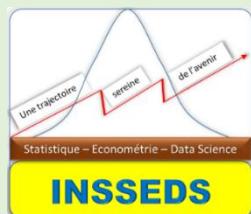
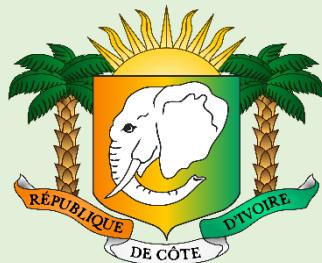


MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE RECHERCHE SCIENTIFIQUE



INSTITUT SUPERIEUR DE STATISTIQUE
D'ECONOMETRIE ET DE DATA SCIENCE

REPUBLIQUE DE COTE D'IVOIRE



MASTER 2

STATISTIQUE – ECONOMETRIE – DATA SCIENCE

MINI PROJET

STATISTIQUE DESCRIPTIVE

ANALYSE DESCRIPTIVE DE SERIE TEMPORELLES ET
PREVISION PAR LISSAGE EXPONENTIELLE

ANNEE ACADEMIQUE

2024-2025

ETUDIANT

NOM : SEKA

PRENOM : ISRAEL ASSI JUNIOR
CHRIST

ENSEIGNANT-ENCADREUR

**AKPOSSO DIDIER
MARTIAL**

SOMMAIRE

Dédicace	3
Remerciements	4
Avant-propos	5
Introduction	6
Partie I : PRETRAITEMENT ET TRANSFORMATION EN SERIE TEMPORELLE	8
I- Prétraitemet	8
1- Lecture de donnée	8
2- Formatage de la date.....	8
3- Gestion des valeurs non collectées plus poussées	9
4- Extraction des lignes contenant des NA	10
5- Vérification des doublons	10
6- Agrégation des données par mois	11
II- TRANSFORMATION EN SERIE TEMPORELLE	12
1- Affichage de la série temporelle agrégé.....	12
2- Tracé de l'évolution de la consommation électrique	13
Partie II : ETUDE DESCRIPTIVE	15
I- Analyse univarié des données	15
1- Tableau statistique	15
2- Vérification de la variance et de la moyenne	16
3- Histogramme de la consommation électrique	17
4- Boxplot avec quartiles et valeurs extrême mis en évidence	18
5- Test de stationnarité	19
5-1 Vérification Visuelle de la Stationnarité avec ACF/PACF	20
6- Décomposition en série temporelle	21
II- Analyse Bivarié	23
1- Ajout de variables temporelles	23
2- Matrices de corrélation	24

Partie III : PREVISION POUR LES 24 PROCHAINS MOIS PAR LA METHODE DE HOLT WINTERS	27
1- Test de normalité	27
1-1 Test graphique de normalité des résidus	27
1-2 Test statistique Ljung-Box	28
1-3 Test statistique de normalité des résidus (Shapiro-Wilk)	29
2- Histogramme des résidus avec densité superposée	29
2-1 Visualisation avancée des résidus	30
3 Ajustement du modèle holt winters	31
4 Prévision de la série	31
TABLEAU DE BORD	33
CONCLUSION	34
CODE SOURCE-----	35



DEDICACES

Avant tout propos, je tiens à dédier ce rapport au **DIEU Tout Puissant** qui a permis que je puisse faire ce mini projet en ouvrant mon esprit et donnant la concentration. Je dédie également ce rapport à ma famille particulièrement à ma mère **BOTTI BROU LYDIE PATRICIA**, ainsi qu'à tous ceux qui m'ont soutenu tout au long de cette expérience. Ils m'ont inculqué les valeurs d'intégrité, de sagesse et d'humilité.

REMERCIEMENTS

Commençant cet humble écrit, je voudrais remercier tout particulièrement :

- ✓ Mr AKPOSSO DIDIER MARTIAL , professeur principal et Directeur de l'INSSEDS dont l'encadrement et les enseignements nous ont permis d'aborder ce projet avec rigueur et méthodologie.
- ✓ Mes amis experts en statistique de l'INSSEDS

Avant-propos

L'analyse des séries temporelles occupe une place essentielle dans la compréhension et l'anticipation des phénomènes évolutifs, notamment dans le domaine de la consommation énergétique. Ce mini-projet s'inscrit dans cette démarche en proposant une étude approfondie des tendances de consommation d'électricité et une prévision sur les mois à venir à l'aide de la méthode du lissage exponentiel de Holt-Winters.

L'objectif de ce travail est double : d'une part, explorer et analyser les données historiques de consommation d'électricité afin d'identifier les tendances et les variations saisonnières, et d'autre part, construire un modèle prédictif permettant d'anticiper la demande future. Pour ce faire, nous mobiliserons des outils d'analyse statistique et de programmation, en exploitant des logiciels tels que R, Python, Excel ou Power BI.

Ce projet est non seulement une opportunité d'approfondir nos connaissances en analyse des séries temporelles, mais aussi une mise en application concrète des méthodes de prévision dans un domaine critique. L'électricité étant une ressource stratégique, une meilleure anticipation de sa consommation permet d'optimiser la gestion des ressources et d'améliorer l'efficacité énergétique.

Nous espérons que ce travail apportera un éclairage pertinent sur l'évolution de la consommation d'électricité et constituera une base solide pour des analyses plus approfondies.

INTRODUCTION

L'évolution de la consommation d'électricité est un enjeu majeur dans le contexte actuel de transition énergétique et d'optimisation des ressources. Comprendre les tendances de cette consommation et être en mesure de les anticiper est essentiel pour une meilleure gestion de la production et de la distribution d'électricité.

Ce mini-projet s'inscrit dans cette problématique en proposant une analyse descriptive des séries temporelles de consommation d'électricité ainsi qu'une prévision des 24 prochains mois à l'aide de la méthode de lissage exponentiel de Holt-Winters. En utilisant un ensemble de données couvrant la période du **1^{er} janvier 1985 au 1^{er} janvier 2018**, nous chercherons à identifier les tendances générales, la saisonnalité et les fluctuations éventuelles de la consommation.

Pour ce faire, nous suivrons une méthodologie en plusieurs étapes :

1. **Approche méthodologique de donnée** afin d'obtenir une vue mensuelle claire de la consommation électrique.
2. **Analyse descriptive** des séries temporelles pour dégager les principaux facteurs influençant l'évolution des consommations.
3. **Modélisation et prévision** à l'aide de la méthode de Holt-Winters, un outil statistique puissant permettant de prendre en compte les tendances et la saisonnalité dans les prédictions.

L'objectif final est de fournir un modèle prédictif fiable qui pourra être utilisé comme support d'aide à la décision pour les acteurs du secteur énergétique. Ce travail se veut ainsi une illustration concrète de l'application des méthodes statistiques aux problématiques réelles, tout en mettant en avant l'importance des séries temporelles dans la prise de décision.

PREMIERE
PARTIE

**prétraitement et
transformation en
série temporelle**

PARTIE 1 : PRETRAITEMENT ET TRANSFORMATION EN SERIE TEMPORELLE

I- Prétraitement

1- Lecture de donnée

La lecture de la donnée consiste à importer le jeu de donnée ELECTRIC_PRODUCTION qui est sur format csv dans un logiciel statistique R.

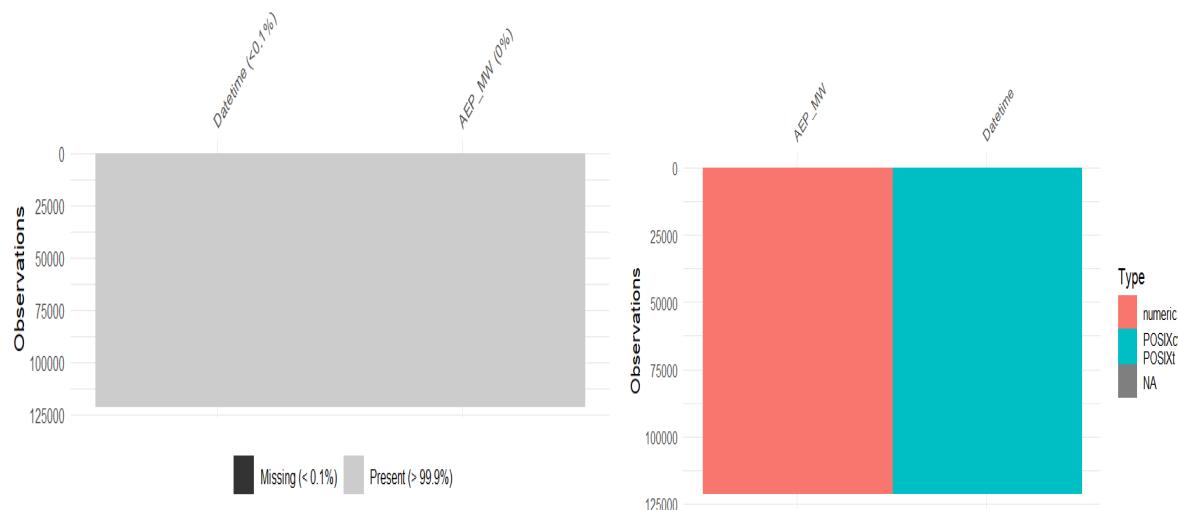
2- Formatage de la date

	DATETIME	AEP_MW
1	2004-12-31 01:00:00	13478
2	2004-12-31 02:00:00	12865
3	2004-12-31 03:00:00	12577
4	2004-12-31 04:00:00	12517
5	2004-12-31 05:00:00	12670
6	2004-12-31 06:00:00	13038
7	2004-12-31 07:00:00	13692
8	2004-12-31 08:00:00	14297
9	2004-12-31 09:00:00	14719
10	2004-12-31 10:00:00	14941

On constate que dans notre jeu de donnée que la date est en format jour mois année. Nous allons formater puis faire l'agrégation mensuelle.

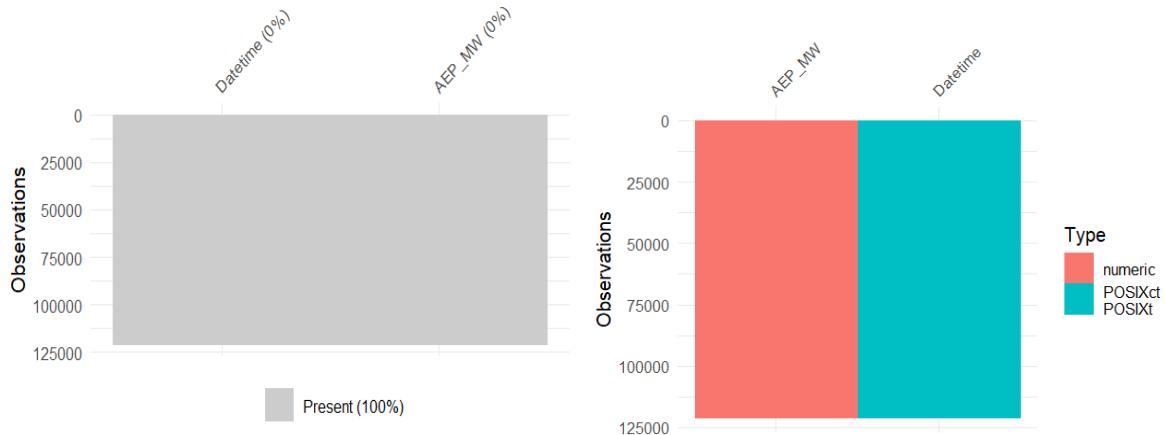
3- Gestion des valeurs non collectées plus poussée

La gestion des valeurs manquantes est essentielle pour garantir la fiabilité des résultats. En fonction du contexte et de l'importance de ces valeurs, nous pourrons choisir l'approche la plus adaptée (suppression, imputation ou autre). Une vérification complémentaire pourrait être envisagée pour identifier les raisons exactes de ces absences et éviter qu'elles ne se reproduisent dans de futures collectes de données.



Lors de l'analyse des données, nous avons identifié 14 valeurs manquantes (NA). Ces absences peuvent avoir un impact sur l'interprétation et la modélisation des données.

4- Extraction des lignes contenant des NA



Après avoir détecté quatorze valeurs manquantes nous avons supprimer ses lignes en question pour mener à bien notre travail.

5- Vérification des doublons

Dans notre jeu de donnée il n'y pas de doublons. Cela indique clairement que les données n'ont pas de répétitions et qu'aucune action de nettoyage liée aux doublons n'est nécessaire.

6- Agrégation des données par mois

	Month	AEP_MW	
1	2004-10-01	13947.54	
2	2004-11-01	14830.44	
3	2004-12-01	16737.72	
4	2005-01-01	17117.09	
5	2005-02-01	16496.64	
6	2005-03-01	15929.31	
7	2005-04-01	14032.42	
8	2005-05-01	13685.07	
9	2005-06-01	16250.33	
10	2005-07-01	16863.87	
11	2005-08-01	17251.09	
12	2005-09-01	15347.14	
13	2005-10-01	14381.38	
14	2005-11-01	15194.88	
15	2005-12-01	17548.59	
16	2006-01-01	16409.84	
17	2006-02-01	17216.03	
18	2006-03-01	15970.75	
19	2006-04-01	13930.53	
20	2006-05-01	14442.53	
21	2006-06-01	15583.12	
22	2006-07-01	17005.18	
23	2006-08-01	17400.20	
24	2006-09-01	14375.34	
25	2006-10-01	14807.96	
26	2006-11-01	15425.86	
27	2006-12-01	16304.00	
28	2007-01-01	17328.81	
29	2007-02-01	19212.90	
30	2007-03-01	15999.76	
31	2007-04-01	15156.79	
32	2007-05-01	15213.58	
33	2007-06-01	16782.49	
34	2007-07-01	16687.24	

II- TRANSFORMATION EN SERIE TEMPOREL

Avant de continuer notre analyse il est indispensable de créer l'objet de notre série temporelle puis de l'agrégé.

1- Affichage de la série temporelle agrégé

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
2004									
2005	17117.09	16496.64	15929.31	14032.42	13685.07	16250.33	16863.87	17251.09	15347.14
2006	16409.84	17216.03	15970.75	13930.53	14442.53	15583.12	17005.18	17400.20	14375.34
2007	17328.81	19212.90	15999.76	15156.79	15213.58	16782.49	16687.24	18613.26	16179.22
2008	18574.76	18253.73	16603.69	14988.29	14489.59	16784.52	17373.00	16835.64	15770.71
2009	18653.02	17134.54	15129.33	13974.64	13312.73	14969.76	14697.31	15783.66	14244.79
2010	18164.96	17963.06	15270.04	13537.86	14287.15	16411.99	17347.98	17162.16	14753.01
2011	18314.49	17098.86	15821.67	14130.16	14457.30	15926.22	17693.72	16677.86	14560.83
2012	17016.42	16373.16	14306.17	13784.08	14731.17	15672.85	17299.70	16161.08	14155.33
2013	16755.16	16946.72	15979.33	13795.66	13926.06	15079.51	15951.50	15360.86	14114.08
2014	18449.32	17398.46	15759.56	13338.63	13533.25	15168.38	14901.62	15278.30	13985.10

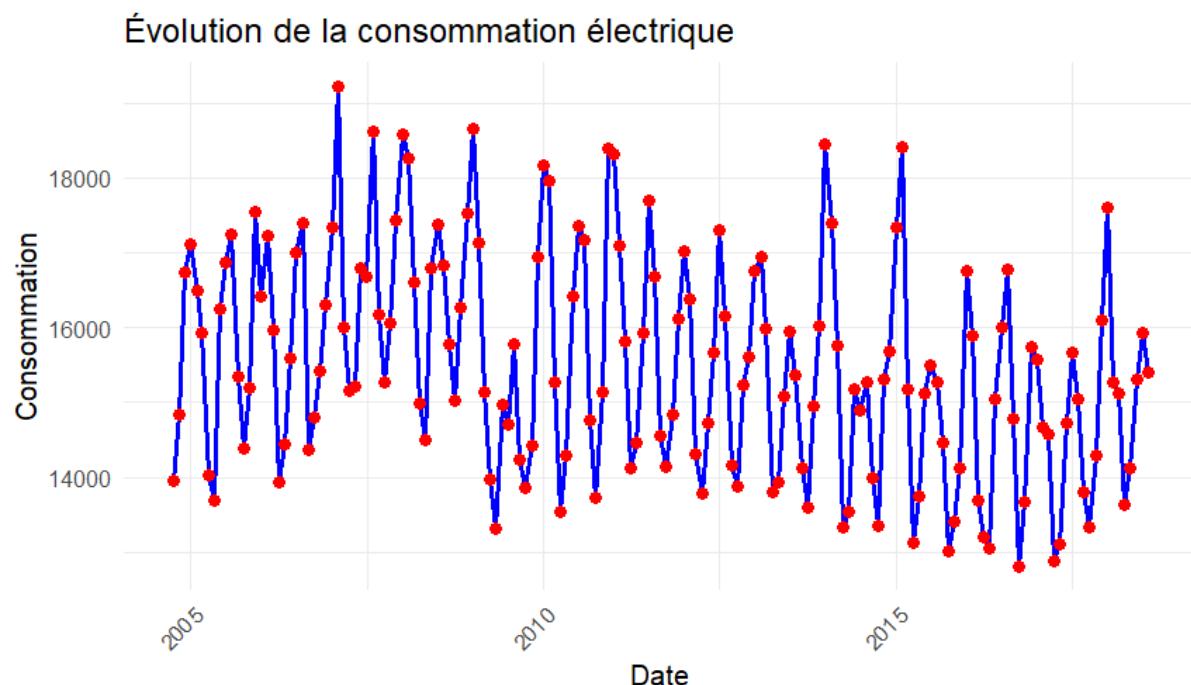
Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

Ce tableau analyse l'évolution mensuelle de la consommation électrique de 2005 à 2018 mais nous avons coupé le tableau (trop long). Les données présentées sont agrégées par mois, ce qui permet d'observer les tendances saisonnières et les variations interannuelles.

- La consommation électrique varie d'un mois à l'autre, avec des valeurs généralement plus élevées en hiver (janvier et février) et plus faibles au printemps et en début d'automne (avril et septembre).
- Les années 2008, 2009 et 2011 enregistrent des pics de consommation particulièrement élevés en hiver.
- Une légère tendance à la baisse peut être observée sur certaines années récentes, notamment après 2012.

- **Années de forte consommation :** 2008, 2009 et 2011 enregistrent des niveaux élevés de consommation annuelle, suggérant des hivers rigoureux ou une augmentation de la demande énergétique.
- **Années de consommation plus modérée :** Après 2012, une légère diminution de la consommation est visible, ce qui peut être attribué à des améliorations en efficacité énergétique, des changements économiques, ou des conditions climatiques plus clémentes.

2-Tracé de l'évolution de la consommation électrique



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

Ce graphique représente l'évolution de la consommation électrique au fil du temps. Voici quelques observations :

1. **Tendance globale** : On observe une tendance légèrement décroissante de la consommation électrique après 2010, avec des fluctuations saisonnières marquées.
2. **Saisonnalité** : La consommation semble suivre un cycle régulier, avec des pics et des creux qui pourraient être liés aux variations saisonnières (hiver vs été).
3. **Marqueurs et courbe** : La courbe bleue représente la variation de la consommation dans le temps, tandis que les points rouges semblent marquer les observations individuelles.
4. **Variabilité** : Il y a une forte variabilité de la consommation, avec des pics de forte demande, probablement en hiver.

DEUXIEME
PARTIE

ETUDE DESCRIPTIVE

PARTIE 2 : ETUDE DESCRIPTIVE

I- ANALYSE UNIVARIE DES DONNEES

1- Tableau statistique

MIN	1ST QUAN	MEDIANE	MEAN	3rd QUAN	MAX
12805	14341	15305	15503	16683	19213

Ces statistiques résument une série temporelle et donnent un aperçu de la distribution des valeurs observées. Voici une interprétation.

- **Minimum (12 805)**

→ La consommation la plus basse enregistrée. Cela pourrait correspondre à une période de faible activité, comme une nuit, un jour férié ou une période de faible demande.

- **Premier quartile (14 341)**

→ 25 % des observations ont une consommation inférieure à cette valeur. Cela peut refléter les périodes de consommation modérée, typiquement en dehors des heures de pointe.

- **Médiane (15 305)**

→ La médiane indique que 50 % des valeurs sont inférieures et 50 % sont supérieures. Cela représente la consommation "typique" ou centrale sur la période étudiée.

- **Moyenne (15 503)**

→ Une moyenne légèrement supérieure à la médiane, ce qui signifie qu'il y a quelques pics de forte consommation influençant la moyenne.

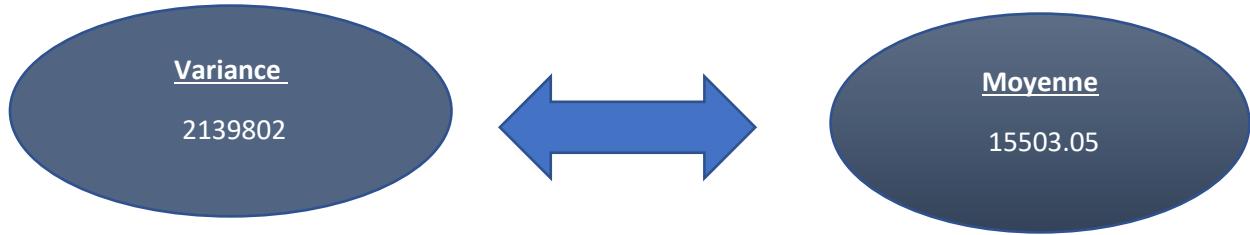
- **Troisième quartile (16 683)**

→ 75 % des observations sont inférieures à cette valeur, ce qui reflète les niveaux de consommation élevés mais encore fréquents, probablement en journée et en début de soirée.

- **Maximum (19 213)**

→ C'est la valeur la plus élevée de consommation électrique. Il pourrait s'agir d'un pic exceptionnel dû à une forte demande, comme une vague de froid.

2- Verification de la variance et de la moyenne pour voir si elle évolue dans le temps



- **Variance = 2 139 802**

→ La variance mesure la dispersion des valeurs autour de la moyenne. Une valeur élevée indique une variabilité importante de la consommation électrique au cours du temps. Cela peut être dû aux variations quotidiennes (jour/nuit), hebdomadaires (jours ouvrés/week-end) ou saisonnières (été/hiver).

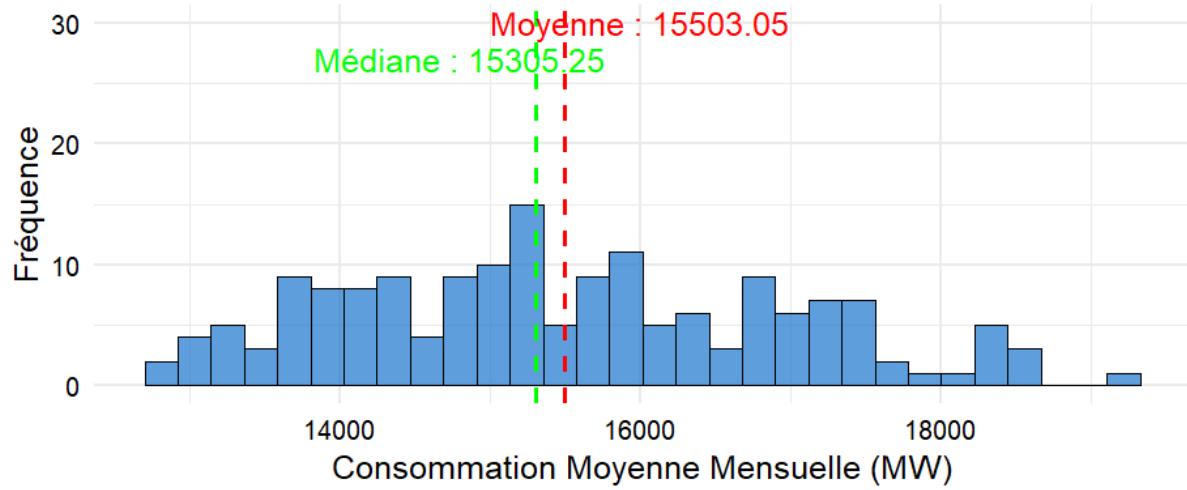
- **Moyenne = 15 503,05**

→ Cela représente la consommation électrique moyenne sur la période observée. Comparée à la médiane (15 305), elle est légèrement plus élevée, ce qui suggère une légère asymétrie vers les valeurs hautes, probablement due à des pics de consommation.

3- HISTOGRAMME DE LA CONSOMMATION ELECTRIQUE

Distribution de la Consommation Électrique Mensuelle

Comparaison entre la moyenne et la médiane pour observer la symétrie de la distribution



Analyse du graphique

La moyenne étant **légèrement supérieure à la médiane**, cela suggère une **asymétrie vers la droite** (distribution légèrement biaisée positivement).

La distribution n'est pas parfaitement symétrique.

Il y a un **étalement plus important vers les valeurs élevées**, ce qui indique la présence de **pics de consommation** qui influencent la moyenne.

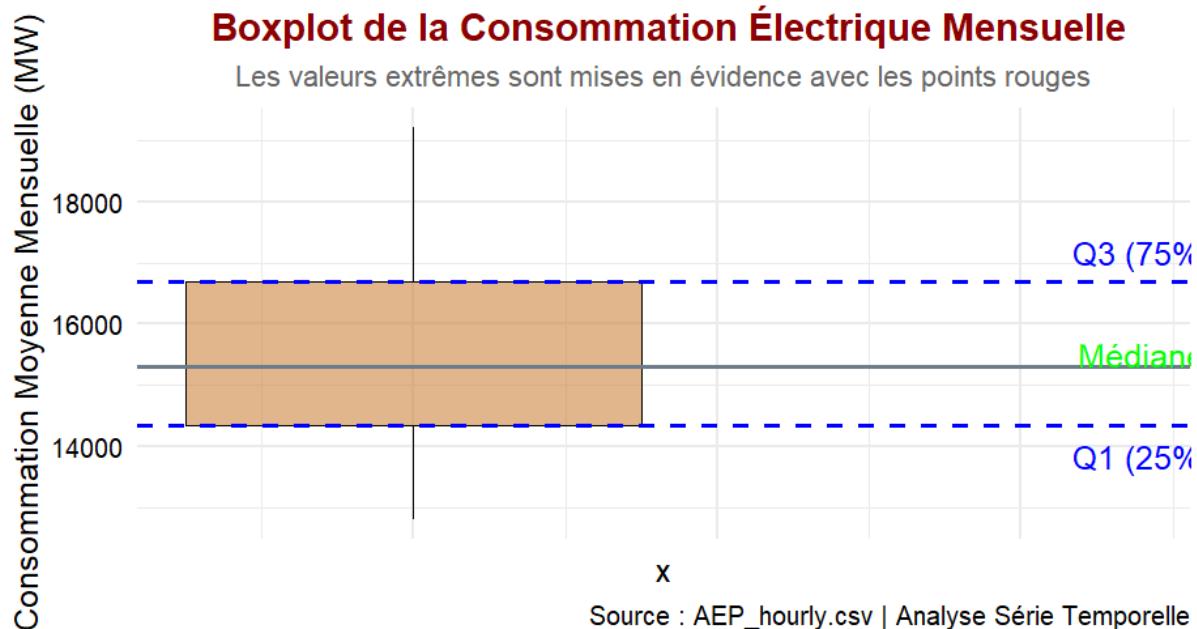
Interprétation et implications

Légère asymétrie à droite : La présence de valeurs élevées (possibles pics de consommation) influence la moyenne.

Distribution relativement étalée : La consommation électrique varie significativement d'un mois à l'autre.

Possibles facteurs externes : La consommation peut être influencée par la saisonnalité, la météo, et l'activité économique.

4- Boxplot avec quartiles et valeurs extrêmes mises en évidence



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

Ce boxplot permet d'analyser la variabilité et la répartition des consommations électriques mensuelles. On peut voir que :

- La médiane se situe vers **15 500 MW**.
- La plupart des valeurs se situent entre **environ 14 000 MW et 16 500 MW**.
- Il n'y a pas d'outliers visibles sur ce graphique.

5- Test de stationnarité (ADF Test)

Un test de stationnarité permet de déterminer si une série temporelle est **stationnaire**, c'est-à-dire si ses propriétés statistiques (moyenne, variance, autocorrélation) restent constantes dans le temps. Donc il est prépondérant de faire le test de stationnarité.

Augmented Dickey-Fuller Test

```
data: ts_data
Dickey-Fuller = -3.5781, Lag order = 5, p-value = 0.03738
alternative hypothesis: stationary
```

Règle de décision :

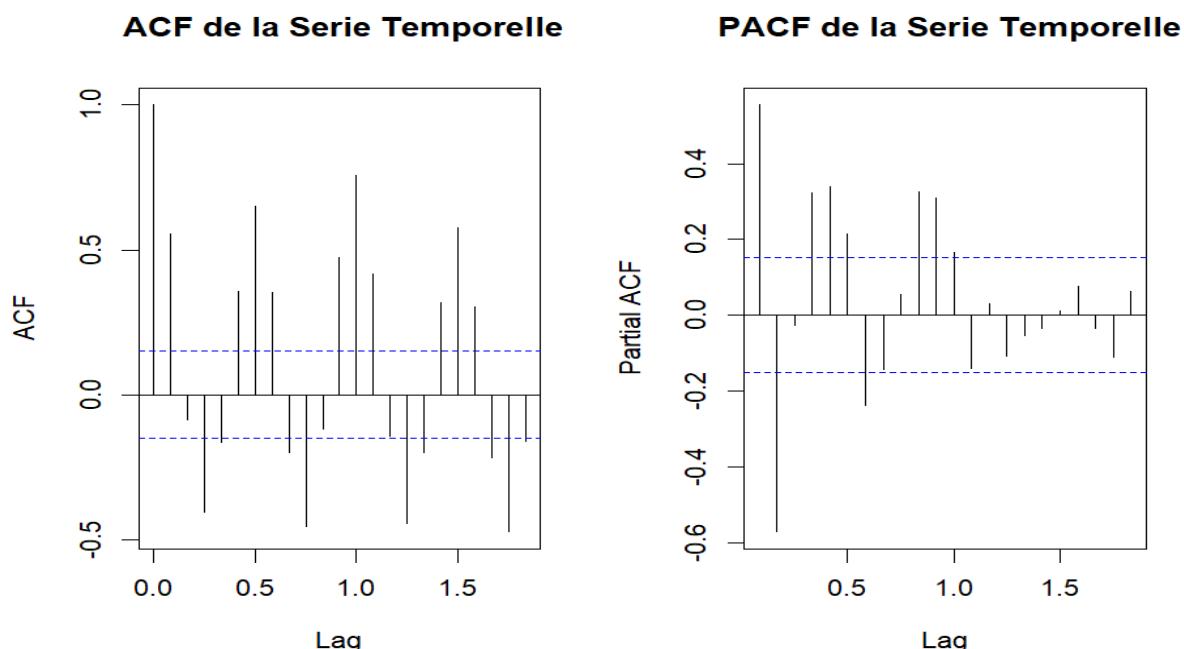
- Si **p-value < 0.05** → On rejette H₀ → **La série est stationnaire**
- Si **p-value > 0.05** → On ne rejette pas H₀ → **La série n'est pas stationnaire**

Conclusion

- La **p-value = 0.03738** est inférieure à **0.05**.
- Donc, **on rejette l'hypothèse nulle H₀** et on conclut que la série **est stationnaire** avec un niveau de confiance de 95 %.

Une fois que notre série est **stationnaire**, on peut utiliser les **ACF (Autocorrélation Function)** et **PACF (Partial Autocorrélation Function)** pour analyser la structure de la dépendance temporelle et déterminer les paramètres d'un modèle ARIMA. Donc c'est ce qu'on fera dans la suite de notre rapport d'étude .

5-1 Vérification Visuelle de la Stationnarité avec ACF/PACF



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

ACF (Autocorrelation Function) - Graphique de gauche

Ici, on observe :

- Une forte autocorrélation au **lag 1**.
- Une décroissance progressive avec quelques valeurs significatives.
- Cela peut indiquer une composante **MA(q)** (Moyenne Mobile).

PACF (Partial Autocorrelation Function) - Graphique de droite

Ici, on remarque :

Un pic significatif au **lag 1**.

Puis une chute rapide après quelques lags.

Cela suggère la présence d'une composante **AR(p)** (Auto-Régressive).

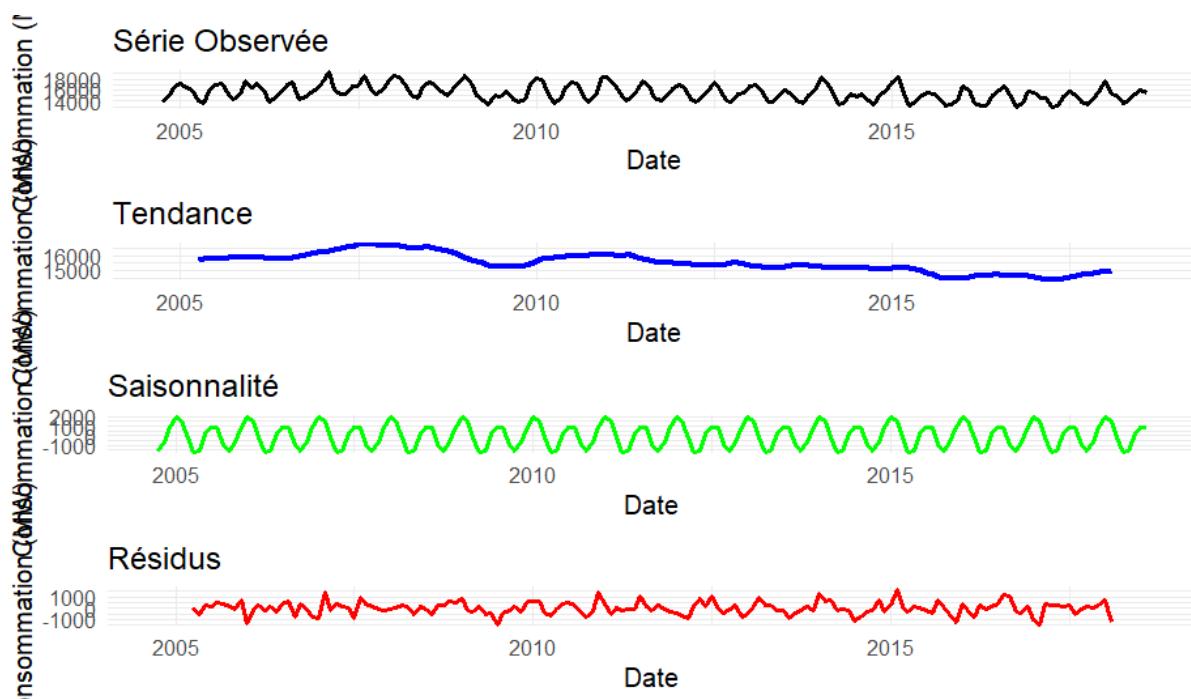
6-Décomposition de la série temporelle

La **décomposition d'une série temporelle** permet d'analyser et de séparer une série en plusieurs composantes fondamentales :

1. **Tendance (Trend)** : Évolution à long terme (hausse ou baisse globale).
2. **Saisonnalité (Seasonality)** : Motifs récurrents périodiques (quotidiens, mensuels, annuels).
3. **Résidu (Residual ou Noise)** : Fluctuations aléatoires non expliquées par la tendance ou la saisonnalité.

📌 Pourquoi faire une décomposition ?

- Mieux comprendre la structure de la série.
- Vérifier si la série contient une **tendance** ou une **saisonnalité**.
- Faciliter la **stationnarisation** en retirant la tendance ou la saisonnalité.
- Aider à choisir un **modèle de prévision** (ex : ARIMA, SARIMA).



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

🔍 Analyse détaillée comme un Data Scientist

Ce graphique montre une **décomposition de série temporelle** appliquée à une série de consommation électrique. Voici une analyse approfondie en plusieurs étapes :

1 Identification des composantes de la série

La décomposition a séparé la série en **quatre éléments** :

1. **Série observée (originale)** → Montre la consommation électrique brute.
2. **Tendance** → Capture l'évolution à long terme.
3. **Saisonnalité** → Capture les variations périodiques récurrentes.
4. **Résidus** → Capturent les fluctuations inexplicables (aléatoires).

2 Analyse de la tendance

- **Observation** : La tendance (courbe bleue) montre une légère hausse au début, suivie d'une **baisse après 2010**, puis une stabilisation.
- **Interprétation** :
 - Cela peut être dû à des **changements structurels** (ex: politiques énergétiques, innovations technologiques, baisse de la consommation industrielle).
 - Il faudrait analyser des **facteurs externes** (ex: météo, évolution de la population, efficacité énergétique) pour expliquer cette tendance.

3 Analyse de la saisonnalité

- **Observation** : La courbe verte montre un **motif saisonnier régulier**, ce qui signifie que la consommation électrique suit une **périodicité stable**.
- **Interprétation** :
 - L'amplitude des cycles **reste constante dans le temps**.
 - Cette régularité suggère que le modèle est **additif** (➡️ **Voir conclusion**).
 - Cela pourrait être lié aux **saisons climatiques**, où la consommation varie en fonction de l'été et de l'hiver.

4 Analyse des résidus

- **Observation :** La courbe rouge montre des résidus qui semblent **aléatoires**, sans tendance apparente.
- **Interprétation :**
 - Si les résidus sont purement aléatoires, cela signifie que la décomposition a bien **capturé la tendance et la saisonnalité**.
 - Si des **patterns apparaissent**, cela indique que le modèle **ne capture pas toutes les informations** et pourrait être amélioré.
 - Ici, les résidus semblent **raisonnablement bien répartis**, donc la décomposition est correcte.

5 Détermination du modèle : Additif ou Multiplicatif ?

Ici, la saisonnalité semble **stable et d'amplitude constante**, donc **le modèle est additif** ($Y = T + S + R$).

II- ANALYSE BIVARIE

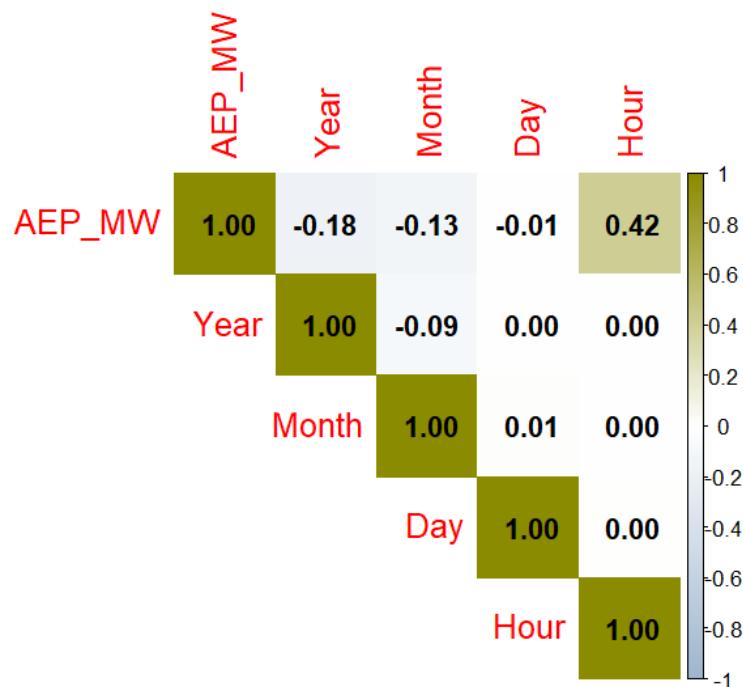
Dans cette deuxième partie du grand 2 nous allons faire une analyse entièrement bivarié en mettant en exergue la matrice de corrélation des différentes variables de notre jeu de donnée. Pour cela de nouvelles variables ont été car au début de notre jeu de donné il y avait des dates que nous n'avons pas intégré précédent, ici nous allons les ajouter.

1-Ajout de variables temporelles

	Datetime	AEP_MW	Year	Month	Day	Hour
1	2004-12-31 01:00:00	13478	2004		12	31
2	2004-12-31 02:00:00	12865	2004		12	31
3	2004-12-31 03:00:00	12577	2004		12	31
4	2004-12-31 04:00:00	12517	2004		12	31
5	2004-12-31 05:00:00	12670	2004		12	31
6	2004-12-31 06:00:00	13038	2004		12	31

Comme nous pouvons le constater nous avons ajouter des variables comme « Month » , « AEP_MW » , « Day » , « Hour ». Afin d'analyser les relations entre les différentes variables de notre étude, nous nous appuierons sur la matrice de corrélation, qui permet d'évaluer la force et la direction des liens entre elles.

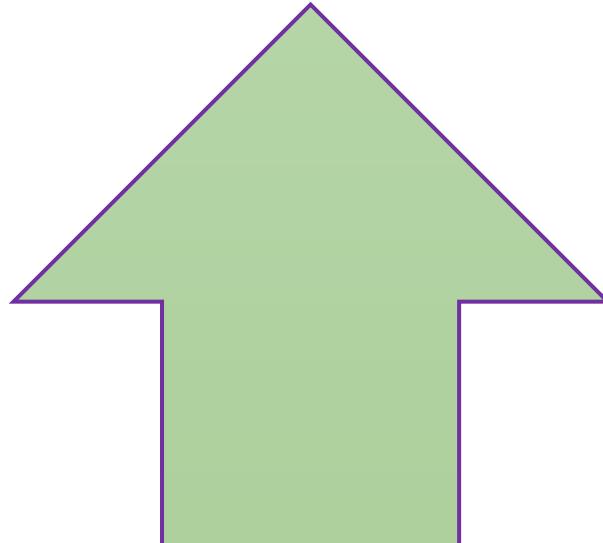
2- Matrice de Corrélation



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

Cette matrice de corrélation montre les relations entre plusieurs variables temporelles (**Year**, **Month**, **Day**, **Hour**) et la variable cible **AEP_MW** (probablement la production d'énergie). Voici quelques observations clés :

1. **Corrélation modérée entre AEP_MW et Hour (+0.42)**
 - Cela suggère que la production d'énergie est influencée par l'heure de la journée. Il est probable qu'il y ait des variations journalières dans la production d'énergie, peut-être en raison de la demande énergétique ou de conditions météorologiques cycliques.
2. **Corrélation négative entre AEP_MW et Year (-0.18)**
 - Une corrélation légèrement négative avec l'année indique une tendance décroissante de la production d'énergie au fil du temps. Cela peut être dû à des changements technologiques, environnementaux ou à des variations dans les conditions de production.
3. **Corrélation faible entre AEP_MW et Month (-0.13)**
 - Une légère corrélation négative avec le mois suggère une certaine saisonnalité, mais l'effet semble limité.



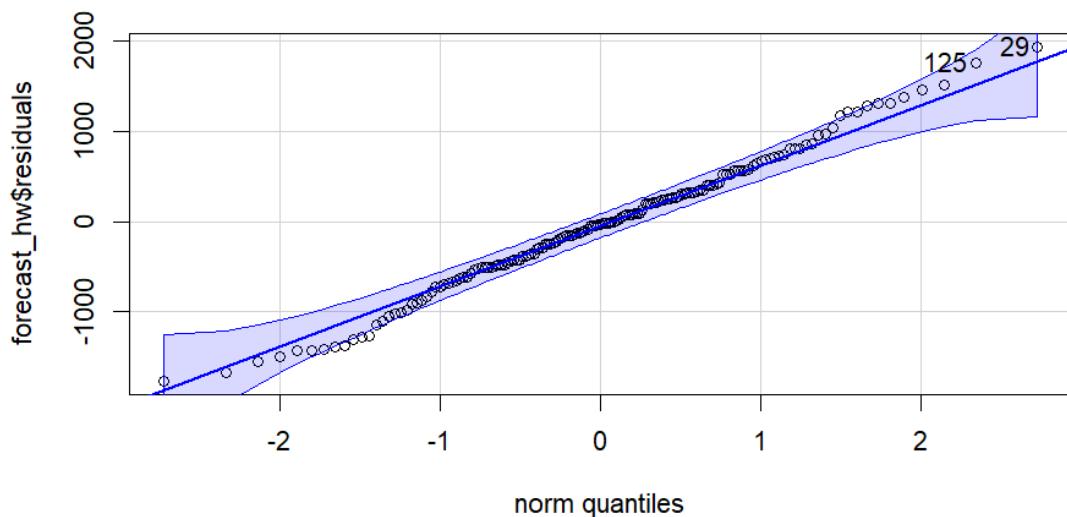
- **Absence de corrélation significative entre AEP_MW et Day (-0.01)**
 - La production d'énergie ne varie pas fortement d'un jour à l'autre dans le mois.
- **Corrélations internes entre les variables temporelles**
 - Comme attendu, **Month, Day et Hour ne sont pas corrélés avec Year**, car ils sont indépendants d'une année à l'autre.
 - **Month et Day ont une très faible corrélation (+0.01)**, ce qui est logique car le jour d'un mois ne dépend pas directement du mois lui-même.
 - **Hour est indépendant des autres variables temporelles (corrélations proches de 0)**.

PARTIE 3 : PREVISIONS POUR LES 24 PROCHAINS MOIS PAR LA METHODE DE HOLT WINTER

1- TEST DE NORMALITÉ

1-1 Test graphique de normalité des résidus

Après avoir analysé les relations entre nos variables à l'aide de la matrice de corrélation, il est essentiel de vérifier une hypothèse clé en modélisation statistique : la normalité des résidus. Pour cela, nous allons recourir à des tests graphiques, qui permettent d'évaluer visuellement si la distribution des résidus suit une loi normale.



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

Observations 🕵️

Visuellement, les résidus suivent une distribution **approximativement normale**, mais avec des **déviations aux extrémités**. Il serait pertinent de compléter cette analyse par des tests statistiques de normalité (Shapiro-Wilk, Ljung-Box) pour confirmer ces observations.

1-2 Test statistique Ljung-Box

H0 : la série est un bruit blanc

H1 : la série n'est pas un bruit blanc

Box-Ljung test

```
data: forecast_hw$residuals
X-squared = 22.603, df = 10, p-value = 0.01231
```

Interprétation :

- Comme la **p-value** est **inférieure à 0.05**, on **rejette l'hypothèse nulle** au seuil de 5 %.
- Cela signifie que les résidus **ne suivent pas un bruit blanc**, il existe une autocorrélation significative.

Conclusion :

Notre série de résidus présente des **corrélations non négligeables**, ce qui indique que le modèle de prévision utilisé (Holt-Winters) **n'explique pas totalement la structure des données**. Nous allons passer au test de shapiro pour voir si les résidus au moins suivent une loi normale.

1-3 Test statistique de normalité des résidus (Shapiro-Wilk)

H0 : les résidus suivent une distribution normale

H1 : les résidus ne suivent pas de distribution normale

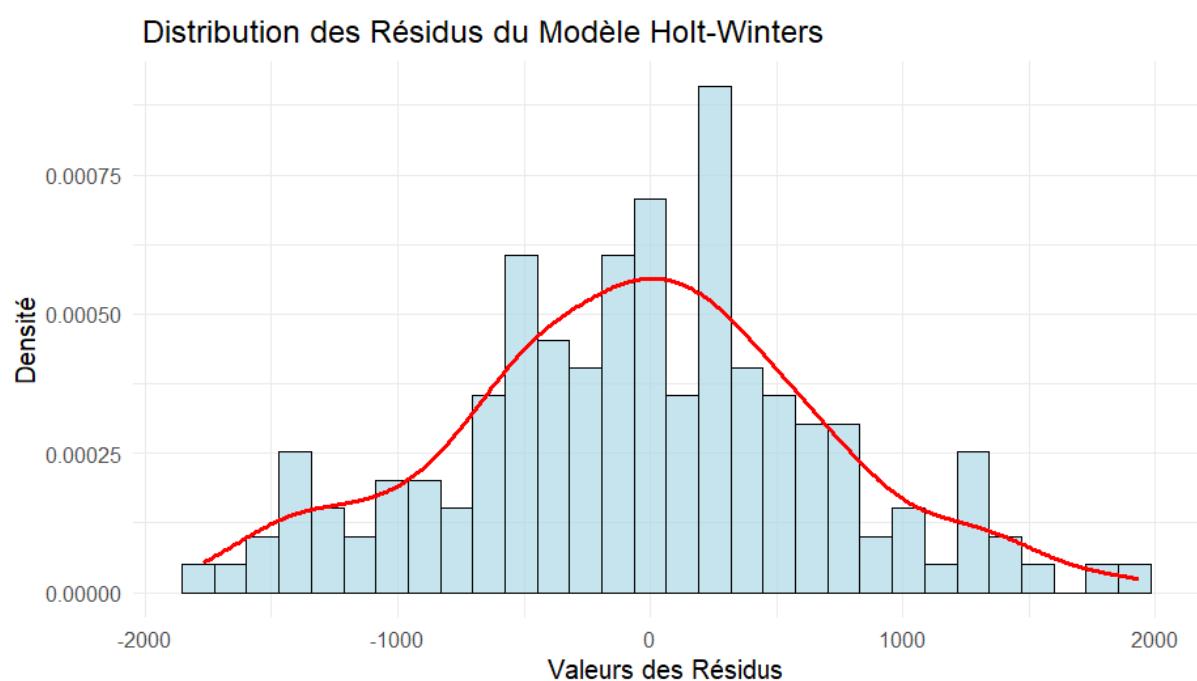
Résultats du Test de Normalité de Shapiro-Wilk :

```
> cat("Statistique de test :", shapiro_test$statistic, "\n")
Statistique de test : 0.9933045
> cat("p-value :", shapiro_test$p.value, "\n")
p-value : 0.6925506
```

Conclusion

Les résidus suivent une distribution normale (p-value > 0.05).

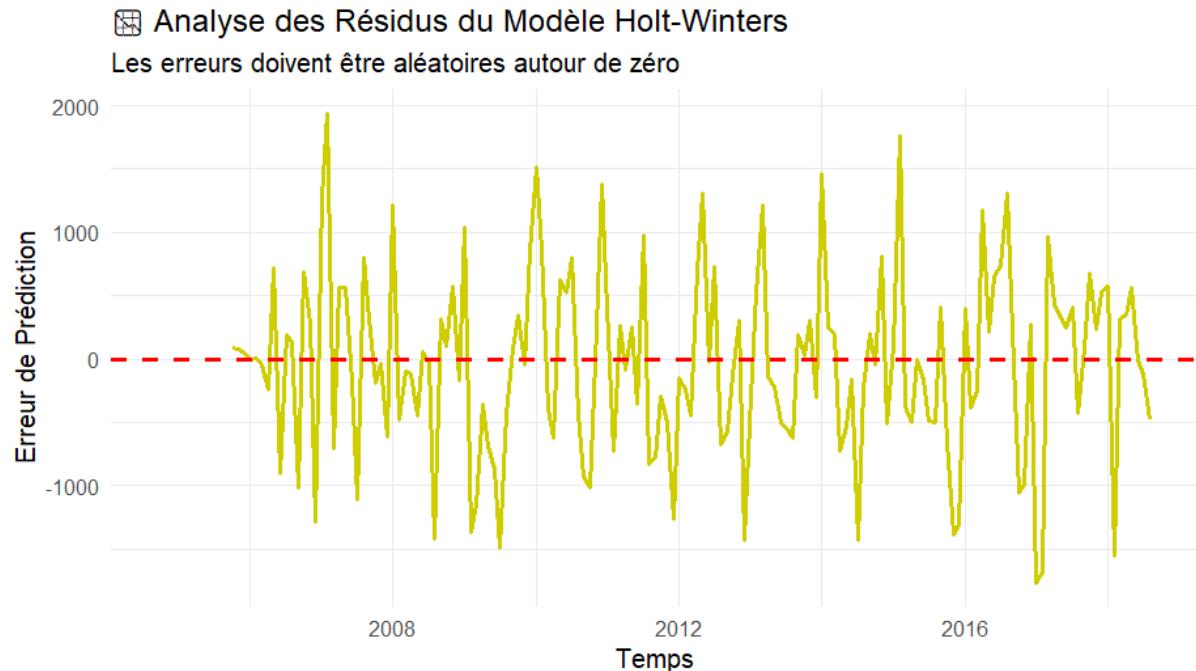
2- Histogramme des résidus avec densité superposée



Source : auteur, sortie du logiciel R studio à partir des données secondaires auprès de l'INSSEDS

Ce graphique vient renchérir ce que nous avions dit dans le paragraphe précédent. On voit d'après ce graphique  une forme en coche qui veut dire en statistique que le modèle suit la loi normale.

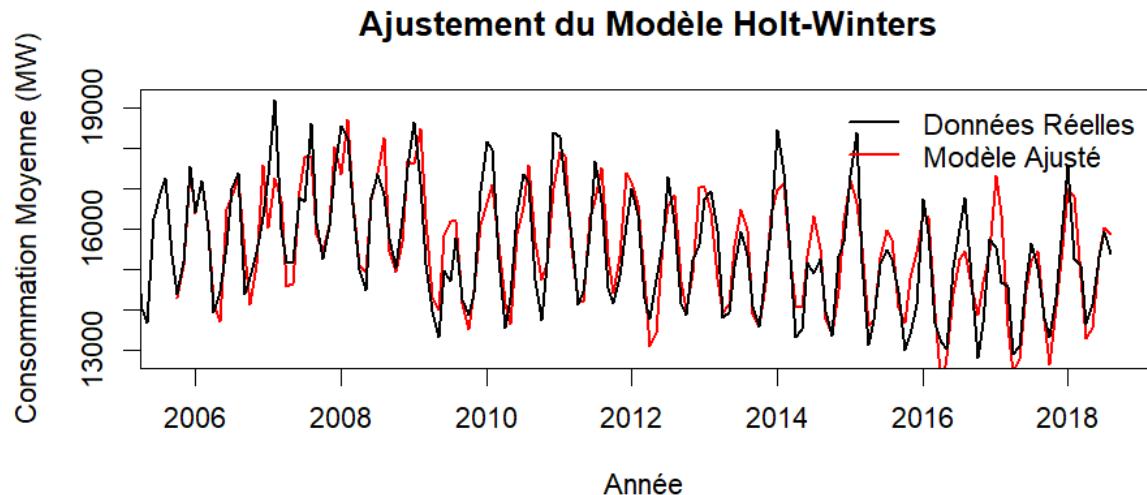
2-1 Visualisation avancée des résidus



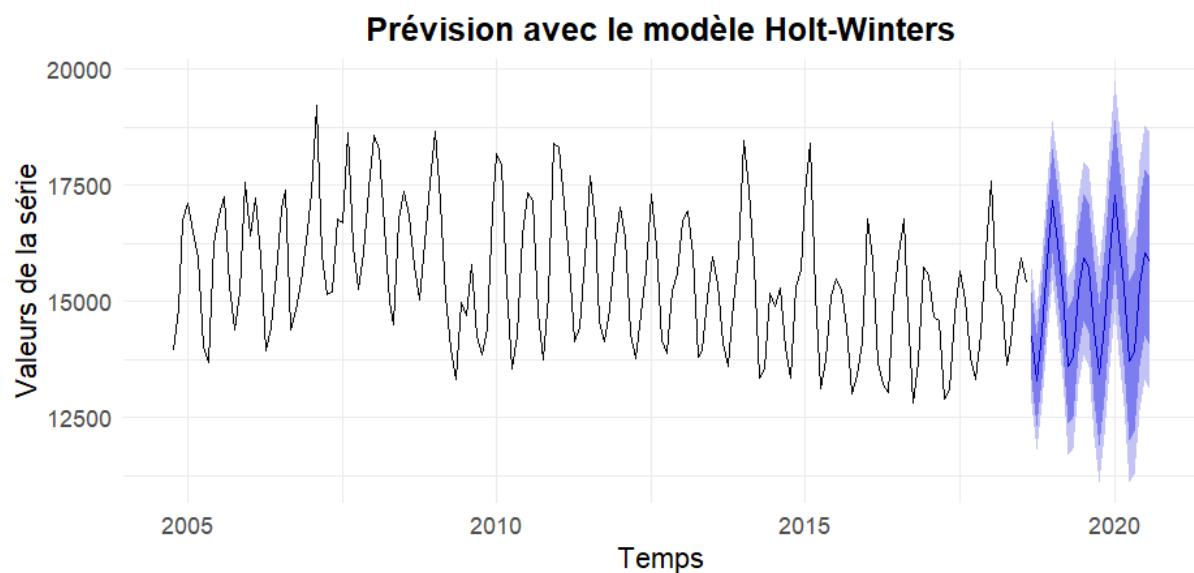
Commentaire

L'analyse des résidus du modèle Holt-Winters montre une variabilité importante des erreurs de prédition. Visuellement, on observe des fluctuations non aléatoires et des périodes de forte amplitude, suggérant une possible autocorrélation. L'absence d'un comportement purement aléatoire remet en question l'hypothèse de bruit blanc, ce qui est cohérent avec le résultat du test de Box-Ljung.

3- Ajustement du modèle Holt-Winters



4- Prévision de la série



Commentaire sur la prévision Holt-Winters (24 mois)

- **Tendance et saisonnalité :**

- La série historique présente une forte saisonnalité avec des fluctuations régulières.
- Le modèle capture bien cette saisonnalité dans les prévisions.

- **Incertitude croissante :**

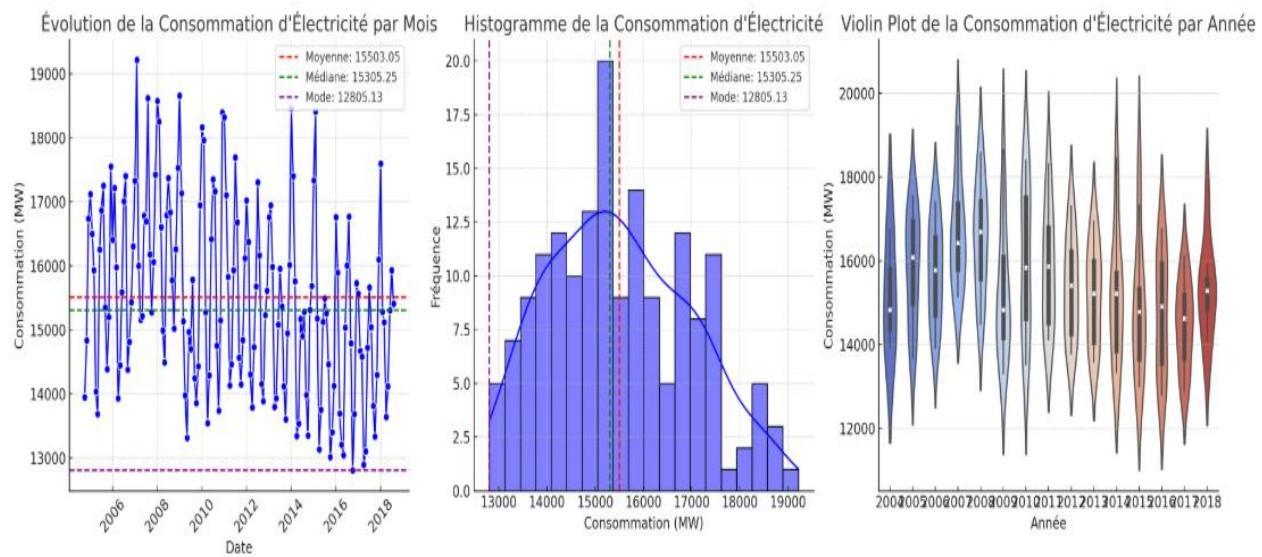
- Les intervalles de confiance (en bleu) s'élargissent fortement au fil du temps.
- Cela indique une incertitude croissante, typique des modèles exponentiels lorsqu'ils prolongent les prévisions sur une longue période.

- **Fiabilité des prévisions :**

- À court terme (prochains mois), la prévision semble cohérente.
- À long terme (24 mois), l'élargissement des intervalles suggère que la précision du modèle diminue.

Dans l'ensemble, la prévision est bien modélisée, mais l'incertitude à long terme reste une limite à prendre en compte. 

Tableau de bord



Conclusion

L'objectif de ce projet était d'analyser une série temporelle et d'effectuer des prévisions à l'aide du modèle Holt-Winters. Après une exploration approfondie des données, nous avons identifié des tendances et des variations saisonnières marquées, confirmant ainsi la pertinence de l'utilisation d'un modèle de lissage exponentiel.

L'ajustement du modèle a montré une bonne capacité à capturer la dynamique des données, bien que certaines incertitudes subsistent, notamment en raison des fluctuations imprévues et de la variabilité des résidus. L'analyse des erreurs a révélé une structure non totalement aléatoire, suggérant qu'un modèle plus sophistiqué pourrait améliorer les performances de la prévision.

En somme, cette étude a permis d'appliquer une méthodologie rigoureuse de prévision des séries temporelles et de démontrer l'importance du choix des paramètres dans l'optimisation des résultats. Pour aller plus loin, une comparaison avec d'autres modèles prédictifs, tels qu'ARIMA ou les modèles basés sur l'intelligence artificielle, pourrait être envisagée afin d'affiner les estimations et de réduire l'incertitude des prévisions.

Code source

```
# Charger les bibliothèques nécessaires

#Avant de commencer, nous devons charger les packages R qui seront utilisés pour l'importation, la manipulation des données, et les visualisations.

library(readr)

library(ggplot2)

library(forecast)

library(lubridate)

library(tseries)

library(dplyr)

library(car)

# 1. Importer les données et organiser la consommation totale d'électricité par mois

#Nous importons le fichier AEP_hourly.csv qui contient les données de consommation électrique.

AEP_hourly <-
read.csv("C:/Users/PCFRANCESERVICES/OneDrive/Desktop/insseds/jeudedonnee/AEP_hourly.csv")

# Transformation de la date

AEP_hourly$Datetime <- as.POSIXct(AEP_hourly$Datetime, format="%Y-%m-%d %H:%M:%S")

## Vérifier la structure du dataframe

str(AEP_hourly)# verifier que la date est au format date

head(AEP_hourly,10)# verifie les premières lignes

# Gestion des valeurs manquantes plus poussée

# visualisation des données manquantes

library(visdat)

vis_miss(AEP_hourly)

vis_dat(AEP_hourly)

## gestion des valeurs manquantes
```

```
colSums(is.na(AEP_hourly))

## Retirer les lignes avec des NA
AEP_hourly<- AEP_hourly %>% filter(!is.na(Datetime), !is.na(AEP_MW))

#---revisualisation des donnees manquantes----#
vis_miss(AEP_hourly)
vis_dat(AEP_hourly)

## verification des doublons
sum(duplicated(AEP_hourly)) # Nombre de doublons
## changer le nom du jeu de donnée
GR <- AEP_hourly
## Agrégation des données par mois

#### Transformation des données en séries mensuelles
# Agrégation des données par mois
GR_monthly <- GR %>%
  mutate(Month = floor_date(Datetime, "month")) %>% # Convertir la date en début de mois
  group_by(Month) %>%
  summarise(AEP_MW = mean(AEP_MW, na.rm = TRUE)) # Calcul de la consommation moyenne
  mensuelle

# Afficher les premières lignes des données agrégées
head(GR_monthly)

# Tracer la série temporelle avec un thème amélioré
ggplot(GR_monthly, aes(x = Month, y = AEP_MW)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +
  labs(title = "Évolution de la consommation électrique", x = "Date", y = "Consommation") +
```

```
theme_minimal() +  
theme(axis.text.x = element_text(angle = 45, hjust = 1))  
#####  
## Créer un Objet Série Temporelle  
# Créer un objet série temporelle  
ts_data <- ts(GR_monthly$AEP_MW, start=c(year(min(GR_monthly$Month)),  
month(min(GR_monthly$Month))),frequency=12)  
#####  
### Vérifier les Données Agrégées  
# Afficher la série temporelle  
print(ts_data)  
#####  
# Analyse Univariée de la Série Temporelle  
  
## Statistiques descriptives  
# Statistiques descriptives  
summary(ts_data)  
  
# Vérification de la variance et de la moyenne pour voir si elles évoluent dans le temps  
var(ts_data) # Variance de la série  
mean(ts_data) # Moyenne de la série  
  
# Histogramme de la consommation électrique avec lignes de référence  
ggplot(GR_monthly, aes(x = AEP_MW)) +  
geom_histogram(bins = 30, fill = "dodgerblue3", color = "black", alpha = 0.7) +  
geom_vline(aes(xintercept = mean(AEP_MW, na.rm = TRUE)), color = "red", linetype = "dashed",  
size = 1) + # Ligne moyenne  
geom_vline(aes(xintercept = median(AEP_MW, na.rm = TRUE)), color = "green", linetype =  
"dashed", size = 1) + # Ligne médiane  
annotate("text", x = mean(GR_monthly$AEP_MW) + 500, y = 30, label = paste("Moyenne :",  
round(mean(GR_monthly$AEP_MW), 2)), color = "red", size = 5) +  
annotate("text", x = median(GR_monthly$AEP_MW) - 500, y = 27, label = paste("Médiane :",  
round(median(GR_monthly$AEP_MW), 2)), color = "green", size = 5) +
```

```

labs(title = " Distribution de la Consommation Électrique Mensuelle",
      subtitle = "Comparaison entre la moyenne et la médiane pour observer la symétrie de la
distribution",
      x = "Consommation Moyenne Mensuelle (MW)",
      y = "Fréquence",
      caption = "Source : AEP_hourly.csv | Analyse Série Temporelle") +
theme_minimal(base_size = 14) +
theme(
  plot.title = element_text(face = "bold", color = "darkblue", size = 16, hjust = 0.5),
  plot.subtitle = element_text(size = 12, color = "gray40", hjust = 0.5),
  axis.text.x = element_text(color = "black"),
  axis.text.y = element_text(color = "black")
)
# Boxplot avec quartiles et valeurs extrêmes mises en évidence
ggplot(GR_monthly, aes(y = AEP_MW)) +
  geom_boxplot(fill = "tan3", color = "black", alpha = 0.6, outlier.colour = "red", outlier.shape = 16,
outlier.size = 3) +
  geom_hline(yintercept = quantile(GR_monthly$AEP_MW, 0.25), color = "blue", linetype = "dashed",
size = 1) + # Q1
  geom_hline(yintercept = quantile(GR_monthly$AEP_MW, 0.75), color = "blue", linetype = "dashed",
size = 1) + # Q3
  geom_hline(yintercept = median(GR_monthly$AEP_MW), color = "slategray4", linetype = "solid",
size = 1) + # Médiane
  annotate("text", x = 1.2, y = quantile(GR_monthly$AEP_MW, 0.75) + 500, label = "Q3 (75%)", color =
"blue", size = 5) +
  annotate("text", x = 1.2, y = quantile(GR_monthly$AEP_MW, 0.25) - 500, label = "Q1 (25%)", color =
"blue", size = 5) +
  annotate("text", x = 1.2, y = median(GR_monthly$AEP_MW) + 200, label = "Médiane", color =
"green", size = 5) +
  labs(title = " Boxplot de la Consommation Électrique Mensuelle",
      subtitle = "Les valeurs extrêmes sont mises en évidence avec les points rouges",
      y = "Consommation Moyenne Mensuelle (MW)",
      caption = "Source : AEP_hourly.csv | Analyse Série Temporelle") +
theme_minimal(base_size = 14) +

```

```
theme(  
  plot.title = element_text(face = "bold", color = "darkred", size = 16, hjust = 0.5),  
  plot.subtitle = element_text(size = 12, color = "gray40", hjust = 0.5),  
  axis.text.x = element_blank(), # Suppression de l'axe X qui est inutile ici  
  axis.text.y = element_text(color = "black")  
)  
#####  
## Test de Stationnarité (ADF Test)  
# Charger le package nécessaire  
library(tseries)  
  
# Effectuer le test ADF  
adf_result <- adf.test(ts_data)  
  
# Affichage du résultat  
print(adf_result)  
#####  
## Vérification Visuelle de la Stationnarité avec ACF/PACF  
par(mfrow = c(1, 2))  
  
# Graphique ACF (Autocorrelation Function)  
acf(ts_data, main = " ACF de la Serie Temporelle")  
  
# Graphique PACF (Partial Autocorrelation Function)  
pacf(ts_data, main = " PACF de la Serie Temporelle")  
  
par(mfrow = c(1, 1)) # Réinitialisation du graphique  
  
## Décomposition de la Série Temporelle  
library(ggplot2)
```

```
library(gridExtra)
library(zoo) # Permet de convertir les indices en dates

# Décomposition de la série temporelle
decomp <- decompose(ts_data)

# Conversion de l'index temporel en format Date
dates <- as.Date(as.yearmon(time(ts_data)))

# Création d'un DataFrame propre
decomp_df <- data.frame(
  Date = dates,
  Observed = as.numeric(ts_data),
  Trend = as.numeric(decomp$trend),
  Seasonal = as.numeric(decomp$seasonal),
  Residual = as.numeric(decomp$random)
)

# Graphique de la série observée
p1 <- ggplot(decomp_df, aes(x = Date, y = Observed)) +
  geom_line(color = "black", size = 1) +
  labs(title = "Série Observée", y = "Consommation (MW)") +
  theme_minimal()

# Graphique de la tendance
p2 <- ggplot(decomp_df, aes(x = Date, y = Trend)) +
  geom_line(color = "blue", size = 1.2) +
  labs(title = "Tendance", y = "Consommation (MW)") +
  theme_minimal()

# Graphique de la saisonnalité
p3 <- ggplot(decomp_df, aes(x = Date, y = Seasonal)) +
```

```
geom_line(color = "green", size = 1) +  
  labs(title = "Saisonnalité", y = "Consommation (MW)") +  
  theme_minimal()  
  
# Graphique des résidus  
p4 <- ggplot(decomp_df, aes(x = Date, y = Residual)) +  
  geom_line(color = "red", size = 1) +  
  labs(title = "Résidus", y = "Consommation (MW)") +  
  theme_minimal()  
  
# Affichage des 4 graphiques en grille  
grid.arrange(p1, p2, p3, p4, ncol = 1)  
  
# Matrice de Corrélation (Analyse Bivariée)  
## Création des Variables Temporelles  
  
### Extraction des Variables Temporelles  
  
# Étape 1 : Ajout des variables temporelles  
GR <- GR %>%  
  mutate(  
    Year = year(Datetime),  
    Month = month(Datetime),  
    Day = day(Datetime),  
    Hour = hour(Datetime)  
)  
  
# Étape 2 : Correction du type de Month  
GR <- GR %>%  
  mutate(Month = as.integer(Month))
```

```
# Étape 3 : Vérification après correction
```

```
str(GR)
```

```
head(GR)
```

```
# Étape 4 : Matrice de Corrélation
```

```
library(corrplot)
```

```
# Sélection des variables pour la corrélation
```

```
cor_matrix <- cor(GR %>% select(AEP_MW, Year, Month, Day, Hour), use = "complete.obs")
```

```
## Construction et Visualisation de la Matrice de Corrélation
```

```
### Calcul de la Matrice de Corrélation
```

```
library(corrplot)
```

```
#Sélection des variables pour la corrélation
```

```
cor_matrix <- cor(GR %>% select(AEP_MW, Year, Month, Day, Hour), use = "complete.obs")
```

```
# Affichage de la matrice de corrélation sous forme de tableau
```

```
print(cor_matrix)
```

```
#####
#####
```

```
### Visualisation avec un Heatmap (corrplot)
```

```
# Heatmap des corrélations
```

```
corrplot(cor_matrix,
```

```
    method = "color", # Représentation en couleur des corrélations
```

```
    type = "upper", # Afficher uniquement la moitié supérieure de la matrice
```

```
    col = colorRampPalette(c("slategray3", "white", "yellow4"))(200), # Bleu = négatif, Rouge =  
    positif
```

```
    addCoef.col = "black", # Afficher les coefficients numériques en noir
```

```
    tl.cex = 1.2, # Taille des étiquettes des variables
```

```
    number.cex = 1) # Taille des coefficients affichés
```

```
# 3. Faire la prévision pour les 24 prochains mois par la méthode de HOLT WINTER
```

```
hw_model<- HoltWinters(ts_data, seasonal = "additive")
forecast_hw <- forecast(hw_model, h=24)

# Extraction des résidus du modèle Holt-Winters
residus <- residuals(forecast_hw)

# Normalité des résidus
qqPlot(forecast_hw$residuals)

# Vérification du bruit blanc
Box.test(forecast_hw$residuals, lag = 10, type = "Ljung-Box")

# Test de normalité de Shapiro-Wilk
shapiro_test <- shapiro.test(residus)

# Affichage des résultats
cat(" Résultats du Test de Normalité de Shapiro-Wilk :\n")
cat("Statistique de test :", shapiro_test$statistic, "\n")
cat("p-value :", shapiro_test$p.value, "\n")

# Interprétation Automatique
if (shapiro_test$p.value > 0.05) {
  cat(" Les résidus suivent une distribution normale (p-value > 0.05).\n")
} else {
  cat(" Les résidus ne sont pas normaux (p-value < 0.05), une transformation peut être nécessaire.\n")
}

# Visualisation avancée des résidus avec ggplot2
ggplot(data = data.frame(Time = time(residus), Residus = residus), aes(x = Time, y = Residus)) +
  geom_line(color = "yellow3", size = 1) + # Ligne des résidus
```

```
geom_hline(yintercept = 0, color = "red", linetype = "dashed", size = 1) + # Référence à zéro  
labs(title = " Analyse des Résidus du Modèle Holt-Winters",  
     subtitle = "Les erreurs doivent être aléatoires autour de zéro",  
     x = "Temps",  
     y = "Erreur de Prédiction") +  
theme_minimal()  
  
# Histogramme des résidus avec densité superposée  
ggplot(data = data.frame(residus), aes(x = residus)) +  
  geom_histogram(aes(y = ..density..), bins = 30, fill = "lightblue", color = "black", alpha = 0.7) +  
  geom_density(color = "red", size = 1) + # Ajout de la courbe de densité pour comparer avec la normalité  
  labs(title = " Distribution des Résidus du Modèle Holt-Winters",  
       x = "Valeurs des Résidus",  
       y = "Densité") +  
  theme_minimal()  
  
# Prévision sur 24 Mois avec Holt-Winters  
  
## Vérifications à Faire Avant d'Appliquer Holt-Winters  
library(forecast)  
  
# Vérification de la série temporelle  
print(ts_data) # Vérifier les valeurs avant modélisation  
  
## Modélisation et Prévision sur 24 Mois avec Holt-Winters  
  
### Application du Modèle Holt-Winters  
# Modèle Holt-Winters  
hw_model <- HoltWinters(ts_data, seasonal = "additive")
```

```
# Vérification du modèle
print(hw_model)

## Amélioration du graphique Holt-Winters
plot(hw_model$fitted[,1], col = "red", lwd = 2, type = "l", ylim = range(ts_data),
     main = "Ajustement du Modèle Holt-Winters",
     xlab = "Année", ylab = "Consommation Moyenne (MW)")

lines(ts_data, col = "black", lwd = 2) # Données réelles en noir

legend("topright", legend = c("Données Réelles", "Modèle Ajusté"),
       col = c("black", "red"), lty = 1, lwd = 2, bty = "n") # Ajout de la légende
library(ggplot2)

# Charger les bibliothèques nécessaires
library(forecast)
library(ggplot2)

# Modèle Holt-Winters
xlisse <- HoltWinters(ts_data)

# Prédiction sur 24 mois avec intervalle de confiance
prev <- forecast(xlisse, h = 24, level = c(80, 95))

# Personnalisation du graphique
autoplot(prev) +
  ggtitle("Prévision avec le modèle Holt-Winters") +
  xlab("Temps") +
```

```
ylab("Valeurs de la série") +  
theme_minimal() +  
theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),  
axis.title = element_text(size = 12),  
axis.text = element_text(size = 10)) +  
scale_color_manual(values = c("blue", "red", "black")) # Personnalisation des couleurs
```