitué de 3000888 observations et de 6 variables.

Tableau 2: Structure du jeu de donnée

```
'data.frame':   3000888 obs. of  6 variables:
 $ id         : int  0 1 2 3 4 5 6 7 8 9 ...
 $ 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 ...
```

3-Visualisation du jeu de donnée

Le tableau ci-dessous présente les cinq (5) premières et dernières observations de notre jeu de données.

Tableau 3: Visualisation des 5 premières des observations

	id	date	store_nbr	family	sales	onpromotion
1	0	2013-01-01	1	AUTOMOTIVE	0	0
2	1	2013-01-01	1	BABY CARE	0	0
3	2	2013-01-01	1	BEAUTY	0	0
4	3	2013-01-01	1	BEVERAGES	0	0
5	4	2013-01-01	1	BOOKS	0	0

Tableau 4: Visualisation des 5 dernières observations

	id	date	store_nbr	family	sales	onpromotion
3000884	3000883	2017-08-15	9	POULTRY	438.133	0
3000885	3000884	2017-08-15	9	PREPARED FOODS	154.553	1
3000886	3000885	2017-08-15	9	PRODUCE	2419.729	148
3000887	3000886	2017-08-15	9	SCHOOL AND OFFICE SUPPLIES	121.000	8
3000888	3000887	2017-08-15	9	SEAFOOD	16.000	0

II-ORGANISATION DES VENTES TOTALES PAR MOIS

L'organisation des ventes totales par mois consiste à agréger les données de ventes en fonction de la date et à regrouper ces données par mois. Cette démarche permet d'obtenir une vue d'ensemble des ventes réalisées chaque mois, ce qui facilite l'analyse des tendances de vente sur une période donnée.

Tableau 5: Répartition des ventes par mois

	mois	ventes_totales
	<date>	<dbl>
1	2013-01-01	10327625.
2	2013-02-01	9658960.
3	2013-03-01	11428497.
4	2013-04-01	10993465.
5	2013-05-01	11597704.
6	2013-06-01	11689344.
7	2013-07-01	11257401.
8	2013-08-01	11737789.
9	2013-09-01	11792933.
10	2013-10-01	11775620.

Les ventes semblent fluctuer légèrement d'un mois à l'autre, mais globalement, elles maintiennent une tendance générale à la hausse au fil de l'année. De plus Les mois de mai, juin et août semblent être des mois où les ventes atteignent des niveaux relativement élevés, dépassant les 11 millions.

DEUXIEME PARTIE: ANALYSE DESCRIPTIVE DES VENTES

L'analyse descriptive des ventes offre une vision globale des performances commerciales d'une entreprise en examinant les tendances temporelles, la répartition des ventes par catégorie et par emplacement, l'impact des promotions, et en identifiant les tendances saisonnières ou cycliques. Ces informations sont essentielles pour prendre des décisions stratégiques informées visant à optimiser les performances commerciales et à répondre efficacement à la demande du marché.

I-VISUALISATION DU JEU DE DONNEE DES SERIES TEMPORELLES

1-Transformation du jeu de donnée en série temporelle

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
2013	10327625	9658960	11428497	10993465	11597704	11689344	11257401	11737789	11792933	11775620
2014	18911641	12038353	20365584	12861251	13379785	13319958	19421891	13885176	20022416	20396101
2015	14896922	13742396	15598608	14955068	17730368	21615360	22209619	22963674	23240882	23878268
2016	23977805	21947409	23131781	25963025	24779432	22209219	23462672	22452414	22417448	24030390
2017	26328160	23250112	26704018	25895308	26911847	25682822	27011478	12433323		

	Nov	Dec
2013	12356559	15803117
2014	20531635	24340454
2015	22804953	27243982
2016	24642640	29640288
2017		

Les données montrent une évolution des ventes au fil des mois et des années. Il est clair qu'il y a des variations saisonnières dans les ventes, avec des mois de pointe et des mois plus calmes. Pour chaque année, les mois de novembre et décembre semblent souvent être des périodes de forte activité commerciale, ce qui est cohérent avec les périodes de fêtes de fin d'année.

2- Test de stationnarité

Un test de stationnarité est une procédure statistique utilisée pour déterminer si une série chronologique est stationnaire ou non. Ces tests sont utiles pour évaluer si une série chronologique satisfait aux critères de stationnarité, tels que la constance de la moyenne, de la variance et de l'autocorrélation.

Il existe plusieurs tests de stationnarité couramment utilisés, parmi lesquels les plus populaires sont :

Test de Dickey-Fuller augmenté (ADF) : Ce test évalue l'hypothèse nulle selon laquelle une série chronologique possède une racine unitaire, ce qui signifie qu'elle n'est pas stationnaire. Si la valeur p calculée est inférieure à un certain seuil (généralement 0.05), on rejette l'hypothèse nulle et on conclut que la série chronologique est stationnaire.

Test de Phillips-Perron (PP) : Ce test est similaire au test ADF mais utilise une méthode différente pour corriger l'autocorrélation et l'hétéroscédasticité dans les résidus. Il est également largement utilisé pour tester la stationnarité des séries chronologiques.

Test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) : Contrairement aux tests ADF et PP, le test KPSS évalue l'hypothèse nulle selon laquelle une série chronologique est stationnaire. Si la valeur p calculée est supérieure à un certain seuil (généralement 0.05), on rejette l'hypothèse nulle et on conclut que la série chronologique n'est pas stationnaire.

- Testons avec le test de Dickey-Fuller Augmenté (ADF)

Tableau 6: Test de Dickey-Fuller Augmennté (ADF)

```
        Augmented Dickey-Fuller Test

data:  serie_temporelle
Dickey-Fuller = -2.4817, Lag order = 3, p-value = 0.3803
alternative hypothesis: stationary
```

Dans notre cas, la statistique de test est -2.4817. Sans connaître les valeurs critiques spécifiques pour ce test, nous ne pouvons pas dire avec certitude si la série est stationnaire ou non. Cependant, la valeur p associée au test est de 0.3803, ce qui est supérieur au seuil couramment utilisé de 0.05. Cela signifie que nous ne pouvons pas rejeter l'hypothèse nulle selon laquelle la série n'est pas stationnaire.

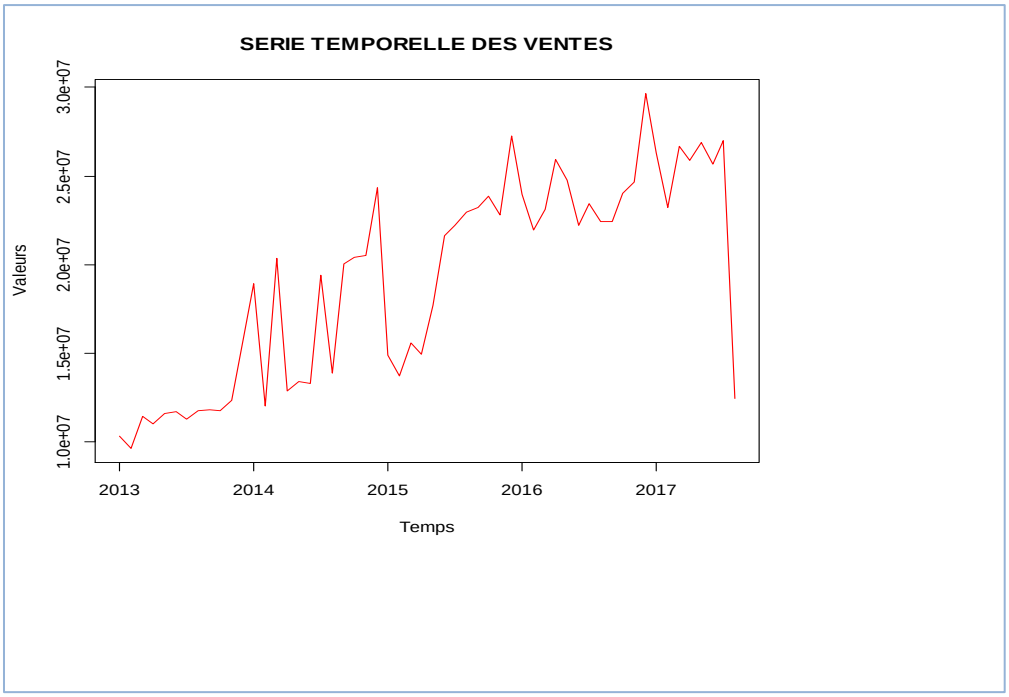
II-INDICES DESCRIPTIFS D'UNE SERIE TEMPORELLE

L'indice descriptif d'une série temporelle est une mesure synthétique utilisée pour résumer les caractéristiques principales de la série. Il vise à fournir une vision globale et concise de l'évolution et du comportement de la série temporelle au fil du temps.

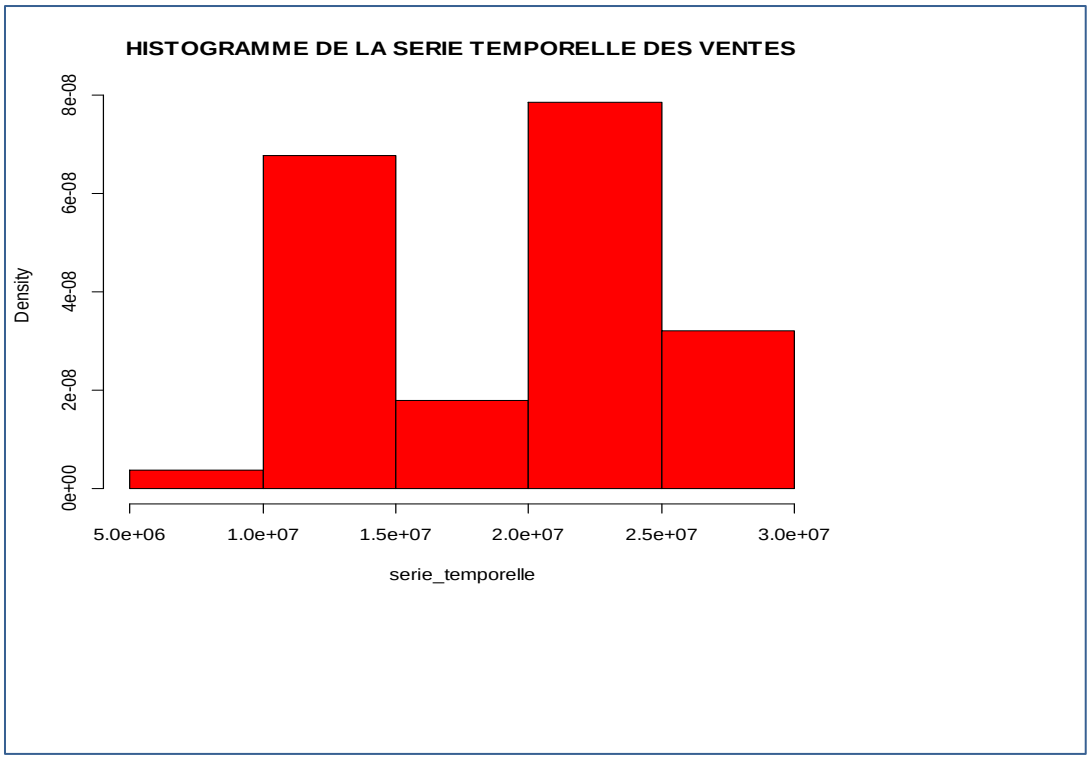
1-Visualisation temporelle des ventes

Une visualisation temporelle des ventes est une représentation graphique qui illustre l'évolution des ventes au fil du temps. Elle permet de visualiser visuellement les tendances, les schémas saisonniers, les fluctuations et les anomalies dans les données de vente sur une période donnée.

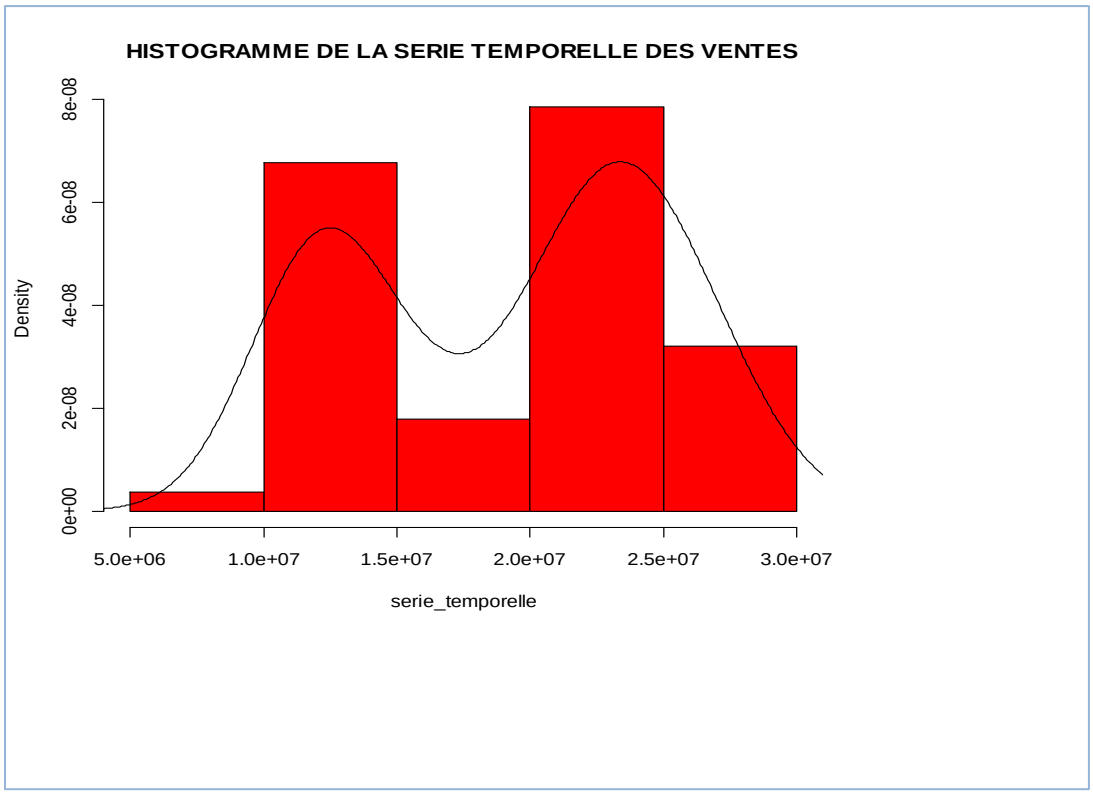
Graphique1: Série temporelle des ventes



Gaphique2: Histogramme de la série temporelle des ventes

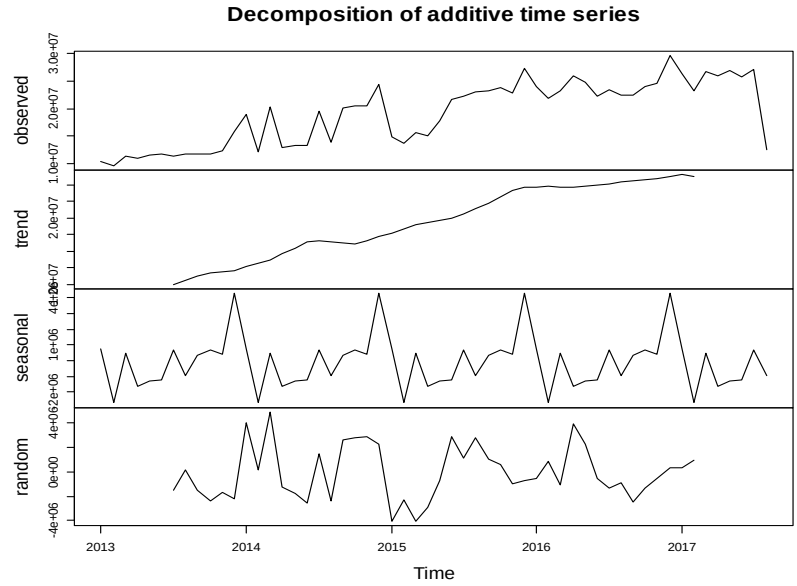


Graphique3: Histogramme de la distribution des ventes



2-Décomposition des ventes de la série temporelle

Graphique 4: Décompositions des ventes



3-Rappel du fondement théoriques des indicateurs de statistique

- **Indicateur de tendance centrale de position**

Les indicateurs de position centrale sont des mesures statistiques qui permettent de résumer la tendance centrale d'un ensemble de données. Les indicateurs de tendance centrale couramment utilisés sont les suivants :

➢ MINIMUM : la plus petite valeur des observations

➢ MAXIMUM : la plus grande valeur des observations

➢ MODE : Le mode est la valeur de la variable statistique qui a le plus grand effectif ou la plus grande fréquence (il peut y avoir plusieurs modes).

➢ MEDIANE : La médiane d'une série statistique est le nombre qui partage cette série en deux séries de même effectif. La moitié des effectifs (50 %) a donc une valeur du caractère en dessous de la valeur médiane et l'autre moitié (50 %) au-dessus.

Q1 : Le quartile Q1, la médiane Me et le quartile Q3 partagent les valeurs ordonnées de la série en quatre parties égales. Q1 est la plus petite donnée de la série pour laquelle au moins 25 % des données (soit 1/4 des données) sont égales ou inférieures à Q1 (Q1 est obligatoirement une donnée de la série).

Q2 : Q2 correspond à la médiane.

Q3 : Q3 est la plus petite donnée de la série pour laquelle au moins 75 % des données (soit 3/4
des données) sont égales ou inférieures à Q3 (Q3 est obligatoirement une donnée de la série).

➢ MOYENNE: c'est la valeur moyenne de la série statistique. Ces indicateurs de position centrale sont utilisés pour résumer et comprendre la distribution des données. Ils offrent des informations sur la concentration des valeurs autour d'une valeur centrale et aident à interpréter les tendances générales des données. Il est souvent recommandé d'utiliser plusieurs indicateurs de position centrale ensemble pour obtenir une vue complète de la tendance centrale des données, car chaque mesure peut avoir ses propres forces et limitations en fonction de la distribution des données.

- **Indicateur de dispersion**

  Les indicateurs de dispersion sont des mesures statistiques qui permettent de quantifier la variabilité ou la dispersion des données autour d'une mesure centrale et aident à comprendre la dispersion des données. Voici quelques indicateurs de dispersion couramment utilisés :

  ➢ VARIANCE: La variance est une mesure de dispersion qui quantifie la moyenne des carrés des écarts par rapport à la moyenne. Elle est calculée en prenant la moyenne des carrés des écarts entre chaque valeur et la moyenne. Une variance élevée indique une dispersion plus importante des données.

  ➢ ECART-TYPE: L'écart-type est une mesure de dispersion qui indique dans quelle mesure les valeurs d'un ensemble de données sont dispersées autour de la moyenne. Plus l'écart-type est élevé, plus les valeurs sont dispersées. Il est calculé en prenant la racine carrée de la variance.

  ➢ COEFFICIENT DE VARIATION: Généralement exprimé en %, Il est notamment utile dans le cas où l'on souhaite comparer deux groupes d'observations. Un coefficient de variation faible indique une distribution homogène Un cv élevé indique une distribution hétérogène.

  ➢ECART INTERQUARTILE: L'écart interquartile est la différence entre le troisième quartile (Q3) et le premier quartile (Q1) d'un ensemble de données. Il représente la dispersion des valeurs autour de la médiane et est moins sensible aux valeurs aberrantes que l'écart-type.

  Ces indicateurs de dispersion sont essentiels pour comprendre la variabilité des données et pour évaluer la cohérence ou la fiabilité des mesures centrales comme la moyenne ou la médiane. Ils permettent de quantifier la dispersion des valeurs et d'identifier les points atypiques ou les tendances importantes dans un ensemble de données.

- **Indicateurs de forme**

  Les indicateurs de forme sont des mesures statistiques qui fournissent des informations sur la distribution ou la forme d'un ensemble de données. Voici quelques indicateurs de forme couramment utilisés :

➢SKEWNESS : La skewness mesure l'asymétrie de la distribution des données par rapport à la moyenne. Une skewness positive indique une distribution asymétrique avec une queue plus longue du côté droit de la moyenne, tandis qu'une skewness négative indique une distribution asymétrique avec une queue plus longue du côté gauche de la moyenne.

➢ KURTOSIS(APLATISSEMENT) : La kurtosis mesure la forme de la distribution des données par rapport à une distribution normale. Une kurtosis élevée indique une distribution plus concentrée avec des queues plus épaisse (leptokurtique), tandis qu'une kurtosis faible indique une distribution plus aplatie avec des queues plus minces (platykurtique).

Ces indicateurs de forme sont utilisés pour caractériser la distribution des données et fournir des informations sur la symétrie et l'aplatissement e. Ils aident à comprendre les propriétés des données et peuvent être utiles pour prendre des décisions statistiques appropriées, en particulier lors de l'analyse des données financières, économiques ou biologique.

Tableau 7: Résumé numérique

INDICATEURS	VALEURS	INTERPREETATION
INDICATEURS DE TENDANCE CENTRALE ET DE POSITION		
MINIMUM	9658960	La valeur minimale des ventes est de 9658960
MAXIMUM	29640288	La valeur maximale des ventes est de 29640288
MODE	23408911	La majorité des produits vendus dans les magasins de l'épicerie ont connu une vente de 23408911
MOYENNE	19172231	La vente moyenne est de 19172231
1ER QUARTILE Q1	13205281	25% des produits vendus dans les magasins de l'épicerie ont connu des ventes maximum de 13205281
2Eme QUARTILE Q2	20463868	La moitié des produits vendus dans les magasins de l'épicerie ont connu des ventes maximum de 20463868
3ème QUARTILE Q3	23903152	75% des produits vendus dans les magasins de l'épicerie ont connu des ventes maximum de 23903152
INDICATEURS DE DISPERSION		
VARIANCE	3.372074e+13	On interprètera l'écart type qui est la racine carrée de la variance
ECART TYPE	5806956	La vente moyenne est de 19172231. L'écart-type de 5806956 indique que les ventes sont dispersées et en moyenne, elles varient entre [19172231-

		5806956 ; 19172231+5806956] soit [13365275 ;24979187]
COEFFICIENT DE VARIATION	30.28837	Le coefficient de variation des ventes est de 30.28837
INDICATEUR DE FORME		
SKEWNESS	-0.1546965	Le Skewness étant négative, cela indique que la distribution des ventes est étalée à gauche
KURTOSIS	1.573615	Le Kurtosis est faible, cela indique que la distribution des ventes et plus aplatie avec des queues plus minces (platikurtique)

4-Indices de dépendance

Un indice de dépendance est une mesure statistique utilisée pour quantifier et évaluer la force ou la nature de la relation entre deux variables. Il est également connu sous le nom de coefficient de corrélation ou de coefficient de dépendance. Les indices de dépendance permettent de déterminer dans quelle mesure les variations dans une variable sont associées aux variations dans une autre variable.

- **Autocorrélation simple**
  L'autocorrélation simple, également appelée autocorrélation à retard d'une série temporelle, mesure la corrélation entre les valeurs de la série temporelle à différents moments dans le temps. Plus précisément, elle mesure la corrélation entre une observation et une autre observation à un certain nombre de périodes de décalage (**lag**) précédentes ou suivantes.
  Par exemple, pour une série temporelle ***Yt***, l'autocorrélation à un retard (**lag**) ***k*** est calculée comme la corrélation entre ***Yt*** et ***Yt−k.*** Si cette autocorrélation est significativement différente de zéro, cela indique que les observations sont corrélées avec elles-mêmes à un retard de ***k*** périodes.

Tableau 8: Autocorrélation simple

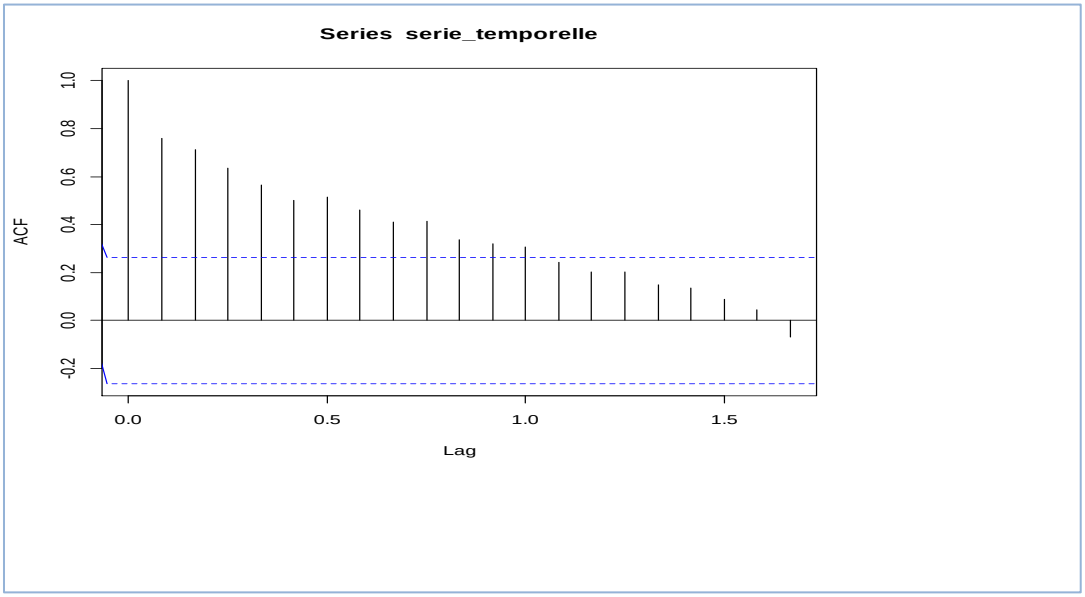
```
Autocorrelations of series 'serie_temporelle', by lag

0.0000 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000
 1.000  0.759  0.714  0.634  0.566  0.503  0.514  0.460  0.409  0.413  0.337  0.319  0.308
1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667
 0.242  0.202  0.203  0.149  0.137  0.089  0.044 -0.070
```

Par exemple :

- À un retard de 0.0833, l'autocorrélation est de 0.759, ce qui suggère une corrélation positive significative entre les observations décalées d'un mois.

- À un retard de 1.6667, l'autocorrélation est de -0.070, ce qui suggère une corrélation négative mais faible entre les observations décalées de 20 mois.

Graphique 5: Autocorrélation simple



On constate que les autocorrélations semblent diminuer à mesure que le retard augmente, ce qui est cohérent avec une décroissance de la corrélation entre les observations à mesure qu'elles sont plus éloignées dans le temps.

- **Autocorrélation partielle**
  L'autocorrélation partielle mesure la corrélation entre deux observations d'une série temporelle après avoir éliminé les effets des observations intermédiaires. Autrement dit, elle mesure la corrélation entre deux points de données une fois que l'influence des autres points de données situés entre eux a été contrôlée.

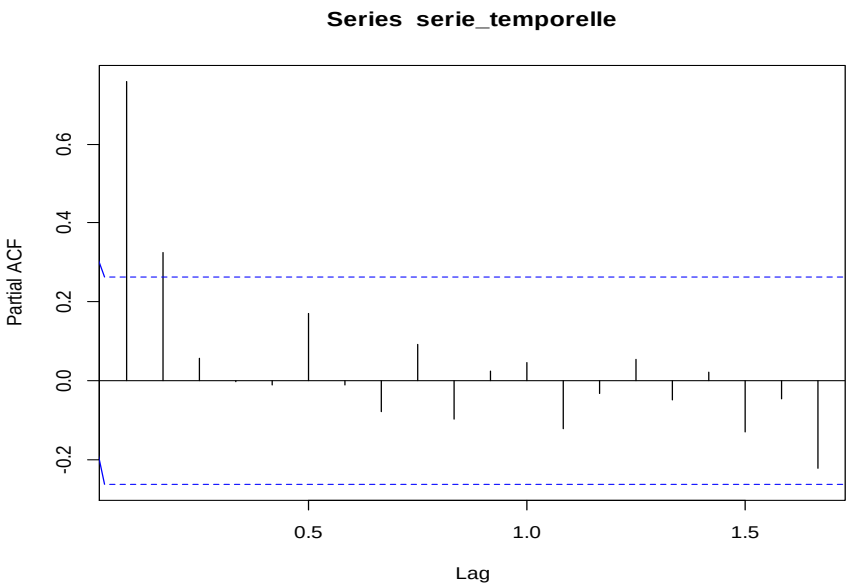
Tableau 9: Autocorrélation partielle

```
Partial autocorrelations of series 'serie_temporelle', by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833
 0.759  0.326  0.057 -0.003 -0.012  0.172 -0.012 -0.079  0.092 -0.098  0.024  0.045 -0.123
1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667
-0.033  0.054 -0.050  0.023 -0.130 -0.045 -0.223
```

Par exemple :

- À un retard de 0.0833, l'autocorrélation partielle est de 0.759. Cela signifie qu'après avoir contrôlé les effets des autres retards, il y a une corrélation positive significative entre les observations décalées d'un mois.
- À un retard de 0.5833, l'autocorrélation partielle est de -0.012. Cela indique qu'après avoir contrôlé les effets des autres retards, il y a une corrélation faible et négative entre les observations décalées de 7 mois.

Graphique 6: Autocorrélation partielle



TROISIEME PARTIE: PREVISION POUR LES 12 PROCHAINS MOIS PAR LA METHODE DE HOLT WINTER

La méthode de **Holt-Winter** est une technique couramment utilisée pour la prévision des séries temporelles, en particulier lorsque les données présentent à la fois une tendance et une saisonnalité.

I-PREVISION DE LA SERIE TEMPORELLE AVEC INTERVALLE DE CONFIANCE

Le lissage simple est une méthode de prévision simple et intuitive pour les séries temporelles, mais elle ne prend pas en compte la saisonnalité ou d'autres structures complexes des données. Les intervalles de confiance fournissent une mesure de l'incertitude associée aux prévisions et aident à évaluer la fiabilité des prévisions du modèle de lissage simple.

1-Choix des paramètres

Le choix des paramètres dans la modélisation des séries temporelles est essentiel pour obtenir des prévisions précises et fiables. Dans les méthodes de lissage, comme le lissage simple ou le lissage exponentiel, le paramètre de lissage contrôle la pondération des observations passées dans la génération des prévisions. La formule est la suivante :

$$\text{Prévision} = \alpha \times \text{Observation} + (1-\alpha) \times \text{Prévision précédente}$$

Ou α est le paramètre de lissage.

Un paramètre de lissage plus élevé donne plus de poids aux observations les plus récentes, tandis qu'un paramètre de lissage plus faible donne plus de poids à l'ensemble des données historiques.

Tableau 10: Choix de alpha et beta

Colonne1
Holt-Winters exponential smoothing with trend and additive seasonal component.
Call:
HoltWinters(x = serie_temporelle)
Smoothing parameters:
alpha: 0.5117962
beta : 0
gamma: 0.8780483
Coefficients:
[,1]
a 20518833.5
b 421794.1
s1 -620141.8
s2 204914.1
s3 -364801.4
s4 3447562.4
s5 -184223.4

s6	-2840079.7
s7	741357.2
s8	430721.1
s9	207346.7
s10	-1509484.4
s11	-283183.9
s12	-7244515.9

- Le paramètre alpha (α) contrôle le lissage de la tendance. Dans ce cas, alpha est estimé à environ 0.512, ce qui signifie qu'environ 51,2% du poids est attribué à la dernière observation pour la prédiction de la tendance.
- Le paramètre beta (β) contrôle le lissage de la composante de la tendance. Ici, beta est estimée à 0, ce qui suggère qu'il n'y a pas de lissage de la tendance saisonnière.
- Le paramètre gamma (γ) contrôle le lissage de la saisonnalité. Dans ce cas, gamma est estimé à environ 0.878, ce qui indique une forte pondération sur les données saisonnières pour la prédiction de la saisonnalité.

2-Lissage exponentielle simple sans intervalle de confiance

Tableau 11: Lissage exponentielle simple sans intervalle de confiance

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
2017									20320486	21567336
2018	22443581	20209519	24212750	24323908	24522327	23227290	24875385	18335847		
	Nov	Dec								
2017	21419414	25653572								
2018										

Pour l'année 2017 : Les ventes semblent être relativement élevées, atteignant un pic en décembre avec un montant de 25 653 572.

Pour l'année 2018 : Les ventes semblent augmenter progressivement de janvier à avril, atteignant un pic en avril avec un montant de 24 323 908. Ensuite, il y a une légère baisse des ventes en mai et juin, suivie d'une fluctuation mais globalement des ventes stables jusqu'en novembre.

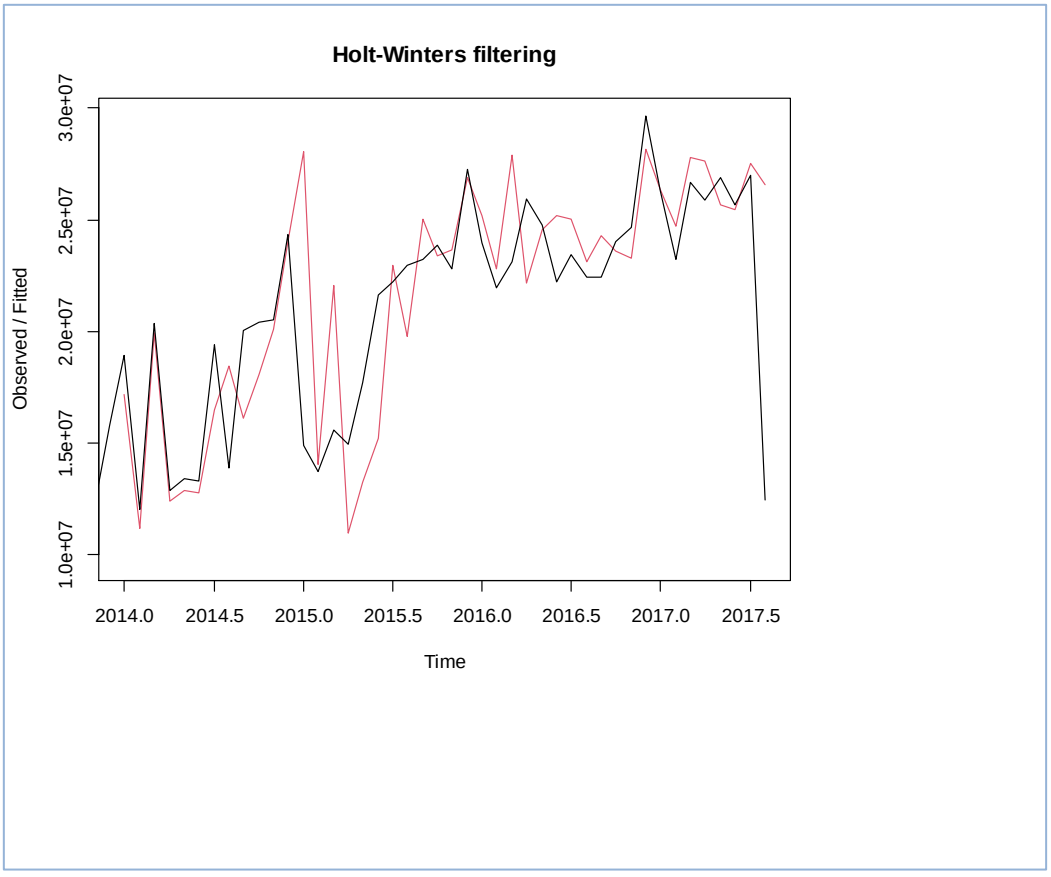
3-Lissage exponentielle simple avec intervalle de confiance

Tableau 12: Avec intervalle de confiance1

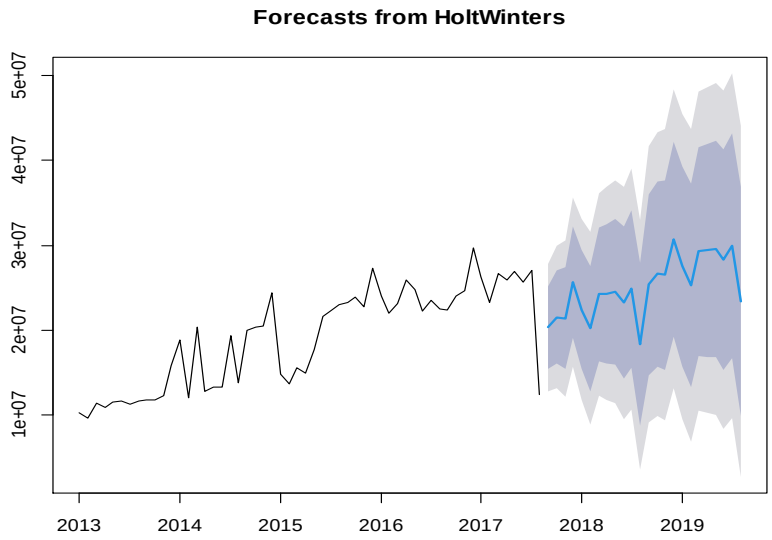
	fit	upr	lwr
Sep 2017	20320486	27762741	12878231
Oct 2017	21567336	29927660	13207012
Nov 2017	21419414	30606518	12232311
Dec 2017	25653572	35598959	15708185
Jan 2018	22443581	33093397	11793765
Feb 2018	20209519	31519975	8899062
Mar 2018	24212750	36147333	12278166
Apr 2018	24323908	36851563	11796252
May 2018	24522327	37616219	11428436
Jun 2018	23227290	36863927	9590654
Jul 2018	24875385	39033976	10716794
Aug 2018	18335847	32997824	3673870

Ces données fournissent des informations sur les prévisions de ventes pour chaque mois, ainsi que des intervalles de confiance associés.

Graphique 7: Visualisation de la prévision



Graphique 8: Visualisation de la prévision



II-VALIDATION DU MODELE DE LISSAGE PAR L'ANALYSE DU RESIDU

La validation du modèle de lissage par l'analyse des résidus est une étape essentielle pour s'assurer de la précision et de la fiabilité des prévisions produites par le modèle. Cela permet de détecter les violations des hypothèses du modèle et d'identifier les ajustements potentiels nécessaires pour améliorer les performances du modèle de lissage.

1-Bruit blanc

Un processus de bruit blanc est une suite de variables aléatoires ($\boldsymbol{Xt}$), d'espérance et de variance constante (c'est à dire ($\boldsymbol{Xt}$) et ($\boldsymbol{Xt}$) ne dépendent pas de ($\boldsymbol{t}$) et tel que $\boldsymbol{Co(Xt\ ,Xt+h)} = \boldsymbol{0}$

- $\boldsymbol{E(Xt)} = \boldsymbol{\mu}\ \forall \boldsymbol{t}$
- $\boldsymbol{V(Xt)} = \boldsymbol{\sigma}\ \boldsymbol{2}\ \forall \boldsymbol{t}$
- $\boldsymbol{Cov(Xt\ ,Xt+h)} = \boldsymbol{0}\ \boldsymbol{si}\ \boldsymbol{h} \neq \boldsymbol{0}$

Si en plus on a :

$(\boldsymbol{Xt}) = \boldsymbol{0}$ alors le bruit blanc est dit centré

Les variables sont gaussiennes (suivent une loi normale) alors le bruit blanc est dit gaussien. Les bruits blancs sont des processus stationnaire.

2-Recuperation des résidus

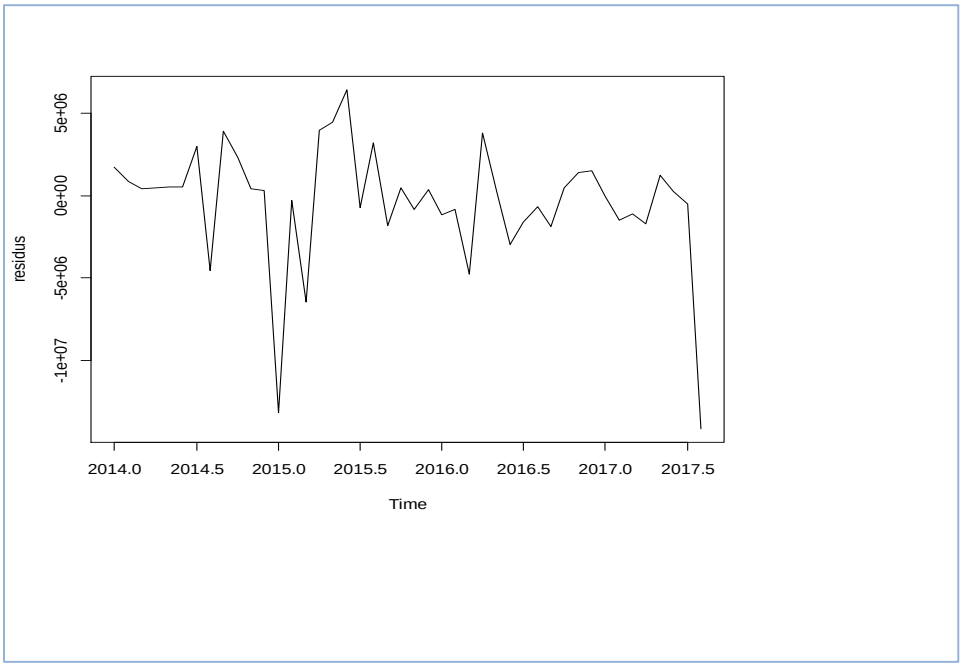
Dans le contexte de la modélisation des séries temporelles, la récupération des résidus est une étape importante dans l'analyse des modèles. Une fois qu'un modèle a été ajusté aux données, les résidus sont calculés en soustrayant les valeurs prédites par le modèle des valeurs observées. Ces résidus représentent l'erreur de prédiction du modèle pour chaque observation.

Tableau 13: Récupération des résidus

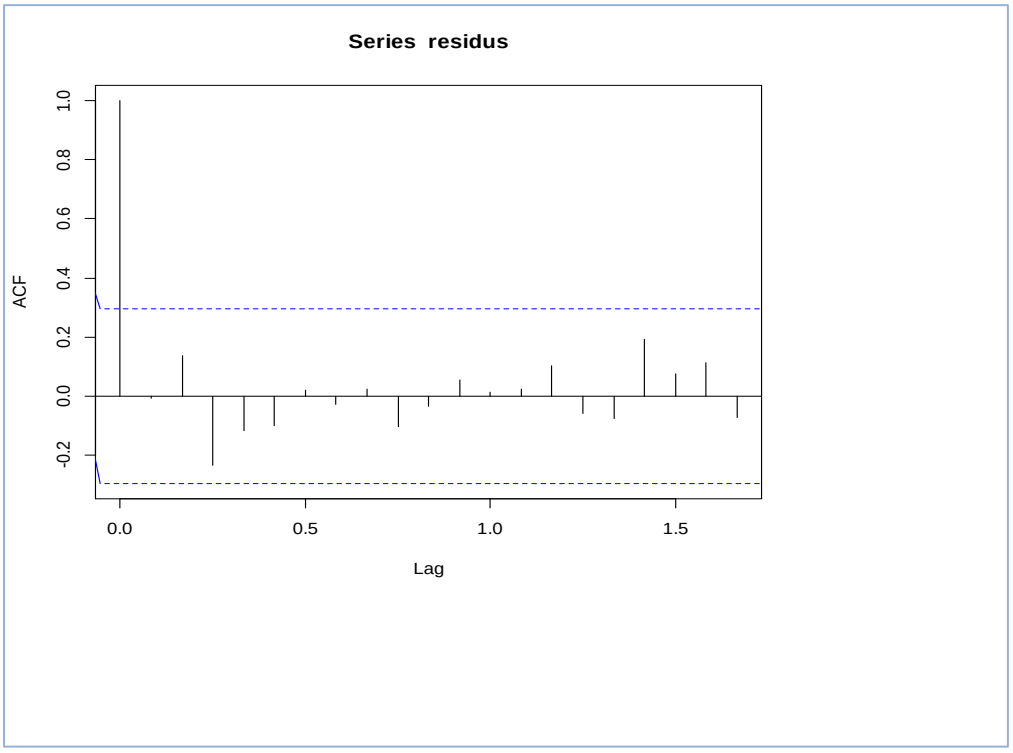
	Jan	Feb	Mar	Apr	May	Jun
2014	1719432.44	847300.91	424231.00	487398.90	515970.74	526455.31
2015	-13161584.45	-325052.59	-6437780.24	3971054.61	4461222.76	6392867.13
2016	-1191121.59	-859488.45	-4750211.77	3807796.52	212507.96	-2964608.62
2017	-24558.68	-1479779.51	-1115684.17	-1726298.33	1252053.27	246013.40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

Les nombres positifs représentent des montants de ventes positifs ou des valeurs croissantes par rapport à une référence, tandis que les nombres négatifs représentent des montants de ventes négatifs ou des valeurs décroissantes par rapport à une référence. De plus Il semble y avoir des variations significatives dans les montants de ventes d'un mois à l'autre et d'une année à l'autre.

Graphique 9: Visualisation des résidus



Graphique 10 : Test de bruit blanc et autocorrélation



3-Test de bruit blanc par la méthode de LJUNG-BOX

Le test de bruit blanc par la méthode de **Ljung-Box** est une technique statistique utilisée pour évaluer la présence de corrélations significatives dans les résidus d'un modèle de séries temporelles. Il est largement utilisé pour valider les modèles et s'assurer que les résidus sont indépendants et identiquement distribués, comme le requiert l'approche classique de la modélisation des séries temporelles.

H0 : la série est un bruit blanc

H1 : la série n'est pas un bruit blanc

Tableau 14: LJUNG-BOX

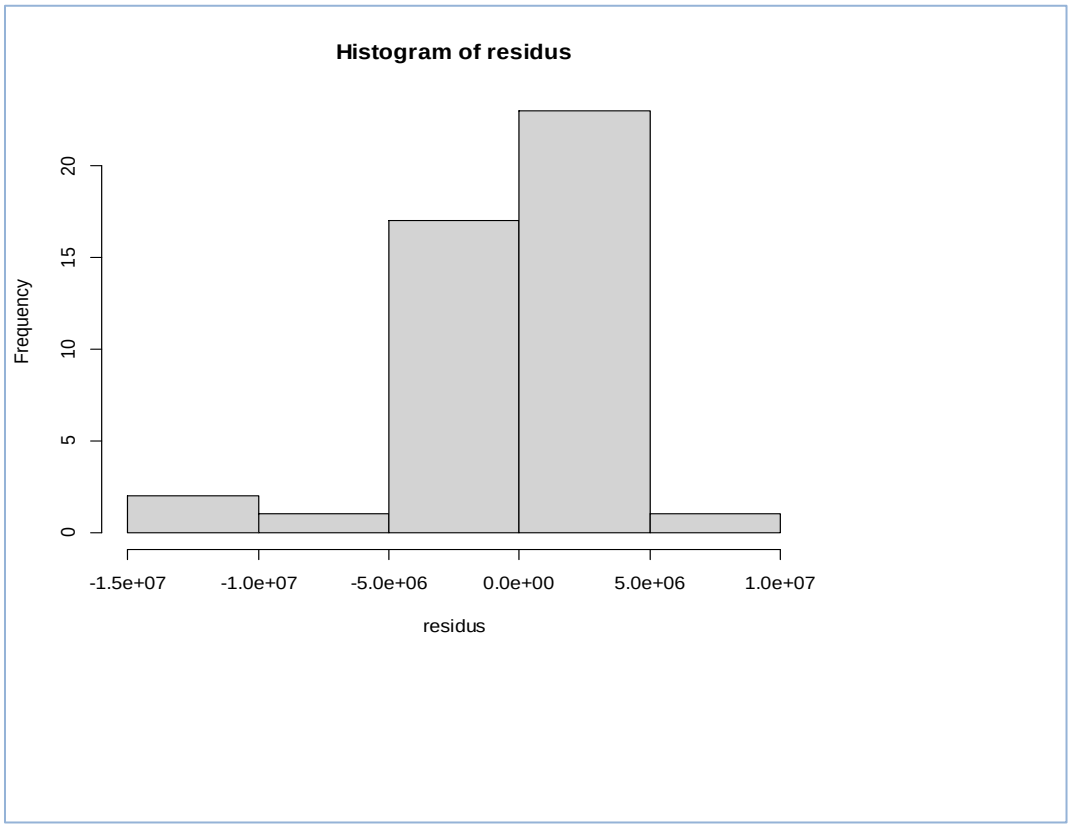
Colonne1
Box-Ljung test
data: residus
X-squared = 5.846, df = 11, p-value = 0.8834

p-value >0.8834 , on ne peut rejeter l'hypothèse nulle, Les résidus du modèle ne présentent pas de corrélation significative à différents retards, donc la série est un bruit blanc. Ici, la statistique de test de Ljung-Box est de 5.846, et la valeur p est de 0.8834. Pour être sûr que le modèle prédictif ne peut pas être amélioré, il est également 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.

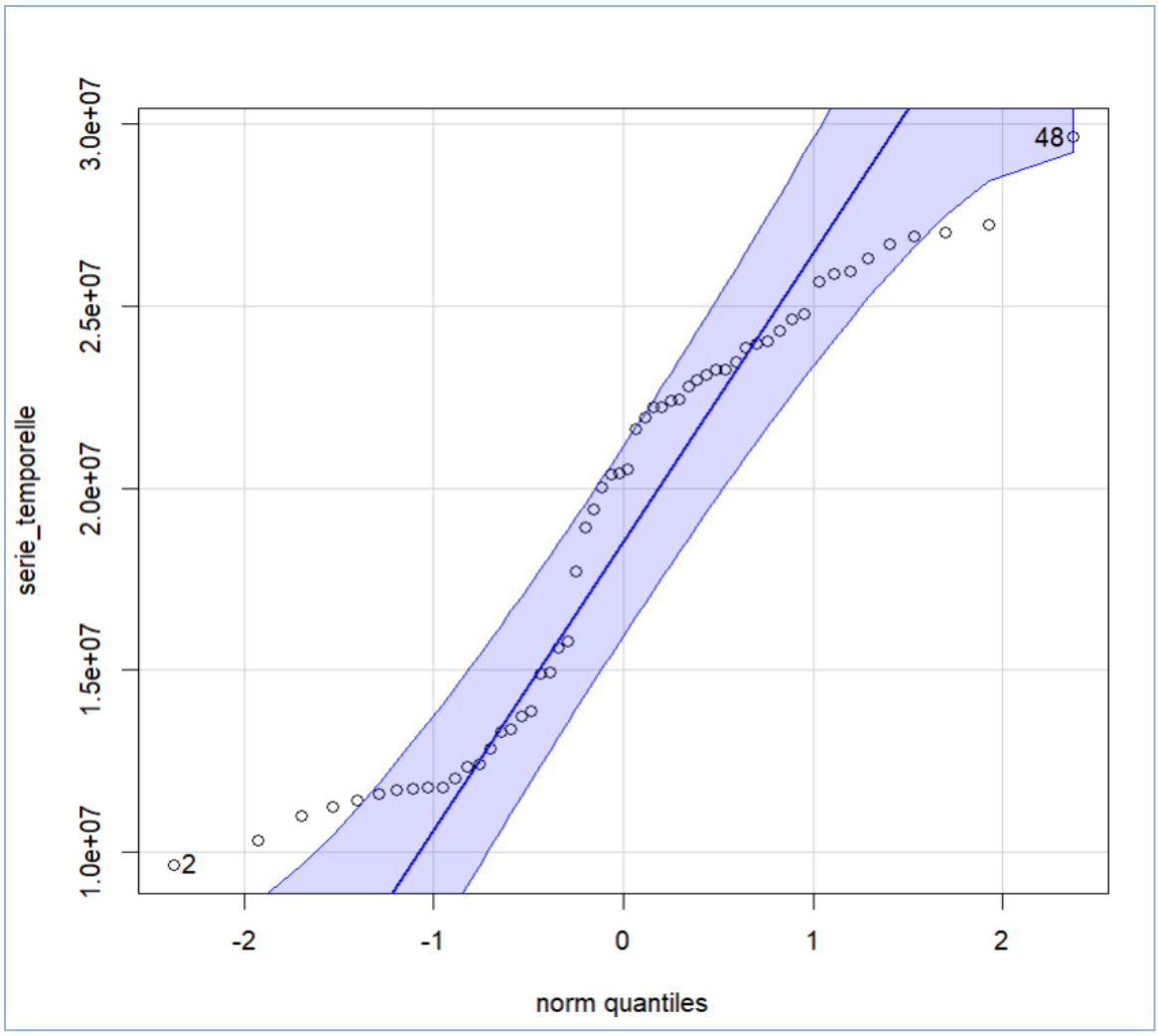
4-Test de normalité

Pour vérifier si les erreurs de prévision sont normalement réparties avec le zéro moyen, nous pouvons tracer un histogramme des erreurs de prévision.

Graphique 11: Histogramme des résidus



Graphique 12: Test de Graphique de normalité



5-Test de SHAPIRO WILK

Le test de Shapiro-Wilk est un test de normalité largement utilisé pour évaluer si un échantillon de données suit une distribution normale ou non. Il est considéré comme l'un des tests les plus puissants pour détecter les écarts par rapport à la normalité, en particulier pour les échantillons de petite à moyenne taille.

Tableau 15: Test de SHAPIRO

```
        Shapiro-Wilk normality test

data:  residus
W = 0.82233, p-value = 9.283e-06
```

H0 : la série suit une loi normale

H1 : la série ne suit pas une loi normale

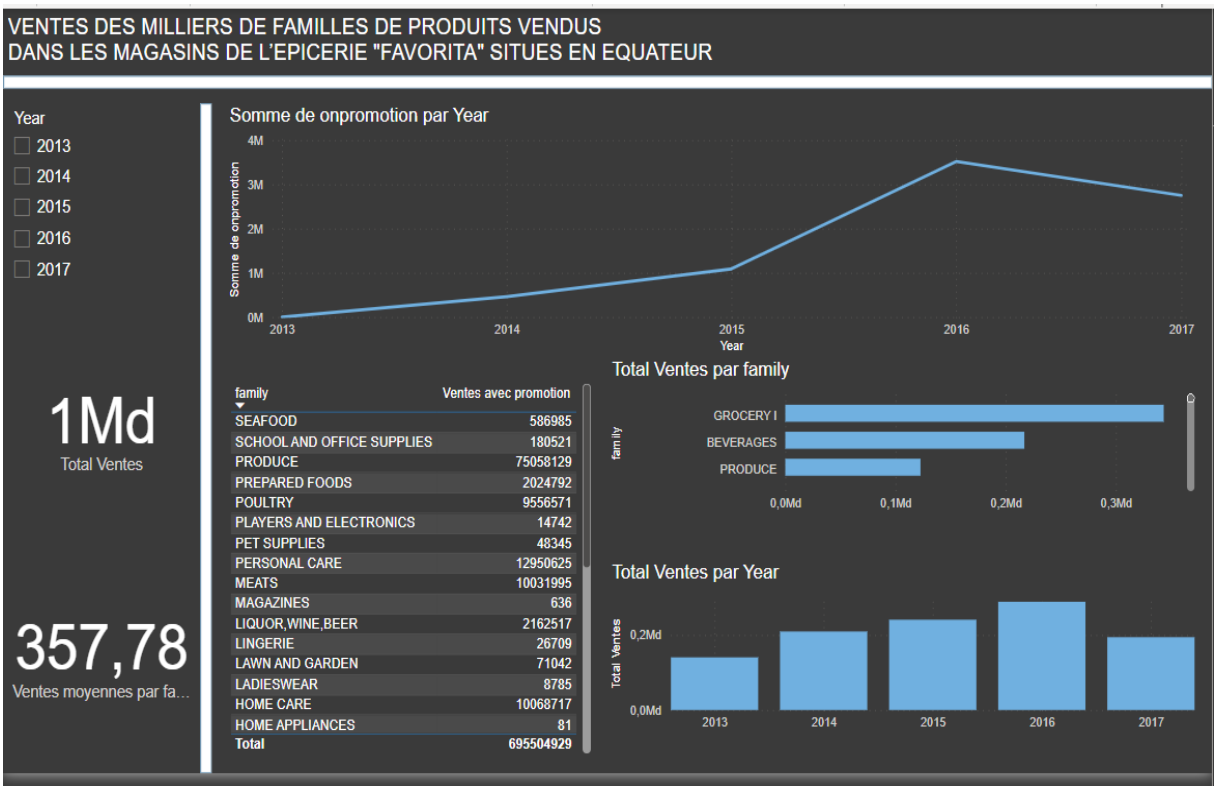
La P-value<5%, On rejette l'hypothèse nulle donc les résidus ne suivent pas la loi normale.

6-Moyenne des résidus

Une moyenne de résidus de -424918.6 indique qu'il y a un biais systématique dans les prédictions du modèle.

Une valeur négative de la moyenne des résidus suggère que le modèle sous-estime généralement les valeurs observées. Cela signifie que, en moyenne, les valeurs prédites par le modèle sont inférieures aux valeurs réelles observées.

POWER BI



CONCLUSION

En conclusion, l'utilisation de techniques de modélisation pour prévoir les ventes de chaque famille de produits dans chaque magasin de l'épicerie "Favorita" en Équateur offre un potentiel significatif pour améliorer la gestion des stocks et la planification opérationnelle. En développant des modèles précis de prévision des ventes, les détaillants peuvent mieux anticiper la demande, éviter les pénuries de produits, réduire les coûts liés au stockage excessif et optimiser les niveaux de stock pour maximiser les ventes tout en minimisant les pertes.

Ces prévisions de ventes peuvent également aider à orienter les décisions en matière de tarification, de promotions et de stratégies de marketing, en fournissant des informations précieuses sur la manière de stimuler les ventes et de fidéliser les clients. En fin de compte, l'utilisation de modèles de prévision des ventes contribue à améliorer l'efficacité opérationnelle et la rentabilité des magasins d'épicerie "Favorita", renforçant ainsi leur position concurrentielle sur le marché.

SOURCE DU CODE R

```
# IMPORTATION DES LIBRARIES

Library (tseries)

Library (dplyr)

Library (ggplot2)

Library (forecast)

Library (car)

# IMPORTATION DES DONNEES

epicerie <- read.csv("C:/Users/wawa/Desktop/INSSEDS/MINI PROJET/epicerie.csv",
stringsAsFactors=TRUE,row.names = 1)

# EXPLORATION DES DONNEES

head(epicerie, 5)

tail(epicerie, 5)

str(epicerie)

summary(epicerie)

# AGREER LES VENTES PAR MOIS

ventes_par_mois <- epicerie %>%

  group_by(mois = as.Date(cut(date, breaks = "month"))) %>%

  summarise(ventes_totales = sum(sales, na.rm = TRUE))

print(ventes_par_mois)


# CONVERTIR LE DATAFRAME EN SERIE TEMPORELLE

serie_temporelle <- ts(ventes_par_mois$ventes_totales, start = c(2013,1),frequency = 12)

print(serie_temporelle)


#TESTER LA STATIONNARITE

adf.test(serie_temporelle)
```

```
# ANALYSE DESCRIPTIVE

plot(serie_temporelle,col="red",main="SERIE TEMPORELLE DES VENTES",xlab="Temps",ylab="Valeurs")

hist(serie_temporelle,prob=TRUE,col="red",main="HISTOGRAMME DE LA SERIE TEMPORELLE DES
VENTES")

lines(density(serie_temporelle,na.rm=TRUE))

wawa.qt.resume(serie_temporelle)


# IDENTIFIER LES TENDANCES ET SAISONS

decomposition <- decompose(serie_temporelle)

plot(decomposition)


#AUTOCORRELATION SIMPLE

acf(serie_temporelle,lag.max =20,plot=FALSE)

acf(serie_temporelle,lag.max =20,plot=TRUE)


#AUTOCORRELATION PARTIELLE

pacf(serie_temporelle,lag.max = 20,plot=FALSE)

pacf(serie_temporelle,lag.max =20,plot=TRUE)


#CHOIX DES PARAMETRES

xlisse <- HoltWinters(serie_temporelle)

xlisse

plot(xlisse)


#LISSAGE EXPONENTIELLE SIMPLE SANS INTERVALLE DE CONFIANCE

xlisse <- HoltWinters(serie_temporelle)

prev<-predict(xlisse,n.ahead=12,prediction.interval=FALSE)

prev


#LISSAGE EXPONENTIELLE SIMPLE AVEC INTERVALLE DE CONFIANCE

xlisse <- HoltWinters(serie_temporelle)
```

```
prev<-predict(xlisse,n.ahead=12,prediction.interval=TRUE)

prev

plot(xlisse)


# REPRESENTATION GRAPHIQUE DE LA PREVISION

plot(forecast(xlisse))


# RECUPERER LES RESIDUS

residus <- residuals(xlisse)

residus


#VISUALISATION DU RESIDU

plot(residus)

acf(residus, lag.max=20, na.action = na.pass)


# TEST DE BRUIT BLANC PAR LA METHODE DE L JUNG-BOX

Box.test(residus, lag=11, type="Ljung-Box")


#HISTOGRAMME DU RESIDU

hist(residus)


#TEST GRAPHIQUE DE NORMALITE

library(car)

qqPlot(serie_temporelle)


#TEST DE SHAPIRO WILK

shapiro.test(residus)


# MOYENNE DES ERREURS DE PREDICTION

mean(residus, na.rm=TRUE)
```