UNIVERSITÉ DE LORRAINE

eazee link

# PROJET LONG : APPROCHES STATISTIQUES POUR UNE STARTUP

**KOWOU Kossi Julien**

Dirigé par **Mme Anne GEGOUT-PETIT**

PROJET TUTORÉ

**Master 2 en Ingénierie Mathématique pour la Science des Données**
**Faculté des Sciences et Technologies (FST)**

FST Faculté des Sciences et Technologies Nancy

# Sommaire 2

# Sommaire

Vazee, une startup française, actuellement appelé Eazee-Link œuvre dans le secteur de la consommation. Son objectif c'est de faciliter la relation entre les marques, les établissements de restauration et les consommateurs .

Pour améliorer les solutions fournies aux marques, Eazee-Link veut revoir sa méthodologie d'analyse des prix de boissons par marque, intégrant un calcul de prix moyen par contenant.

# Objectifs du projet

- Proposer une technique permettant de reconnaitre le contenant normalisé d'une boisson.
- Déterminer dans quel contexte la moyenne pourrait être employée comme un indicateur de prix d'une boisson (la marque DESPERADOS).
- Proposer une analyse de prix de boisson dans le temps (la marque DESPERADOS).

# Sommaire

# Sélection des boissons

- Une boisson = Un produit de la table products de la base de données vazeeadd-preprod avec un selene-product-id non nul.
- Pricei-label = variables qui contiennent les contenants personnalisés.
- Un contenant normalisé = volume + unité (ex : 25cl)
- Une observation désigne une boisson d'une marque avec une valeur donnée de contenant personnalisé, vendue dans un établissement à un prix et à une date donnée.

# Extraction des données

- Utilisation des expression régulières pour l'extraction de contenants normalisés.
- Le vocabulaire des bières a été utilisé.
- Exemples : bouteille/bte/belles (75cl), pinte (50cl), galopin (125ml)

# Traitement des contenants de boissons

- 300555 observations au total.

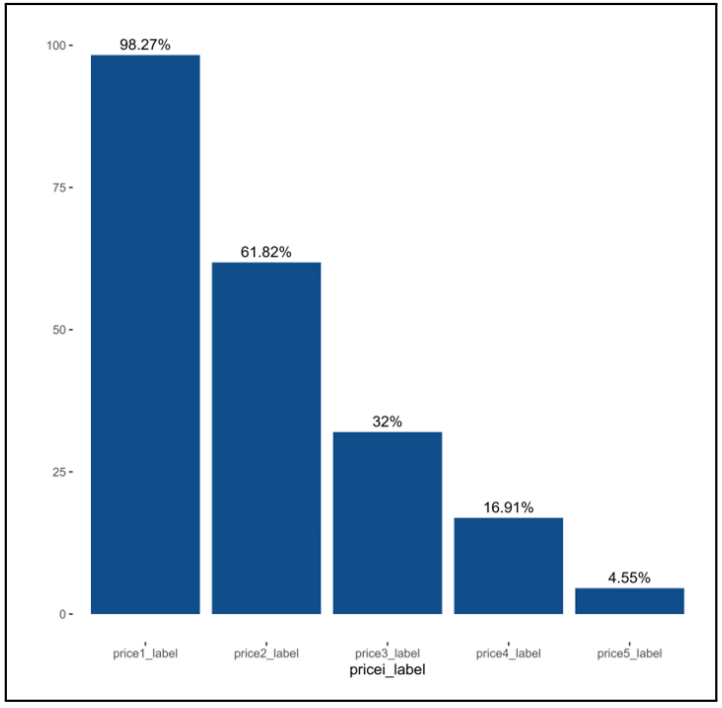


Figure – Taux de pricei-label renseignés

# Traitement des contenants de boissons (suite)

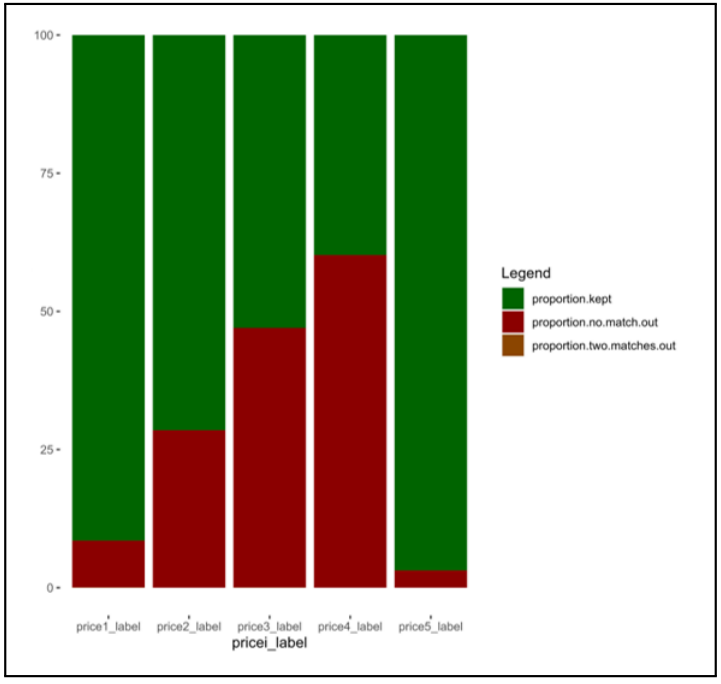


Figure – Proportion d'observations gardées par pricei-label, i$\in [1, 5]$

- 248851 observations ont pu être normalisés (soit 82.08% des contenants personnalisés)

# Sommaire

Etude de la distribution des prix de boissons de la marque DESPERADOS pour les contenants normalisés 25cl, 50cl.

# DESPERADOS 25cl

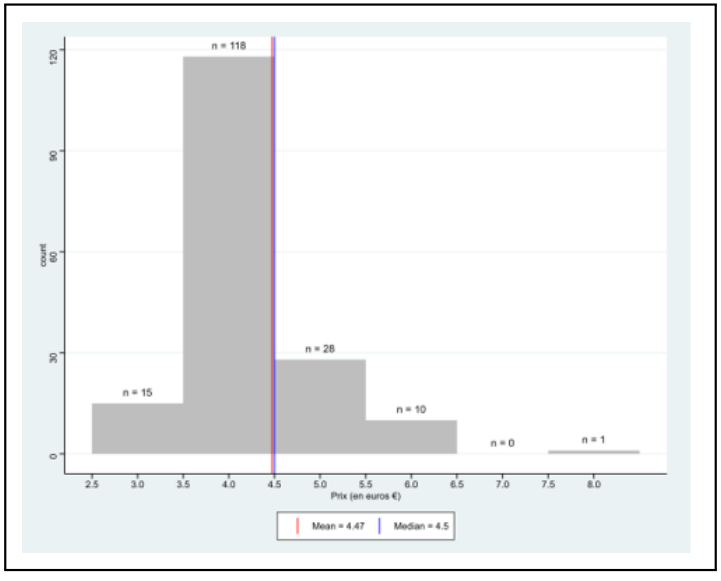


Figure – Histogramme des prix du produit DESPERADOS 25cl.

Une moyenne n'est pas adaptée comme indicateur de prix du produit DESPERADOS 25cl.
50% des établissements vendent le produit DESPERADOS 25cl pour un prix inférieur ou égal à 4.5 euros.

# DESPERADOS 50cl

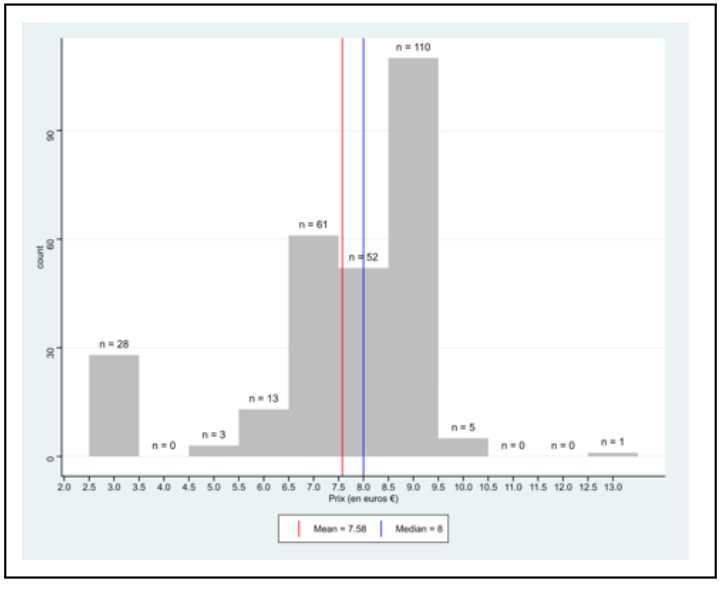


Figure – Histogramme des prix du produit DESPERADOS 50cl.

Une moyenne n'est pas adaptée comme indicateur de prix du produit DESPERADOS 50cl.
50% des établissements vendent le produit DESPERADOS 50cl pour un prix inférieur ou égalou égal à 8 euros.

# conclusion partielle

- La moyenne ne peut être utilisée comme indicateur des prix pour aucun des produits de la marque DESPERADOS ayant pour contenant 25cl, 33cl ou 50cl.
- Utiliser la médiane ou les quartiles.

# Modèles

→ Expliquer le prix du produit DESPERADOS 33cl à l'aide de certaines covariables.
→ Deux modèles de régression classiques : forêt aléatoire et XGBoost.
→ Séparation des données en deux : un échantillon d'apprentissage (70% des observations) et un échantillon de test (30%)

# Forêt aléatoire

En moyenne, le modèle explique 73.62% des informations du prix, et fait une erreur de 0.3 en termes de MSE.

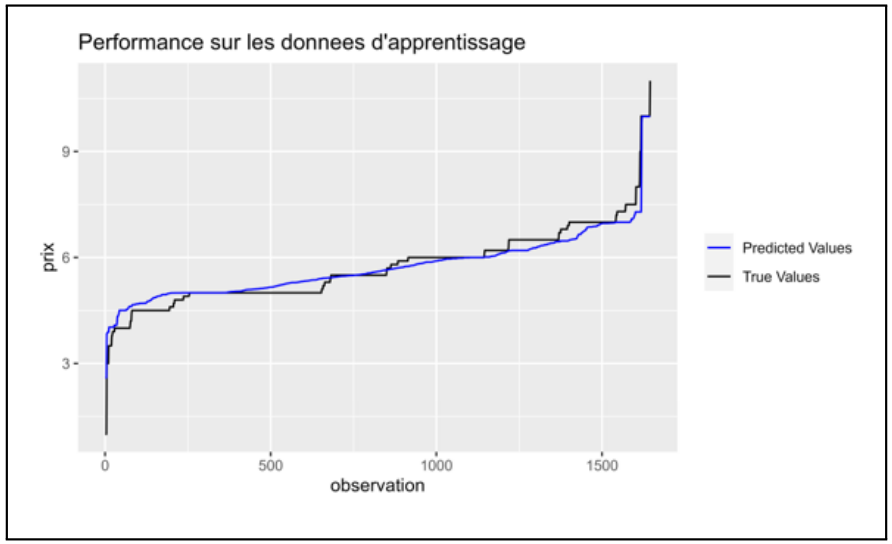


Figure – Performances de la forêt aléatoire sur les données d'apprentissage.

# Forêt aléatoire (suite)

Après la prédiction sur l'ensemble test pour évaluer la performance du modèle, le MSE vaut 1.014 et $R^2$ est égal à 19% .

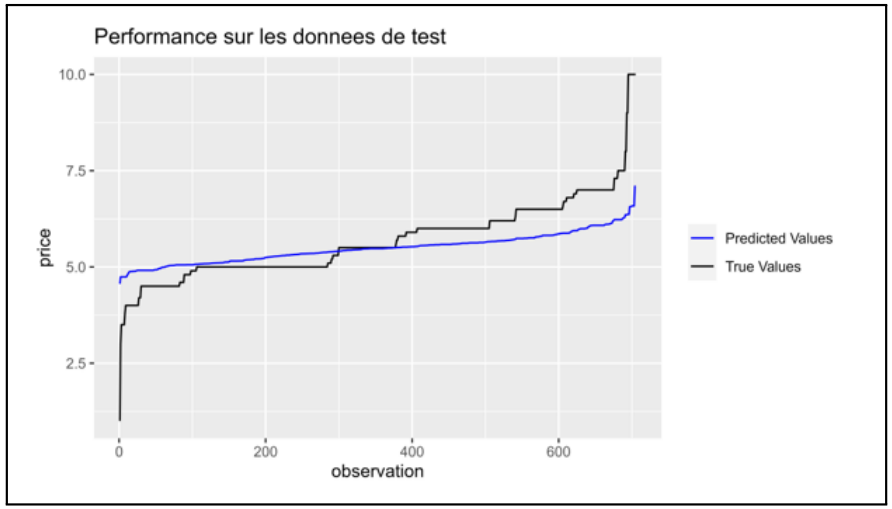


Figure – Performances de la forêt aléatoire sur les données de test.

- Ce modèle surapprend, il est donc rejeté.

# XGBoost

→ Recodage des variables en utilisant la méthode de One-hot encoding qui consiste à créer une colonne binaire pour chaque modalité possible de chaque variable qualitative.

| | Ens. d'apprentissage | Ens. test |
|---|---|---|
| MSE | 0.06 | 0.26 |
| $R^2$ | 95% | 79% |

→ Le modèle souffre d'un surapprentissage.

→ Paramétrages nécessaires sur le modèle avant l'entrainement, son erreur de généralisation en termes de MSE vaut 0.33, tandis qu'à l'apprentissage le MSE vaut 0.21.

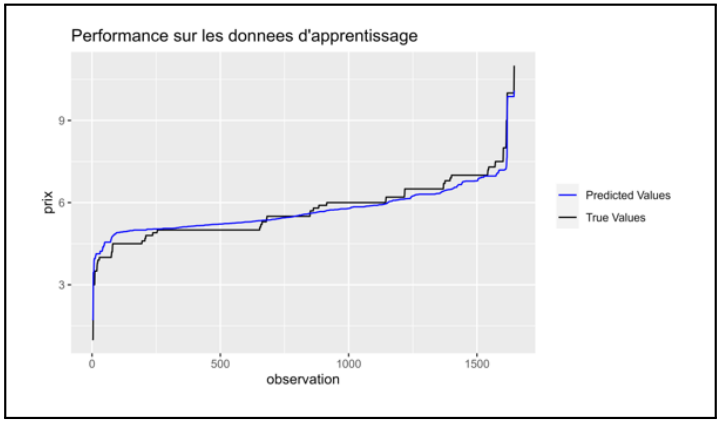


Figure – Performance du XGBoost sur les données d'apprentissage.

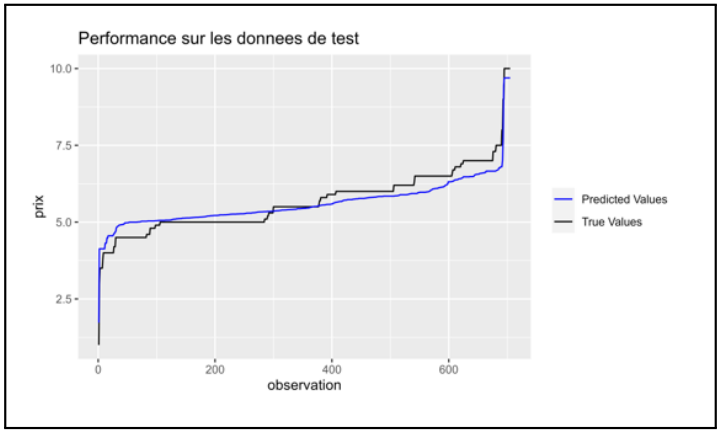


Figure – Performance du XGBoost sur les données de test.

# comparaison

Le modèle du XGBoost surapprend moins que celui de la forêt aléatoire :

| | MSE apprentissage | MSE test |
|---|---|---|
| XGBoost | 0.21 | 0.33 |
| Forêt aléatoire | 0.3 | 1.014 |

$\rightarrow$ XGBoost arrive à régler un peu le problème de surapprentissage.

# Sommaire

# Modèle de série temporelle

- 30 observations de prix moyen hebdomadaire de Septembre 2021 à Mai 2022.
- Visualisons ces données :

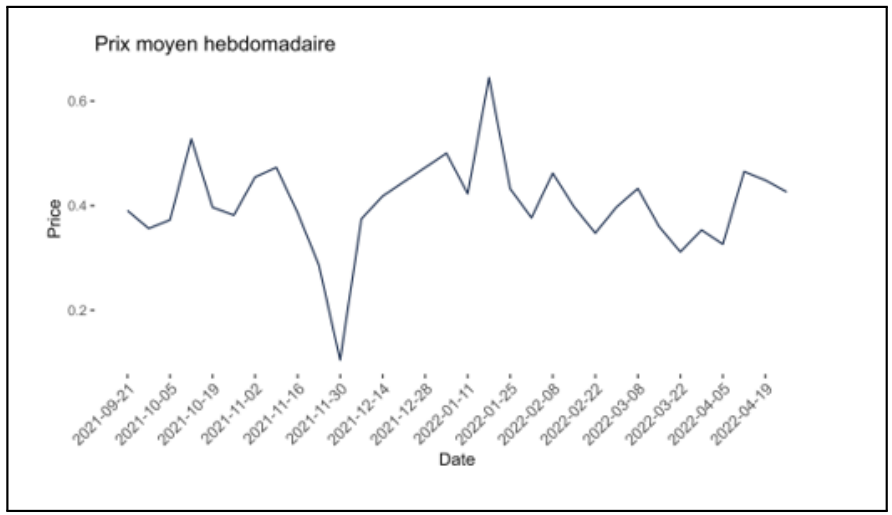


Figure – Evolution du prix moyen hebdomadaire de Septembre 2021 à Mai 2022.

- Modèle SARIMA$_s[(p, d, q)(P, D, Q)]$.

Présence d'une saisonnalité dans la série. Nous désaisonnalisons la série brute en utilisant la fonction d'autocorrélation.

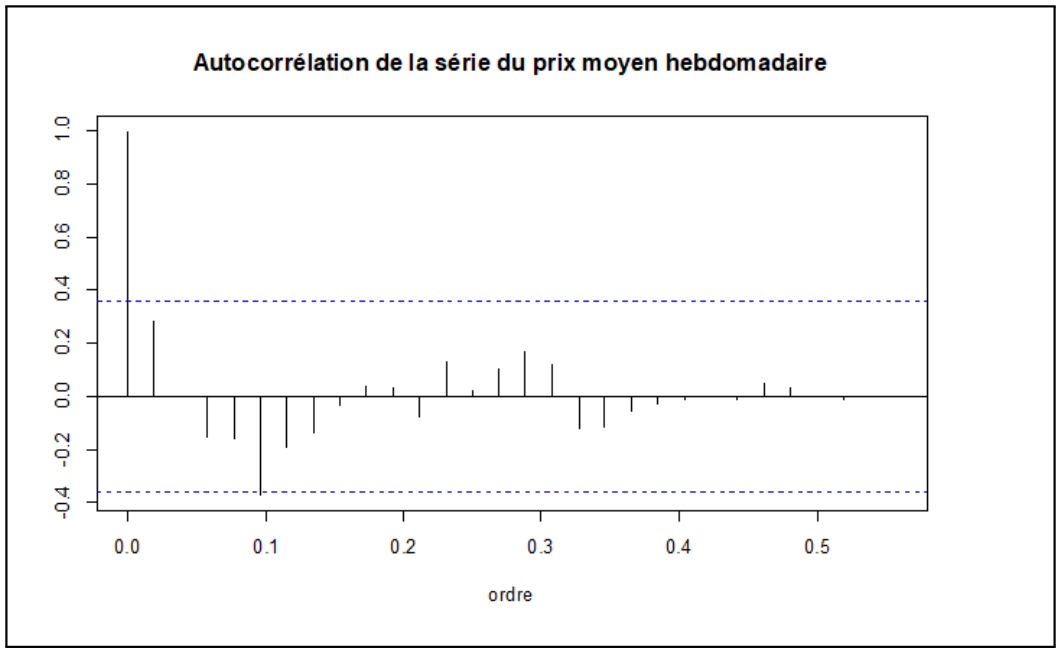


Figure – Autocorrélation de la série du prix moyen hebdomadaire.

Nous construisons la fonction d'autocorrélation et la fonction d'autocorrélation partielle de la série désaisonnalisée.
On remarque la série ne présente plus de tendance.

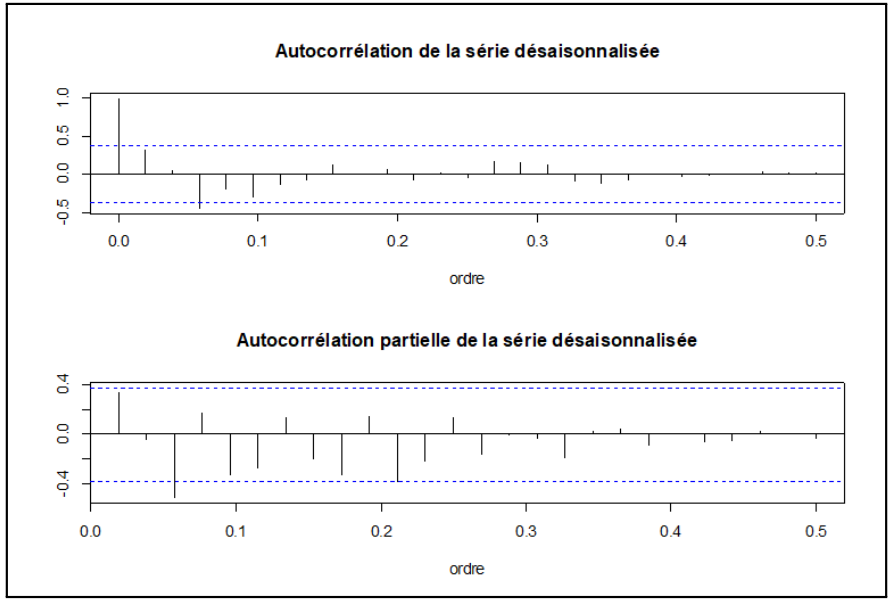


Figure – Autocorrélation et autocorrélation partielle pour la série du prix moyen hebdomadaire désaisonnalisée.

On détermine les ordres de p, P, q et Q .
Pmax = 1, pmax = 3, Qmax = 1, qmax = 3

Le modèle qui a l'AIC le plus petit est $\text{SARIMA}_3[(0,0,1)(0,0,1)]$.

Pour valider ce modèle étudions les résidus :
$\rightarrow$ Le test de Box-Pierce de niveau de confiance 95% a été réalisé (pv =0.62), les résidus sont non correlés.
$\rightarrow$ Le test de Shapiro-wilk de niveau de confiance 95% a été réalisé (pv = 0.01). Les résidus ne sont pas de loi normales.

Nous validons ce modèle et visualisons la prévision du modèle sur les 30 observations de Septembre 2021 à Mai 2022.

# Prévision

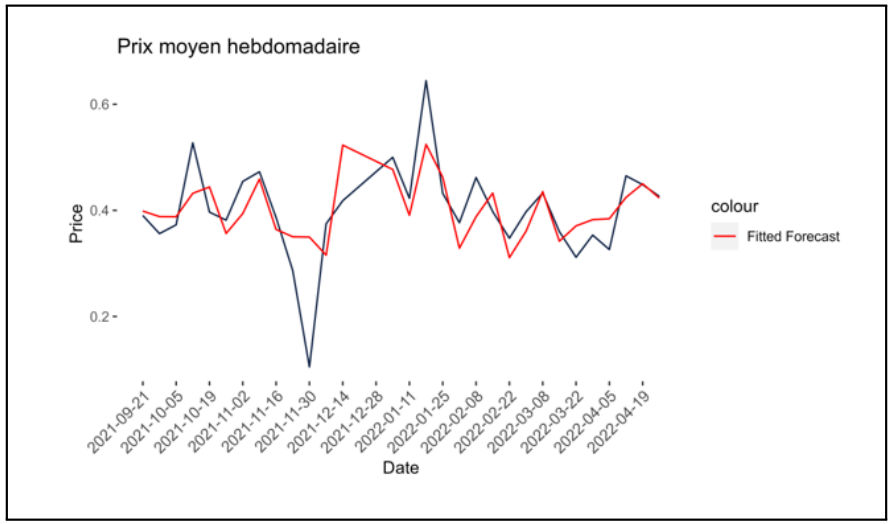


Figure – Autocorrélation et autocorrélation partielle pour la série du prix moyen hebdomadaire désaisonnalisée.

Pour modéliser l'évolution du prix moyen hebdomadaire de la marque DESPERADOS, le modèle $\text{SARIMA}_3[(0,0,1)(0,0,1)]$ *est choisi*.

# Sommaire

# Conclusion

- Sur les observations disponibles , on pu faire correspondre uniquement 82.08% d'entre elles à un contenant normalisé.
- L'analyse des histogrammes montre que la moyenne ne peut être utilisée comme indicateur de prix.

- Pour modéliser l'évolution du prix moyen, bien qu'il commet quelques erreurs on garde le modèle $\text{SARIMA}_3[(0,0,1)(0,0,1)]$.

**Merci pour votre attention ! ! !**