Master 1 Econométrie, Statistiques

PRÉVISION DES VENTES D'UNE COMPAGNIE PHARMACEUTIQUE

Réalisé par :

SAIDI EL OUALI Amal
MISSINHOUN Johannes
RUBEN CHARLES Reginald

Encadré par :

BAGORO ISMAËL

2023-2024

Table des matières

Graphiques :

Tableaux :

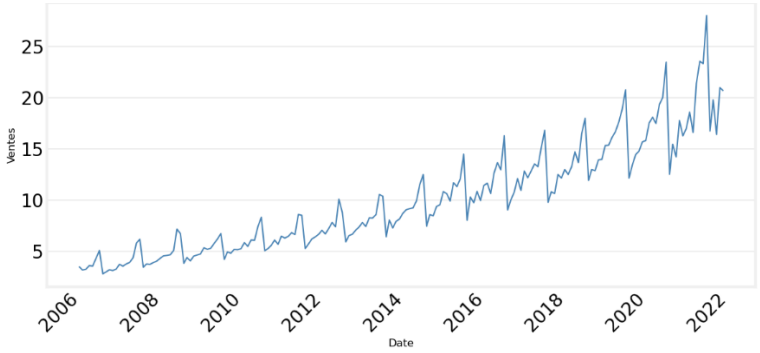
1. Introduction

Dans le cadre de l'analyse des données de vente mensuelles de notre étude, nous avons entrepris une étude détaillée pour identifier le modèle le plus approprié pour prédire les tendances futures basées sur l'historique des données collectées de juillet 2006 à juin 2022.

Dans un environnement commercial de plus en plus compétitif et dynamique, la capacité d'une entreprise à anticiper avec précision les tendances futures devient fondamental pour sa stratégie et sa planification opérationnelle. Les prévisions de ventes, en particulier, jouent un rôle essentiel en permettant aux organisations de gérer efficacement les inventaires, d'optimiser les allocations de ressources. C'est dans ce contexte que le présent rapport se propose d'analyser l'efficacité de trois méthodes de prévision statistique : le modèle SARIMA, la méthode de lissage exponentiel de Holt-Winters et le modèle Prophet, afin de déterminer laquelle de ces techniques offre la meilleure précision pour prévoir les ventes mensuelles de l’entreprise.

2. Analyse descriptive

Le graphique suivant présente l'évolution des ventes sur la période de 2006-2022.



Graphique 1: Ventes mensuelles de juillet 2006 à juin 2022

La série des ventes présente une tendance générale à la hausse des ventes au fil des années. Cela indique une croissance continue de l'activité.

Des variations régulières et récurrentes sont observées sur le graphique. On peut voir des pics et des creux qui se répètent chaque année, ce qui suggère un effet saisonnier sur les ventes.

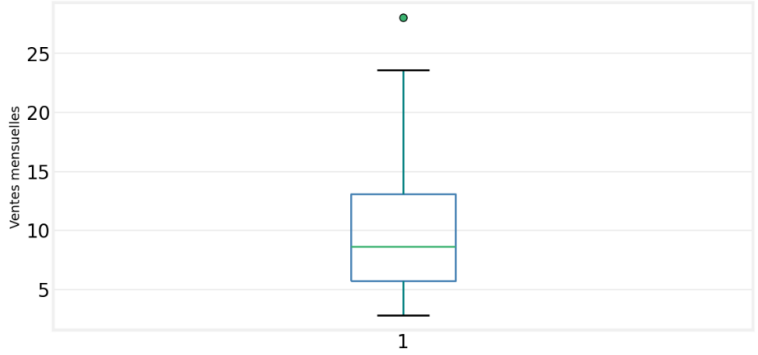
- Statistiques descriptives

Les statistiques descriptives permettent de détecter la présence d'éventuelles valeurs aberrantes.

Tableau 1 : Statistiques descriptives de la série Monthly Sales

Moyenne	Ecart-type	Minimum	Maximum	25%	50%	75%
9.9146	*5.1788*	*2.81452*	*28.038*	*5.7080*	*8.6621*	*13.06*

En moyenne les ventes mensuelles s'élève à 9.9146 unités monétaires. Le maximum extrême (28.038), très éloigné de la moyenne suggère la présence de valeurs aberrantes élevées.



Graphique 2 : Box-plot de la série de données

On note ici la présence d'un montant de ventes assez élevé de 28000 unités monétaires environs. Ce montant a été enregistré en 2022 soit durant la pandémie du covid-19. Il n'est donc pas nécessaire de traiter cette valeur de manière particulière.

2.1 Division de l'échantillon en train-test

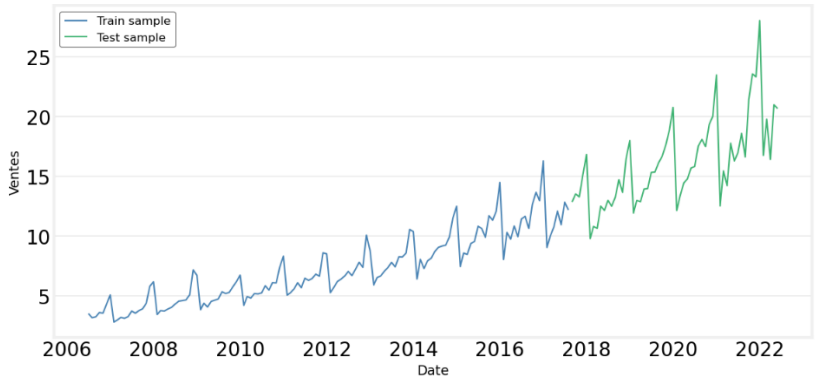
La division train-test est une technique populaire pour partitionner un ensemble de données en deux parties : un ensemble d'apprentissage et un ensemble de test.

De cette division découle deux phases pour la construction du modèle prédictif :

Une phase d'entraînement : On utilise l'échantillon d'apprentissage pour créer le modèle. Dans le cas d'un modèle paramétrique, par exemple, c'est dans cette phase que nous estimons les paramètres. En général, on entraîne plusieurs modèles ou plusieurs variantes d'un même modèle.

Une phase de validation : On utilise l'échantillon test (ou de validation) pour évaluer la performance du modèle sur des données qui n'ont pas servi à l'entraînement, de façon à éviter le surapprentissage. La performance du modèle peut se baser sur différents indicateurs. On choisit le modèle qui obtient la meilleure performance (l'EQM la plus faible) sur l'échantillon de validation.

L'échantillon d'apprentissage contient 134 observations et l'échantillon test 58 observations.



Graphique 3 : Echantillon train et test

2.2 Analyse sur l'échantillon d'apprentissage

2.2.1 La décomposition tendance-cycle

L'analyse des séries temporelles repose le plus souvent sur une décomposition tendance saisonnalité et composante résiduelle de la série.

Toutefois, cette décomposition, si elle est très utilisée en pratique, ne repose pas sur une construction théorique unique.

La tendance : Elle traduit l'évolution à long terme du phénomène. On parle aussi de mouvement conjoncturel ou mouvement extra-saisonnier.

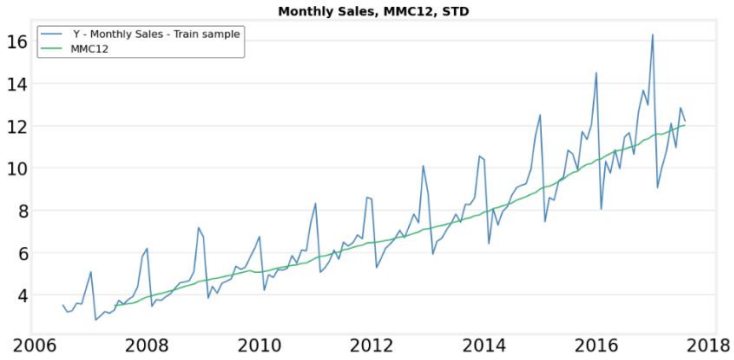
La saisonnalité : Elle correspond à des fluctuations périodiques qui se reproduisent à intervalle de temps régulier.

La composante résiduelle : Elle rassemble tout ce que les autres composantes n'ont pu expliquer du phénomène observé. Elle contient donc de nombreuses fluctuations, en particulier accidentelles, dont le caractère est imprévisible (catastrophes naturelles, guerres, grèves...).

2.2.2 Détection de la tendance

Les moyennes mobiles permettent d'effacer tout le bruit, toutes les variations dans la série des ventes afin d'avoir une vue plus claire sur son évolution.

Le lissage de la série par la méthode des moyennes mobiles fait apparaitre une tendance générale.



Graphique 4 : Lissage de la série par moyenne mobile

La série de moyennes mobiles est croissante ce qui implique que notre série n'est pas stationnaire en moyenne. Nous discuterons plus loin de la nature de la non-stationnarité du processus.

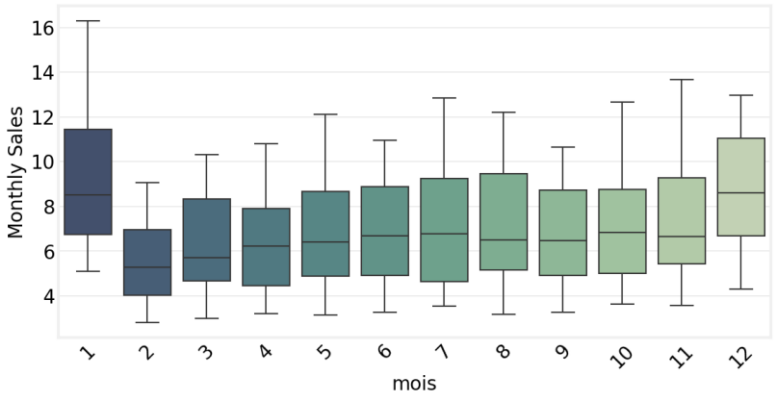
2.2.3 Détection de la saisonnalité

L'étude de la saisonnalité est un préalable au traitement d'une série chronologique. En effet, lorsque cette composante existe, il convient de l'isoler afin de pouvoir analyser les autres caractéristiques. Une désaisonnalisation systématique, sans tester l'existence de cette composante, peut créer un « bruit » parasite nuisible à l'analyse de la chronique et donc dégrader la qualité de la prévision.

- Analyse graphique

Un box-plot peut aider à détecter la saisonnalité dans une série temporelle en comparant les distributions des données sur différentes périodes récurrentes, comme les mois ou les trimestres.

Si la série temporelle présente une saisonnalité, les box-plots de ces périodes récurrentes montreront des différences systématiques dans leurs distributions.



Graphique 5 : Box-plots mensuels des ventes

Le box-plot indique une présence marquée de saisonnalité avec des ventes nettement plus élevées et variables en janvier et en décembre, contrastant avec des mois plus stables et moins variables.

- Analyse de la variance et test de Fisher

L'examen visuel du graphique ou du tableau ne permet pas toujours de déterminer avec certitude l'existence d'une saisonnalité, de surcroît, il interdit l'automatisme de traitement qui peut s'avérer nécessaire dans le cas d'un nombre important de séries à examiner. Le test de Fisher à partir de l'analyse de la variance permet de pallier ces deux inconvénients.

Ce test suppose la chronique sans tendance ou encore sans extra-saisonnalité. Dans le cas contraire cette composante est éliminée par une régression sur le temps (extra-saisonnalité déterministe), ou par une procédure de filtrage (extra-saisonnalité aléatoire).

Les données d'apprentissages sont regroupées par années (le facteur ligne) et par mois (le facteur colonne).

Le test de Fisher pour la saisonnalité consiste donc à tester l'influence du facteur colonne. Les hypothèses sont les suivantes :

$$\begin{cases} H_0 : Pas\ d'influence\ du\ facteur\ colonne \\ H_1 : Influence\ du\ facteur\ colonne \end{cases}$$

Tableau 2 : Test de Fisher

F-statistic	p-value	Décision
99.15266	*8.6830e-55*	*Rejet de H_0*

On conclut donc que la série présente bien une composante saisonnière.

3. Choix du modèle de décomposition

3.1 Les types de modèles de décomposition

Il en existe essentiellement trois grands types :

Le schéma additif : Il s'agit du modèle classique de décomposition. La chronique X_t est la somme de la tendance T_t, de la saisonnalité S_t et de l'erreur E_t.

On a:

$$X_t = T_t + S_t + E_t \quad (1.1)$$

Le schéma multiplicatif (mixte) : dans lequel la composante saisonnière est liée à l'extra-saisonnier.

$$X_t = T_t * S_t + E_t \quad (1.2)$$

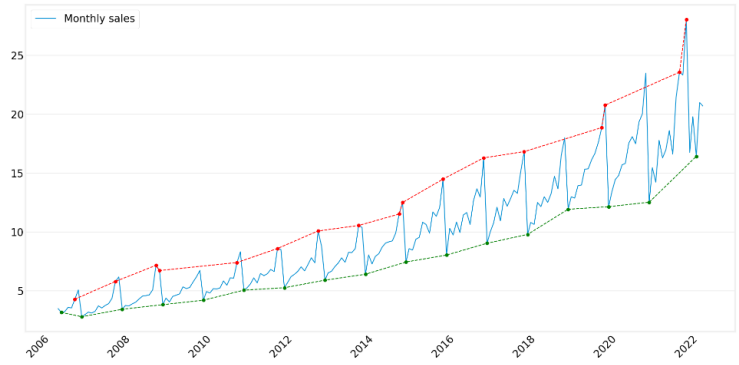
Le schéma multiplicatif complet : présentant une interaction entre les trois composantes.

$$X_t = T_t * S_t * E_t \quad (1.3)$$

3.2 Les tests de décomposition

- La procédure de la bande

Le « test de la bande » consiste à partir de l'examen visuel du graphique de l'évolution de la série brute à relier, par une ligne brisée, toutes les valeurs « hautes » et toutes les valeurs « basses » de la chronique. Si les deux lignes sont parallèles, la décomposition de la chronique peut se faire selon un schéma additif ; dans le cas contraire, le schéma multiplicatif semble plus adapté.



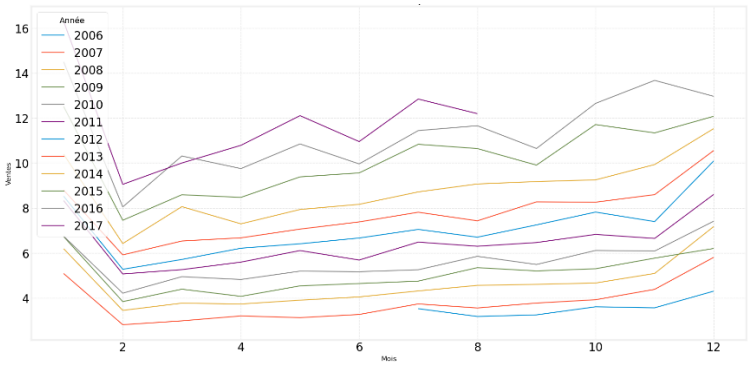
Graphique 6 : Illustration de la procédure de la bande

Les deux lignes ne sont pas parallèles, le schéma multiplicatif semble donc le plus adapté.

- La méthode du profil

Elle consiste à découper la série d'observations en sous-séries annuelles puis à les représenter sur un même graphe.

Si les différentes courbes de ces sous-séries annuelles sont parallèles, le modèle est additif ; sinon le modèle est multiplicatif.



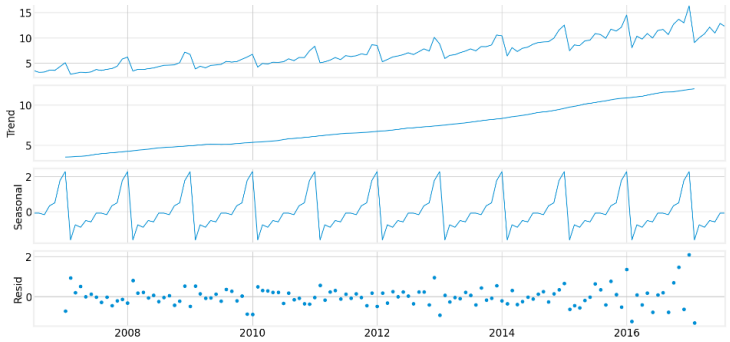
Graphique 7 : Tendance des ventes par mois

On peut constater sur le graphique que les différentes courbes des sous-séries annuelles ne sont pas parallèles. On conclut donc le modèle multiplicatif est le plus adapté.

Pour enrichir le travail, nous utilisons une dernière méthode basée sur les modèles de décomposition. Il s'agira décomposer la série suivant les modèles additif et multiplicatif et d'étudier le comportement des résidus.

- Décomposition de la série des ventes mensuelles

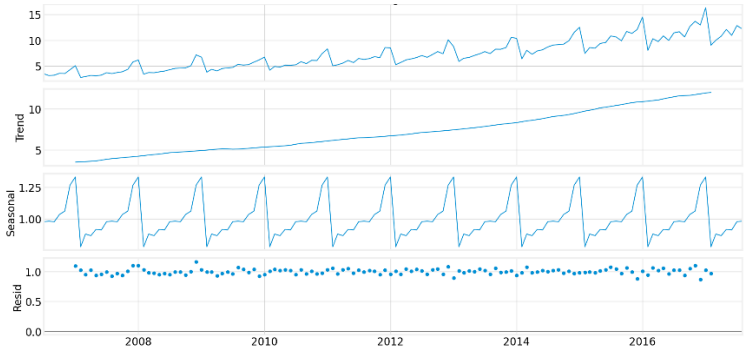
Le modèle additif



Graphique 8 : Décomposition additive de la série Monthly sales

Les résidus sont autocorrélés et semblent contenir une composante saisonnière. Cela suppose donc que le schéma additif n'est pas le plus adapté à notre jeu de données. Nous passons donc au schéma multiplicatif.

Le modèle multiplicatif



Graphique 9 : Décomposition multiplicative de la série Monthly sales

Sous ce schéma, les résidus sont non autocorrélés. On conclut donc que le modèle multiplicatif est le bon schéma de décomposition.

Il est possible d'appliquer une transformation logarithmique à la série pour retomber sur un modèle additif.

3.2 Choix du modèle de prévision

Il existe une grande variété de modèles permettant la prévision de série temporelles.

On distingue parmi elles deux grandes méthodes :

- Les lissages exponentiels
- Les modèles de types ARIMA

3.2.1 Les lissages exponentiels

Les techniques de lissages exponentiels ont été introduits par Holt en 1957 mais surtout par Brown en 1962. Elles consistent à prévoir l'évolution d'une série par extrapolation.

Le lissage regroupe l'ensemble des techniques empiriques qui ont pour caractéristiques communes d'accorder un poids plus important aux valeurs récentes de la chronique.

Nous pouvons citer :

- Le lissage exponentiel simple
- Le lissage exponentiel double
- Lissage exponentiel double amorti de Gardber & McKenzie
- Le lissage exponentiel triple de Holt-Winters.

Les trois premières méthodes sont inefficaces pour des séries temporelles avec des structures complexes comme la saisonnalité car elles n'intègrent pas dans leur spécification, des paramètres permettant de prendre en compte les composantes saisonnières des séries. De plus, leurs performances sont assez médiocres pour les prévisions à long terme.

Pour la suite, nous focalisons donc sur la méthode de Holt-Winters. Elle est basée sur trois équations de lissage : une pour le niveau de la série, une pour la tendance et une pour la saisonnalité.

Il existe deux variantes de cette méthode qui diffèrent par la nature de la composante saisonnière. La méthode additive est préférable lorsque les variations saisonnières sont à peu près constantes tout au long de la série, tandis que la méthode multiplicative est préférable lorsque les variations saisonnières changent proportionnellement au niveau de la série.

La modèle de décomposition de la série de ventes étudiée est multiplicatif. La méthode de Holt-Winters semble donc la plus adaptée aux données.

Nous présentons donc la méthode de lissage multiplicative de Holt-Winters ainsi que les résultats obtenus pour notre échantillon test.

- Le schéma multiplicatif de Holt-Winters

Les équations relatives au Lissage Exponentiel Triple Multiplicatif de Holt-Winter nous sont données par :

- Equation associée au niveau :

$$l_t = \alpha\left(\frac{y_t}{s_{t-m}}\right) + (1-\alpha)(l_{t-1} + T_{t-1}) \quad (2.1)$$

- Equation associée à la tendance (variation) :

$$T_t = \beta(l_{t-1} + T_{t-1}) + (1-\beta)T_{t-1} \quad (2.2)$$

- Equation associée à la saisonnalité :

$$S_t = \lambda\left(\frac{y_t}{l_t + T_{t-1}}\right) + (1-\lambda)S_{t-m} \quad (2.3)$$

- Equation de prévision :

$$\hat{y}_{t+h|t} = (l_t + T_t h) * S_{t-m+h_m} \quad (2.4)$$

Le niveau de la série est estimé à travers l'équation (2.1) comme une moyenne pondérée entre la dernière observation corrigée de variation saisonnière ($yt - st{-}m$) ainsi que de ($l_{t-1} + T_{t-1}$), la prévision faite en (t-1) de la série corrigée de toute variation saisonnière. Nous retrouvons donc dans cette 1ère équation la forme classique d'un lissage exponentiel dans lequel la prévision à date dépend de la valeur actuelle de la série ainsi que de la dernière prévision disponible.

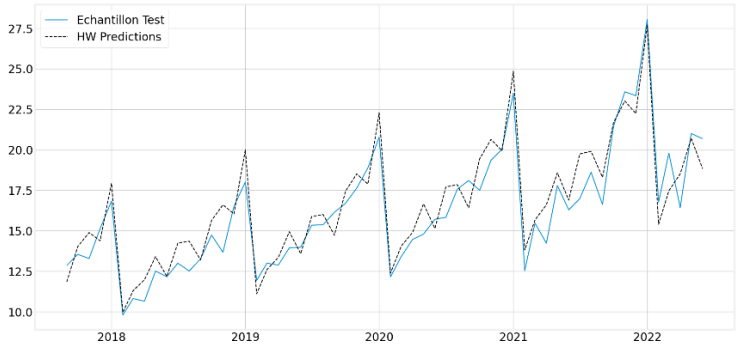
Cette fois ci, le Lissage Exponentiel se fait sur la série Corrigée de Variation Saisonnière. L'équation (2.3) correspond au lissage de la composante saisonnière obtenue à travers une moyenne pondérée entre la valeur actuelle du coefficient ($Y_t - l_{t-1} - T_{t-1}$) et son estimation pour la période actuelle mais faite à la période précédente (p périodes avant).

L'équation (2.4) renseigne les prévisions qui sont obtenues en combinant les 3 composantes lissées estimées dans les équations (2.1), (2.2) et (2.3). Remarquons qu'il s'agit là de la méthode de Holt mais avec un nouveau terme S_{t-m+h_m} relatif à la composante saisonnière.

- Prévision avec le lissage exponentiel triple de Holt-Winters

Sous python, la prévision se fait en utilisant la fonction ExponentialSmoothing du module statsmodels.

Les résultats obtenus sur l'ensemble test peuvent être visualisés sur le graphique suivant :



Graphique 10 : Prévision avec la méthode de Holt-Winters

On peut constater que l'algorithme de Holt-Winters ajuste assez bien les données de l'échantillon de validation.

Pour évaluer la performance du modèle, il est indispensable de calculer quelques mesures de performances. Nous nous limitons ici à trois mesures : la Mean Square Error (MSE), la Mean Absolute Percentage Error (MAPE) et la Mean Absolute Error (MAE).

Tableau 3 : Critères de performance

MSE	MAPE	MAE
1.580	*6.61%*	*1.042*

Les mesures sont assez faibles témoignant de la qualité de la prévision. Pour ne pas se limiter à ce modèle, les prévisions ont également réalisées sur la base d'un modèle SARIMA et de l'algorithme Prophet de Facebook.

Le modèle retenu est celui qui minimise les critères précédemment cités.

4. Modèle SARIMA

Les modèles SARIMA, ou Seasonal Autoregressive Integrated Moving Average, sont une extension des modèles ARIMA (Autoregressive Integrated Moving Average) qui sont spécialement conçus pour modéliser et prédire des séries temporelles qui présentent des comportements saisonniers. Les modèles SARIMA sont très utilisés dans les analyses temporelles où les données montrent non seulement des tendances et/ou des cycles mais aussi des variations saisonnières claires, comme dans le cas de ventes mensuelles, de températures journalières, ou d'autres indicateurs économiques et environnementaux.

Un modèle SARIMA est généralement noté comme SARIMA(p, d, q)(P, D, Q)[s], où

(p, d, q) sont les paramètres non saisonniers:

p : L'ordre des termes autorégressifs (AR). Il s'agit du nombre de dépendances retardées incluses dans le modèle.

d : Le degré de différenciation. Cela indique combien de fois les données doivent être différenciées pour rendre la série stationnaire, c'est-à-dire pour stabiliser la moyenne de la série temporelle.

q : L'ordre des termes de moyenne mobile (MA). Il s'agit du nombre de termes de moyenne mobile dans le modèle.

(P, D, Q)[s] sont les paramètres saisonniers:

P : L'ordre des termes autorégressifs saisonniers.

D : Le degré de différenciation saisonnière. Il est similaire à d mais appliqué aux composantes saisonnières de la série.

Q : L'ordre des termes de moyenne mobile saisonniers.

S : Le nombre de périodes dans chaque saison. Par exemple, *s* serait 12 pour des données mensuelles avec une saisonnalité annuelle, ou 4 pour des données trimestrielles.

4.1 Fonctionnement des Modèles SARIMA

La première étape dans l'ajustement d'un modèle SARIMA est souvent de rendre la série temporelle stationnaire, car les modèles ARIMA et SARIMA nécessitent que la série ne montre ni tendance ni saisonnalité dans la variance ou la moyenne. Cela est souvent accompli en différenciant la série une ou plusieurs fois (le paramètre d). Une fois la série rendue stationnaire, le composant AR modèle la régression de la variable dépendante sur ses propres valeurs retardées. Cela signifie que les valeurs futures sont prédites en fonction de leurs valeurs passées. Le terme MA dans un modèle SARIMA modélise l'erreur de la prédiction comme une combinaison linéaire des erreurs des prédictions passées. Cela ajoute de la souplesse au modèle en permettant de lisser les fluctuations aléatoires dans la série temporelle. Quant aux composants saisonniers (P, D, Q) permettent au modèle de tenir compte des variations qui se répètent à intervalles réguliers.

- Tests Statistiques

Afin de choisir les paramètres appropriés pour notre modèle SARIMA, notre première étape a été de déterminer si la série temporelle des ventes était stationnaire. La stationnarité d'une série est une exigence préalable pour l'analyse avec de nombreux modèles de séries temporelles, y compris SARIMA. Pour ce faire, nous avons utilisé le test de Dickey-Fuller augmenté, un outil statistique conçu pour tester la présence d'une racine unitaire.

Tableau 4 : Test de Dickey-Fuller

Test de Dickey-Fuller			
	Statistique	Pvalue	Ordre d'intégration
$\Delta Monthly\ Sales$	-3.462	*0.009016*	I (0)

Les résultats du test de Dickey-Fuller augmenté ont révélé que la série n'était pas stationnaire à son niveau initial, indiquant la nécessité de différenciations pour stabiliser la moyenne de la série. Sur cette base, nous avons décidé d'intégrer un terme de différenciation (d=1) pour rendre la série stationnaire.

Après avoir effectué une analyse préliminaire approfondie, nous avons ensuite utilisé la fonction auto_arima de la bibliothèque Python pmdarima. Cette méthode automatise la sélection des paramètres du modèle ARIMA en testant une plage de combinaisons différentes pour les termes AR, I et MA, ainsi que pour les composantes saisonnières.

Tableau 5 : Modèle SARIMA optimal

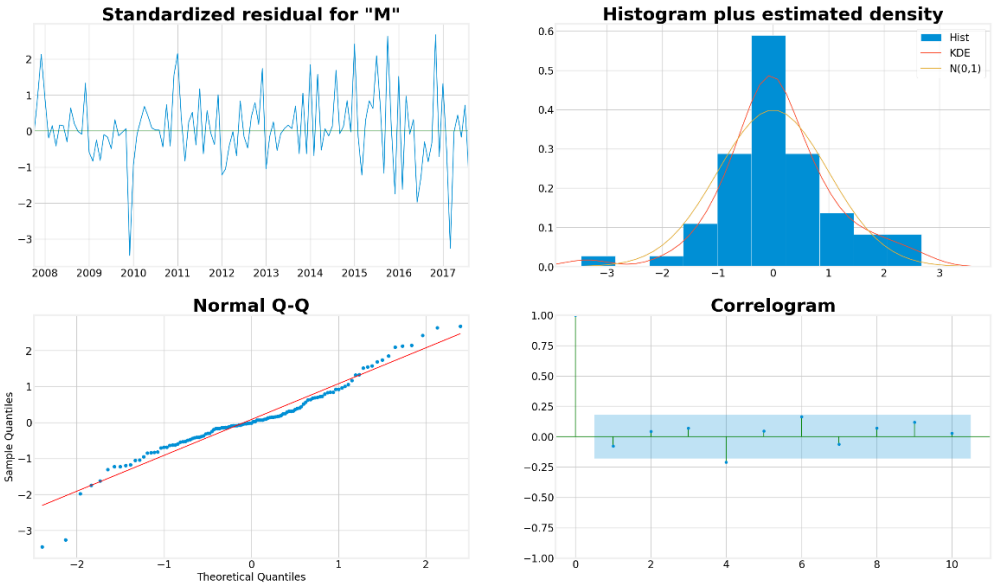
Modèle SARIMA
SARIMA(0,1,1)x(0,1,0)[12]

La fonction auto_arima a conclu que le modèle optimal pour nos données est un SARIMA(0, 1, 1)x(0, 1, 0)[12]. Voici une explication approfondie des paramètres choisis par auto_arima:

(0, 1, 1) : Ces paramètres indiquent qu'aucun terme autorégressif (AR=0) n'est nécessaire, un terme de différenciation (I=1) pour rendre la série stationnaire, et un terme de moyenne mobile (MA=1) suffit pour capturer l'autocorrélation dans les résidus de la série.

(0, 1, 0)[12] : La partie saisonnière du modèle indique également qu'aucun terme AR saisonnier n'est nécessaire, une différenciation saisonnière (D=1) est requise, et qu'aucun terme MA saisonnier est nécessaire, avec une périodicité de 12 mois reflétant la saisonnalité annuelle des données.

Diagnostic

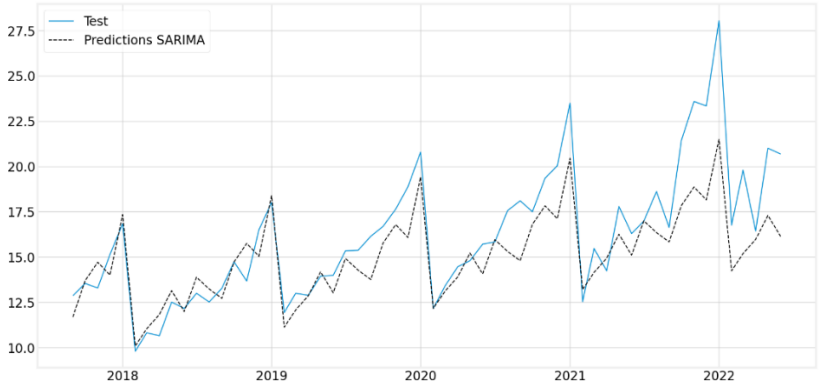


Graphique 11 : Diagnostics du modèle

L'analyse du graphique des résidus standardisés et du corrélogramme permettent de conclure quant à l'absence d'autocorrélation à divers décalages dans les résidus. Cela signifie que le modèle a réussi à capturer adéquatement l'information temporelle dans les données sans laisser de structures autocorrélées inexpliquées. Cependant le modèle souffre d'un problème d'hétéroscédasticité et de normalité des résidus. Dans un modèle à finalité de prédiction, les problèmes d'hétéroscédasticité et de normalité des résidus ne constituent pas nécessairement des obstacles critiques, car l'objectif principal est l'exactitude des prévisions plutôt que l'inférence statistique.

Le modèle SARIMA choisi a été ajusté sur l'ensemble train. Les prévisions ont été effectuées en utilisant la fonction get_forecast, qui permet de générer des prévisions à partir du modèle ajusté pour un nombre spécifié de périodes à l'avance.

- Prévision avec le modèle SARIMA



Graphique 12 : Prévision avec le modèle SARIMA

L'analyse des prévisions de ventes mensuelles générées par le modèle SARIMA, comparées aux données réelles de test sur la période de 2018 à 2022, révèle plusieurs points clés concernant la performance du modèle. Globalement, les prédictions du modèle s'ajustent bien aux données de l'échantillon de validation, ce qui indique que le modèle capture de manière efficace les tendances générales et les variations saisonnières observées dans les données de vente.

Cependant, certaines périodes, notamment vers la fin de 2021 et le début de 2022, montrent des divergences entre les valeurs prédites et les données réelles. Ces écarts peuvent refléter des limites dans le modèle actuel, possiblement dues à des facteurs externes (comme la covid) non pris en compte ou à des anomalies dans les données qui n'ont pas été modélisées.

4. 2 Performance du Modèle

Pour quantifier l'exactitude des prédictions, nous avons calculé trois mesures d'erreur standard :

Tableau 6 : Mesures de performance

Mesures d'erreur		
MSE	MAE	MAPE
4.3458	1.4845	8.21%

Le MSE obtenu est de 4.3458, une mesure qui reflète la moyenne des carrés des écarts entre les prédictions et les valeurs réelles. Le MAE est de 1.4845, indiquant la moyenne des valeurs absolues des écarts, ce qui donne une idée de l'erreur en termes d'unités de ventes. Quant au MAPE s'élève à 8.21%, une métrique qui exprime l'erreur en pourcentage par rapport aux valeurs réelles, facilitant l'interprétation des erreurs dans un contexte relatif.

5. Méthode Prophet de Facebook

Prophet est une bibliothèque open-source développée par Facebook pour la prévision de séries temporelles. Elle est particulièrement conçue pour traiter les séries temporelles avec des tendances non linéaires qui évoluent selon des saisons annuelles, hebdomadaires, des jours fériés et des événements irréguliers. Cette méthode possède les caractéristiques principales suivantes :

Facilité d'utilisation : Prophet est conçu pour être facilement utilisable par des non-experts en statistiques.

Gestion des composantes saisonnières et des jours fériés : La bibliothèque peut inclure automatiquement des effets saisonniers et des jours fériés spécifiques à différents domaines.

Robustesse face aux valeurs manquantes et aux outliers : Prophet est capable de gérer les valeurs manquantes et les valeurs aberrantes sans impact significatif sur les prévisions.

- Détail et Structure de l'Algorithme Prophet

Le modèle Prophet est composé de plusieurs composantes :

Tendance (g(t)) : Capture la tendance générale des données, qui peut être linéaire ou logistique.

Saisonnalité (s(t)) : Capture les effets saisonniers (annuels, hebdomadaires) à l'aide de modèles de Fourier.

Effets des jours fériés (h(t)) : Capture les effets des jours fériés spécifiques.

Terme d'erreur (ε(t)) : Captures les erreurs résiduelles.

Le modèle complet par défaut est défini de manière additive comme:

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \tag{2.5}$$

Lorsque la tendance et la saisonnalité sont multiplicatives, le modèle Prophet doit refléter cette interaction proportionnelle entre les différentes composantes. Le modèle multiplicatif s'exprime par l'équation suivante :

$$y(t) = g(t) * s(t) * h(t) + \epsilon(t) \tag{2.6}$$

Il existe principalement deux types de modèles de croissance : la croissance linéaire et la croissance logistique.

- La croissance linéaire

La croissance linéaire est représentée par une pente constante. Ce modèle suppose que les changements dans la tendance sont proportionnels au temps. La formule de la croissance linéaire est :

$$g(t) = kt + m \tag{2.7}$$

Où :

$g(t)$ est la valeur de la tendance au temps t.

k est le taux de croissance, représentant la pente de la tendance.

m est l'ordonnée à l'origine, représentant la valeur de la tendance au temps zéro.

Ce modèle est utile lorsque les données montrent une tendance stable et constante au fil du temps sans saturation ou inflexions significatives.

- Croissance Logistique

La croissance logistique est utilisée pour modéliser des séries temporelles qui montrent une saturation après une période de croissance. Ce modèle est utile lorsque la série temporelle atteint un plateau ou une capacité maximale. La formule de la croissance logistique est :

$$g(t) = \frac{C}{1 + e^{-k(t-m)}} \tag{2.8}$$

Où :

$g(t)$ est la valeur de la tendance au temps t.

C est la capacité de saturation, représentant la valeur maximale que peut atteindre la série temporelle.

k est le taux de croissance, représentant la rapidité de la montée vers la saturation.

m est le temps auquel la tendance atteint la moitié de sa saturation.

La composante de saisonnalité dans Prophet capture les effets périodiques récurrents dans les séries temporelles. Prophet utilise des modèles de Fourier pour représenter les composantes saisonnières. Il existe principalement deux types de saisonnalité : la saisonnalité annuelle et la saisonnalité hebdomadaire.

- Saisonnalité Annuelle

La saisonnalité annuelle capture les variations périodiques qui se répètent chaque année telles que les effets saisonniers dus aux changements climatiques ou aux cycles économiques. Prophet utilise un modèle de Fourier pour représenter cette saisonnalité. La formule est :

$$s(t) = \sum_{n=1}^{N} (a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \cos\left(\frac{2\pi nt}{P}\right)) \quad (2.9)$$

Où :

$s(t)$ est la valeur de la saisonnalité au temps tt.

N est le nombre de termes de Fourier.

$a_n$et $b_n$sont les coefficients de Fourier.

P est la période de la saisonnalité, qui est de 365.25 jours pour les données annuelles.

- Saisonnalité Hebdomadaire

La saisonnalité hebdomadaire capture les variations périodiques qui se répètent chaque semaine. Similaire à la saisonnalité annuelle, un modèle de Fourier est utilisé, mais avec une période de 7 jours :

$$s(t) = \sum_{n=1}^{N} (a_n \cos\left(\frac{2\pi nt}{7}\right) + b_n \cos\left(\frac{2\pi nt}{7}\right)) \quad (3)$$

Où :

$s(t)$ est la valeur de la saisonnalité au temps tt.

N est le nombre de termes de Fourier.

a_n et $b_n$sont les coefficients de Fourier.

P est la période de la saisonnalité, qui est de 7 jours pour les données hebdomadaires.

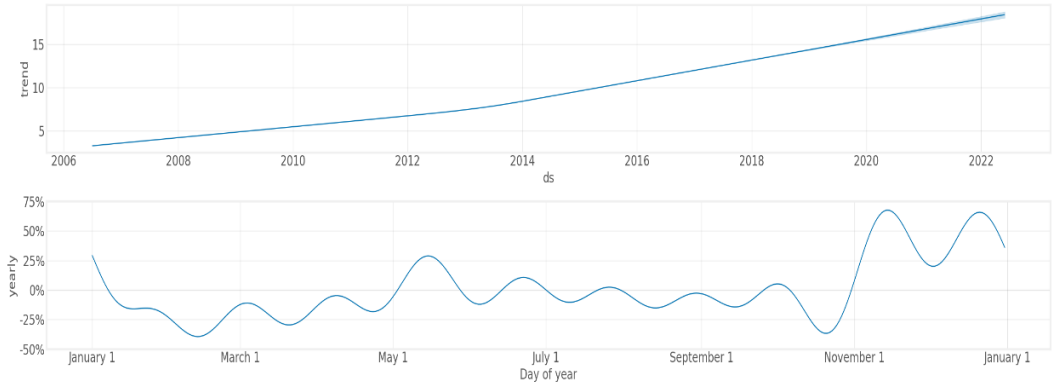
Les jours fériés peuvent aussi avoir des effets significatifs sur certaines séries temporelles, comme les ventes au détail ou le trafic web. Prophet permet d'inclure ces effets spécifiques en ajoutant des variables indicatrices pour les jours fériés.

Pour chaque jour férié, une variable indicatrice est ajoutée au modèle pour représenter l'impact de ce jour sur la série temporelle. Par exemple, si une journée spécifique (comme Noël) a un effet significatif sur les données, cette journée peut être incluse dans le modèle avec une variable qui

prend la valeur 1 ce jour-là et 0 sinon. Cette approche permet de capturer les effets spécifiques des jours fériés et d'améliorer la précision des prévisions.

5.1 Prévision avec la méthode de Prophet

Nous allons maintenant faire la prévision avec Prophet sur Python. Une étape primordiale est la préparation des données de série temporelle dans un DataFrame pandas avec des colonnes nommées impérativement **'ds'** pour les dates et **'y'** pour les valeurs. Ensuite, un modèle Prophet est initialisé et ajusté à nos données historiques de l'échantillon d'apprentissage. Dans notre cas, nous spécifions la saisonnalité multiplicative puisqu'elle correspond mieux (cf. 3.2.2). Nous pouvons visualiser les différentes composantes de notre modèle à travers la figure ci-dessous.



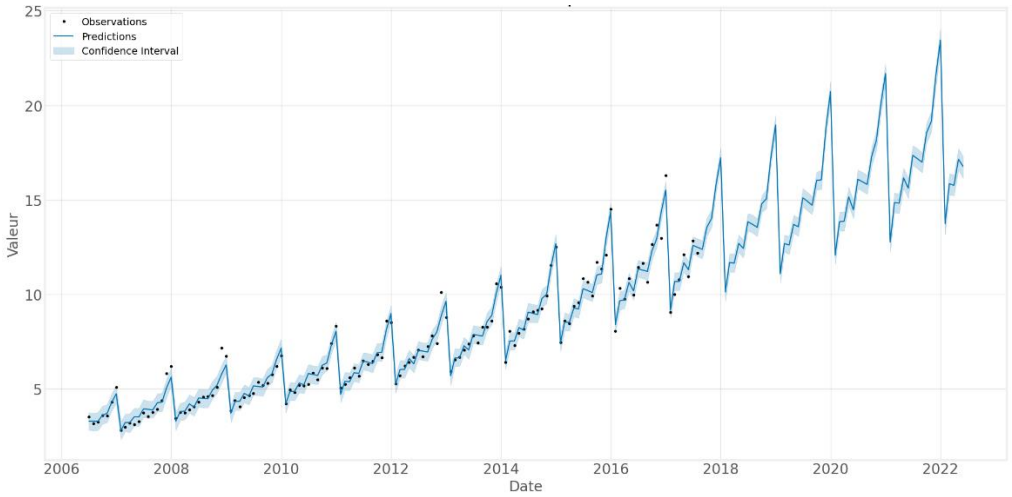
Graphique 13 : Décomposition multiplicative de la série par la méthode Prophet

Le graphique ci-dessus présente une vue détaillée des deux principaux aspects qui influencent nos prévisions : la tendance (trend) et la saisonnalité annuelle. Ces composantes sont essentielles pour comprendre les dynamiques sous-jacentes des données temporelles et pour identifier les facteurs clés qui affectent le phénomène étudié.

La première composante, la tendance, est représentée dans le premier sous-graphe. Cette courbe montre une croissance continue et linéaire des valeurs de 2006 à 2022. La courbe ascendante indique une augmentation régulière au fil du temps, sans fluctuations majeures ou ruptures dans la tendance.

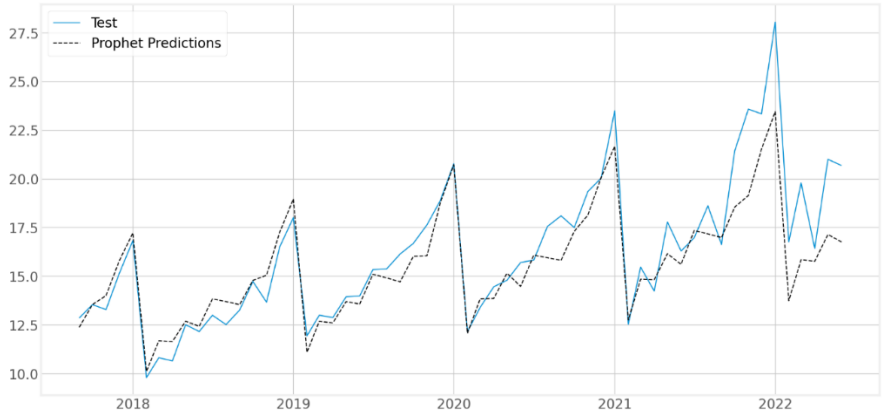
Le second sous-graphe illustre la composante de saisonnalité annuelle, qui capture les variations périodiques récurrentes au cours d'une année typique. La courbe de saisonnalité montre quelques pics et des creux à des moments spécifiques de l'année. En effet, on observe un pic significatif juste avant le début de l'année, autour de janvier, ce qui indique une activité accrue pendant cette période. Des pics moins prononcés apparaissent également autour de juillet et novembre, suggérant d'autres périodes d'activité accrue. Ces variations saisonnières peuvent être attribuées à des événements récurrents tels que des saisons de vente spécifiques.

Une fois le modèle ajusté, un DataFrame contenant les dates futures pour lesquelles les prévisions sont souhaitées est créé. Prophet génère alors des prévisions en appliquant les composantes de tendance, de saisonnalité. Ces prévisions nous permettront de vérifier la capacité du modèle à prédire en faisant une comparaison avec notre échantillon de validation. Les résultats incluent non seulement les valeurs prédites mais aussi les intervalles de confiance, fournissant ainsi une estimation de l'incertitude. Les prévisions sont ensuite visualisées à l'aide de graphiques qui montrent les données historiques, les prévisions futures, ainsi que les différentes composantes du modèle.



Graphique 14 : Représentation des données originales et des prévisions

Les résultats obtenus sur l'ensemble test peuvent être visualisés sur le graphique suivant :



Graphique 15 : Prévision avec la méthode Prophet

On peut constater que l'algorithme de Prophet Facebook ajuste assez bien les données de l'échantillon de validation.

Les performances du modèle sont présentées comme suit :

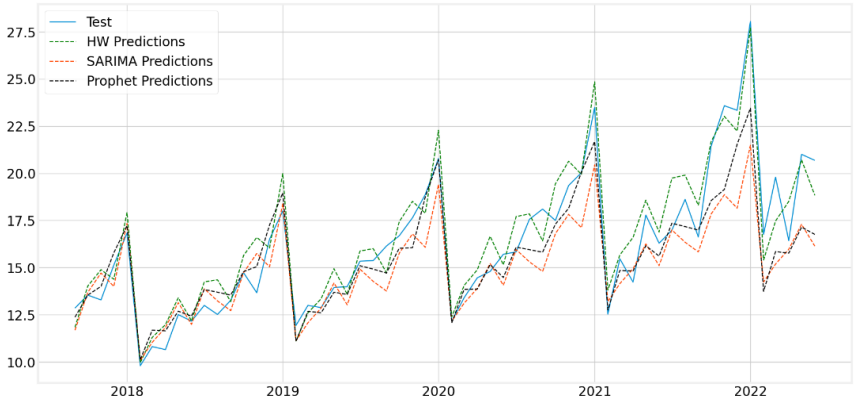
Tableau 7 : Mesures de performance

MSE	MAPE	MAE
2.48	*6.03%*	*1.07*

Les résultats obtenus sont généralement fiables et restent proches de ceux produits par la méthode de lissage exponentiel triple de Holt-Winters, à l'exception de la Mean Squared Error (MSE).

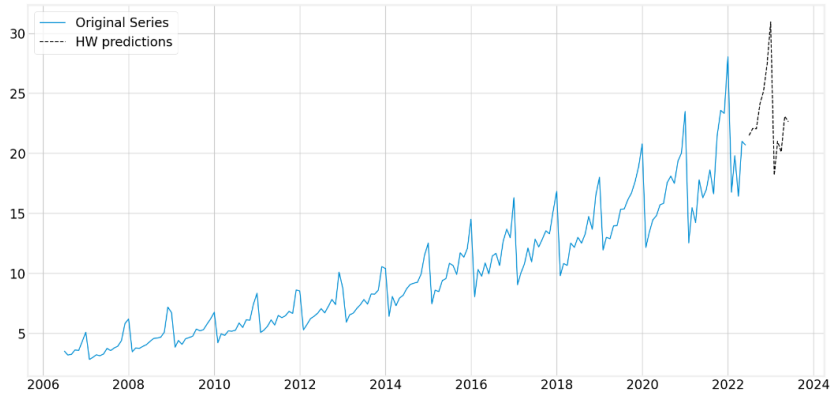
6. Méthode retenue pour la prédiction

Dans le cadre de notre étude visant à améliorer la précision des prévisions de ventes, nous avons évalué trois méthodes de prévision différentes : SARIMA, Holt-Winters et Prophet. Après une analyse comparative basée sur la MSE, MAPE et MAE, la méthode Holt-Winters démontre une meilleure adéquation aux données historiques, ce qui se traduit par une capacité supérieure à minimiser les erreurs de prévision sur notre ensemble de données. Ce résultat souligne l'efficacité de la méthode Holt-Winters, particulièrement adaptée à nos séries temporelles comportant des composantes saisonnières marquées.



Graphique 16 : Comparaison Holt-Winters SARIMA et Prophet

La prédiction des ventes pour la période juillet 2022 à juin 2023 basé sur la méthode de Holt-Winters se présente comme suit :



Graphique 17 : Prédiction des ventes de juillet 2022 à juin 2023

Tableau 8 : Prédictions réalisées avec la méthode Holt-Winters

Années	*Prédictions*
2022-07-01	*21,49173*
2022-08-01	*22,07004*
2022-09-01	*22,04203*
2022-10-01	*24,08863*
2022-11-01	*25,17671*
2022-12-01	*27,36163*
2023-01-01	*30,94066*
2023-02-01	*18,18151*
2023-03-01	*21,00422*
2023-04-01	*20,11337*
2023-05-01	*23,08448*
2023-06-01	*22,63392*

6.2 Synthèse

Ce document détaille l'application de trois méthodes de prévision statistique majeures—SARIMA, Holt-Winters, et Prophet pour déterminer laquelle fournit la meilleure précision sur les données de vente mensuelles de juillet 2006 à juin 2022.

La première section, l'analyse descriptive, met en évidence une tendance croissante des ventes avec des fluctuations saisonnières significatives, illustrée par des graphiques détaillés et des analyses de tendance et de saisonnalité. L'importance des résidus, leurs analyses via des tests statistiques comme le test de Fisher, et l'identification des valeurs aberrantes montrent la rigueur de l'approche analytique.

La section suivante discute des différentes structures de modèles de décomposition (additif, multiplicatif, et mixte) et sélectionne le modèle multiplicatif comme le plus adéquat, basé sur les caractéristiques des données traitées. Les résultats sont enrichis par des analyses graphiques, offrant une vue claire des tendances et saisonnalités.

L'évaluation des modèles de prévision révèle que la méthode de Holt-Winters, avec la plus faible erreur quadratique moyenne (MSE), surpasse les autres méthodes dans l'exactitude des prévisions. Le rapport conclut que, parmi les méthodes testées, Holt-Winters est la plus appropriée pour intégrer dans les processus décisionnels de l'entreprise pour la prévision des ventes futures.

Ce travail met non seulement en avant l'efficacité comparative des méthodes de prévision mais sert également de guide pour leur implémentation pratique, assurant ainsi que les stratégies de vente de l'entreprise sont fondées sur des données prédictives solides et fiables.